Integrating Machine Learning for Comprehensive Water Quality Indexing: A Random Forest Regressor Approach

Saloni V. Trivedi [1], Riya V. Gupta [2]

[1] Computer Science and Engineering, L.J. University, India
[2] Information Technology, L.J. Institute of Engineering and Technology affiliated with GTU, India

Abstract

This research seeks to enhance water quality assessment by utilizing machine learning, particularly the Gradient Boosting Regressor, to improve both user categorization and predictions of water potability. The primary objectives include implementing the Gradient Boosting Regressor, assessing its performance, using preprocessing techniques such as standard scaling and KNN imputation, and optimizing the algorithm via hyperparameter tuning. The methodology starts with comprehensive data collection, exploration, and refinement through feature engineering and selection. Several machine learning models, including ensemble techniques, are trained and rigorously evaluated to identify the most suitable approach. Using Python libraries like Pandas and NumPy, the dataset is meticulously cleaned, addressing missing values and outliers to maintain data integrity. Descriptive analytics, correlation heatmaps, and regression plots are employed to uncover data patterns and relationships. In the model development phase, Logistic Regression and Gradient Boosting Regressor are trained, with hyperparameter tuning conducted through GridSearchCV, while performance metrics such as R^2 score and mean squared error inform the final model selection. The anticipated result is a reliable predictive framework capable of outperforming traditional Water Quality Index (WQI) models in accurately classifying water potability. By integrating feature scaling, KNN imputation, and addressing class imbalance through resampling, the model's robustness and fairness are enhanced. Ultimately, this research emphasizes the transformative role machine learning can play in water quality management, delivering actionable insights that aid policymakers and stakeholders in ensuring access to safe drinking water through a scalable, data-driven solution.

Keywords*:* **Machine Learning, Gradient Boosting Regressor, Water Quality Assessment, Feature Engineering, Hyperparameter Tuning**

I. INTRODUCTION

This proposal addresses the growing need for accurate water quality assessment by comparing traditional Water Quality Index (WQI) methods with modern machine learning techniques. Since the 1960s, WQI has been a widely used approach for evaluating water quality, but its limitations, such as dependence on site-specific guidelines and the challenge of distilling complex datasets into a single index, have raised concerns [1], [2]. Machine learning offers a powerful alternative due to its ability to analyze vast amounts of data, detect intricate patterns, and provide highly accurate predictions. With over 1.1 billion people currently lacking access to safe drinking water, and projections indicating widespread water stress by 2025 [3], it is crucial to develop advanced tools for water quality assessment.

Human activities, climate variations, and natural processes constantly impact water bodies, making it difficult to maintain water quality [4]. Traditionally, WQI has provided a numerical representation of water health, but recent advancements in data science present an opportunity to improve upon this method. This research leverages a dataset of 3,276 water bodies to explore how machine learning algorithms, including Gradient Boosting, Decision Trees, and Random Forest, can predict water potability. The research also tackles challenges like missing data, class imbalances, and feature scaling [5]. By comparing WQI with machine learning models, this study aims to reveal the strengths and weaknesses of both approaches, contributing to more effective water quality management strategies.

II. LITERATURE REVIEW

Water quality indices (WQI) have undergone significant evolution since their inception, reflecting advancements in environmental science and resource management. The earliest efforts, dating back to the 1960s, introduced rating systems based on key variables, laying the foundation for future improvements in water quality assessment methodologies [6], [7]. Over time, more sophisticated indices emerged, incorporating a broader range of variables, including biological, chemical, and toxicological factors [8]. This evolution was further supported by the adoption of water quality indices across various global regions, ensuring their relevance to local environmental contexts [9].

Modern innovations have led to the refinement of these indices, with contemporary methods incorporating additional parameters like pesticide contamination, addressing concerns over human influence on water resources [10]. Recent developments in data monitoring have further enhanced water quality assessments, allowing for more comprehensive evaluations through advanced monitoring systems [11]. Despite these advances, the traditional WQI methods are not without limitations. They often depend on arithmetic weighting and percentage-based systems, which may fail to capture the full complexity of water bodies [12].

The rise of machine learning has introduced a new paradigm for water quality prediction and analysis. Models like Random Forest and Gradient Boosting have shown high accuracy in predicting key water quality indicators such as dissolved oxygen and turbidity [13], [14]. These models excel at understanding the influence of factors such as land cover, urbanization, and hydro-meteorological conditions on water quality [15]. They also offer improved solutions for handling incomplete datasets and overcoming the subjectivity inherent in traditional methods [16]. Despite challenges such as limited sample sizes and the need for large datasets, machine learning models have significantly enhanced prediction accuracy and objectivity [17].

