Water Quality Analysis and Treatment using AI

Potnuru Jayanth, Varra Bandhan Reddy, Revanth Atluri
CSE, Vellore Institute of Technology, Chennai, Tamil Nadu, India
potnuri.jayanth2021@vitstudent.ac.in
CSE, Vellore Institute of Technology, Chennai, Tamil Nadu, India
varrabandhan.reddy2021@vitstudent.ac.in
CSE, Vellore Institute of Technology, Chennai, Tamil Nadu, India
atluri.revanth2021@vitstudent.ac.in

ABSTRACT

Water quality study is an important step toward ensuring community safety and well-being. In this study, we used machine learning (ML) and deep learning (DL) algorithms to evaluate water quality and distinguish between places with clean and contaminated water. Our goal was to create an accurate and efficient model for evaluating water quality. We gathered a dataset of 100 rows of raw data containing various parameters related to water quality. Two ML algorithms, Decision Tree and Lasso, were used to extract significant patterns and correlations from the data. Furthermore, three deep learning methods, namely LSTM, Bi-LSTM, and ANN, were used to capture complicated temporal and spatial connections. The root mean square (RMS) value was used as a metric of prediction accuracy to assess each algorithm's performance. We were able to distinguish between places with pure and contaminated water based on our findings. Notably, out of all the algorithms, the Decision Tree method had the lowest RMS value, demonstrating its higher prediction accuracy in the analysis of water quality. For the purpose of creating decision rules and finding important aspects for water quality study, the Decision Tree method worked well. Furthermore, the Lasso method showed strong feature selection capabilities that made it possible to identify important factors that are responsible for water contamination. The temporal and geographical dependencies that the LSTM, Bi-LSTM, and ANN algorithms demonstrated allowed them to better forecast the quality of the water. Our study effectively assessed water quality data and gave insights into places with pure and contaminated water by utilizing the capabilities of ML and DL algorithms. The fact that the Decision Tree algorithm is the most accurate model emphasizes how suitable it is for evaluating water quality. These results have enormous promise for putting targeted actions and policies into place to enhance the quality of the water in certain areas. Furthermore, our methodology acts as a basis for other studies and applications related to the analysis of water quality. We conclude by showing the effectiveness of ML and DL algorithms in the analysis of water quality. The decision tree approach was found to be the most suitable model for an accurate assessment of water quality because of its low RMS value in comparison to other algorithms. These findings advance our knowledge of water quality and help guide the development of public health and water resource management plans.

Keywords: Water quality study, Machine Learning, Deep Learning, Decision Tree, Lasso, LSTM, BiLSTM, ANN, Prediction Accuracy, Feature Selection.

I. INTRODUCTION

Water quality analysis and treatment are critical processes aimed at ensuring the safety and purity of water from lakes or seas. With the advent of Artificial Intelligence (AI), these procedures have witnessed significant advancements, enhancing efficiency and precision. Analyzing water quality entails determining a number of factors, including turbidity, dissolved oxygen, pH, and the presence of pollutants.

Artificial Intelligence plays a pivotal role by automating the monitoring process through the use of sensors and data analytics. This enables real-time analysis, allowing for swift detection of anomalies or pollutants in the water. Treatment of lake or sea water involves the removal of impurities and contaminants to meet safety standards. By anticipating changes in water quality and modifying treatment procedures accordingly, artificial intelligence is used to optimize treatment plants.

Machine learning algorithms can learn from historical data, providing insights into potential issues and suggesting proactive measures. The integration of AI in water quality management contributes to the development of smart water systems. These systems can dynamically adapt to changing conditions, improving overall efficiency and reducing operational costs. Additionally, AI aids in the early detection of waterborne diseases, safeguarding public health. The ability of AI systems to analyze vast datasets and adapt to dynamic conditions empowers water authorities to make informed decisions for sustainable resource management. Furthermore, AI-driven predictive models contribute to long-term planning, aiding in the development of robust strategies to tackle emerging water quality issues. This synergy between advanced technology and environmental stewardship underscores the potential for AI to play a pivotal role in ensuring the resilience and longevity of our vital water ecosystems, ultimately benefiting both human populations and the diverse aquatic life dependent on these ecosystems. Artificial Intelligence (AI) has revolutionized water management by introducing innovative solutions to enhance the efficiency and accuracy of these processes. In the context of lake or sea water, AI technologies are employed for real-time monitoring, analysis, and treatment optimization. In summary, the synergy between water quality analysis, treatment, and AI marks a significant advancement in environmental management. By leveraging AI technologies, we can achieve a more sustainable and resilient approach to safeguarding the quality of lake and sea water, promoting the well-being of ecosystems and human communities that rely on these vital resources.

