IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Water Quality Prediction For Agriculture

Dr.R. Arthy,G. Siva Prakash,A. Rakesh,A. Gowri Supramanian

Assistant Professor, Department of Information Technology
, Student of Department of Information Technology
, Student of Department of Information Technology, Student of Department of Information Technology
Kamaraj College of Engineering and Technology
Virudhunagar, India

Abstract : Water is one of the most essential element for the existence of life. The safety and accessibility of drinking-water are major concerns throughout the globe. Health risks may arise from consumption of water contaminated with infectious agents, toxic chemicals etc. In this paper a system is proposed to check the water quality and warn the user before water gets contaminated .There are different parameters that can contaminate the water. These parameters are taken into account and used for predicting when to clean the water. The system uses technologies such as IoT and Machine Learning. It consist of the physical and chemical sensor to measure pH, turbidity ,colour, DO, conductivity etc. to check the parameters .The data obtained from the sensors are recorded in the database and further sent for analysis. The neural network algorithm is used for predicting the result. It is used to obtain non-linear relationship for predicted output. The system sends the alert message to user when any of the parameters are lower than the standard values. This helps the user to know beforehand about the contamination of water in their residential tanks. This technique can not only be limited up to residential tanks but can also be used in water treatment plants and industries.

***IndexTerms* -** Parameter, Accuracy, Heatmap Generation, Time series analysis,Prediction..

I INTRODUCTION

A subfield of artificial intelligence (AI) and computer science called machine learning focuses on using data and algorithms to simulate how humans learn, gradually increasing the accuracy of the system.IBM has a long history with artificial intelligence. One of its own, Arthur Samuel, is credited with creating the term "machine learning" with his research on the game of checkers (PDF, 481 KB) (link lives outside IBM). In 1962, Robert Nealey, a self-described checkers master, competed against an IBM 7094 computer, but he was defeated. This achievement nearly looks insignificant in comparison to what is currently possible, but it is regarded as a significant turning point in the development of artificial intelligence. The subsequent two decades' worth of technological advancements will result in Certain cutting-edge technologies that we already know and appreciate, like Netflix's recommendation engine or self-driving cars, will be made possible by increased storage and processing capacity.The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are trained to generate classifications or predictions using statistical techniques, revealing important insights in data mining operations. The decisions made as a result of these insights influence key growth indicators in applications and enterprises, ideally. Data scientists will be more in demand as big data continues to develop and grow, necessitating their assistance in determining the most pertinent business questions and ultimately the data to answer them.

II LITERATURE SURVEY

[1] Forecast of Water Quality This work presents the modelling and forecasting of water quality using artificial intelligence algorithms developed . The Water Quality Index (WQI) and Water Quality Classification (WQC) algorithms, as used in today's advanced technology, have been used in this research. Long Short-Term Memory (LSTM) and Nonlinear Autoregressive Neural Network (NARNET) are two examples of deep learning methods. In order to categorise the WQI, machine learning methods like SVM, KNN, and Nave Bayes are also used.

[2]Jitha P. Nair and M. S. Vijaya's Prediction Models for River Water Quality Using Machine Learning and Big Data Approaches. This Paper discusses how water resources are becoming more contaminated**.** Industrial waste, human waste, vehicle waste, agricultural runoff from farmlands carrying chemical elements, undesired nutrients, and other pollutants from point and non-point sources flow to water bodies, which affects the quality of the water resources, etc. The amount and quality of water are affected by the rise in pollution, which puts human health and other living things on the earth at serious risk. In light of this, it is now important and relevant to do research on how to assess, monitor, and predict the quality of water. Traditional methods have been employed by many researchers; currently, they are evaluating and predicting water quality utilising technologies like ML and big data analytics.