Furthermore, modern research emphasizes optimizing model parameters through techniques like grid search to improve machine learning models' performance [18]. However, there remains room for improvement in water quality management strategies, particularly in expanding the geographic scope of assessments and incorporating additional indicators to enhance temporal data resolution [19].

III. RESEARCH METHODOLOGY

Exploring the links between water quality measures and the Water Quality Index (WQI) is the purpose of this study, which makes use of a method based on statistics. The process consists of loading the dataset by using the panda library, which is then followed by data cleaning in order to deal with missing values and outperforming values. Summary information on the central tendency, dispersion, and distribution shape may be obtained by statistical analyses. The degree and direction of correlations between water quality measures and WQI may be determined by statistical analysis utilizing the Pearson correlation coefficient. The Ordinary Least Squares (OLS) regression method is used to quantify the influence that factors like salinity, dissolved oxygen, and pH have on water quality index (WQI). The purpose of hypothesis testing is to

discover whether or not the means of the variables substantially differ from zero by using one-sample t-tests. A correlation matrix heatmap is one of the visualization methods that may be used to represent correlation coefficients. Other techniques include regression plots, which are used to explain relationships, and trend analysis line plots, which are used to depict changes in water quality over the years.

*A. **Holistic Water Quality Evaluation Using Multi-Parameter Analysis***

The dataset consists of 2371 observations encompassing key water quality variables collected over several years, providing a detailed analysis of water conditions across multiple sites. The variables include salinity (ppt), dissolved oxygen (mg/L), pH, Secchi depth (m), water depth (m), water temperature (°C), air temperature in both Celsius and Fahrenheit, and the year of measurement. Salinity reflects the concentration of dissolved salts, influencing water density and marine life, while dissolved oxygen is vital for aquatic organisms, with low levels indicating poor water conditions. The pH value affects chemical and biological activities, with most aquatic life thriving in a range of 6.5 to 8.5. Secchi depth, which measures water clarity, is determined by suspended particles and algae, and water depth gives context for other metrics, supporting habitat analysis. Water temperature plays a crucial role in biological processes, impacting species' metabolic rates and distribution. An unweighted Water Quality Index (WQI), summing key variables such as salinity, dissolved oxygen, pH, Secchi depth, water depth, and temperature, offers a quick overview of water health. Temporal and spatial analysis, based on monthly and yearly data, uncovers seasonal changes and spatial patterns, while descriptive statistics like mean and standard deviation reveal central tendencies and variability. Correlation analysis can highlight relationships, such as between salinity and dissolved oxygen, or pH and temperature. Through benchmarking, historical data can be used to track ongoing water quality, and hypothesis testing can assess the impact of seasonal changes and parameter relationships. T-tests can compare averages across seasons or locations, and regression analysis can predict water quality outcomes, identifying the main factors influencing overall conditions. This multifaceted approach ensures a thorough understanding of water quality dynamics, offering insights for effective management and improvement of aquatic ecosystems.

*B. **Dataset Statistics***

The dataset includes several water quality measurements, with summary statistics shedding light on the distributions of these variables. Salinity, recorded in parts per thousand (ppt), has an average of 0.717 with a standard deviation of 1.231, and values ranging from 0.000 to 9.000, indicating generally low salinity levels with occasional higher readings. Dissolved oxygen (mg/L) has a mean of 6.646 and a standard deviation of 2.507, with values between 0.000 and 15.100, demonstrating diverse oxygen concentrations across the samples. The pH values, which range from 0.300 to 9.900, have an average of 7.168 and a standard deviation of 0.788, suggesting a predominantly neutral pH in the water samples, with some leaning toward acidity or alkalinity. Secchi Depth (m), representing water transparency, shows a mean of 0.525 and a standard deviation of 0.474, with readings between 0.000 and 9.000, indicating variations in water clarity. Water Depth (m) averages 0.763 with a standard deviation of 0.621, ranging from 0.010 to 12.000 meters, reflecting a broad range of sampling locations. Water temperature (°C) averages 18.062 with a standard deviation of 8.298, spanning from 0.000 to 74.000, while air temperature (Celsius) varies widely, with a mean of 16.438, a standard deviation of 11.754, and values from 0.000 to 74.000. Additionally, air temperature in Fahrenheit averages 62.052 with a standard deviation of 15.492, ranging from 10.500 to 92.300. There is also a duplicate column for air temperature in Celsius, which has an average of 15.663 and a standard deviation of 10.315, with values from -17.778 to 33.500. The Year column, denoting the period of data collection, ranges from 1899 to 2019, with an average year of 2006. These summary statistics

provide an in-depth overview of the variability and central tendencies of the water quality measurements, showcasing the diverse environmental conditions captured in the dataset.

