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

Sally Elghamrawy
Department of Computer Science and Engineering
American University in Cairo
Cairo, Egypt
Sally_elghamrawy@ieee.org

Shaymaa Mahmoud*
Department of Mathematics and Actuarial Science
American University in Cairo
Cairo, Egypt
shaymaa123@aucegypt.edu

Sama Gouda*
Department of Mathematics and Actuarial Science
American University in Cairo
Cairo, Egypt
samaamr@aucegypt.edu

Mona Ibrahim*
Department of Mathematics and Actuarial Science
American University in Cairo
Cairo, Egypt
monamahmoud@aucegypt.edu

***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.**

***Index Terms*—Sustainability, Machine Learning, Deep Learning, Classification, Big Data Technologies, H2O**

I. Introduction

A. Background Information

Water is one of the most essential resources for sustaining life, economic activities, and ecosystem balance. 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. 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.

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

B. 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. . 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.

C. Research Objectives

This paper explores the application of big data and machine learning methodologies in predicting water quality. By analyzing existing literature and datasets, our aim is 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.

D. Research Methodology

This research is structured into four main sections: literature review, methodology, proposed method description, and conclusion. The literature review thoroughly investigates the

present setting of water potability classification. Specifically, it delves into the currently existing methodologies and approaches, showcasing the evolution of water quality prediction and their applicability in the dynamic real life events. In the methodology section, the study offers a brief description on the pre-processing process utilized along with the different algorithms that were brought into use. Lastly, the proposed method section provides an overview on the selection criteria for these algorithms and their effectiveness in accurately classifying the water quality, with the aim to identify the best suitable model for classifying water as potable or not.

E. Research Question

- How does the performance of traditional machine learning algorithms compare to that of deep learning models in classification of water quality?
- How does feature importance analysis reveal the most critical water quality parameters for prediction?
- How does prediction performance compare between conventional water quality indices and machine learning predictions?

II. Literature Review

A. Introduction

According to Wu et al [14], 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.

B. Deep Learning (DL) Models

Yao et. al [16]have predicted long-term water quality using deep learning models integrating the (WQI) water quality index to categorize the data into 5 categories(based on its quality) conducted on the water of Chaohu Lake, China. They utilized the long sequence time series forecasting method (LSTF) to predict the selected water quality parameter WQI. Then to model the data they proposed the Transformer-based models (Informer and FEDformer) which uses both the LSTF and Encoder-Decoder model. The performance was compared to other deep learning models: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), where the Informer model reached the best performance with an MSE = 0.2455, and MAE = 0.2449.

C. Machine-Learning (ML) Models

Karunanidhi et. al. [8], utilized different ML models along with WQI to predict the groundwater quality in the a River Basin in South India. They have compared Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost). In addition to integrating the Geographic Information System (GIS), combining the spatial data with predictive modeling techniques, producing a more accurate model that takes into account the complex spatial data. According to $R\hat{2}$ validation (Coefficient of determination), LR showed the maximum fitness of 95%, being the most effective with high accuracy and low error rates. The authors have also studied the Gibbs dominance diagram to identify the correlation between the water chemistry and controlling mechanisms [8]. The diagram indicated that rock-water interaction was the dominant process influencing groundwater chemistry, with minor effects from evaporation and crystallization.

D. Both ML & DL Models

In another research by Aldhyani et. al. [4], they compared both ML models and DL models. The study was conducted on data collected from different Indian states. 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. As for the performance of models, both NARNET and LSTM nearly perform equally demonstrating high accuracy for the WQI prediction. As for WQC prediction, the SVM performed the best.

III. Methodology

A. Introduction

This research leveraged H2O's machine learning framework to conduct a comprehensive water potability prediction model. The data was imported from Kaggle. The data was first analyzed to better understand its structure, and then it was passed into a series of preprocessing steps to deal with the null values, and outliers. The cleaned version was then passed into SMOTE to deal with class imbalance, and then it was standardized using the Z-score.

