A Comprehensive Machine Learning Approach for Water Quality Prediction with the H2O Framework

Shaymaa Mahmoud[1,*]
Sama Gouda[1,*]
Mona Ibrahim[1,*]
Sally Elghamrawy[2]

[1]Department of Mathematics and Actuarial Science, American University in Cairo
[2]Department of Computer Science and Engineering, American University in Cairo

Sally_elghamrawy@ieee.org,
[shaymaa123,samaamr,monamahmoud]@aucegypt.edu

April 18, 2025

Abstract

This paper explores the use of machine learning (ML) and deep learning (DL) methods to predict water quality, focusing on classifying water as potable or non-potable. Using a comprehensive dataset, we pre-process and analyze data with Big Data tools such as Apache Spark and H2O. Our framework integrates ML and DL models to improve prediction accuracy, addressing challenges such as data imbalance and missing values. This research aims to improve the monitoring of water quality and environmental sustainability.

[1]The authors marked with an asterisk (*) contributed equally to this work.

1 Introduction

Water is one of the most essential resources for sustaining life, economic activities, and ecosystem balance. The degradation of water quality has become a major global environmental issue that affects the stability of ecosystems, human health, and economic growth [2]. The management of water resources and the preservation of the environment depend on the efficient identification and control of the impact factors on water quality. Potable water refers to water suitable for human consumption. Due to climate change, rapid urbanization, rising food demand, and the unsustainable exploitation of natural resources, nearly 40% of the global population now faces water scarcity and are unable to access potable water [10]. Contaminants such as heavy metals, nitrates, phosphates, and microbial pollutants pose severe health and environmental risks [11]. As a result, accurate and timely water quality prediction has become a critical area of research to support sustainable water management, pollution control, and public health safety. According to Wu et al [13], the current research is roughly divided into three types: traditional statistical method prediction, artificial intelligence method prediction and hybrid model method prediction. Traditional models rely on statistical analysis and deterministic equations, but they often struggle with the nonlinear and complex nature of water pollution patterns. In contrast, artificial intelligence (AI) techniques, including machine learning (ML) and deep learning (DL), have shown great promise in capturing intricate relationships between various water quality indicators. Hybrid models aim to leverage the strengths of both approaches, improving predictive accuracy and generalizability. This paper explores the application of big data and machine learning methodologies in predicting water quality. By analyzing existing literature and datasets, we aim to assess the effectiveness of different predictive models and highlight the advantages and limitations of each approach. Through this research, we seek to contribute to the development of more accurate and efficient water quality monitoring systems, ultimately promoting public health and environmental sustainability.

2 Problem statement

Water quality is one of the most pressing global issues, with implications for human health, environmental sustainability. Access to clean and safe drinking water is a fundamental human right, yet millions of people worldwide are exposed to polluted water sources, which causes severe health problems such as malnutrition, and even death(Nayan, et.al,2020). Additionally, poor water quality can significantly impact the aquatic ecosystems, disrupting agricultural productivity.

Most of the previous traditional and statistical methods that have been used for water quality prediction problems include many flaws(_). Additionally, employing machine and deep learning methods for this problem will be effective and provide accurate prediction given the complexity of the model. ML methods have been tested in different datasets varying in parameters and size. In this project, we aim to address this critical issue by leveraging machine learning techniques in addition to the Big Data framework tools such as Apache Spark and H2O to manage and preprocess the dataset in order to use

it to classify water quality as either potable (1) or non-potable (0) based on a dataset containing various water quality parameters. By developing an accurate classification model, we can assist in monitoring water quality, identifying contamination sources, and implementing timely interventions to ensure the safety of drinking water. This project has the potential to contribute significantly to public health, environmental conservation, and sustainable development.

3 Literature Review

3.1 DEEP LEARNING (DL) MODELS

Yao et. al [15]have predicted long-term water quality using deep learning models and they've also integrated the (WQI) water quality index (Refer to Appendix 9 for more details) into their study conducted on the water of Chaohu Lake, China. The WQI was then used to categorize the data into 5 categories - Excellent, Good, Medium, Average, and poor- as follows in *Figure 1*:

Comprehensive Water Quality Index Interval	Water Quality Conditions
91–100	Excellent
71–90	Good
51–70	Medium
26–50	Average
0–25	Poor

Figure 1: WQI Range

The dataset they were using consisted of daily monitored data from four sites in the Lake from the years 2019 to 2022. They were divided into 10 different groups where each group consisted of different combinations of predicted input variables, represented in *Figure 2*.

Groups	Predicted Input Variables
S1	DO
S2	DO NH_4^+
S3	DO TUR
S4	DO COD
S5	DO COD NH_4^+
S6	DO COD NH_4^+ TN
S7	DO COD NH_4^+ TUR
S8	DO COD NH_4^+ TN TUR
S9	DO COD NH_4^+ TN TUR TP
S10	DO COD NH_4^+ TN TUR TP T

Figure 2: Grouping of water quality variables, dissolved oxygen (DO),ammonia nitrogen (NH4-N), CODMn, turbidity (TUR), total phosphorus (TP).

They used the long sequence time series forecasting method (LSTF) to predict the selected water quality parameter WQI. Then to model the data they compared Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), and proposed Transformer-based models (Informer and FEDformer) which uses both the LSTF and model, their architecture is represented in *Figure 3*.

To evaluate these models they mainly used the Mean Squared Error (MSE) and Mean Absolute Error (MAE). The best model to perform was the Informer model with the best prediction achieved in the case of group 8 with a MSE = 0.2455, and MAE = 0.2449.

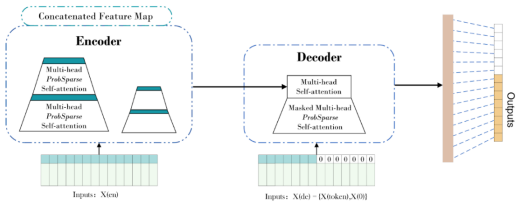

Figure 3: Informer's model architecture

Thus when ordering models according to their adaptability to improve the accuracy of WQI prediction in solving the LSTF Chaohu Lake water quality prediction problem, the model affects Informer >FEDformer >MLP >Transformer>RNN >LSTM. *Figure 4* represents the results of the different models.