II. LITERATURE SURVEY

The growing frequency of adulterants in lake and ocean water makes water quality analysis and treatment—especially for these sources of water—an essential field of study. We'll go over the results of a number of studies on artificial intelligence (AI) modeling-grounded water quality analysis and treatment.

Escher et al. (2014) showed that a tailored panel of bioassays should be used for routine monitoring, as certain cell-based bioassays can be used to standardize water quality. The importance of sophisticated monitoring techniques for precise water quality assessment is stressed by this study. This points to a possible avenue for future exploration in creating AI-supported bioassay platforms to improve and automate the monitoring of water quality. According to Palansooriya et al. (2020), biochar is a feasible option for addressing the problems associated with treating common pollutants that are frequently detected in drinking water. This includes specifics like endocrine-disrupting chemicals (EDCs), volatile organic compounds (VOCs), heavy metals, microbiological contaminants, and personal care products (PPCPs). This finding raises the possibility of optimizing biochar-based water treatment systems by combining AI and machine learning algorithms.

Vegetated treatment systems (VTSs) have been linked by Stehle et al. (2011) as a feasible and effective threat mitigation approach for agricultural nonpoint source pesticide pollution of surface waters. This presents an opportunity to explore AI-based VTS optimization to improve their efficacy in reducing pesticide pollution in aquatic environments.

Osta et al. (2022) emphasized the growing use of real-time sensor monitoring to efficiently collect data on water quality; they also highlighted that in order to extract valuable insights from the data, AI-driven data analytics and interpretation are needed.

Park et al. (2020) talked about how crucial water and wastewater treatment operations can be automated and optimized using artificial intelligence methods and machine learning models. This implies that future exploration could focus on creating AI-driven decision support systems to improve the operation and energy efficiency of water treatment plants. Also, Lowe et al. (2022) and Wang et al. (2016) emphasized the need for applicable treatment to provide safe water, suggesting a possible avenue for exploration to develop AI-driven smart systems for monitoring and treating water in both urban and rural settings. Further emphasizing the value of modeling and simulating integrated urban wastewater systems, Benedetti et al. (2013) suggested that future studies may explore AI-supported modeling and simulation tools for urban water systems.

Eventually , there's a lots of the potential for improving the effectiveness, efficiency, and sustainability of water treatment operations through the integration of AI technology for water quality analysis and treatment. Future studies in this field should focus on creating AI-driven modeling, treatment, and monitoring systems to deal with the emerging challenges in water quality assessments.

III. METHODS AND MATERIAL

A).Data Acquisition : Water quality data was obtained from [source of data, e.g., historical records from a water treatment plant, sensor readings from a monitoring system]. The data encompassed various parameters such as pH, temperature, conductivity, dissolved oxygen. Missing values were handled..The data was then normalized to ensure all features were on a comparable scale.

B).Data Exploration : Exploratory data analysis (EDA) was conducted using statistical methods and visualization tools. This analysis aimed to understand the data distribution, identify potential relationships between parameters, and uncover any challenges that might impact model performance.

C).Model Selection and Development :
Decision Tree: A Decision Tree model was chosen due to its interpretability, ability to handle non-linear relationships, and suitability for tabular data like ours. The model's hyperparameters, such as maximum depth and minimum samples per split, were optimized using grid search to achieve the best performance.

Lasso Regression: Lasso Regression was selected for its ability to perform feature selection and handle high-dimensional data. The hyperparameter, regularization parameter (alpha), was tuned using cross-validation to balance model complexity and prediction accuracy.

