International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Mar-2024)
http://dx.doi.org/10.12785/ijcds/150184

A Comprehensive Comparative Study of Machine Learning Algorithms for Water Potability Classification

Fuad Ahmad Musleh

Department of Civil Engineering, College of Engineering, University of Bahrain, Isa Town, Bahrain

fuadakm@hotmail.com

Received 13 Sep. 2023, Revised 24 Jan. 2024, Accepted 11 Feb. 2024, Published 1 Mar. 2024

Abstract: Water quality (WQ) prediction is of utmost importance due to the scarcity of uncontaminated water resources. In this study, six machine learning (ML) algorithms, including Bagging classifier, Logistic regression (LR), J48, Random Forest (RF), IBk, and AdaBoostM1, were employed to assess water potability. Evaluation metrics such as accuracy, recall, precision, F-measure, false positive (FP) rate, receiver operating characteristic (ROC) area, and precision-recall curve (PRC) area were used to compare the capability of the models. The outcomes of the comparative analysis revealed that RF and J48 achieved the highest accuracy values of 0.993, followed closely by the Bagging classifier with an accuracy of 0.992. The AdaBoostM1 algorithm achieved an accuracy of 0.971, while the LR algorithm achieved an accuracy of 0.958. The IBK algorithm showed a lower accuracy of 0.714. The comparative analysis of the FP rate metric demonstrated that RF achieved the lowest rate of 0.006, followed closely by the Bagging classifier and J48, both with a rate of 0.007. AdaBoostM1, LR, and IBK had higher rates of 0.026, 0.041, and 0.289, respectively. Regarding precision, RF and J48 achieved the highest precision rates of 0.993, followed by the Bagging classifier at 0.992. The AdaBoostM1 algorithm achieved a precision rate of 0.972, and LR achieved 0.958. IBK showed less precision rate of 0.714. For the recall metric, RF and J48 achieved the highest recall values of 0.993, followed closely by the Bagging classifier with a recall value of 0.992. The AdaBoostM1 algorithm obtained a recall value of 0.971, while LR and IBK achieved values of 0.958 and 0.714, respectively. The study highlights the effectiveness of RF, J48, and the Bagging classifier in predicting water potability. These findings contribute valuable insights for the implementation of accurate prediction models, supporting the sustainable management of water resources.

Keywords: *Artificial Intelligence, Neural Network, Water Potability, Machine Learning, Civil Engineering, Environmental Engineering.*

1. INTRODUCTION

Clean uncontaminated water is a precious supply critical for the well-being of both humanity and the preservation of ecosystems. With the ever-increasing demand and scarcity of freshwater worldwide, ensuring access to safe and potable water has become a critical challenge. Civil and Environmental engineering are pivotal in tackling this challenge by innovating water management and treatment techniques. Their crucial role involves developing and implementing novel approaches that ensure efficient and sustainable solutions for addressing WQ issues and safeguarding precious water resources [1].

In recent years, the field of WQ classification has witnessed advancements through the application of various techniques. These include chemical analysis, remote sensing, and statistical models. However, the integration of ML techniques has shown itself to be a commanding tool in the field of assessment and prediction [2].

ML has proven to be valuable in various aspects of WQ control (WQC), including predicting the potability of water and optimizing wastewater treatment processes. By leveraging large datasets and complex algorithms, ML models can analyze and interpret WQ parameters to provide accurate predictions and valuable insights. These advancements have revolutionized the field, enabling more efficient and effective management of water resources [3].

One of the key techniques used in ML for WQ prediction is the application of NNs. NNs are ML models designed to replicate the structure and operation of the human brain. Comprising interconnected nodes or "neurons," these models effectively process and transmit information. In the context of WQ, neural networks excel at recognizing Patterns and interdependencies within complex datasets, allowing for accurate forecasting of WQ parameters [4].

Quality assessment is the supervised machine learning (SML) classifier. SML classifiers utilize labeled training data to observe regularities and generate forecasts based on

E-mail: fuadakm@hotmail.com

new, unseen data. By training the classifier on historical WQ data, it can learn to classify water samples as potable or non-potable based on specific criteria. This enables rapid and automated assessment of WQ, reducing the need for time-consuming and costly laboratory analyses [5].

In addition to predicting water potability, ML techniques are also instrumental in developing comprehensive WQ indices. These indices provide a holistic assessment of WQ by combining multiple parameters, such as chemical, physical, and biological indicators. ML models can analyze large datasets and determine the effectiveness of each parameter, enabling the creation of accurate and reliable WQ indices. These indices serve as valuable tools for policymakers, researchers, and water management professionals in making informed decisions and prioritizing interventions [6].

In summary, the integration of ML techniques in the sector assessing and predicting the quality of water emerged as a game-changer. Civil and environmental engineers, along with ML practitioners, are at the forefront of developing innovative approaches to guarantee the availability of clean and safe water. NN and SML classifiers play vital roles in predicting water potability and automating WQ assessment. Furthermore, ML models play a part in the development of comprehensive WQ indices, enabling effective management and monitoring of water resources. As water scarcity continues to be a global concern, the advancements in ML offer promising solutions for maintaining the availability and quality of this precious resource.

In this research, a comprehensive comparison of multiple classification algorithms was conducted to evaluate their performance in the sector of forecasting WQ. The algorithms assessed in this study encompassed the Bagging classifier, LR, J48, RF, IBk, and AdaBoostM1. Various performance metrics, including recall, precision, ROC area, F-measure, PRC area, TP rate, FP rate, were utilized to assess the effectiveness of these algorithms in WQ classification tasks. The outcomes of this comparative analysis provide valuable information into the suitability and performance of these algorithms for accurate and reliable WQ prediction and assessment.

2. LITERATURE REVIEW

In recent years, the involvement of ML algorithms in WQ assessment has garnered considerable focus. Numerous studies have explored the use of ML algorithms and models to analyze WQ parameters, enabling accurate predictions and efficient monitoring. This literature review aims to summarize and evaluate the findings of these studies, highlighting the advancements and potential of ML in WQ assessment.

This research conducted by Ahmed, et al. [2] explores the usage of different SML algorithms to estimate the Water Quality Index (WQI) and WQC using input features: pH, turbidity, temperature, and total dissolved solids. Gradient boosting achieves the highest efficient WQI prediction with an MAE of 1.9642, while MLP exhibits the highest WQC classification accuracy of 0.8507. These findings show the potential of real-time water quality detection systems with minimal parameters. validate

Another study conducted by Lu and Ma [7] proposes two novel hybrid decision tree-based ML models, integrating The combination of complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) alongside extreme gradient boosting (XGBoost) and RF has been utilized for improved short-term WQ prediction. Using data from polluted sites, six WQ indicators were predicted. Results demonstrate superior performance of CEEMDAN-RF and CEEMDAN-XGBoost models in various indicators, with MAPEs ranging from 0.27% to 14.94%. These models exhibit the best overall prediction performance, with average MAPEs of 3.90% and 3.71%, respectively. Furthermore, the stability analysis confirms their higher prediction stability compared to benchmark models.