Models		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
MLP	MSE	0.8678	0.5191	0.8721	0.6705	0.4562	0.5002	0.4569	0.3605	0.3427	0.3319
-	MAE	0.6785	0.5881	0.6776	0.5124	0.3529	0.3961	0.3676	0.3675	0.3555	0.3523
RNN	MSE	0.6441	0.6264	0.6184	0.6005	0.6157	0.6263	0.6168	0.5836	0.5794	0.5821
-	MAE	0.6168	0.6352	0.6142	0.6007	0.6131	0.6294	0.6147	0.5957	0.5863	0.5981
LSTM	MSE	0.7075	0.6247	0.5942	0.6537	0.6764	0.7175	0.6589	0.6533	0.6247	0.5914
-	MAE	0.6552	0.5949	0.5954	0.6196	0.6261	0.6463	0.6152	0.6181	0.6114	0.6041
Transformer	MSE	0.3339	0.3431	0.3175	0.3687	0.3475	0.3669	0.3122	0.3304	0.3591	0.3673
-	MAE	0.4652	0.4686	0.4629	0.4865	0.4863	0.4852	0.4432	0.4598	0.4853	0.4915
FEDformer	MSE	0.4575	0.3867	0.4778	0.5031	0.5432	0.5182	0.5445	0.5272	0.5182	0.7165
-	MAE	0.5382	0.4919	0.5747	0.5789	0.6021	0.5724	0.5991	0.5901	0.5663	0.6914
Informer	MSE	0.3071	0.2601	0.2962	0.2951	0.3033	0.2416	0.2458	0.2455	0.2545	0.2685
-	MAE	0.3536	0.3902	0.3306	0.3455	0.3389	0.2921	0.2563	0.2449	0.2567	0.2615

Figure 4: results of 6 DL models for 10 combinations of input data.

This research by Yao et. al. [15] underscores the potential of different deep learning models, particularly, transformer-based models, in enhancing the accuracy and efficiency of long-term water quality predictions.

Y. Wang, J. Zhou, et.al (2017) addressed the challenge of time-series predictions in water quality by proposing a Long Short-Term Memory (LSTM) neural network. The study utilized raw data from Taihu Lake water quality indicators, specifically Dissolved Oxygen (DO) and Total Phosphorus (TP), collected monthly from 2000 to 2006, as the training set and optimized the LSTM model through a series of analyses and configuration choices. The LSTM model was compared with two other methods: a digital synchronous extreme machine learning system and a backpropagation neural network. The results showed that the LSTM model outperformed the other methods, capturing temporal dependencies in the data. This study highlights the potential of deep learning techniques, particularly LSTM, for water.

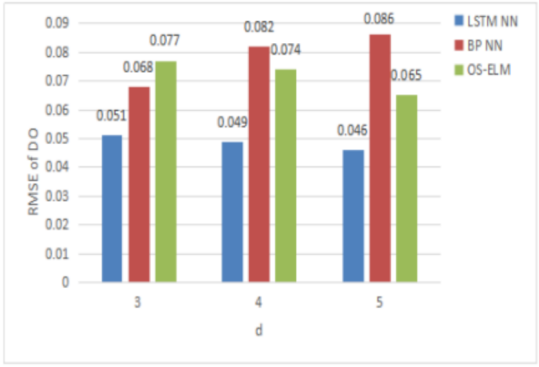

Figure 5: RMSE of Dissolved Oxygen across the models

The LSTM model was optimized through a series of analysis and configuration choices, achieving the best performance with a time step of 5 for DO (RMSE = 0.046) and a time step of 3 for TP (RMSE = 0.041). The LSTM model was compared with two other methods: a backpropagation neural network (BP NN) and an online sequential extreme learning machine (OS-ELM).

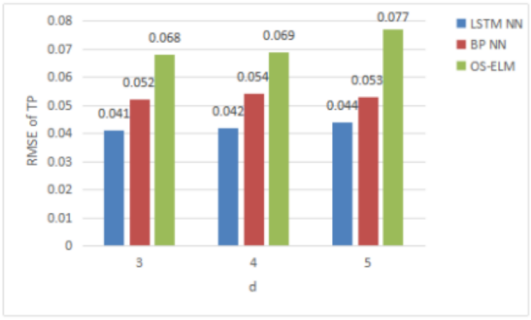

Figure 6: RMSE of Total Phosphorus across the models

A study Wu et. al [13] examined the Yellow River in Lanzhou, China to construct a water quality prediction model. It uses multi-task deep learning, taking the chemical oxygen demand (COD) of the water environment to capture relationships between multiple water quality indicators. The model is trained on a 36-month dataset from four monitoring sections of the Lanzhou portion of the Yellow River. The study focuses on key water quality parameters, such as chemical oxygen demand (COD), pH, dissolved oxygen (DO), ammonia nitrogen (NH_3-N), and total phosphorus (TP). The study empahsizes that different section of the river have similar trends in their water quality, and therefore these patterns and correlations between sections needs to be taken into account when creating a water quality prediction model. By leveraging correlations between four different monitoring sections, the multi-task mode enhances water quality prediction by sharing and learning from water quality information across multiple sections simultaneously. In addition, the model integrates deep convolutional neural networks (CNN), Long Short-Term Memory (LSTM), and multi-task learning (MTL) to achieve both information sharing and section-specific predictions.

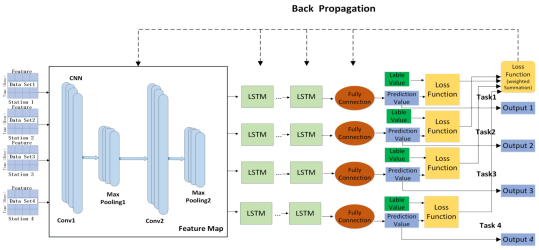

Figure 7: MTL-CNN-LSTM model architecture

This hybrid model effectively identifies nonlinear relationships in complex time-series data while improving predictive accuracy by enabling collaborative learning between different monitoring sections.

Compared to previous models that does not take into account the correlation between different sections of the river, this study was able to construct a model whose mean absolute error (MSE) and root mean square error (RMSE) of the predicted value of the model

Model Component	Kernel Size	Number of Convolution Kernels	Number of Parameters
Convolutional layer 1	7 × 1	6	42
Max Pooling layer 1	2 × 1		0
Convolutional layer 2	7 × 1	14	98
Max Pooling layer 2	2 × 1		0
Convolutional layer 3	7 × 1	28	420
Max Pooling layer 3	2 × 1		0
LSTM		50	15,800
Dense layer		1	51

Figure 8: MTL-CNN-LSTM model hyperparameter settings.

are decreased by 13.2% and 15.5% compared to other models in the literature. The solution in this study lacks hyperparameter tuning and it does not implement regularization techniques to avoid overfitting.

Model	Station									
	1		2		3		4		Mean	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MTL-CNN-LSTM	0.250	0.312	0.294	0.380	0.224	0.288	0.279	0.304	0.262	0.321
CNN-LSTM	0.342	0.421	0.417	0.479	0.235	0.296	0.214	0.323	0.302	0.380
LSTM	0.379	0.452	0.446	0.517	0.386	0.438	0.494	0.509	0.426	0.479
CNN	0.394	0.472	0.584	0.664	0.448	0.491	0.542	0.567	0.492	0.549

Figure 9: Prediction performance evaluation of different models across sections

Another interesting study by Najah et. al [9] uses artificial neural networks as it is capable of identifying complex non-linear relationships between input and output data when compared to other classical modeling techniques. The study was made on Johor River Basin located in Johor state, Malaysia, a river that is heavily affected by urbanization along the river. In this study, several methods were tested to model water quality, including linear regression models (LRM), multilayer perceptron neural networks (MLP-NN) and radial basis function neural networks (RBF-NN).