For the potability classification, we applied several machine learning algorithms and deep learning models, and evaluated their performance into accurately classifying the water quality. These included the H2O's Deep Learning Estimator, Random Forest Estimator, Gradient Boosting Estimator, XG Boost Estimator, and automl (including: Gradient Boosting Machine, Distributed Random Forest, XG Boost, Generalized Linear Model). The performance of such models was evaluated using metrics such as Mean Square Error (MSE), Root Mean Squared Error (RMSE), Log of Loss, Area Under the Curve (AUC), and Gini Coefficient.

This approach not only facilitated a deep analysis of water quality but has also allowed for a comparative study of different machine and deep learning techniques in classification.

The following *figure 1* represents a flow diagram with the steps that have been implemented throughout the whole paper.

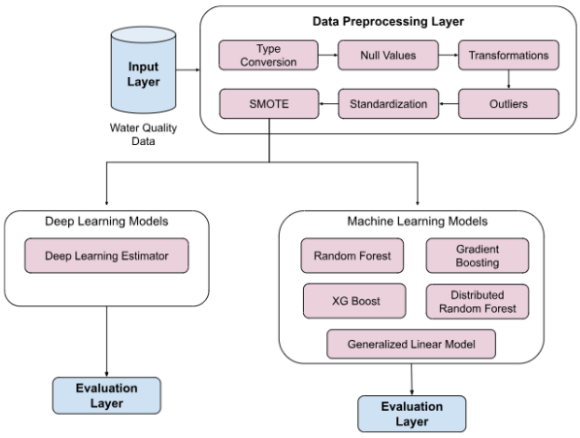

Fig. 1. Flow Chart for the whole project

B. *Dataset*

We used a dataset provided on Kaggle and can be accessed from here [13]. 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. The following *table I* represents a summary of the data.

TABLE I
DATASET DESCRIPTION

#	Column	Non-Null Count	Dtype
0	Index	1,048,575	int64
1	pH	1,028,344	float64
2	Iron	1,041,584	float64
3	Nitrate	1,029,880	float64
4	Chloride	1,017,741	float64
5	Lead	1,043,891	float64
6	Zinc	1,020,900	float64
7	Color	1,047,594	object
8	Turbidity	1,039,881	float64
9	Fluoride	1,015,357	float64
10	Copper	1,013,693	float64
11	Odor	1,017,243	float64
12	Sulfate	1,014,050	float64
13	Conductivity	1,019,772	float64
14	Chlorine	1,038,413	float64
15	Manganese	1,029,236	float64
16	Total Dissolved Solids	1,048,277	float64
17	Source	1,033,040	object
18	Water Temperature	1,018,887	float64
19	Air Temperature	1,043,272	float64
20	Month	1,031,654	object
21	Day	1,031,026	float64
22	Time of Day	1,028,214	float64
23	Target	1,048,575	int64

C. *Pre-processing*

Pre-processing is one of the most important steps to enhance the accuracy of our models. The first step was to apply type conversion to each of the columns that have incorrectly identified types. These were the Source, Color, Month, Day, and Target variables, casting ”categorical” type to each. Digging deeper into each of the variables, we found out, by looking at *Figure 2*, that there are some missing values along the variables with the maximum reached by Copper with 34,882 missing values.

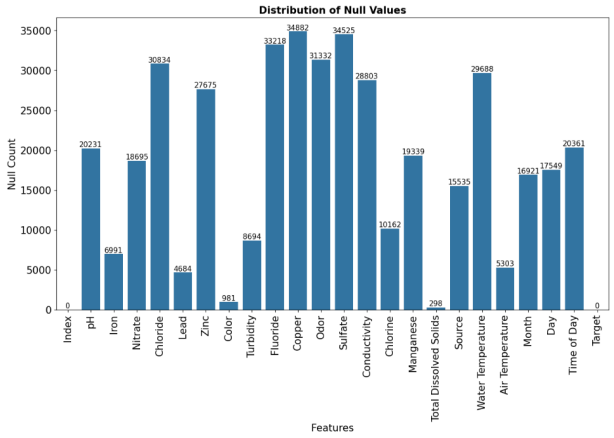

Fig. 2. Distribution of Null Values

Next step was figuring out if the variables need any kind of transformation to be of a better fit. Luckily by looking at the graphs of each of the plotted variables there wasn't any need for the transformation. *Figure 3* represents an example for the ”pH” variable.