Accurate forecasting of WQ time series is of utmost importance for efficient water resource handling. Traditional linear models face challenges due to the complex nature and noise present in WQ data. To address this issue, Bi et al. [5] proposed a hybrid model that combines a Savitzky-Golay filter with a long short-term memory (LSTM)-based encoder-decoder architecture in a neural network (NN). The Savitzky-Golay filter eliminates noise, while the LSTM captures nonlinear characteristics in the water environment. Experimental results using realistic data demonstrate that the integrated model outperforms several state-of-the-art models, providing superior prediction performance. Additionally, another algorithm based on LSTM NN is established and trained by Wang, et al. [8] using monthly WQ indicator data from Taihu Lake (2000-2006). Simulations and parameter selection enhance predictive accuracy. The method is compared with backpropagation neural network and online sequential extreme learning machine, demonstrating superior accuracy and generalization. It offers a more precise and comprehensive approach to WQ prediction.

Moving forward, in another study performed by Azrour, et al. [3] the advantages of ML algorithms are utilized to develop an algorithm with the ability of forecasting the WQI and WQC. The proposed method is built upon the parameters: turbidity coliforms, pH, and temperature. The implementation of multiple regression algorithms is essential and proves to be effective in forecasting the WQI. Additionally, the adoption of artificial neural networks offers a highly efficient approach for accurately classifying WQ. The results highlight the efficacy of this approach in accurately predicting the WQI and class, showcasing the potential of ML in WQ assessment.

In order to forecast harmful algal bloom (HAB) incidents, the study conducted by Deng, et al. [9] utilizes ML methods, including SVM and artificial neural networks (ANN) enhanced by hybrid learning algorithms. With over 30 years of measured data, the accuracy and applicability of both ML methods in predicting algal growth and eutrophication trends in Tolo Harbor are demonstrated. ANN shows quick response and satisfactory results, while SVM accurately identifies optimal models despite longer training time. Moreover, the ML methods effectively capture the intricate interconnections between coastal environmental variables and algal dynamics, accurately identifying significant factors. The findings underscore the potential of ML models to enhance the coastal hydro-environment management water forecasting by providing valuable insights into HAB outbreak mechanisms and evolution.

The WQ forecasting performance of ten ML models was compared in a study conducted by Chen et al. [10] using a large dataset that contains a total of 33,612 data point collected from prominent rivers and lakes across China over the period of 2012 to 2018. Evaluation metrics, including precision, recall, F1-score, and weighted F1-score, were employed to assess the models. The outcomes suggested that the capability of learning models improved as the dataset size increased. Specifically, the DT, RF, and deep cascade forest (DCF) models trained on pH, the dissolved oxygen (DO), CODMn, and NH3-N datasets outperformed other models, accurately predicting all six WQ levels specified in the Chinese governmental guidelines. Additionally, two specific water parameter sets (DO, CODMn, NH3-N; CODMn, NH3-N) were identified as highly effective for WQ prediction. Thus, the study recommended the utilization of DT, RF, and DCF models incorporating these parameters for future WQ monitoring and timely warnings.

To enhance WQ forecasting accuracy, this study performed by Yu, et al. [11] introduces a novel hybrid model that combines data decomposition, fuzzy C-means clustering, and bidirectional gated recurrent unit (BiGRU). The unprocessed WQ dataset undergoes empirical wavelet transform for decomposition into subseries, which are then recombined using fuzzy C-means clustering. Each clustered series is subjected to a bidirectional gated recurrent unit to develop a prediction model. The forecast result is obtained by summing the predictions for the subseries. The effectiveness of the proposed model is tested using Poyang Lake's WQ data from China, demonstrating highly accurate forecasts for all six WQ parameters. MAPE for seven-day ahead predictions are 4.59%. Moreover, the proposed model outperforms other models, reducing MAPE by an average of 32.86% compared to the single BiGRU model. These results highlight the impact fullness of the proposed hybrid model for WQ forecasting.

Haq and Harigovindan [12] conducted a study to enhance WQ prediction in aquaculture through the introduction of hybrid deep learning (DL) models. These models combined convolutional neural network (CNN) with LSTM and gated recurrent unit (GRU) architectures. The CNN component effectively captured the important characteristics of aquaculture WQ, while LSTM and GRU models learned long-term dependencies in the time series data. Extensive experiments were carried out using different WQ datasets, analyzing the influence of hyperparameters on model performance. The proposed hybrid DL models, CNN-LSTM and CNN-GRU, were compared with baseline LSTM, GRU, and CNN models, as well as attention-based LSTM and GRU models. The results revealed that the CNN-LSTM model demonstrated superior performance in terms of predictive accuracy and computational efficiency, surpassing all other models.

Researchers Nair and Vijaya [13] conducted a study aiming to create a prediction model with high efficiency for assessing the quality of river water and categorizing index values based on predetermined WQ standards. The focus was on constructing a robust model capable of accurately predicting WQ and classifying it according to established standards. The study utilized a dataset collected from eleven sampling sites situated along the course of the Bhavani River, which spans across Kerala and Tamil Nadu. The dataset incorporated 27 parameters, including DO, temperature, pH, alkalinity, hardness, chloride, and coliform count, among others, to determine the WQI. To facilitate the development of ML models, the dataset underwent data normalization and feature selection techniques.

Several algorithms, such as linear regression, MLP regressor, RF, and support vector regressor were employed to construct a WQ prediction model. Additionally, classifiers including SVM, naive Bayes, DT, and MLP were utilized to establish a classification model for the WQI. Experimental results highlighted the MLP regressor's effectiveness in accurately predicting the WQI, yielding RMSE of 2.432. Moreover, the MLP classifier achieved an impressive 81% accuracy in classifying the WQI. These findings demonstrate the promising outcomes of the developed models for WQ prediction and classification.

In another study proposed by Juna et al. [4], a novel approach was introduced to handle the issue of absent values in WQ prediction. The approach combined a nine-layer MLP with a K-nearest neighbor (KNN) imputer. The study compared this method with seven other ML models under two conditions: Omitting incomplete data and utilizing a KNN imputer. The outcomes revealed that the nine-layer MLP model, in conjunction with the KNN imputer, achieved exceptional accuracy of 0.99 for forecasting WQ. This high accuracy was further validated through K-fold cross-validation, affirming the robustness and reliability of the proposed approach.

In summary, the mentioned studies focused on constructing efficient prediction and classification models for river WQ. The first study showcased the effectiveness of MLP regressor and classifier models, while the second study introduced a novel approach combining a nine-layer MLP

with a KNN imputer to address null values. Both studies demonstrated promising results in accurately predicting and forecasting WQ, contributing to the field of WQ management.