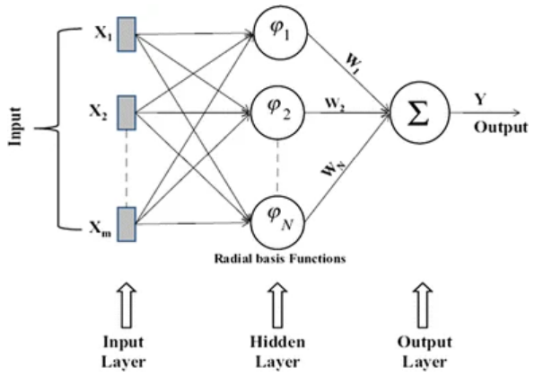

Figure 10: RBF-NN Architecture

The studied variables were Electrical conductivity (EC), total dissolved solids (T.D. solids), and turbidity. The results showed that the use of neural networks can describe the behavior of water quality parameters more accurately than linear regression models. When using the MLP-NN models, the study showed a slow convergence during training because they required a relatively large number of hidden neurons and hidden layers. On

the other hand, RBF-NN models find a solution faster than the MLP, require fewer computational complexities to train, and is the most accurate and reliable tool in terms of processing large amounts of non-linear, non-parametric data. In addition, the RBF-NN proved its potential to incorporate and mimic different stochastic patterns and chemical process of EC, total dissolved solids (T.D. solids), and turbidity at different water body and stream flow levels (main stream and tributary). The solution presented in this study does not mention whether cross-validation was applied to the models. Given that water quality data can vary significantly across different time periods and locations, a model trained on one dataset may fail to generalize to other water bodies. Also, it does not address the idea that that different section of the river have similar trends in their water quality, and focuses on a section my section basis.

3.2 MACHINE-LEARNING (ML) MODELS

Karunanidhi et. al. [8], utilized different ML models along with WQI (Refer to Appendix 9 for more details) to predict the groundwater quality in the Arjunanadi River Basin in South India. They have also utilized the WQI to classify the groundwater quality into 5 categories: Excellent, Good, Poor, Very Poor, and Not Suitable.

Range	Type of Groundwater
< 50	Excellent
50—100	Good
100—200	Poor
200—300	Very Poor
> 300	Not Suitable

Figure 11: Categorization of Water Quality Index used for classifying the groundwater samples

The WQI showed that the data was nearly balanced with 53% of the area having good water quality, while 47% had poor water quality.

As for the models used to predict the water quality, they compared Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost). They integrated these ML models with Geographic Information System (GIS), combining the spatial data with predictive modeling techniques. This produced a more accurate model that takes into consideration the complex spatial data. For evaluation, they used Relative Squared Residual (RSR), Error and Nash–Sutcliffe efficiency (NSE), Mean Absolute Percentage Error (MAPE) and the Coefficient of determination (R2).

According to R2 validation (Coefficient of determination), LR showed the maximum fitness of 0.95 or 95%. Arranging the models by their performance they would be ordered as LR >SVM >Adaboost >XgBoost >RF. The LR model was identified to be the most effective with high accuracy and low error rates.

The authors have also studied the Gibbs dominance diagram that "identified the correlation among the water chemistry and the underground lithological appearances of

Model Errors Rates	LR	RF	SVM	AdaBoost	XgBoost
RSR	0.22	0.52	0.29	0.31	0.31
NSE	0.95	0.72	0.92	0.90	0.90
MAPE	1.3	9.2	2.3	5.26	5.26
R^2	0.95	0.72	0.92	0.90	0.90
Accuracy %	95	72	92	90	90

Figure 12: Results of ML models

an aquifer system by the main controlling mechanisms like rock interaction with water, evaporation and precipitation" [8]. It indicated that rock-water interaction was the dominant process influencing groundwater chemistry, with minor effects from evaporation and crystallization. The following graphs demonstrated this influence.

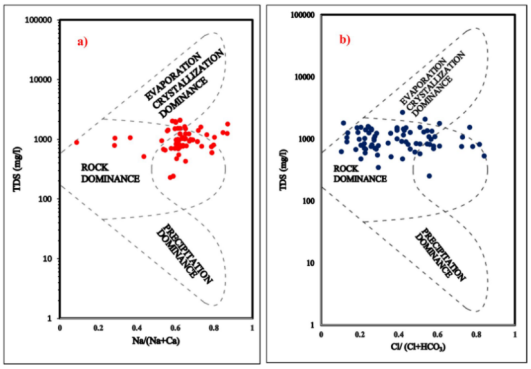

Figure 13:
(a): influence of evaporation and crystallization in the Na / (Na + Ca) vs TDS plot
(b): influence of evaporation and crystallization in Cl / (Cl + HCO3) vs TDS

The study showcases how ML models have a huge potential in predicting water quality, especially in regions with limited data.

3.3 BOTH ML & DL MODELS

In another research by Aldhyani et. al. [4], they compared both ML models and DL models. The study was applied to a dataset containing data from different Indian states for the years 2005 until 2014. The data was then normalized using the z-score, and WQI was calculated using a weighted formula. For WQI prediction they utilised Nonlinear Autoregressive Neural Network (NARNET) and Long Short-Term Memory (LSTM), which are DL models. However, for classifying water quality (WQC), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Naive Bayes algorithms were employed. To evaluate their models' performance they used MSE, R^2, accuracy, sensitivity, specificity, precision, and F-score. For WQI prediction both NARNET and LSTM nearly perform equally demonstrating high accuracy

As for WQC prediction, the SVM performed the highest accuracy, sensitivity, specificity, precision and F-score, followed by KNN and then Naive Bayes.

Models	Training data set		Testing data	
	MSE	R (%)	MSE	R (%)
NARNET	0.2815	95.97	0.1353	96.17
LSTM	0.1316	93.93	0.1028	94.21

Figure 14: NARNET and LSTM performance

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-score (%)
SVM	97.01	99.23	97.78	94.93	98.54
KNN	83.63	84.73	94.93	87.50	85.84
Naive Bayes	75.20	77.76	91.65	78.08	81.51

Figure 15: ML Models' performance

Their results show that both DL models tried were highly effective, while SVM outperformed all other ML models tried.

Chen et al. [5] (2020) has examined the limitations of traditional water quality monitoring techniques and proposed the use of artificial neural networks(ANN) and other ML methods such as random forest , logistic regression, support vector machine (SVM) and naive Bayes. His study analyzed 23 different water quality characteristics and has used Tensorflow in Python to train the models. The authors highlighted that the ML methods and approaches are accessible without any hardware. Thus, this shows the effectiveness of ML methods over traditional methods. The authors emphasized that the results demonstrated by ANNs significantly outperform the traditional methods.