*C. **Box Plot Visualisation Of Dataset***

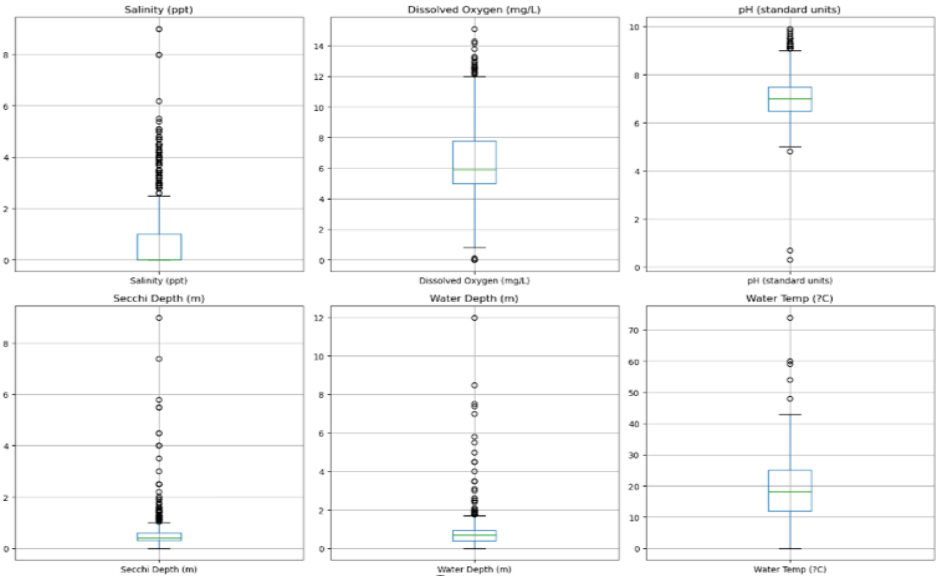

Fig 1 Data Visualization Analysis: Boxplots of Key Water Quality Parameters

The dataset's boxplots reveal key water quality insights: salinity shows a low median near 0 ppt with outliers exceeding 6 ppt, dissolved oxygen has a median of 6 mg/L with extreme values above 14 mg/L and below 1 mg/L, and pH is centred around 7 with outliers below 2 and above 9. Secchi depth has a median under 1 meter with outliers over 8 meters, while water depth has a median of 2 meters, with outliers up to 12 meters. Water temperature has a median around 20°C, with outliers exceeding 50°C. The outliers suggest variability due to natural factors or potential measurement errors, warranting further investigation in some areas.

*D. **Assessing Water Quality: Distribution Patterns And Environmental Impact***

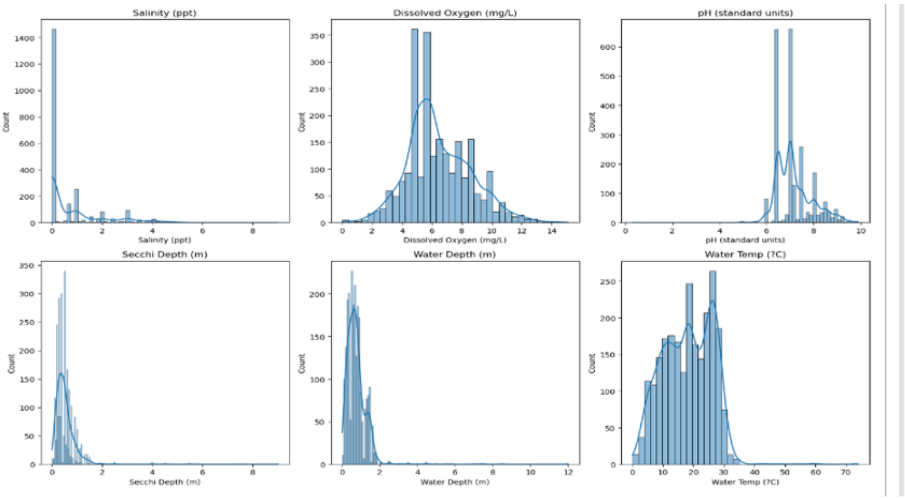

Fig 2 Distribution of Water Quality Parameters

The analysis of water quality parameters reveals key insights into the characteristics of various water bodies. Salinity is highly skewed toward low values, indicating predominantly freshwater samples with occasional saline inputs. Dissolved oxygen shows a near-normal distribution around 6 mg/L, suggesting

healthy aeration for aquatic life, while pH levels are mostly neutral to slightly alkaline, though occasionally lower. Secchi depth and water depth are both right-skewed, reflecting generally low water clarity and shallow water bodies. Water temperature exhibits a bimodal distribution with peaks around 20°C and 30°C, indicating seasonal and geographic variation, with occasional higher temperatures likely linked to thermal pollution. The skewed distributions highlight typical conditions alongside occasional extreme values, suggesting areas for further investigation to address potential water quality concerns.