To enhance WQ forecasting with non-point source (NPS) pollution, a novel SOD-VGG-LSTM DL model was developed by Wan, et al. [14]. Combining the SOD module based on physical processes, VGG module capturing spatial characteristics, and LSTM module utilizing DL, it overcomes the limitations of existing models. Applied to the Lijiang River watershed, it outperformed mechanism models and LSTM in extreme value prediction. The evaluation of the prediction model revealed maximum relative errors of 8.47%, 19.76%, 24.1%, and 35.4% for the parameters DO, CODMn, NH3-N, and TP, respectively. In comparison to alternative methods such as ARIMA, SVR, and RNN, the SOD-VGG-LSTM model displayed superior performance, achieving an R^2 that was 3.2-39.3% higher. As a result, the SOD-VGG-LSTM model presents a promising approach for accurately predicting WQ affected by nonpoint source (NPS) pollution.

In their research, Shah et al. [1] proposed a novel framework that leverages particle swarm optimization (PSO) to optimize the hyperparameters of feed forward neural network (FFNN) and gene expression programming (GEP) models. The primary objective of this framework was to improve the capability of FFNN and GEP models by optimizing their hyperparameters. The optimized models, namely PSO-FFNN and PSO-GEP, were then employed to predict the levels of dissolved oxygen (DO) and total dissolved solids (TDS) in the upper Indus River, utilizing a consistent 30-year dataset.

To recognize the influential input parameters for accurate DO and TDS prediction, principal component analysis (PCA) was employed. The capability of the models was evaluated using five statistical evaluation techniques. The results demonstrated the effectiveness of the PSO algorithm in optimizing the models. Notably, the hybrid PSO-GEP model exhibited superior accuracy, achieving an R value above 0.85, a performance index close to 1, and an RMSE below 3 mg/l. External validation confirmed the generalizability of the models, and cross-validation yielded the best statistical metrics, with an R value of 0.87 and an RMSE of 2.67 for PSO-FFNN, and an R value of 0.895 and an RMSE of 2.21 for PSO-GEP.

This study highlights the potential of leveraging artificial intelligence (AI) algorithms with optimization routines for accurate forecasting of WQ. The proposed framework, incorporating PSO optimization with FFNN and GEP models, showcases promising results in predicting DO and TDS concentrations in the upper Indus River.

The primary focus of this study by Prasad, et al. [15] is to develop a water quality forecasting model that utilizes reliable and accurate data. As the production of big data from IoT-based smart WQ monitoring systems continues to increase, the complexity of WQ data has become more pronounced. To address this complexity, an advanced DL theory was employed, capitalizing on the effectiveness of LSTM DNNs in predicting time-series. This led to the development of a model specifically tailored for predicting drinking WQ. The capability of the model was assessed by utilizing data from the Guazhou Water Source of the Yangtze River in Yangzhou, spanning from January 2016 to June 2018. The findings illustrate the model's precise prediction of WQ trends, thus confirming the practicality and efficacy of LSTM DNNs in forecasting drinking WQ (word count) of this Khan et al. [16] present a WQ forecasting model in their paper, which utilizes the principal component regression technique. This model involves three main steps: calculating the WQI using the weighted arithmetic index method, performing principal component analysis (PCA) on the dataset to extract influential WQI parameters, and using the PCA output along with different regression algorithms for WQI prediction. Finally, the WQ status is classified using the Gradient Boosting Classifier. The Gulshan Lake-related dataset is utilized to experimentally evaluate the proposed system. As a result of the analysis, the principal component regression method achieved a prediction accuracy of 95%, while the GBC method exhibited flawless classification accuracy of 100%. These results establish the model's commendable performance, outperforming contemporary models in the field.

This paper proposed by Raheja, et al. [17] evaluates the capability of three ML algorithms, namely GBM, DNN and XGBoost, in forecasting groundwater indices in Haryana state, India. The study utilizes two WQ indices, Entropy WQ Index (EWQI) and WQI. Results indicate that DNN outperforms the other models, exhibiting lower error values in predicting both EWQI and WQI. For EWQI, DNN achieves a Correlation Coefficient (CC) of 0.989, RMSE of 0.037, Nash–Sutcliffe efficiency (NSE) of 0.995, and Index of agreement (d) of 0.999. For WQI, CC is 0.975, RMSE is 0.055, NSE is 0.991, and d is 0.998. Electrical conductivity (EC) is identified as the most significant input parameter, while pH has the least significance in predicting both indices. These findings can aid in accurately predicting groundwater quality for potability assessments.

In the study conducted by Venkataramana [18], various deep learning (DL) models were explored to forecast the WQI and water quality category (WQC) as indicators of WQ. Water samples from Korattur Lake in Chennai were collected and analyzed for several parameters, including pH, total dissolved salts, turbidity, phosphate, nitrate, iron, chemical oxygen demand, chloride, and sodium. DL models, such ANN, LSTM, and recurrent neural network (RNN), were trained and evaluated for both binary and multi-class classification tasks. Among the DL models, LSTM exhibited the highest accuracy, reaching approximately 94%, while also demonstrating the shortest execution time compared to other DL models.

3. RESEARCH METHODOLOGY AND APPROACH

A. Back Ground of the Research Study

This research utilized the Weka platform, an open-source ML software, to compare and evaluate different classification algorithms for predicting WQ. Six distinct machine learning techniques, including the Bagging classifier, J48 trees, RF, IBk, LR and AdaBoostM1, were employed to analyze the dataset. The study followed the structured Cross-Industry Standard Process for Data Mining (CRISP-DM), which consists of six phases. This systematic approach ensured a comprehensive and well-organized research process for WQ prediction. By adhering to the CRISP-DM methodology, the study-maintained rigor and reliability, thereby enhancing the validity of the research findings [19]. Refer to Figure 1 for an illustration of the stages in the CRISP-DM Methodology.

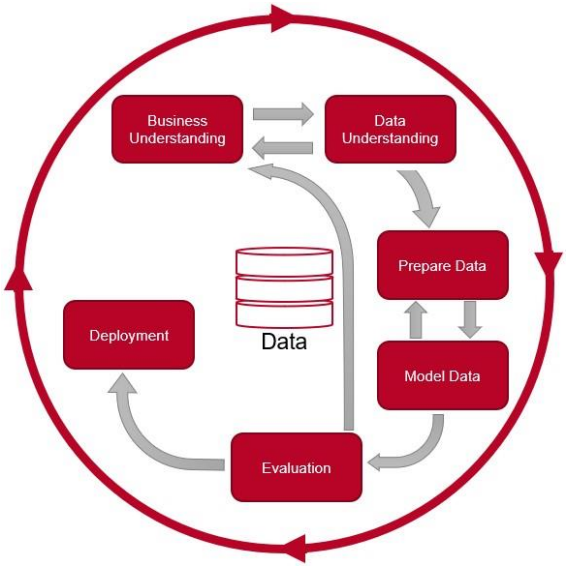

Figure 1. CRISP-DM Methodology stages.