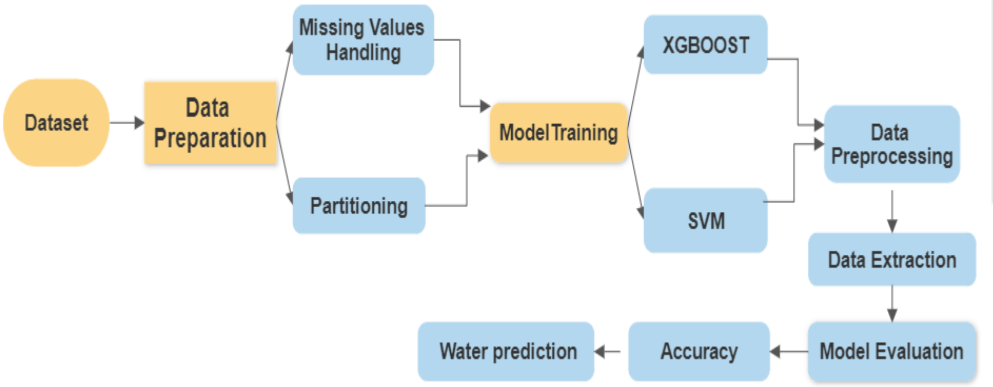
[3] S. Vijay and Dr. K. Kamaraj's publication, "Ground Water Quality Estimation Using Machine Learning Algorithms in R," recounts how the bore wells from which the samples were taken are frequently utilised for drinking. The projected values of the water quality parameters include PH, TDS, EC, chloride, sulphate, nitrate, carbonate, and bicarbonate, as well as metal ions and trace elements. In the Vellore district, there are two main types of water contamination: High and Low. This study focuses on forecasting water quality with high accuracy and efficiency utilising the Machine Learning classifier algorithm C5.0, Naive Bayes, and Random Forest.

[4] Authors Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin sifatul Islam, and Mostofa Kamal Nasir used principal component regression and gradient boosting classifier technique to predict and classify water quality. This work describes a main component regression-based water quality prediction model. First, the weighted arithmetic index approach is used to determine the water quality index (WQI). Second, the dataset is subjected to principal component analysis (PCA), and the most important WQI parameters have been retrieved. Thirdly, various regression techniques are applied to the PCA result in order to forecast the WQI. The Gradient Boosting Classifier is then used to assign a status to the water quality. A dataset connected to Gulshan Lake is used to experimentally evaluate the suggested system. The outcomes show that the principal's predictions were 95% accurate. As compared to cutting-edge models, the component regression approach and the Gradient Boosting Classifier method exhibit respectable performance.

[5] Anil Kumar Bisht, Ravendra Singh, Rakesh Bhutiani, and Ashutosh Bhatt's work, "Application of Predictive Intelligence in Water Quality Forecasting of the River Ganga Using Support Vector Machines," discusses how to use The bodies in charge of water management may profit from a precise prediction of river water quality. However, making a prediction is a difficult task because of the complicated interaction that exists between many components. Here, the scientists tried to create a model that could forecast or anticipate the water quality of the Ganga River using a machine learning-based predictive intelligence technique called a support vector machine (SVM). Five sample stations were used for the monthly data sets of five water quality indicators from 2001 to 2015, and they were located from Dev Prayag to Roorkee in the Indian state of Uttarakhand. The radial basis function (RBF) is used in the experiments in Python 2.7.13 as a kernel for creating a non-linear SVM-based classifier that serves as a model for predicting water quality. The results demonstrated the

SYSTEM METHODOLOGY

The suggested system's main goal is to assess potability. It is split into two sections: a testing phase and a training phase. The subsequent steps are completed in both sections. Information on training pH and hardness tests There are several words that can be used to describe something, including solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. This is how the data set was selected: The selection of the water quality data set, which is a requirement for model construction, is based on the collection of crucial parameters that affect water quality, the estimation of the number of data samples, and the definition of the class labels for each data sample that is present in the data. The ten indicator parameters that make up the data sets used in this study. Hardness and pH value are two examples of these elements. A substance's characteristics can be described using the phrases solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability, among others. However, neither the number of parameters nor the choice of parameters restricts the proposed approach. The learning and testing framework in this study is established using a k-fold cross validation technique and corresponds to each data sample in the data set. This method divides the dataset into k-disjointed sets of equal size and about equal class distribution. The remaining subsets of this split are used as the training set, and the remaining subsets are used as the test set.