Haghiabi et al. [7](2018) evaluated the effectiveness of artificial neural(ANNs) and support vector machines (SVM) for predicting water quality components in the Tireh River in southwest Iran. This study has tested various transfer and kernel functions to optimize SVm models. Results indicate that both models performed well, with the svm slightly outperforming the ANN in terms of accuracy. The study demonstrated the effectiveness of ML in forecasting water quality components such as pH and dissolved oxygen.

Another paper by Akhlaq et. al [3] studies the water quality classification at the Alpine glacial lakes and rivers in three districts of Pakistan. Nine water quality parameters (cadmium, chromium, lead, nickel, iron, arsenic, and total dissolved solids) in mg/L, power of hydrogen (Ph), and electrical conductivity (Ec) µS/Cm were used to compute the Water Quality Index (WQI) for classification. This study employs supervised machine learning models, including a decision tree, the k-nearest neighbor method, a neural network model (multi-layer perceptron), a support vector machine, and a random forest, to predict and validate the water quality class. The accuracy rates for the validation set were observed to be 83% for the decision tree model, 75% for the K-nearest neighbor method, 83% for the neural network, 88% for the support vector machine, and 88% for the random forest model.

The Random Forest effectively handles complex, nonlinear relationships between water quality indicators and provides high predictive power and so it outperformed all other models. The solution presented in this study gives equal weights to all variables and does not explore feature importance to determine which variables are highly affecting the model's predictions.

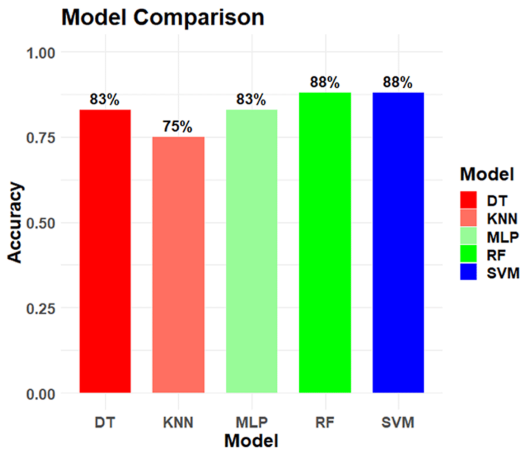

Figure 16: Evaluating accuracy metrics across various supervised machine learning models.

4 Dataset

For our analysis we plan on using a dataset provided on Kaggle and can be accessed from here [12]. This dataset represents different indicators including its nutrients, color, air and water temperature for measuring the water quality. All these variables are aligned with the date and time these indicators were getting measured. The dataset has 1,048,575 observations with 22 variables to be observed. Out of the 22 there are 3 categorical variables, while the rest are numeric. In addition to that, there is a target variable that identifies whether the water is potable or not. Digging deeper into each of the variables, we found out, by looking at *Figure 17*, that there are some missing values along the variables with the maximum reached by Copper with 34,882 missing values.

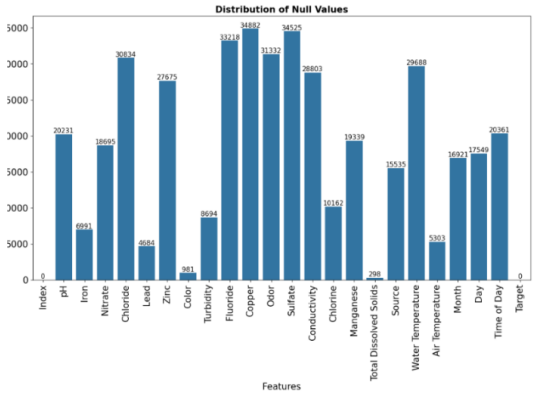

Figure 17: Distribution of Null Values

However, luckily no duplicates were found in the dataset. Looking at the categorical variables, for the source they all are somehow equally distributed with around the same number of observations from each source, this is illustrated in *Figure 18*.

As for the color variable, represented in *Figure 19*, they're unequally distributed with the colorless and near colorless observations having the highest number of observations.

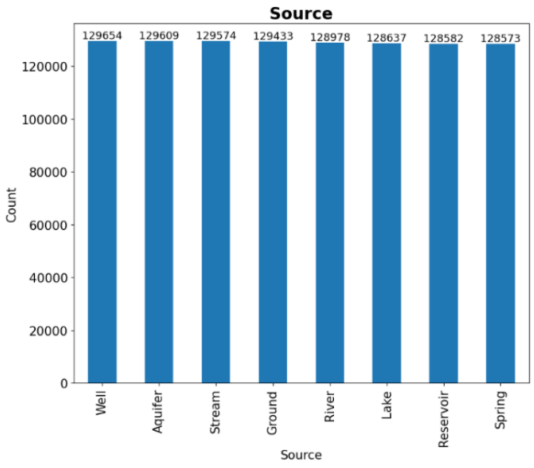

Figure 18: Source Variable unique values distribution

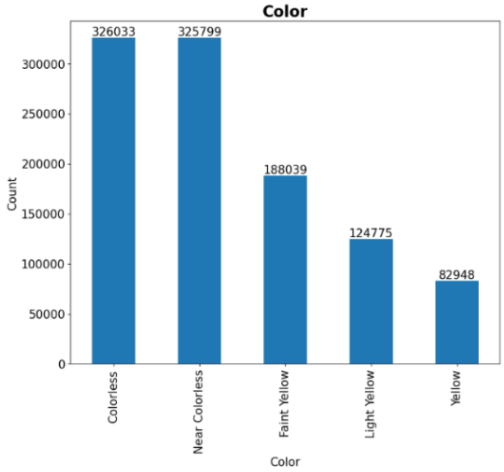

Figure 19: Color Variable unique values distribution

Lastly, as for our target variables, shown in *Figure 20* there seems to be a high imbalance between the two classes urging for an intervention by any of the statistical techniques to increase the size of our minority class, such as Synthetic Minority Oversampling Technique (SMOTE).

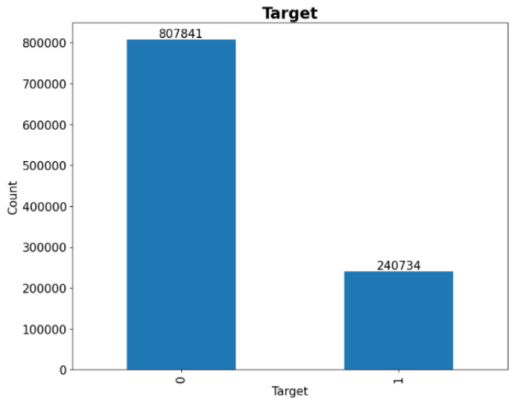

Figure 20: Target Variable unique values distribution

Moving on to the numerical variables, when getting to measure the percentage of outliers by just getting the percentage of values outside the Inter-quartile range (IQR), we found out that lead had the highest percentage across all, followed by Iron and Manganese, as represented by *Figure 21*.