B. Data Set Description

The dataset utilized in this study comprises various parameters associated WQ in India, collected between 2012 and 2021. These parameters were obtained from an official website associated with the Government of India [20]. The dataset consists of 7339 instances, each characterized by six attributes and a single outcome. Out of approximately 12,000 instances, 7339 were selected as non-null instances. Table 1 provides a definition of the attributes, and Figure 2 illustrates the balanced distribution of potable water (1) and non-potable water (0) within the dataset. Among the attributes, DO signifies the level of free and non-compound oxygen within the water, playing a crucial role in assessing WQ. pH indicates the presence of acidic or basic compounds, while EC measures the water's ability to conduct electricity. Biochemical Oxygen Demand (BOD) quantifies the oxygen consumption by aerobic microorganisms during organic matter decomposition, reflecting the impact of wastewater discharge on the recipient area. Elevated BOD values suggest a higher availability of organic compounds for oxygen-consuming bacteria. Nitrate (NA) is formed by the combination of oxygen or ozone with nitrogen, and while nitrogen is beneficial to living organisms, excessive NA levels can be harmful. Total Coliform (TC) serves as an indicator of bacteria found in animals, including humans. Although coliform itself does not cause diseases, certain types like E. coli can pose significant health risks [21].

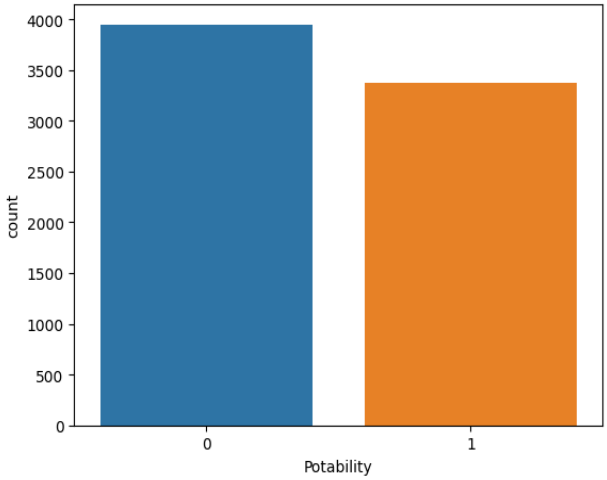

Figure 2. CRISP-DM Methodology stages.

C. Classification of Water Potability Using WQI

To determine the potability of water based on its quality, several important factors are considered, including pH value, EC, DO, total coliforms (TC), and BOD. In order to calculate the WQI, the data is utilized to derive new parameters such as npH, ndo, nco, nbdo, nec, and nna, which are obtained from the original measurements (pH, EC, DO, BOD, NA, and TC) using the classification provided in [22]. These newly derived parameters are then used to compute weighted averages for pH (wph), DO (wdo), BOD (wbdo), EC (wec), NA (wna), and TC (wco), as described by the formulas (1) to (6) elucidated in [21]. The WQI is subsequently calculated using a predefined formula (7), as outlined in the procedure detailed by [22].

$$wph = nph \times 0.165 \quad (1)$$

$$wdo = ndo \times 0.281 \quad (2)$$

$$wndo = nbdo \times 0.234 \quad (3)$$

$$wec = nec \times 0.009 \quad (4)$$

$$wna = nna \times 0.028 \quad (5)$$

$$wco = nco \times 0.281 \quad (6)$$

$$WQI = wph + wdo + wbdo + wec + wna + wco \quad (7)$$

Based on the resulting WQI value, the water sample is classified as potable (1) if the WQI exceeds 75, or non-potable (0) if the WQI is less than 75 [23]. This classification method allows for an assessment of WQ based on measured concentrations and corresponding criteria.

D. Correlation Matrix

The heatmap correlation matrix provides valuable understanding of the correlation between the six chosen features and the quality of water in terms of its potability (Figure 3). Examination of the analysis reveals that DO and BOD exhibit the most significant predictive power [24]. DO shows a positive correlation of 25%, indicating that as the level of DO increases, the probability of water being potable also rises. Conversely, BOD demonstrates a moderate negative correlation of -18%, suggesting that higher BOD levels are associated with decreased potability due to elevated organic matter content.

The feature pH shows a weak positive correlation of 9% with water potability. This implies a slight inclination for water potability to increase as pH levels rise. Alternatively, the features EC, Na, and TC display weaker relationships with water potability. EC exhibits a small negative correlation coefficient of -7%, indicating that higher EC levels are marginally associated with reduced potability. Similarly, Na and TC demonstrate small negative correlations of -1% and -3% respectively, implying that higher concentrations of nitrate and TC are weakly linked to a slight decrease in water potability.

It is important to note that while DO and BOD exhibit stronger correlations with water potability, the relationships for pH, EC, Na, and TC are comparatively weaker. These findings highlight the varying degrees of influence that different WQ parameters have on the potability assessment. The positive correlations of DO and the negative correlation of BOD underscore their importance in determining water potability due to their direct impact on oxygen levels and organic matter decomposition.

Overall, the examination of the correlation matrix reveals valuable understanding regarding the connections between the chosen features and the potability of water. DO and BOD emerge as the most influential factors, while pH, EC, NA, and TC exhibit weaker associations. These findings contribute to a deeper understanding of the key parameters affecting WQ and can aid in developing more accurate predictive models for assessing water potability.

TABLE I. DATASET DESCRIPTION

Attribute	Definistion	Datatypes
Dissolved Oxygen (DO)	The dataset includes measurements of the concentration of dissolved oxygen (DO) in water at different time points. The desired or ideal level of DO is considered to be 10 mg/L	float64
pH	The dataset encompasses the variation in hydrogen ion concentrations within water across different time periods. The optimal benchmark for pH, representing the hydrogen ion concentration, is identified as 8.5.	float64
Conductivity (EC)	The dataset encompasses the temporal changes in water conductivity measurements (EC). The preferred or desired value for conductivity is established at 1,000 µS/cm.	float64
BOD	The dataset consists of the temporal measurements of the Biological Oxygen Demand (BOD) in water. The target or preferred value for BOD is set at 5 mg/L.	float64
Nitrate (NA)	The dataset comprises the temporal observations of the nitrate content (NA) in water. The desired or ideal value for nitrate is identified as 45 mg/L.	float64
Total coliform (TC)	The dataset includes the comprehensive quantification of coliform bacteria in water (TC). The target or optimal value for the total coliform count is established at 100 per 100 mL.	float64
Potability	The dataset provides an indication of whether the water is suitable for human consumption. A value of 1 represents potable water, indicating it is safe for human use, while a value of 0 signifies non-potable water, suggesting it is not safe for human consumption.	object

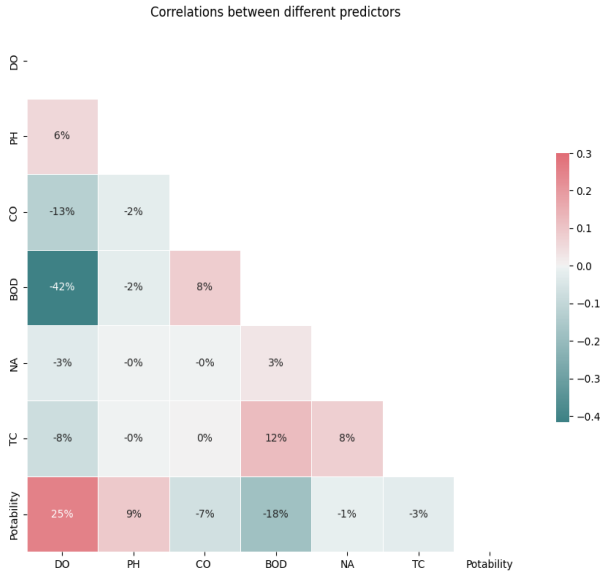

Figure 3. Heatmap Correlation Matrix.