Data Preparation

Data gathering: Gathering pertinent information on water quality from a variety of sources, including governmental organisations, water treatment facilities, and academic institutions. Data cleaning: Eliminating inaccuracies, inconsistent data, and incomplete data. This include eliminating redundant entries, adding values when they are missing, and fixing mistakes. Data normalisation is the process of putting data into a format that is commonly utilised for analysis. This entails transforming categorical variables into numerical values and scaling the data to a common range. Choosing the features that are most important to improving the water quality. In order to do this, the association between various features must be examined, and the most crucial features must then be chosen. Dividing the data into training and testing sets is known as data splitting. The testing set is used to assess the model's performance after it has been trained using the training set. Data augmentation: Adding noise to current data or changing it to produce new data. When the dataset is tiny and the model need more information to perform better, this is helpful. Visualizing the data to understand its distribution, the relationships between its many properties, and outliers. This aids in adjusting the model and choosing the best machine learning method.

Model Training

A data scientist can start building a model after preprocessing the acquired data and separating it into train and test sets. This procedure comprises "feeding" training data to the algorithm. Predictive analysis uses an algorithm to evaluate data and produce a model that can locate a target value (attribute) in fresh data. To create a model is the goal of model training.

Data Pre-Processing

Pre-processing is done to transform raw data into a format that can be used by machine learning. A data scientist can use an applied machine learning model to obtain more accurate findings by using structured and clean data. Data formatting, cleansing, and sampling are all part of the method.

Data Extraction

Training, test, and validation sets should be divided into three subsets when using a dataset for machine learning. Set for training. In order to train a model and determine its ideal parameters, a data scientist needs a training set. test system. For an assessment of the trained model's generalizability, a test set is required. The latter refers to a model's capacity to find patterns in fresh, unexplored data after being trained on training data. The inability for generalisation we described earlier, known as model over fitting, must be avoided by using distinct subsets for training and testing.

Model Evaluation

To assess how well the model can forecast the output variable based on the input variables, metrics like as accuracy, precision, recall, F1 score, and others are measured. For model evaluation, a number of methods are employed, including as cross-validation, holdout validation, and bootstrapping. The dataset is split up into different subgroups for cross-validation, and the model is trained and evaluated on each subset separately. The dataset is split into two portions for training and testing in holdout validation. Bootstrapping is the process of creating many samples for training and testing by randomly sampling the dataset with replacement. In machine learning, model evaluation is essential since it reveals a model's advantages and disadvantages. One can adjust the parameters and raise the model's accuracy by comprehending how it performs on the provided dataset. Moreover, model evaluation aids in deciding which model, out of several that were trained on the same dataset, is the best.

III RESULTS ANALYSIS

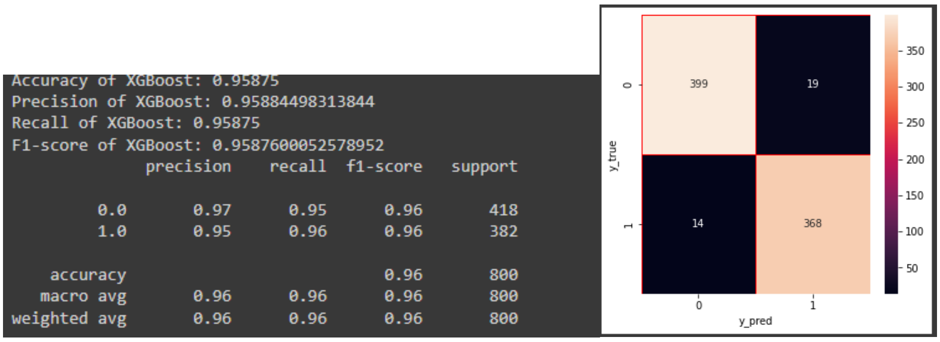
Saving time and resources in lab analysis has been made possible by modelling and the prediction of water quality. As a substitute technique for estimating and forecasting water quality, artificial intelligence algorithms were investigated. This study makes the case for the integrated artificial intelligence methods put forth in the current study as a promising tool for precisely simulating water quality and level. The created model enables quick and low-cost prediction of water quality and index, and subsequently high-accurate classification of water quality. The XGBOOST and SVM demonstrate greater accuracy predictions in this study. The performance accuracy of machine learning algorithms is generally higher.