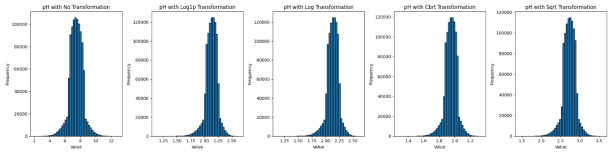

Fig. 3. Transformation applied to pH variable

Futhermore, we had to check for any present outliers within the dataset, by measuring 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 4*.

In an attempt to deal with Null values and outliers, we tried out several combination. 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 dataset size reaching 701k observations. Second, we tried replacing Null values by the median, and dropping outliers. Yet, this has resulted in a significant decrease in dataset reaching around

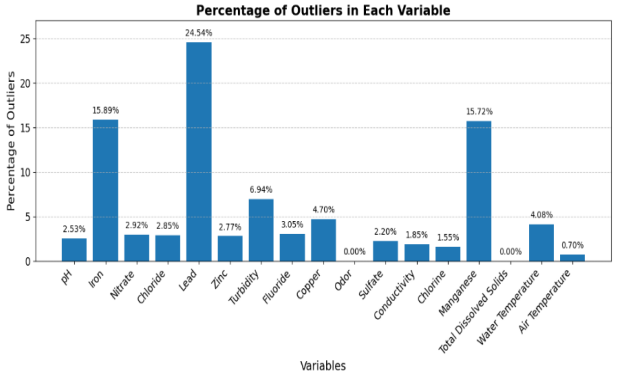

Fig. 4. Percentage of Outliers

446k observations. Thus, we decided to try both sets of data and we'd continue our analysis using the one that would give higher results.

In order to decrease the effect of the outliers on the data, we decided to standardize the data using the Z-Score. This is implemented as a parameter within the function of each of the models in H2O.

One important factor that we can't turn a blind eye to is identifying whether there's a class imbalance or not. This was done by plotting a bar graph showing the count of both classes. As demonstrated by *figure 5*, there's a huge class imbalance within the data, urging for the use of Synthetic Minority Oversampling Technique (SMOTE). This is also a parameter within the function of each of the models.