E. Data Preparation.

Following the data exploration stage, the data preparation phase commenced, involving various procedures to address missing data, perform data scaling, encode categorical variables, and split the dataset. These processes were

implemented to ensure data integrity and optimize the dataset for subsequent analysis.

- Missing Data

  The 7339 instances used were chosen out of around 12000 instances.

- Data Scalarization

  By utilizing the MinmaxScaler function, the data is rescaled to a specified range, typically between 0 and 1. This scaling process ensures that the values are uniformly transformed while preserving the original distribution shape. The MinmaxScaler function effectively adjusts the values without distorting the inherent characteristics of the data.

- Encoding Categorical Data

  NominalToBinary filter was implemented to convert potable class to 1 and not potable class to 0 on the class variable.

- Splitting Data

  To ensure accurate evaluation and mitigate bias, a 10-fold cross-validation technique was implemented in this study to partition the data. This methodology enabled a rigorous assessment of the classification algorithms used for the forecasting of the quality of water, thereby boosting the credibility and validity of the research outcomes.

F. Modelling

The six ML algorithms Bagging classifier, LR, J48, RF, IBk, and AdaBoostM1 are implemented to evaluate their performance in WQ prediction.

Bagging classifier is a widely adopted ensemble learning technique designed to enhance prediction accuracy and reduce variance. It achieves this by combining multiple base classifiers, each trained on a subset of the training data that is distinct and obtained through bootstrap sampling. The final prediction is determined through a majority voting mechanism. Due to its effectiveness in dealing with complex classification tasks and improving overall model performance, the Bagging classifier has found extensive application in diverse domains, including WQ prediction [25].

Logistic regression classifier is a widely utilized statistical modeling approach that is particularly suitable for binary classification tasks. It estimates the probability of an event occurrence by utilizing a logistic function and considering input variables. LR is appreciated for its interpretability, simplicity, and computational efficiency. In domains such as WQ prediction and others, LR has been extensively applied due to its capability to model the relation between predictors and the probability of a specific outcome [26].

J48 tree, also referred to as C4.5, is a popular decision tree algorithm employed for classification purposes. It creates a tree structure by recursively dividing the data based on attribute values, with the objective of minimizing entropy or maximizing information gain at each node. The resulting tree is easily interpretable and can be used for making predictions. J48 is widely utilized in various fields, including WQ prediction, due to its effectiveness, simplicity, and capacity to handle both categorical and numerical attributes [27] .

Random Forest classifier is a powerful ensemble learning technique that improves classification accuracy and mitigates overfitting by combining multiple decision trees. By randomly selecting subsets of features and instances for each tree, it ensures diversity within the ensemble. The RF demonstrates robustness and can effectively handle high-dimensional data. Its ability to capture complex relationships and provide reliable classification results has made it widely adopted across diverse domains, including WQ prediction [4].

IBk classifier, a non-parametric algorithm also referred to as k-Nearest Neighbors (k-NN), is extensively employed for classification purposes. It assigns a class to an instance by considering the largest count of votes of its k nearest neighbors within the training data. IBk is renowned for its simplicity and versatility in handling both numerical and categorical data. It has been successfully utilized in diverse domains, including WQ prediction, where instance-based learning methods are particularly effective [28].

AdaBoostM1 classifier is an ensemble learning technique that constructs a powerful classifier by combining multiple weak classifiers. It assigns weights to training instances, placing greater emphasis on misclassified instances during subsequent iterations. AdaBoostM1 excels in handling intricate classification tasks and enhancing overall accuracy. Its effectiveness in improving the capability of weak classifiers has led to extensive application in various domains, including WQ prediction [29].

G. Performance Evaluation

The capability assessment of six ML algorithms was conducted based on the following metric parameters:

Accuracy is a widely employed metric in artificial intelligence (AI) that evaluates the capability and efficacy of an ML model. It assesses the model's ability to correctly predict or classify data. Accuracy is determined by dividing the count of correctly predicted instances by the total count of instances in the dataset. It indicates the ratio of correct predictions made by the model to the total number of predictions. A high accuracy score suggests that the model is making precise predictions, whereas a low score indicates lower reliability. However, it is important to note that accuracy alone might not provide a comprehensive assessment of a model's performance, particularly when dealing with imbalanced datasets or when certain types of

errors are more significant than others. Hence, it is often necessary to consider additional evaluation metrics and contextual factors to obtain a holistic understanding of a model's performance [29].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Precision is a performance measure that evaluates the ratio of accurately forecasted instances that are positive to the total instances forecasted as positive. It assesses the accuracy of positive predictions, focusing on the precision of identifying true positives while disregarding false positives. This metric quantifies the effectiveness of correctly identifying positive cases, without considering the number of false positives [29].

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

The False Positive rate is a performance measure that calculates the ratio of incorrectly classified instances as positive, which are negative, to the overall count of true negative instances. It quantifies the classifier's propensity to erroneously predict negative instances as positive. This metric provides insight into the rate of FP predictions, highlighting the classifier's tendency to make such incorrect classifications [28].

$$FPR = \frac{FP}{FP+TN} \quad (10)$$

Recall metric, also known as sensitivity, evaluates the classifier's ability to correctly identify positive instances by measuring the ratio of accurately classified positive instances to the total number of actual positive instances. This performance measure provides valuable insights into the classifier's effectiveness in detecting and capturing positive cases from the available data [28].

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

F-Measure is an evaluation metric that combines precision and recall providing a comprehensive assessment of a classifier's effectiveness. By calculating the harmonic mean, it achieves a balance between accurately identifying positive instances and minimizing false positives. The F-Measure offers a holistic evaluation of the classifier's performance by considering both precision and recall simultaneously. This unified measure provides insights into the classifier's overall accuracy and its ability to effectively manage the trade-off between true positives and false positives [29].

$$F - Measure = \frac{2TP}{2TP+FP+FN} \quad (12)$$

ROC area is a performance measure that assesses the discriminative power of a classifier. The classifier's quantification provides an estimate of the likelihood that a positive instance will have a higher rank than a negative instance. The ROC area summarizes the classifier's performance across different classification thresholds, providing a comprehensive indication of its discriminative ability. It serves as a valuable metric for evaluating the overall performance and ranking capabilities of the classifier [12].