S.NO	DATASET	ACCURACY	
		SVM	XGBOOST
1	3277 rows and 10 columns	0.93	0.96
2	4177 rows and 10 columns	0.92	0.94
3	6000 rows and 10 columns	0.94	0.98

XGBOOST

Popular machine learning techniques like XGBoost are employed for supervised learning tasks, particularly for classification and regression issues. Extreme Gradient Boosting is what it's called, and Tianqi Chen and Carlos Guestrin created it in 2016. Because XGBoost is a tree-based ensemble approach, its final forecast is a combination of predictions from various decision trees. Decision trees are sequentially added to the model as part of the algorithm, with each tree attempting to correct the flaws of the one before it. The handling of missing data and regularisation approaches by XGBoost, which helps avoid overfitting, is one of its important characteristics. Additionally, it offers an effective implementation that makes it possible to train models quickly even when the datasets are big. For a variety of uses, including machine learning contests, finance, healthcare, and more, XGBoost has grown to be a popular option. It can be programmed in a number of languages, including Python, R, Java, and C++.

$$F2(x) = H0(x)+eta(H1(x)) +eta(H2(x))$$

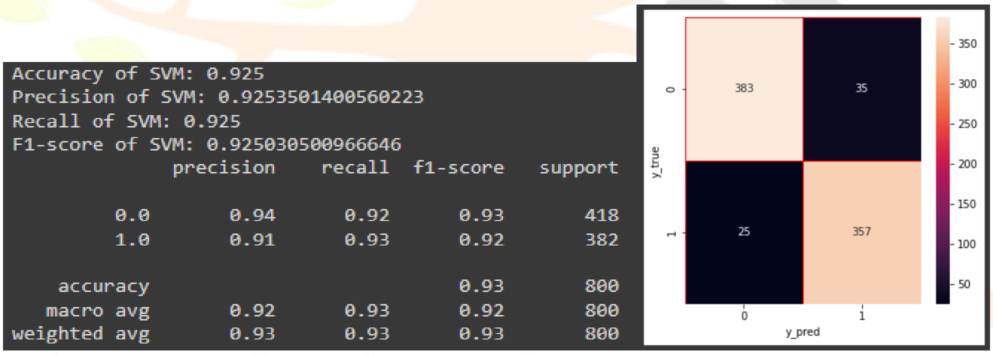


```
Accuracy of XGBoost: 0.95875
Precision of XGBoost: 0.95884498313844
Recall of XGBoost: 0.95875
F1-score of XGBoost: 0.9587600052578952
              precision    recall  f1-score   support

         0.0       0.97      0.95      0.96       418
         1.0       0.95      0.96      0.96       382

    accuracy                           0.96       800
   macro avg       0.96      0.96      0.96       800
weighted avg       0.96      0.96      0.96       800
```

Fig 1,XGBOOST.