Long Short-Term Memory Networks (LSTMs): LSTMs were considered for their capability in capturing long-term dependencies within sequential data, which might exist in certain water quality parameters.

Bidirectional Long Short-Term Memory Networks (bi-LSTMs): To leverage potential information from both past and future data points, bi-directional LSTMs were also explored. The model architecture was similar to LSTMs, but with a bi-directional layer that processes information in both directions.

Artificial Neural Networks (ANNs): ANNs were evaluated for their ability to learn complex, non-linear relationships. A feed-forward ANN architecture and was employed.

D).Model Training and Evaluation: Using a [split ratio, 70:30] ratio, the data was divided into training, validation, and testing sets. The validation set was used to fine-tune the models' hyperparameters after they had been trained on the training set. The unseen testing set was used for the final performance assessment. To assess the performance of the model, criteria like recall, accuracy, precision, and F1-score were used, depending on the prediction task (classification vs. regression). For water quality parameter prediction (regression tasks), metrics like mean squared error (MSE) and R-squared were used.

Metrics Derived from Confusion Matrix:

1. **Accuracy:**

Accuracy = (TP + TN) / (Total Number of Instances)

2. **Precision:**

Measures the proportion of predicted positives that were actually positive.

Precision = TP / (TP + FP)

3. **MSE(Mean Square Error):**

A metric used to evaluate the performance of regression models, which predict continuous values.

Formula : MSE = $(\Sigma (yi - \hat{y}\_i)^2) / n$

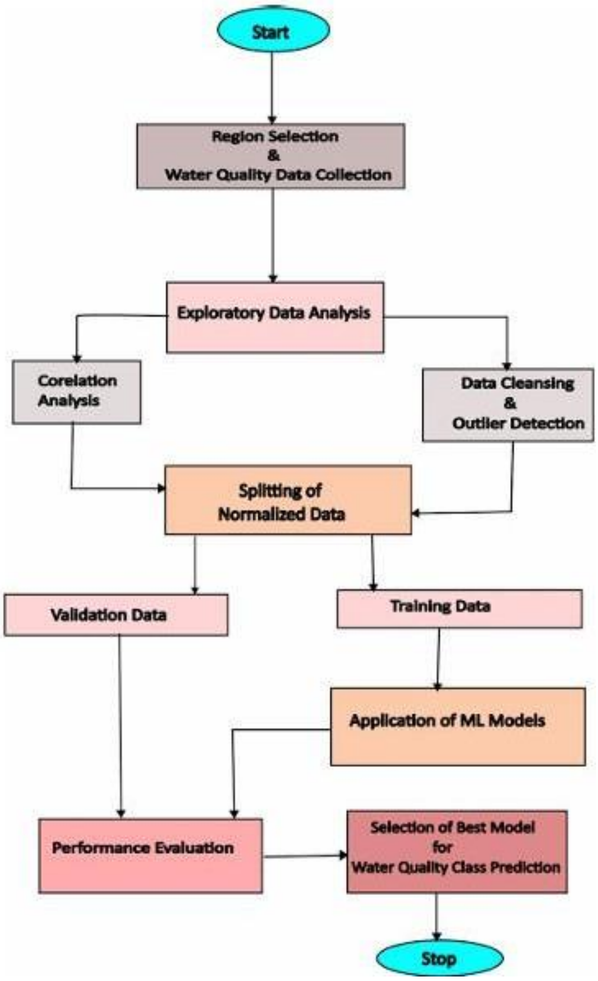


Figure 1: The architecture diagram of machine learning and deep learning models.