PRC area, also known as the area under the precision-recall curve, is a measure of the model's effectiveness or capability that evaluates the classifier's effectiveness in balancing precision and recall. It quantifies the overall performance of the classifier by calculating the integral of precision as recall varies. This metric reflects the classifier's ability to provide accurate positive predictions across different levels of recall. The PRC area offers valuable insights into the classifier's ability to strike a balance between precision and recall, providing an assessment of its performance and the quality of its positive predictions [17].

4. RESULTS AND DISCUSSION

The evaluation of the capability of the algorithms with respect to multiple metrics, including accuracy, precision, TP rate, F-measure, ROC area, FP rate, and PRC area, was utilized to assess the effectiveness of these algorithms in WQ classification tasks. Table II presents the metrics-based results attained by each model.

TABLE II. DATASET DESCRIPTION

	Accuracy	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Bagging	0.992	0.007	0.992	0.992	0.992	0.999	0.999
LR	0.958	0.041	0.958	0.958	0.958	0.991	0.999
J48	0.993	0.007	0.993	0.993	0.993	0.994	0.992
RF	0.993	0.006	0.993	0.993	0.993	1	1
IBk	0.714	0.289	0.714	0.714	0.714	0.78	0.78
AdaBoostM1	0.971	0.026	0.972	0.971	0.971	0.992	0.991

The comparison of ML algorithms based on the accuracy metric showed that the RF algorithm and J48 achieved the highest accuracy value of 0.993. The Bagging classifier closely followed with an accuracy of 0.992, and the AdaBoostM1 algorithm obtained an accuracy of 0.971. The LR algorithm obtained an accuracy of 0.958, followed by the IBK algorithm with an accuracy of 0.714. The results are presented in Figure 4.

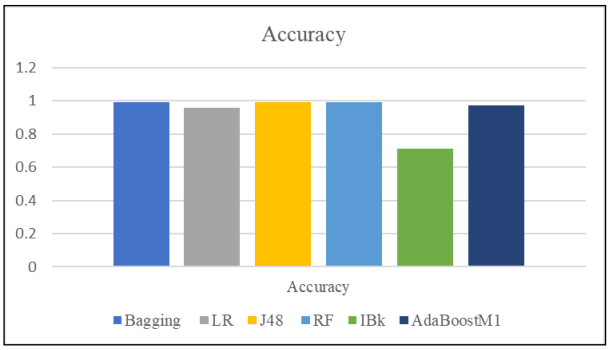

Figure 4. Accuracy plot for the proposed Algorithms

In terms of FP rate metric, the comparison of ML algorithms demonstrated that RF achieved the lowest rate of 0.006. The Bagging classifier and J48 closely followed with an FP rate of 0.007 each. On the other hand, AdaBoostM1, LR, and IBk exhibited higher rates of 0.026, 0.041, and 0.289, respectively. These results are visually presented in Figure 5, providing a clear overview of the varying FP rates across the different algorithms.

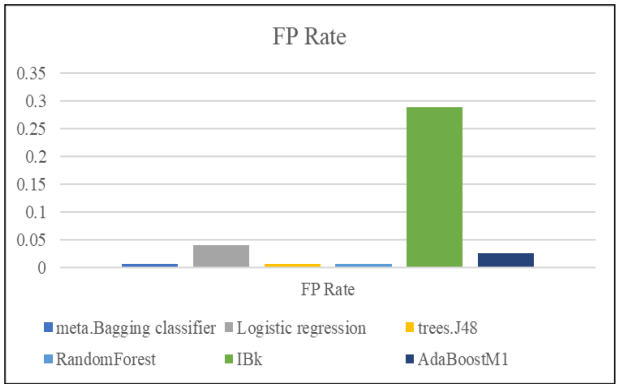

Figure 5. FP Rate plot for the proposed Algorithms

When comparing the precision metrics of various ML algorithms, it was observed that RF and J48 exhibited the highest precision rate of 0.993. Following closely behind was the Bagging classifier with a precision rate of 0.992. The AdaBoostM1 algorithm achieved a precision rate of 0.972, while LR showed a slightly lower precision rate of 0.958. In contrast, IBK displayed a lower precision rate of 0.714. Figure 6 visually represents these findings, highlighting the varying precision rates of the different algorithms.

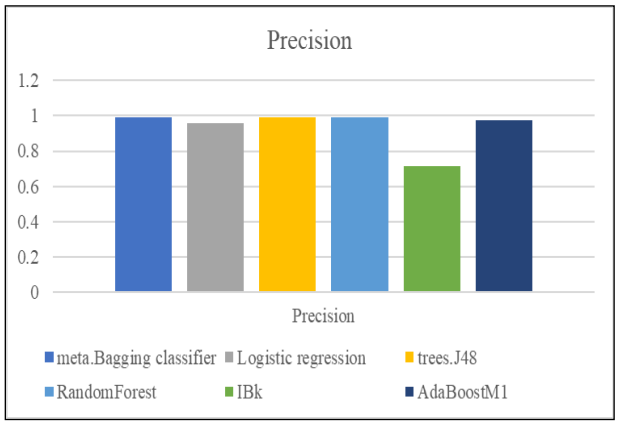

Figure 6. Precision plot for the proposed Algorithms

The comparison of ML algorithms for the recall metric demonstrated that the RF algorithm and J48 achieved the highest recall value of 0.993. The Bagging classifier closely followed with a recall value of 0.992, and the AdaBoostM1 algorithm obtained a recall value of 0.971. The LR algorithm obtained a recall value of 0.958, followed by the IBK algorithm with a value of 0.714. The results are presented in Figure 7.

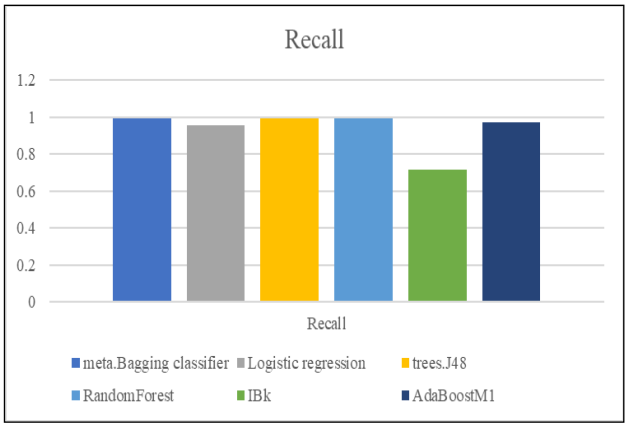

Figure 7. Recall plot for the proposed Algorithms

F-Measure metric comparison across ML algorithms revealed the RF algorithm and J48 achieved the highest F-measure value of 0.993. The Bagging classifier closely followed with a score of 0.992, and the AdaBoostM1 algorithm obtained a score of 0.971. The LR algorithm obtained a metric of 0.958, followed by the IBK algorithm with a metric of 0.714. The results are shown in Figure 8.