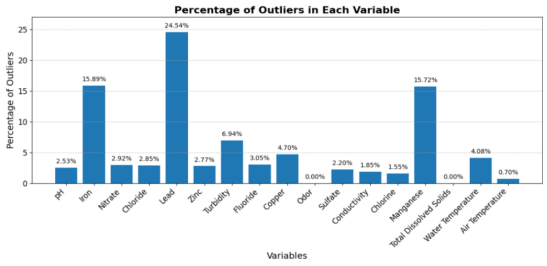

Figure 21: Percentage of outliers in each variable

You can access the Exploratory data analysis (EDA) done on the dataset here [6].

5 Visualization

After cleaning the data, we decided to do some visualizations to observe the data.

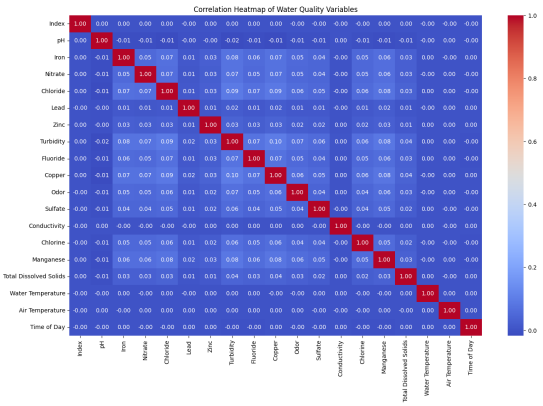

Figure 22: Correlation Matrix

The correlation matrix in *Figure 22* shows that there are no strong correlation between the variables.

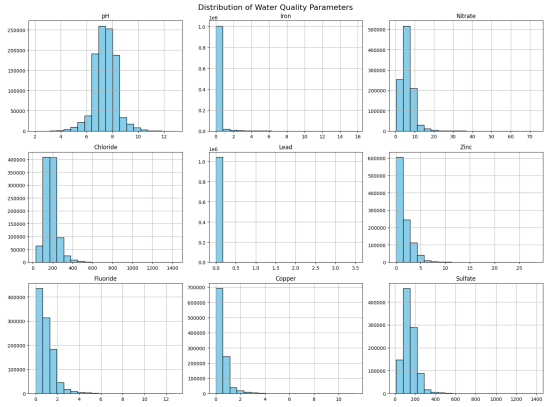

Figure 23: Distribution of water quality parameters

As shown in *Figure 23* Iron, Nitrate, Zinc, Cholride, Lead, Copper, Sulfate, Fluride are skewed to the right, while pH is normally distributed.

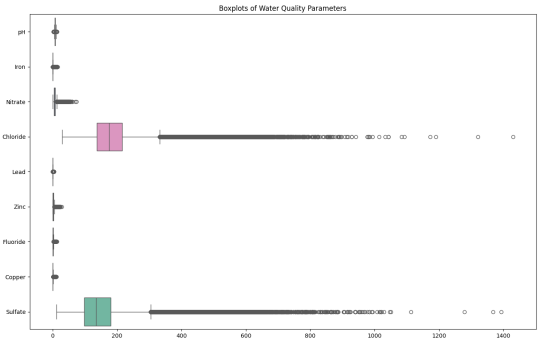

Figure 24: Boxplots of water quality parameters

After creating boxplots for each parameter, *Figure 24* shows that Chloride and Sulfate are the only parameters that has a wide range if values, while the rest of the varaibles values are very close to each other.

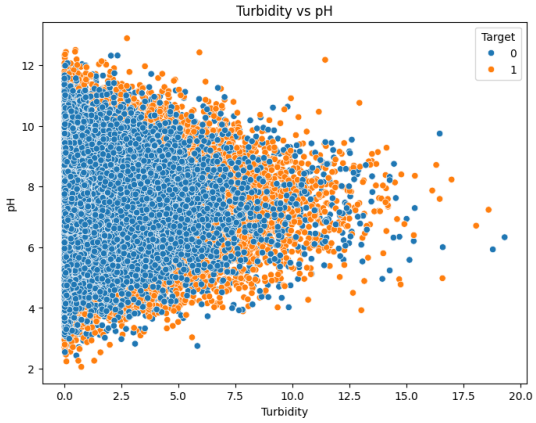

Figure 25: Turbidity vs. pH

Figure 25 shows a positive relation between turbidity and pH, where the higher the turbidity the higher is the pH levels. For the target 0, it seems that values are inclined more towards having lower turbidity, while for target 1, values of the turbidity are higher.

6 Data Manipulation

In an attempt to deal with Null values and outliers, we tried several combination to then reach the best option at the end. The first attempt was trying to drop Null values from the data and replace the outliers by median values since the median is less sensitive to outliers; however, this has resulted in a reduction in the data reaching 701k observations. Then, this time we tried replacing Null values by the median, and dropping outliers. Yet, this has resulted in a significant decrease in dataset reaching around 446k observations. Thus, we decided to only replace Null values by the median and leave the outliers as is. By that we preserve the dataset size.

Second step was to standardize the data and solve the class imbalance that was previously identified. Hence, within the different models we have tried, we added "balance_classes=True" and "standardize=True" to the parameters. This is mainly due to the huge data size were it could not be done before passing in the dataset, yet it had to be added in the models tried within H2O.

7 Modelling

We explored multiple models using the H2O framework to achieve optimal performance for the water quality prediction task. The models tested span a range of machine learning and deep learning techniques. We also utilized AutoH2O, which generates a leaderboard of the best-performing models based on various evaluation parameters. All models were trained on 80% of the dataset and tested on the remaining 20%. Prior to modeling, we performed imputation for missing values and removed outliers.

7.1 ReLU Neural Network

We experimented with various deep learning architectures, tuning the number of hidden layers and neurons using grid search. The ReLU activation function consistently yielded the best performance compared to other activation functions. The final neural network architecture, shown in Figure 26, consists of five hidden layers with ReLU activations and a softmax output layer.

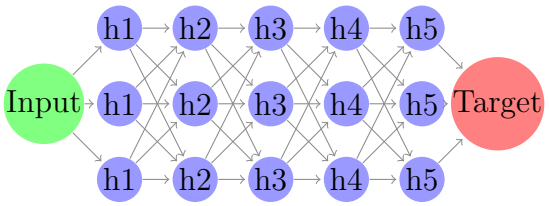

Figure 26: ReLU Neural Network for Water Quality Classification

The final model included five hidden layers with the following number of neurons: 300, 200, 100, 50, and 25, respectively. As shown in Figure 27, the neural network achieved an AUC of 0.95, demonstrating excellent predictive capability based on the ROC curve.