IV. RESULTS AND DISCUSSION

Table1: Performance Metrics

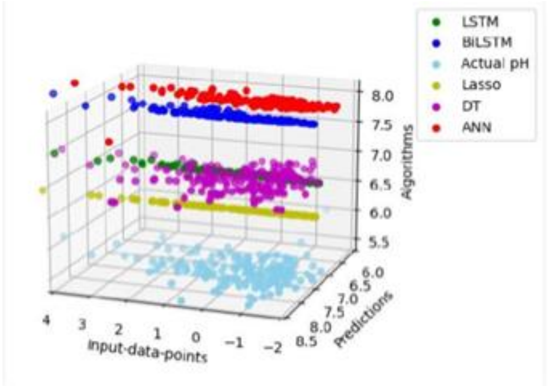
S.NO	Model Name	MSE
1.	Lasso	0.228
2.	Decision tree	0.183
3.	LSTM	0.469
4.	BI-LSTM	0.215
5.	ANN	14.365

Regression models are frequently assessed using the mean squared error as a performance metric. The average squared difference between the expected and actual values is represented by the MSE, and a smaller MSE denotes better predictive accuracy.

In this case, we can observe that the Decision Tree model has the lowest MSE of 0.183, followed by BI-LSTM with an MSE of 0.215. LASSO has an MSE of 0.228, which is slightly higher than the previous two models but still relatively low. LSTM has the highest MSE of 0.469, indicating poorer performance compared to the other models. Finally, the ANN model has a significantly higher MSE of 14.365, suggesting it performs the worst among the models listed.

Based on these results, it appears that the Decision Tree and BI-LSTM models are the most suitable for water quality analysis, as they have the lowest MSE values and therefore provide more accurate predictions.

Figure 2: Charting clarity with machine learning and deep learning algorithms through the lens of a confusion matrix.



The health of the environment depends on the ability to forecast water quality. This work examined the approaches taken by various machine learning algorithms to this problem. Lasso regression, artificial neural networks (ANNs), Decision Trees (DTs), and deep learning models like Long Short-Term Memory and BiLSTMs were compared and studied. As our 3D image illustrates, DT achieved the highest efficiency, maybe in terms of accuracy. This suggests that the most useful tool might be a DT, particularly for our non-sequential water quality data. While there are advantages to deep learning techniques, depending on the task and data, a more straightforward and maybe faster approach, such as a DT, can be similarly effective. This result underscores the significance of comparing different algorithms to determine which one best suits your unique water quality.

Figure 3: Charting clarity with machine learning and deep learning algorithms through the lens of a confusion matrix.

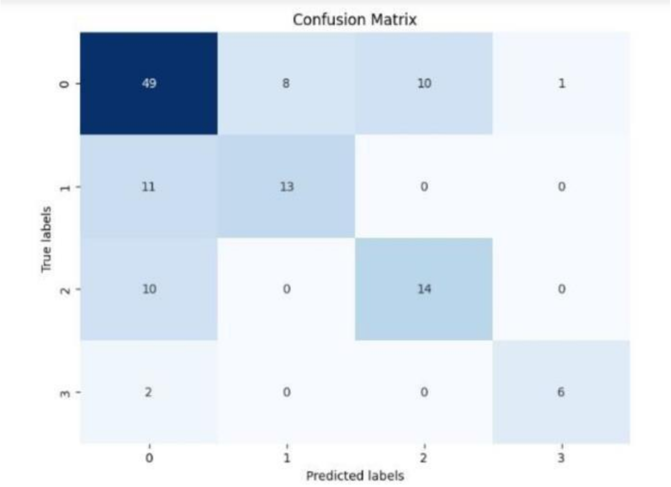
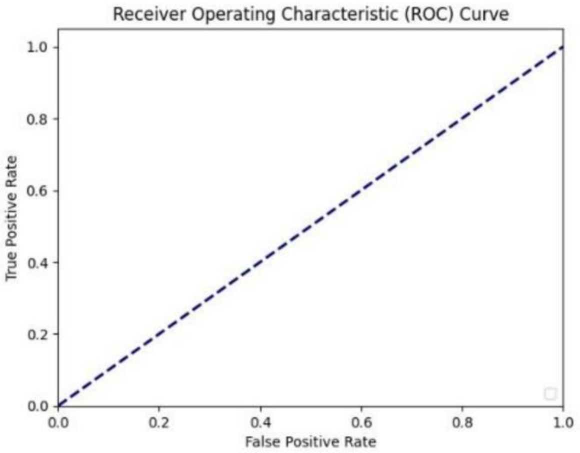


Figure 4: Plotting precision with machine learning and deep learning algorithms, depicted through the ROC curve's ebb and flow.