Figure 8. F-Measure plot of proposed Algorithms

Among the evaluated ML algorithms for the ROC area metric, RF expressed the leading capability with a value of 1, and Bagging classifier followed closely with ROC areas of 0.999. J48 followed with a value of 0.994, while AdaBoostM1 achieved 0.992. LR had a value for ROC area metric of 0.991, and IBK with 0.78, as illustrated in Figure 9.

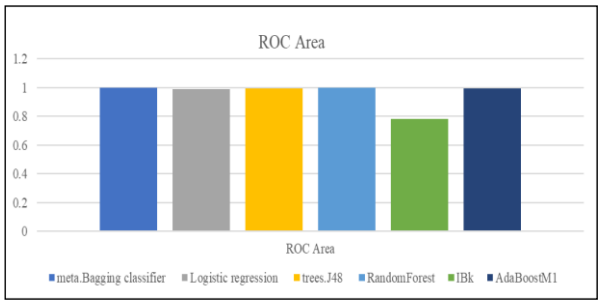

Figure 9. ROC Area plot for the proposed Algorithms

The comparison of ML algorithms for the PRC area metric revealed that RF attained the predominant result with an area of 1. Bagging and LR closely followed with an area of 0.999, while J48 obtained 0.992. AdaBoostM1 had a PRC area of 0.991, and IBK had the lowest value of 0.78. The results are displayed in Figure 10.

The high values achieved by RF, J48, and the Bagging classifier in predicting water potability can be attributed to several factors. RF's ensemble approach combines multiple decision trees, enabling it to capture complex relationships and handle high-dimensional data effectively. J48, a decision tree algorithm, offers simplicity and interpretability while handling diverse attribute types. The Bagging classifier, with its ensemble learning technique and majority voting, effectively handles diverse and potentially noisy data. On the other hand, the lower values obtained by IBk and LR can be attributed to their limitations in capturing complex relationships or handling certain data characteristics. IBk's reliance on local neighborhoods may not adequately address the complexity of water potability prediction, while LR's linear nature might struggle with capturing non-linear patterns. In contrast, It is worth mentioning that the selection of algorithms can also be influenced by factors such as specific dataset characteristics and the trade-off between interpretability and predictive performance.

The attribute evaluator used in this study was InfoGainAttributeEval, which was employed in combination with a Ranker. Its purpose was to rank the attributes according to their effectiveness within the model. The evaluated attributes, listed in order of effectiveness, were as follows: TC, BOD, pH, DO, EC, and NA.

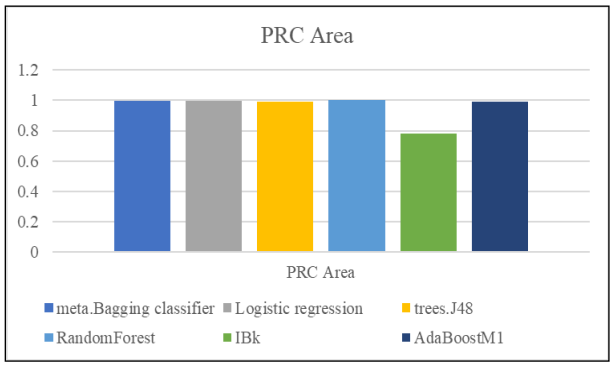

Figure 10. PRC Area plot for the proposed Algorithms

5. Conclusion and Future Direction

The aim of this comparative research analysis was to assess and compare the predictive capabilities of the Bagging classifier, LR, J48, RF, IBk, and AdaBoostM1 models. These models were implemented to evaluate their performance in the domain of WQ prediction, and the assessment included metrics such as precision, F-measure, TP rate, PRC area, FP rate, and ROC area. The findings indicate that the RF model outperformed the other models, demonstrating superior performance.

The J48 and RF algorithms have showcased superior performance with a very slight difference with the Bagging classifier. AdaBoostM1, LR, and IBk followed after that with a significant difference for IBK values.

The J48 model, a decision tree algorithm, holds superiority due to its versatility, interpretability, feature selection capability, robustness to missing values, and computational efficiency. It handles diverse datasets, provides understandable models, and performs well in real-time applications, making it a valuable choice for classification tasks.

RF models provide benefits including ensemble learning to mitigate overfitting, robustness to outliers and noise, feature importance estimation, non-parametric nature, and parallelization capabilities. These advantages contribute to the exemplary performance of J48 and RF models in the study.

The Bagging classifier offers advantages such as improved accuracy and robustness. By combining multiple models trained on subsets of the data, it reduces variance and enhances generalization. It handles complex datasets, provides reliable predictions, and is suitable for applications where accuracy and stability are crucial.

The lower metric values observed for IBK in predicting potable water using features such as pH, BOD, NA, DO, and others could be attributed to several factors. IBK's performance might be impacted by sensitivity to feature scaling, noisy or irrelevant features, imbalanced data, inappropriate choice of k value, and sensitivity to outliers. Inconsistent feature scaling, the presence of noisy or irrelevant features, imbalanced class distributions, suboptimal k values, and the influence of outliers could contribute to IBK's lower accuracy. Addressing these factors through preprocessing, feature selection, class balancing techniques, parameter tuning, and outlier detection could potentially improve IBK's performance in predicting potable water. Overall, the combination of the unique advantages of MLP and RF, such as their ability to handle non-linearity, capture complex patterns, and effectively handle noisy data, contributed to their superior performance in contrast with the different models explored.

In the field of WQ prediction, ML has faced several challenges. One of the main difficulties is the availability and quality of data. WQ datasets are often limited, incomplete, or contain missing values, which can hinder the training and performance of ML models. Another challenge is the inherent complexity of water systems, where numerous factors and interdependencies influence WQ. Developing accurate models that can capture and effectively model these complex relationships presents a significant challenge. Furthermore, the dynamic nature of water systems, influenced by environmental changes and human activities, requires models that can adapt and generalize well to new conditions. Addressing these challenges requires careful data collection, preprocessing techniques, feature engineering, and the development of robust and adaptable ML algorithms tailored to the specific characteristics of WQ prediction tasks.

For subsequent research in the forecasting of WQ sector, three recommendations can be suggested:

- Utilizing other datasets: To obtain more comprehensive and robust results, it is of utmost importance to explore and utilize additional datasets. This approach helps avoid potential biases that may have existed in the dataset used in the current study. Incorporating diverse datasets from different regions, sources, and time periods can provide a broader understanding of WQ patterns and enhance the generalizability of the findings.
- Explore more NN models: Since only NN model was used in the current study, it would be beneficial to explore and compare the capability of different NN architectures. Variations such as CNNs, recurrent neural networks (RNNs), and more advanced architectures like LSTM networks or transformer-based models could be considered. This exploration would help identify the most suitable NN model for WQ prediction tasks and potentially uncover new insights.
- Investigate hybrid models: Hybrid models bring together the strengths and powerful aspects of multiple models and can potentially improve prediction accuracy. Future research could focus on developing and evaluating hybrid models that integrate different ML techniques, such as combining NNs with ensemble methods like RF or gradient boosting. By leveraging the strengths of various models, hybrid approaches can potentially enhance the predictive power and robustness of WQ prediction models.
- Explore different data preprocessing techniques and methodologies: In future research, it is important to investigate and compare various data preprocessing techniques and methodologies for WQ prediction. Different approaches such as feature scaling, outlier detection and handling, missing data imputation, and feature selection can lead to a tremendous uprising in the model's performance. Exploring different preprocessing techniques and methodologies, such as time series analysis, spatial interpolation, or domain-specific preprocessing methods, can help uncover hidden patterns, reduce noise, and enhance the quality of input data for ML models. This exploration will contribute to improving accuracy; in addition to, the reliability of WQ prediction algorithms.