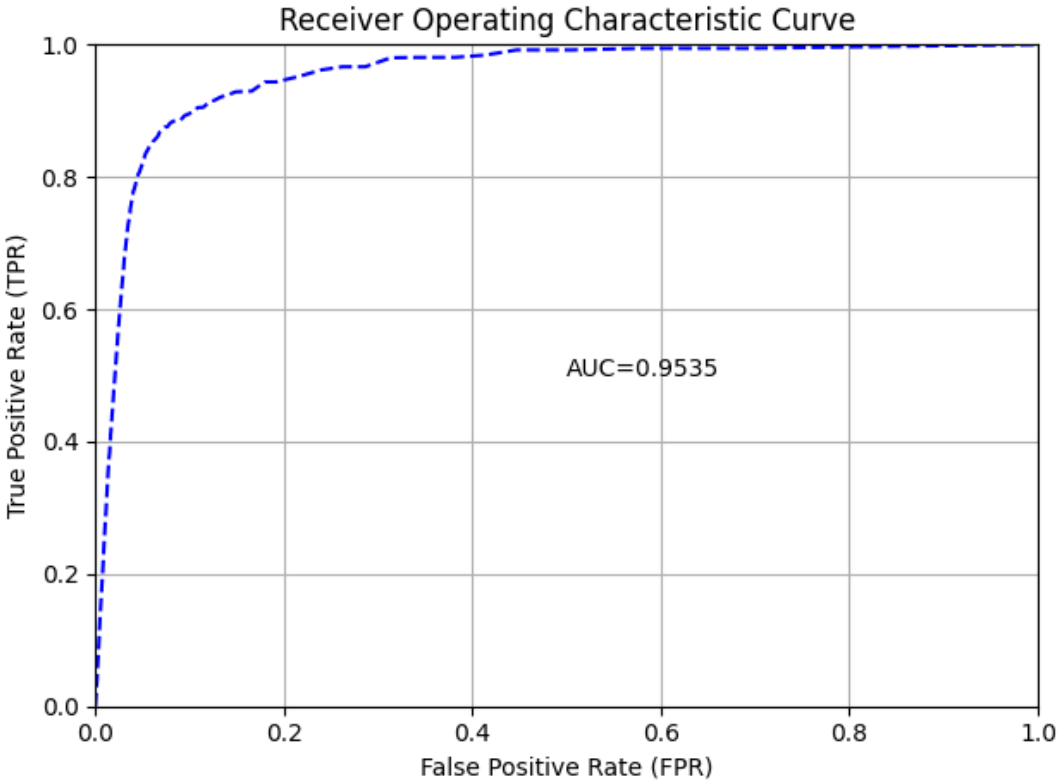

Figure 27: ROC Curve for Neural Network Model

The model achieved an AUC of 0.95 and an accuracy of approximately 94.97%. The log loss was 0.2574, indicating a strong overall performance on the test set.

7.2 Machine Learning Models

We also evaluated several machine learning models to compare and potentially enhance predictive performance and robustness.

7.2.1 Random Forest

The Random Forest model, an ensemble learning technique, was trained using 100 decision trees. This model not only achieved strong predictive performance but also provided valuable insights into feature importance. As illustrated in Figure 28, pH and color emerged as the most influential variables in water quality classification.

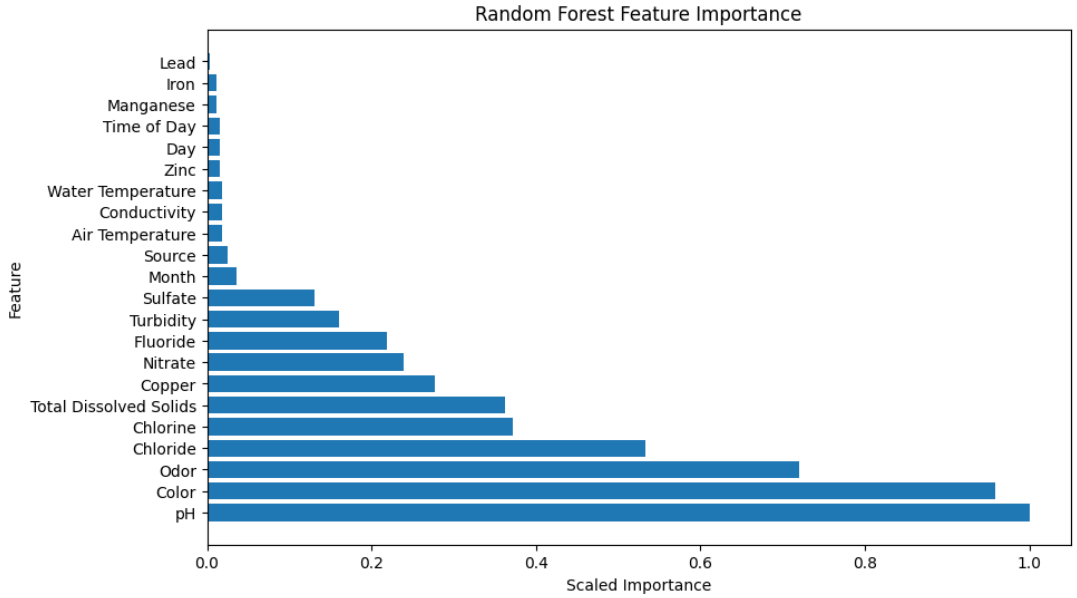

Figure 28: Variable Importance in the Random Forest Model

The model achieved an accuracy of approximately 95.24%, with an AUC of 0.970, indicating strong classification performance. The log loss was 0.133, demonstrating good probability calibration. Additionally, the RMSE was 0.210, and the MSE was 0.044, reflecting a low overall prediction error.

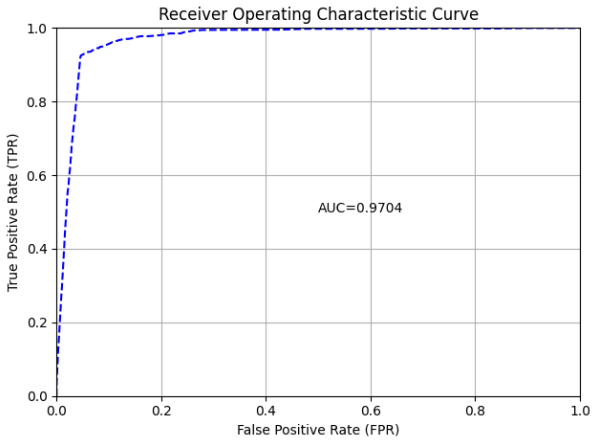

Figure 29: Performance Metrics for the Random Forest Model

7.2.2 Gradient Boosting

We also applied Gradient Boosting, an ensemble method that sequentially builds decision trees, with each tree correcting the errors of the previous one. This model was trained using 100 trees, with a maximum depth of 10. The Gradient Boosting model achieved an accuracy of approximately 95.64% and an AUC of 0.970, indicating strong predictive performance. These results suggest that Gradient Boosting was effective in capturing complex patterns in the data.