It is a graphic depiction of a classification model's performance at different categorization criteria. It displays the results of a model that assigns a positive or negative classification to various entities. The fraction of negative cases that were mistakenly classified as positive is known as the False Positive Rate (FPR), and it is plotted on the x-axis. The True Positive Rate (TPR), or the percentage of positive cases that were accurately categorized as positive, is plotted on the y-axis.

Plotting a ROC curve usually involves placing the FPR on the x-axis and the TPR on the y-axis. A point in the upper left corner of the graph would represent a perfect test, with a TPR of 1 and an FPR of 0. A diagonal line from (0, 0) to (1, 1) would represent a random guess. In general, a model's performance improves the closer the ROC curve is to the upper left corner.

V. CONCLUSION

Using a variety of machine learning (ML) and deep learning (DL) methods, such as Decision Trees (DT), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Artificial Neural Networks (ANN), and Lasso regression, this study provides a thorough examination of water quality. Our investigation not only evaluated algorithm performance within our dataset but also compared our findings with existing research papers employing similar methodologies.

Upon examination, our study found that the Decision Tree algorithm consistently outperformed LSTM, BiLSTM, ANN, and Lasso regression in terms of accuracy for water quality analysis. This observation is particularly significant when considering the performance of these algorithms in previous research. Despite the prevalence of LSTM, BiLSTM, and ANN in water quality analysis literature, our results suggest that Decision Trees offer a more efficient approach for our specific dataset and analysis objectives.

Several factors may contribute to the superior performance of Decision Trees, including dataset characteristics, model interpretability, and the algorithm's ability to capture non-linear relationships effectively. While LSTM, BiLSTM, and ANN remain popular choices in the field, our results underscore the potential advantages of exploring alternative algorithms such as Decision Trees for certain applications.

VI. REFERENCES

[1] Martínez, Ramón., Vela, N.., Aatik, A. el., Murray, Eoin., Roche, Patrick., & Navarro, J. M.. (2020). On the Use of an IoT Integrated System for Water Quality Monitoring and Management in Wastewater Treatment Plants. Water . http://doi.org/10.3390/w12041096

[2] Park, Jungsu., Kim, Keugtae., & Lee, Woo-Hyoung. (2020). Recent Advances in Information and Communications Technology (ICT) and Sensor Technology for Monitoring Water Quality. Water . http://doi.org/10.3390/w12020510

[3] Wang, Weiliang., Liu, Xiaohui., Wang, Yu-fan., Guo, Xiaochun., & Lu, Shaoyong. (2016). Analysis of point source pollution and water environmental quality variation trends in the Nansi Lake basin from 2002 to 2012. Environmental Science and Pollution Research , 23 , 4886-4897 . http://doi.org/10.1007/s11356-015-5625-x

[4] Osta, M. El., Masoud, M.., Alqarawy, Abdulaziz M.., Elsayed, S.., & Gad, M.. (2022). Groundwater Suitability for Drinking and Irrigation Using Water Quality Indices and Multivariate Modeling in Makkah Al-Mukarramah Province, Saudi Arabia. Water . http://doi.org/10.3390/w14030483

[5] Gradilla-Hernández, M.., Anda, J. de., García-González, A.., Meza-Rodríguez, Demetrio., Montes, Carlos Yebra., & Perfecto-Avalos, Y.. (2019). Multivariate water quality analysis of Lake Cajititlán, Mexico. Environmental Monitoring and Assessment , 192 . http://doi.org/10.1007/s10661-019-7972-4

[6] Wang, Xiaoping., & Zhang, Fei. (2018). Multi-scale analysis of the relationship between landscape patterns and a water quality index (WQI) based on a stepwise linear regression (SLR) and geographically weighted regression (GWR) in the Ebinur Lake oasis. Environmental Science and Pollution Research , 25 , 7033-7048 . http://doi.org/10.1007/s11356-017-1041-8