By incorporating these recommendations into future research endeavors, researchers can advance the field of WQ prediction by utilizing diverse datasets, exploring a wider range of NN models, investigating the potential of hybrid models, and exploring different data preprocessing techniques and methodologies. These efforts will contribute to enhancing the accuracy, reliability, and generalizability of WQ prediction systems, ultimately benefiting environmental management, public health, and decision-making in water resource management.

REFERENCES

[1] M. I. Shah, M. F. Javed, A. Alqahtani, and A. Aldrees, "Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data," *Process Safety Environmental Protection*, vol. 151, pp. 324-340, 2021.

[2] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, p. 2210, 2019.

[3] M. Azrour, J. Mabrouki, G. Fattah, A. Guezzaz, and F. Aziz, "Machine learning algorithms for efficient water quality prediction,"

Modeling Earth Systems Environment, vol. 8, no. 2, pp. 2793-2801, 2022.

[4] A. Juna *et al.*, "Water quality prediction using KNN imputer and multilayer perceptron," *Water*, vol. 14, no. 17, p. 2592, 2022.

[5] J. Bi, Y. Lin, Q. Dong, H. Yuan, and M. Zhou, "Large-scale water quality prediction with integrated deep neural network," *Information Sciences*, vol. 571, pp. 191-205, 2021.

[6] K. Chen *et al.*, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," *Water research*, vol. 171, p. 115454, 2020.

[7] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, p. 126169, 2020.

[8] Y. Wang, J. Zhou, K. Chen, Y. Wang, and L. Liu, "Water quality prediction method based on LSTM neural network," in *2017 12th international conference on intelligent systems and knowledge engineering (ISKE)*, 2017, pp. 1-5: IEEE.

[9] T. Deng, K.-W. Chau, and H.-F. Duan, "Machine learning based marine water quality prediction for coastal hydro-environment management," *Journal of Environmental Management*, vol. 284, p. 112051, 2021.

[10] K. Chen *et al.*, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," vol. 171, p. 115454, 2020.

[11] J.-W. Yu, J.-S. Kim, X. Li, Y.-C. Jong, K.-H. Kim, and G.-I. Ryang, "Water quality forecasting based on data decomposition, fuzzy clustering and deep learning neural network," *Environmental Pollution*, vol. 303, p. 119136, 2022.

[12] K. R. A. Haq and V. Harigovindan, "Water quality prediction for smart aquaculture using hybrid deep learning models," *IEEE Access*, vol. 10, pp. 60078-60098, 2022.

[13] J. P. Nair and M. Vijaya, "River water quality prediction and index classification using machine learning," in *Journal of Physics: Conference Series*, 2022, vol. 2325, no. 1, p. 012011: IOP Publishing.

[14] H. Wan, R. Xu, M. Zhang, Y. Cai, J. Li, and X. Shen, "A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods," *Journal of Hydrology*, vol. 612, p. 128081, 2022.

[15] D. V. V. Prasad *et al.*, "Analysis and prediction of water quality using deep learning and auto deep learning techniques," *Science of the Total Environment*, vol. 821, p. 153311, 2022.

[16] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University-Computer Information Sciences*, vol. 34, no. 8, pp. 4773-4781, 2022.

[17] H. Raheja, A. Goel, and M. Pal, "Prediction of groundwater quality indices using machine learning algorithms," *Water Practice Technology*, vol. 17, no. 1, pp. 336-351, 2022.

[18] Y. Venkataramana, "Water quality analysis in a lake using deep learning methodology: prediction and validation," *International Journal of Environmental Analytical Chemistry*, vol. 102, no. 17, pp. 5641-5656, 2022.

[19] M. G. Uddin, S. Nash, A. Rahman, and A. I. J. W. R. Olbert, "A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment," vol. 219, p. 118532, 2022.

[20] Central Pollution Control Board. (2023). NWMP Data 2013. Retrieved from https://cpcb.nic.in/nwmp-data-2013/

[21] Nayan, A. A., Mozumder, A. N., Saha, J., Mahmud, K. R., Azad, A. K. A., & Kibria, M. G. (2021). A machine learning approach for early detection of fish diseases by analyzing water quality. *arXiv preprint arXiv:2102.09390*.

[22] Rustam, F., Ishaq, A., Kokab, S. T., de la Torre Diez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). An Artificial Neural Network Model for Water Quality and Water Consumption Prediction. *Water*, *14*(21), 3359.

[23] Wong, C., & Rylko, M. (2014). Health of the Salish Sea as measured using transboundary ecosystem indicators. *Aquatic Ecosystem Health & Management*, *17*(4), 463-471.

[24] M. U. Maheswari, R. Sudharsanan, M. Arthy, A. Jenefer, L. Oormila, and V. S. Pandi, "Efficient Drinking Water Quality Analysis using Machine Learning Model with Hyper-Parameter Tuning," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2023, pp. 401-406: IEEE.

[25] W. Y. Wong *et al.*, "A Stacked Ensemble Deep Learning Approach for Imbalanced Multi-Class Water Quality Index Prediction," *Computers, Materials Continua*, vol. 76, no. 2, 2023.

[26] M. Koranga, P. Pant, T. Kumar, D. Pant, A. K. Bhatt, and R. Pant, "Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand," *Materials today: proceedings*, vol. 57, pp. 1706-1712, 2022.

[27] J. L. Lerios and M. V. Villarica, "Pattern extraction of water quality prediction using machine learning algorithms of water reservoir," *International Journal of Mechanical Engineering Robotics Research*, vol. 8, no. 6, pp. 992-997, 2019.

[28] A. Sharma, N. Hooda, N. R. Gupta, and R. Sharma, "Efficient RIEV: a novel framework for the prediction of breast cancer cases using ensemble machine learning," *Network Modeling Analysis in Health Informatics Bioinformatics*, vol. 12, no. 1, p. 29, 2023.

[29] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," *Computation*, vol. 11, no. 2, p. 16, 2023.

Fuad Ahmad Musleh, Assistant Professor at the University of Bahrain. Ph.D. and master's degree from the University of Alabama in Huntsville, B.Sc. from Jordan University of Science and Technology in Jordan. Interested in research related to flow through vegetation, water and environmental conservation.