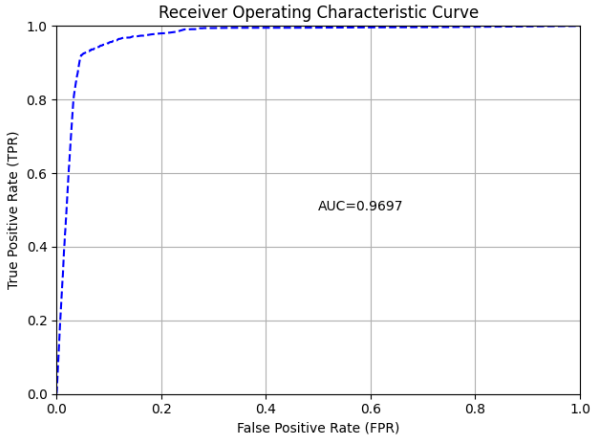

Figure 30: AUC for the Gradient Boosting Model

7.2.3 XGBoost

XGBoost emerged as the top-performing model, offering both accuracy and efficiency in training and inference. Trained with the same parameter configuration as Gradient Boosting (100 trees and a maximum depth of 10), XGBoost achieved an accuracy of approximately 95.78%, along with an AUC of 0.971, demonstrating its superior ability to discriminate between classes. Furthermore, XGBoost exhibited a log loss of 0.101,

indicating well-calibrated predictions. Its RMSE was 0.176, reflecting a low prediction error, making it the most robust model among those tested.

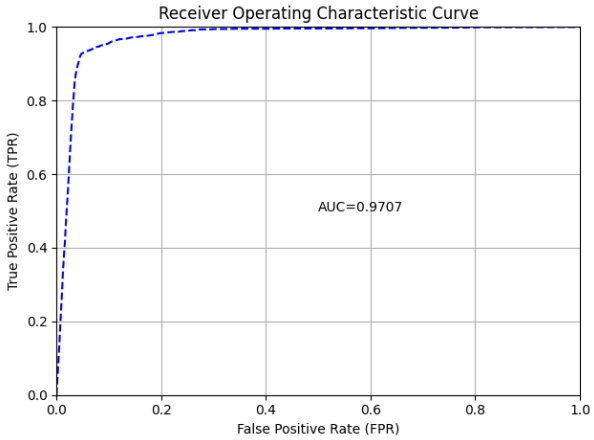

Figure 31: AUC for the XGBoost Model

7.2.4 H2O AutoML and Leaderboard

To automate the model selection process, we utilized H2O AutoML. This tool evaluates a wide range of algorithms and ranks them based on their performance on the validation data. H2O AutoML automatically splits the provided dataset into training and validation sets, typically using an 80-20 split by default, where 80% of the data is used for training the models, and the remaining 20% is reserved for validation. The validation set is used to assess model performance and tune hyperparameters. This split allows the tool to provide a leaderboard of the best-performing models, as shown in Figure 1.

The following code snippet was used to run the H2O AutoML process:

```
aml = H2OAutoML(max_models=20, max_runtime_secs=500, balance_classes=True)
```

The leaderboard reflects the model performance on the validation data set. The best model's performance is also shown in the training and validation graph in Figure 32.

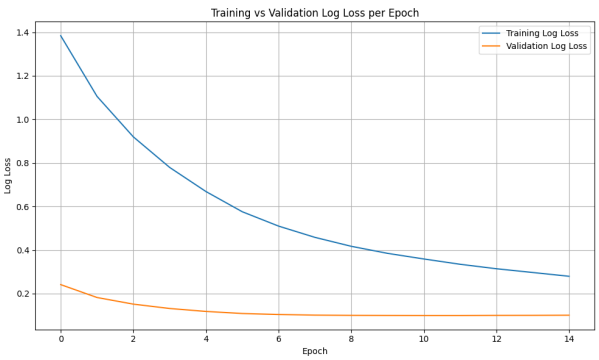

Figure 32: Training and Validation Performance for GBM (Best Model)

Model Name	AUC	Accuracy	RMSE	Parameters
GBM (Best)	0.972174	0.958861	0.1796	No. of Trees: 70, Max Depth: 15, No. of Leaves: 991.5
DRF	0.969660	0.952511	0.2102	No. of Trees: 50, Max Depth: 20, No. of Leaves: 3841.68
GBM	0.968008	0.956873	0.2186	No. of Trees: 10, Max Depth: 7, No. of Leaves: 95.3
XGBoost	0.966078	0.955797	0.1887	No. of Trees: 125
XGBoost	0.965895	0.956018	0.1908	No. of Trees: 175
GLM	0.827229	0.938209	0.2244	Ridge ($\lambda = 3.574 \times 10^{-5}$), No. of Predictors: 44

Table 1: AutoML Model Leaderboard

The performance metrics for the best model (GBM) are reported on the test data as follows:

- **MSE:** 0.0334
- **RMSE:** 0.1828
- **LogLoss:** 0.1052
- **Mean Per-Class Error:** 0.0784
- **AUC:** 0.9711
- **AUCPR:** 0.6346
- **Gini:** 0.9423

Finally, the confusion matrix for the test data is shown in Figure 33, which provides a visualization of model performance in terms of true positives, false positives, true negatives, and false negatives.

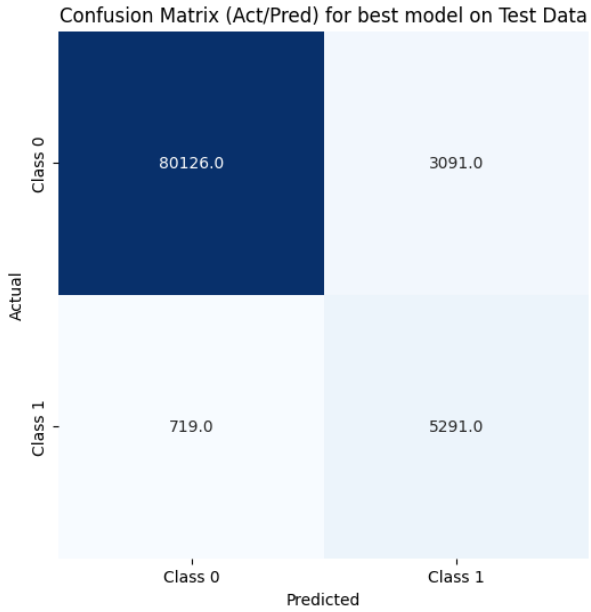

Figure 33: Confusion Matrix for Test Data (GBM Model)

The confusion matrix in 33 shows the performance of the model at the threshold of 0.1405, with an overall error rate of 4.27%, where 3.91% of class 0 and 11.96% of class 1 predictions are incorrect.

8 Results

This section presents the evaluation metrics of the models used, including the deep learning, machine learning, and AutoML approaches. Table 2 summarizes the performance of each model based on AUC, accuracy, RMSE, and key training parameters.