SUPPORT VECTOR MACHINE

A well-liked and effective machine learning approach called Support Vector Machine (SVM) is utilised for both classification and regression analysis. Its foundation is the idea of locating a hyperplane that best distinguishes between the various classes in a given dataset. The core principle of SVM is to transfer the input data into a high-dimensional feature space, where it is simpler to locate a hyperplane that can best distinguish between the various classes. The gap between the nearest data points from each class and the hyperplane is referred to as the margin.The most advantageous hyperplane in SVM is the one that maximises the margin. Support vectors are utilised to define the hyperplane; these are the points that are closest to the hyperplane. In order to select the best hyperplane, the SVM algorithm must solve a quadratic optimization problem. SVM is a strong algorithm that offers a number of advantages over other classification algorithms, including the capacity to handle high-dimensional data, the use of kernel functions to handle non-linearly separable data, and the capacity to prevent overfitting by maximising the margin. SVM is widely utilised in many fields, including bioinformatics, image classification, and text categorization.



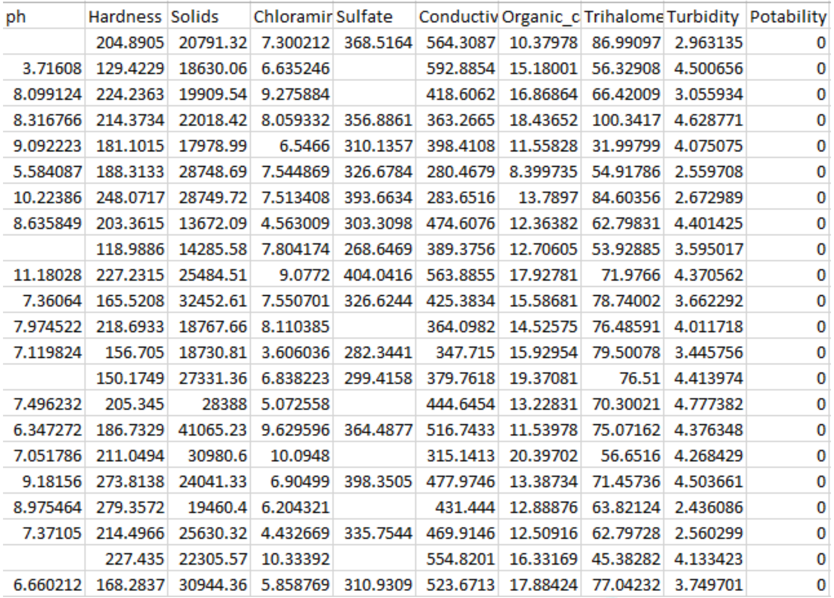
```
Accuracy of SVM: 0.925
Precision of SVM: 0.9253501400560223
Recall of SVM: 0.925
F1-score of SVM: 0.925030500966646
              precision    recall  f1-score   support

         0.0       0.94      0.92      0.93       418
         1.0       0.91      0.93      0.92       382

    accuracy                           0.93       800
   macro avg       0.92      0.93      0.92       800
weighted avg       0.93      0.93      0.93       800
```

Fig 2,SVM.

DATASET FILE

The dataset is in the form of a CSV file and has 3277 rows and 10 columns. It includes elements like conductivity, PH, solids, hardness, chloramine, sulphate, organic carbon, trihalomethane, and turbidity.

IJNRD Research Through Innovation



ph	Hardness	Solids	Chloramir	Sulfate	Conductiv	Organic_c	Trihalome	Turbidity	Potability
	204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
	118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
	150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
8.975464	279.3572	19460.4	6.204321		431.444	12.88876	63.82124	2.436086	0
7.37105	214.4966	25630.32	4.432669	335.7544	469.9146	12.50916	62.79728	2.560299	0
	227.435	22305.57	10.33392		554.8201	16.33169	45.38282	4.133423	0
6.660212	168.2837	30944.36	5.858769	310.9309	523.6713	17.88424	77.04232	3.749701	0

IV Conclusion

Water quality, one of the most crucial resources for life, is determined by portability. Previously, a costly and drawn-out lab analysis was needed to test the quality of water. This work investigated a different machine learning approach for forecasting water quality using just a few straightforward water quality variables. A group of representative supervised machine learning methods were utilised to estimate. Before water was made available for consumption, it would find water of poor quality and alert the proper authorities. It should decrease the number of people who drink water that is of poor quality.