[7] Jan, Farmanullah., Min-Allah, Nasro., & Düştegör, Dilek. (2021). IoT Based Smart Water Quality Monitoring: Recent Techniques, Trends and Challenges for Domestic Applications. Water . http://doi.org/10.3390/w13131729

[8] Santana, Mark V. E.., Zhang, Qiong., & Mihelcic, J.. (2014). Influence of water quality on the embodied energy of drinking water treatment.. Environmental science & technology , 48 5 , 3084-91 . http://doi.org/10.1021/es404300y

[9] Ma, Ruonan., Yu, Shuang., Tian, Ying., Wang, Kaile., Sun, Chongde., Li, Xian., Zhang, Jue., Chen, Kun-song., & Fang, Jing. (2016). Effect of Non-Thermal Plasma-Activated Water on Fruit Decay and Quality in Postharvest Chinese Bayberries. Food and Bioprocess Technology , 9 , 1825-1834 . http://doi.org/10.1007/s11947-016-1761-7

[10] . Palansooriya, K. N.., Yang, Y.., Tsang, Y.., Sarkar, B.., Hou, D.., Cao, Xinde., Meers, E.., Rinklebe, J.., Kim, Ki-Hyun., & Ok, Y.. (2020). Occurrence of contaminants in drinking water sources and the potential of biochar for water quality improvement: A review. Critical Reviews in Environmental Science and Technology , 50 , 549 - 611 . http://doi.org/10.1080/10643389.2019.1629803

[11] Lowe, Matthew., Qin, Ruwen., & Mao, X.. (2022). A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. Water . http://doi.org/10.3390/w14091384

[12] . Benedetti, L.., Langeveld, J.., Comeau, A.., Corominas, L.., Daigger, G.., Martin, Cristina., Mikkelsen, P.., Vezzaro, L.., Weijers, S.., & Vanrolleghem, P.. (2013). Modelling and monitoring of integrated urban wastewater systems: review on status and perspectives.. Water science and technology : a journal of the International Association on Water Pollution Research , 68 6 , 1203-15 . http://doi.org/10.2166/wst.2013.397

[13] . Xu, Baoguo., Zhang, Min., Bhandari, B.., Cheng, Xinfeng., & Sun, Jin-cai. (2015). Effect of Ultrasound Immersion Freezing on the Quality Attributes and Water Distributions of Wrapped Red Radish. Food and Bioprocess Technology , 8 , 1366-1376 . http://doi.org/10.1007/s11947-015-1496-x

[14] Stehle, Sebastian., Elsaesser, D.., Grégoire, C.., Imfeld, G.., Niehaus, E.., Passeport, E.., Payraudeau, S.., Schäfer, R.., Tournebize, J.., & Schulz, R.. (2011). Pesticide risk mitigation by vegetated treatment systems: a meta analysis.. Journal of environmental quality , 40 4 , 1068-80 . http://doi.org/10.2134/jeq2010.0510

[15] Escher, B.., Allinson, M.., Allinson, M.., Altenburger, R.., Bain, P.., Balaguer, P.., Busch, W.., Crago, J.., Denslow, N.., Dopp, E.., Hilscherová, K.., Humpage, A.., Kumar, Anu., Grimaldi, Marina., Jayasinghe, B.., Jarošová, B.., Jia, A.., Makarov, S.., Maruya, K.., Medvedev, Alexander V.., Mehinto, A.., Mendez, J. E.., Poulsen, Anita H.., Procházka, Erik., Richard, J.., Schifferli, A.., Schlenk, D.., Scholz, S.., Shiraishi, F.., Snyder, S.., Su, Guanyong., Tang, J.., Burg, B.., Linden, S. V. D.., Werner, I.., Westerheide, S.., Wong, C.., Yang, Min., Yeung, B.., Zhang, Xiaowei., & Leusch, F.. (2014). Benchmarking organic micropollutants in wastewater, recycled water and drinking water with in vitro bioassays.. Environmental science & technology , 48 3 , 1940-56 . http://doi.org/10.1021/es403899t