Model	**AUC**	**Accuracy**	**RMSE**	**Parameters**
Neural Network (ReLU)	0.9552	0.9496	0.2574	5 Layers
Random Forest	0.9703	0.9524	0.2101	100 Trees
Gradient Boosting	0.9697	0.9564	0.1816	100 Trees
XGBoost	0.9707	0.9566	0.1761	100 Trees
Gradient Boosting(AutoML)	0.9753	0.9567	0.1897	70 Trees

Table 2: Model Performance Evaluation Metrics and Parameters

The results in Table 2 demonstrate the comparative performance of the various models. XGBoost, with an AUC of 0.9707 and accuracy of 0.9566, **initially appeared to be the performing model,** offering high accuracy and strong predictive capabilities. However, when AutoML tested 20 models in 500 seconds, which allowed it to test different combinations of parameters for various models ,and when it happened that the gradient boost, a different result emerged. **The AutoML-optimized Gradient Boosting model achieved a slightly higher AUC of 0.9753** and an accuracy of 0.9567. This

improvement can be attributed to the fine-tuning of hyperparameters in the AutoML process, allowing Gradient Boosting to perform better than its standard configuration.

This result highlights the importance of hyperparameter optimization in improving model performance. Although XGBoost performed excellently, the AutoML approach to Gradient Boosting demonstrated that fine-tuning the number of trees and other parameters could lead to even better results.

9 Conclusion

In conclusion, this study highlights the potential of deep learning and machine learning techniques to predict water quality. We explored various ML methods for classifying water as potable or nonpotable, addressing key challenges such as data imbalance and the preprocessing of large datasets using the H2O framework. Our approach led to the development of a robust classification pipeline.

Although traditional models like XGBoost and manually tuned Gradient Boosting yielded strong performance, the use of H2O AutoML significantly enhanced model optimization through automated hyperparameter tuning and model selection. The AutoML-optimized Gradient Boosting model achieved the highest AUC and accuracy, with an accuracy of 95.67%, outperforming results reported in previous literature.

In general, this work demonstrates the effectiveness of automated machine learning in improving water quality prediction models. It contributes to more efficient and reliable water quality monitoring, ultimately supporting public health and environmental sustainability.

References

[1] A. -A. Nayan, M. G. Kibria, M. O. Rahman and J. Saha, "River Water Quality Analysis and Prediction Using GBM," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 2020, pp. 219-224, doi: 10.1109/ICAICT51780.2020.9333492.

[2] Acheampong A. O. & Opoku E. E. O. (2023) Environmental degradation and economic growth: investigating linkages and potential pathways, Energy Economics, 123, 106734. https://doi.org/10.1016/j.eneco.2023.106734.

[3] Akhlaq, M., Ellahi, A., Niaz, R., Khan, M., Sammen, S. S., & Scholz, M. (2024). Comparative analysis of machine learning algorithms for water quality prediction. Tellus A: Dynamic Meteorology and Oceanography, 76(1).

[4] Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M. Water Quality Prediction Using Artificial Intelligence Algorithms. Appl Bionics Biomech. 2020 Dec 29;2020:6659314. doi: 10.1155/2020/6659314. Retraction in: Appl Bionics Biomech.

2023 Oct 11;2023:9761657. doi: 10.1155/2023/9761657. PMID: 33456498; PMCID: PMC7787777.

[5] Chen Y, Song L, Liu Y, Yang L, Li D. A review of the artificial neural network models for water quality prediction. Applied Sciences. 2020 Aug 20;10(17):5776.

[6] EDA, https://colab.research.google.com/drive/10PCl92qokNIix3TxaLDjtXiRQB96czpf?usp=sharing

[7] Haghiabi, Amir Hamzeh, Ali HeidarNasrolahi, and Abbas Parsaie. "Water quality prediction using machine learning methods." Water Quality Research Journal 53.1 (2018): 3-13.

[8] Karunanidhi, D., Raj, M.R.H., Roy, P.D. et al. (2025) Integrated machine learning based groundwater quality prediction through groundwater quality index for drinking purposes in a semi-arid river basin of south India. Environ Geochem Health 47, 119. https://doi-org.libproxy.aucegypt.edu/ 10.1007/s10653-025-02425-9

[9] Najah, A., El-Shafie, A., Karim, O. A., & El-Shafie, A. H. (2013). Application of artificial neural networks for water quality prediction. Neural Computing and Applications, 22, 187-201.

[10] Scanlon, B.R., Fakhreddine, S., Rateb, A. et al. Global water resources and the role of groundwater in a resilient water future. Nat Rev Earth Environ 4, 87–101 (2023). https://doi.org/10.1038/s43017-022-00378-6

[11] Unigwe, C.O., Egbueri, J.C. Drinking water quality assessment based on statistical analysis and three water quality indices (MWQI, IWQI and EWQI): a case study. Environ Dev Sustain 25, 686–707 (2023).

[12] Water quality prediction. (2023, July 10). Kaggle. https://www.kaggle.com/datasets/vanthanadevi08/water-quality-prediction?resource=download

[13] Wu, X., Zhang, Q., Wen, F., & Qi, Y. (2022). A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China. Water, 14(21), 3408. https://doi.org/10.3390/w14213408

[14] Y. Wang, J. Zhou, K. Chen, Y. Wang and L. Liu, "Water quality prediction method based on LSTM neural network," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, China, 2017, pp. 1-5, doi: 10.1109/ISKE.2017.8258814.

[15] Yao, S., Zhang, Y., Wang, P., Xu, Z., Wang, Y., & Zhang, Y. (2022). Long-Term Water Quality Prediction Using Integrated Water Quality Indices and Advanced Deep Learning Models: A Case Study of Chaohu Lake, China, 2019–2022. Applied Sciences (2076-3417), 12(22), 11329. https://doi.org/10.3390/app122211329

Water Quality Index Calculation:

By Yao et. al [15]:

$$WQI(1) = \frac{\sum_{i=1}^{N} C_i P_i}{\sum_{i=1}^{N} P_i}$$

where:

- N = Total number of parameters
- P_i = Weight assigned to the water quality parameter
- C_i = Normalized value of parameter i

By Karunanidhi et. al [8]:

$$WQI(2) = \sum_{a=1}^{n} \frac{C_a}{\sum_{a=1}^{n} C_a} * q$$

$$C_a = \sigma_a \sum_{a=1}^{n} (1 - r)$$

$$r = \frac{\sum (x_b - \bar{x}_a)(y_b - \bar{y}_a)}{\sqrt{\sum (x_b - \bar{x}_a)^2 \sum (y_b - \bar{y}_a)^2}}$$

$$q = \frac{C_a - C_s}{S_a - C_s} * 100$$

where:

- σ_a = Standard deviation of the a^{th} parameter in dataset
- n = Whole existing parameters
- r = Pearson's correlation coefficient of two parameters
- q = Parameter quality score gage
- C_a = a^{th} variable concentration in a sample
- S_a = Parameter accepted limits proposed by the WHO
- C_s = Optimal quantity of the parameter in pure water