V Future Work

The cost-effective way to prevent water contamination in residential overhead tanks is discussed in the study. IoT devices are used to monitor water quality, and machine learning algorithms are used to anticipate potential water contamination. The suggested system gathers water characteristics using many sensors that are coupled to a controller. Also, the user receives the alert message before the water becomes poisoned. The technology is both economical and useful in protecting the water from contaminants. Further work on this project will focus on identifying diseases brought on by various characteristics and selecting the best method for tank cleaning. In order to improve the quality of the water, biosensors can also be employed to find little bacteria.

References

[1] GIS applications for mapping and spatial modelling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil, P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M. Lima, and C. S. Oliveira, Cadernos de Vida Pblica, vol. 23, no. 4, pp. 875-884, 2007..

[2] A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data is described by C. N. Babu and B. E. Reddy in Applied Soft Computing, volume 23, pages 27–38, 2014.

[3] Evaluation of surface water quality in Lake Gala, Turkey, using satellite image fusion based on PCA approach, IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 5, 2019: 2983-2989. E. Batur and D. Maktav.

[4] Classification based on decision tree method for machine learning, B. Charbuty and A. M. Abdulazeez, Journal of Applied Science and Technology Trends, vol. 2, no. 01, 2021, pp. 20-28.

[5] Classification of water potability using machine learning methods, M. I. K. Haq, F. D. Ramadhan, F. Az-Zahra, L. Kurniaw, and A. Helen, Proceedings of the 2021 International Conference on Artificial Intelligence and Big Data Analytics, October 2021.

[6] Hyperparameter optimization for machine learning models based on bayesian optimization: J. Wu, X.-Y. Chen, H. Zhang, Li-D. Xiong, H. Lei, and Si-H. Deng, Journal of Electronic Science and Technology, vol. 17, no. 1, pp. 26-40, 2019.

[7] Geological resource planning and environmental impact assessments based on GIS, Y. Xie, B. Xie, Z. Wang et al., Sustainability, vol. 14, no. 2, p. 906, 2022.

[8] International Journal for Research in Applied Science & Engineering Technology, December 2022, Anna ji Kuthe, "International Journal for Research"

[9] "Predictive models for river waterquality using machine learning approaches - a survey," by J. P. Nair and M. S. Vijaya, appeared in the proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), held in Coimbatore, India, in March 2021.

[10] IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 5, pp. 2983-2989, 2019, E. Batur and D. Maktav, "Assessment of Surface Water Quality by Using Satellite Images Fusion Based on PCA Method in the Lake Gala, Turkey."

[11] Adaptive neuro-fuzzy inference system (ANFIS) application to estimate the biochemical oxygen demand (BOD) of Surma River by A. A. M. Ahmed and S. M. A. Shah, Journal of King Saud University - Engineering Sciences, vol. 29, no. 3, pp. 237-243, 2017.

[12] A. A. Al-Othman, "Evaluation of Riyadh Mainline Saudi Arabia Surface Water Suitability for a Variety of Uses," Arabian Journal of Chemistry, vol. 12, no. 8, pp. 2104-2110, 2019. View at: Google Scholar | Publisher Website

[13] Intelligent hybrid model to improve time series models for forecasting network traffic, T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, IEEE Access, vol. 8, 2020, pp. 130431–130451

[14] International Journal for Research in Applied Science & Engineering Technology, Anna ji Kuthe, "International Journal for Research", December 2022.

[15] Water, vol. 11, no. 6, p. 1231, 2019 A. S. Abobakr Yahya, A. N. Ahmed, F. Binti Othman et al., "Water quality prediction model based support Vector machine model for Ungauged river Catchment for dual Scenarios."