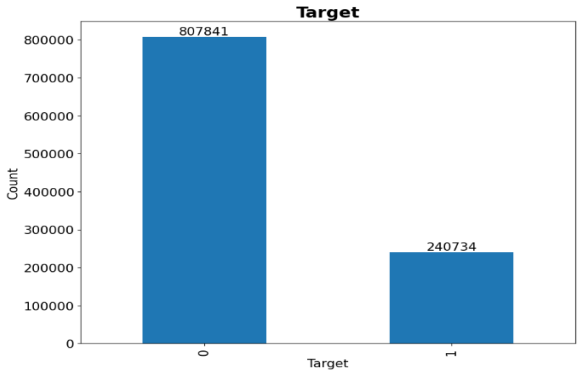

Fig. 5. Distribution of Target variable

D. *Visualizations*

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

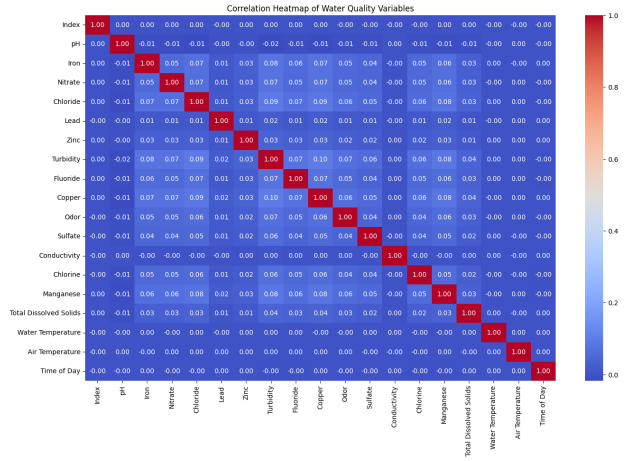

Fig. 6. Correlation Matrix

The correlation matrix in *Figure 6* shows that there are no strong correlation between the variables.

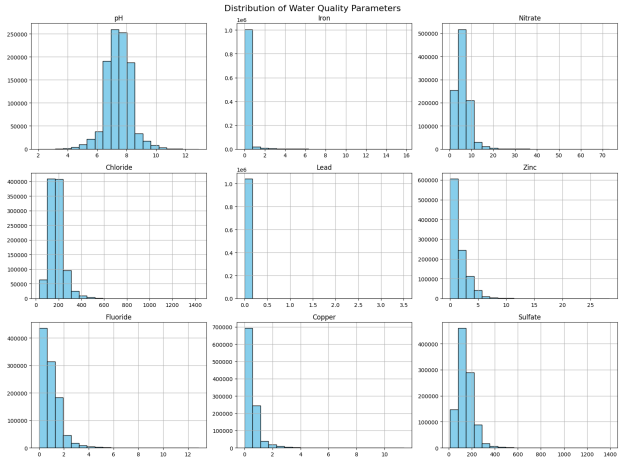

Fig. 7. Distribution of water quality parameters

As shown in *Figure 7* Iron, Nitrate, Zinc, Cholride, Lead, Copper, Sulfate, Fluride are skewed to the right, while pH is normally distributed.

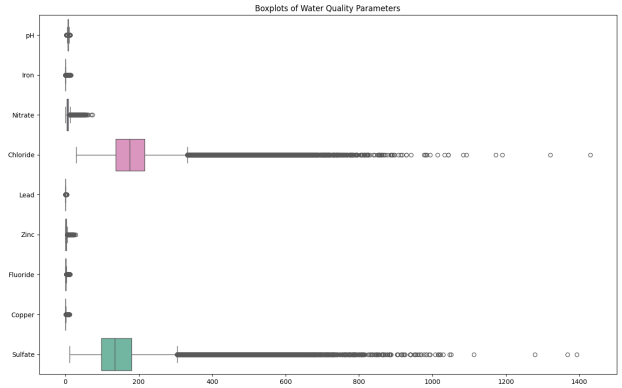

Fig. 8. Boxplots of water quality parameters

After creating boxplots for each parameter, *Figure 8* 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.

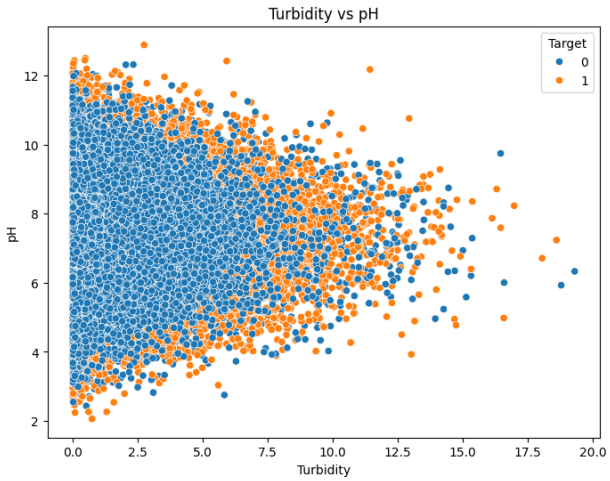

Fig. 9. Turbidity vs. pH

Figure 9 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.

E. Models

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.

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, consists of five hidden layers with ReLU activations and a softmax output layer.

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

Machine Learning Models

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

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 by variable importance, pH and color emerged as the most influential variables in water quality classification. 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.

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.

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.

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 table II. The leaderboard reflects the model performance on the validation data set.

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 II
AUTOML MODEL LEADERBOARD

Finally, the confusion matrix for the test data is shown in Figure 10, providing a visual representation of the model's performance in terms of true positives, false positives, true negatives, and false negatives.

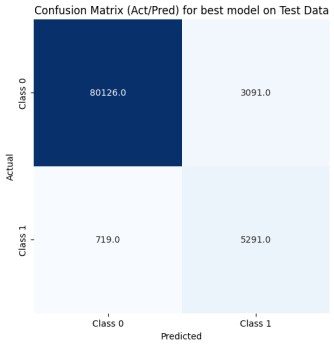

Fig. 10. Confusion Matrix for Test Data (GBM Model)

Figure 10 illustrates the model's performance at a threshold of 0.1405. The overall error rate is 4.27%, with 3.91% of class 0 predictions and 11.96% of class 1 predictions being incorrect. Additionally, the ROC curve for the best-performing model, with an AUC of 0.9711, is presented in Figure 11.

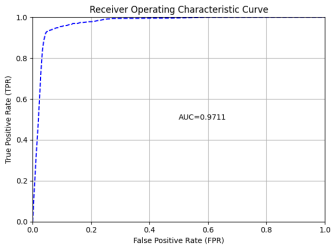

Fig. 11. ROC Curve for the Best Model (GBM)

F. Evaluation

This section presents the evaluation metrics of the models used, including the deep learning, machine learning, and AutoML approaches. Table III summarizes the performance of each model based on AUC, accuracy, RMSE, and key training parameters. The results in Table III demonstrate the performance of the various models, including the AutoML-optimized Gradient Boosting model, and provide insights into how this model compares with others.

Model	F-1	Accuracy	RMSE	Parameters
Neural Network (ReLU)	0.6649	0.9496	0.2574	5 Layers
Random Forest	0.7219	0.9524	0.2101	100 Trees
GBM	0.7200	0.9564	0.1816	100 Trees
XGBoost	0.7331	0.9566	0.1761	100 Trees
GBM (AutoML)	0.7357	0.9567	0.1897	70 Trees

TABLE III
MODEL PERFORMANCE EVALUATION METRICS

a) Comparison of H2O AutoML with the Models:

Initially, XGBoost emerged as the top performer with an AUC of 0.9707 and an accuracy of 0.9566. However, after applying AutoML to test 20 models and optimize hyperparameters, Gradient Boosting achieved a slightly higher AUC of 0.9753 and accuracy of 0.9567. This suggests that the AutoML process was effective in tuning Gradient Boosting, outperforming its standard configuration. While Random Forest and Gradient Boosting (without AutoML) also showed strong performance, with accuracies of 0.9524 and 0.9564, respectively, the AutoML-optimized Gradient Boosting model achieved the best overall performance in terms of AUC.

b) Comparison of Results with Previous Work:

When compared to previous studies, our AutoML-optimized Gradient Boosting model demonstrated strong and competitive performance. It achieved an RMSE of 0.1897, outperforming the Informer model reported in Yao et al. [16], which yielded an RMSE of 0.2455 on a similar water quality prediction task. Furthermore, our model also exceeded the accuracy reported by Karunanidhi et al. [8] for XGBoost, which was approximately 90%. The AutoML-optimized Gradient Boosting model not only achieved a higher accuracy of 95.67% but also maintained consistent performance across other evaluation metrics.

IV. 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] Water quality prediction. (2025, April 18). Github. https://github.com/Sama-Amr/A-Comprehensive-Machine-Learning-Approach-for-Water-Quality-Prediction-with-the-H2O-Framework

[14] 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

[15] 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.

[16] 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