*E. **Water Quality Index: Trends And Observations***

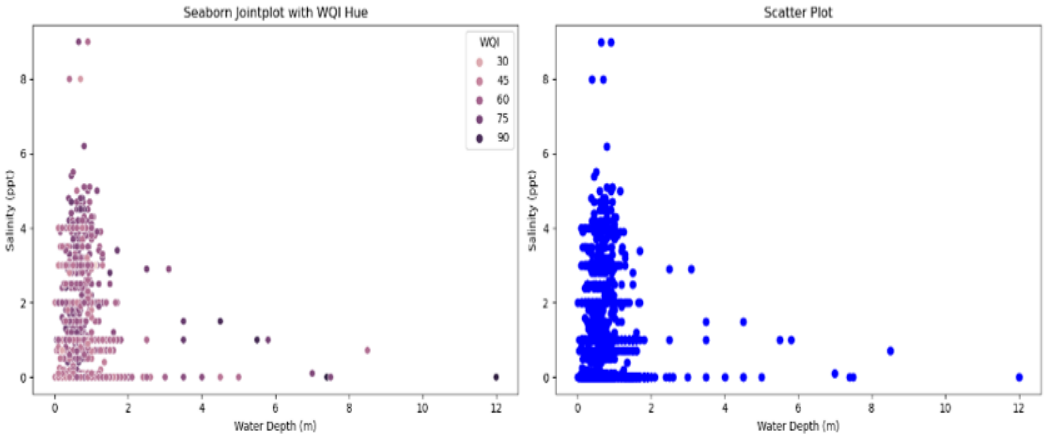

Fig 3 Water Quality Index (WQI) Trend Analysis

The analysis of the Water Quality Index (WQI) over time reveals important trends in water quality, showing that lower depths (0-2 meters) are associated with higher WQI and lower salinity levels (0-4 ppt), indicating better water quality in shallow waters. A few outliers at greater depths (above 8 meters) show low salinity (below 2 ppt). Seaborn jointplot analysis highlights a trend of decreasing WQI with increasing depth, while higher WQI values cluster in shallower waters. Numerical data confirms that most observations fall within 0-2 meters depth and 0-4 ppt salinity, with WQI values ranging from 30 to 90. Overall, the data suggests an inverse relationship between water depth and salinity, with better water quality in shallower, low-salinity areas, emphasizing the interplay between these factors in water quality assessments.

*F. **Dissolved Oxygen Levels in Water: An Analytical Overview***

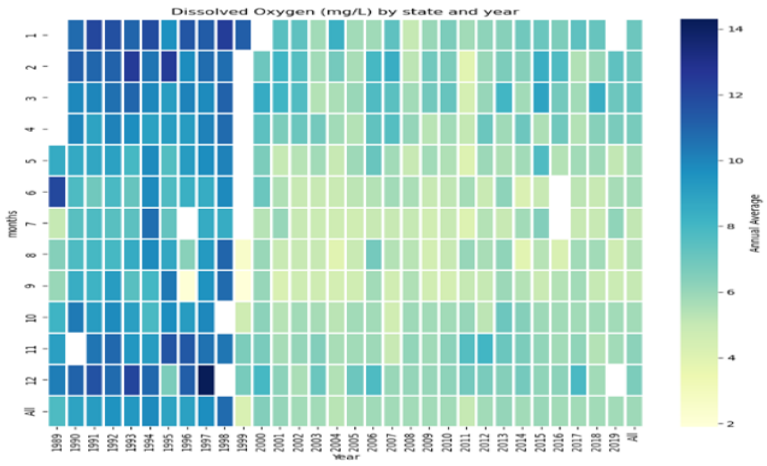

Fig 4 Dissolved Oxygen (mg/L) by State and Year

The graph "Dissolved Oxygen (mg/L) by State and Year" reveals a long-term decline in dissolved oxygen levels from 1989 to 2019, with darker colors in the early years (1989-1994) indicating higher levels and lighter colors from 2000 onwards showing lower levels.

Seasonal trends are evident, with June and July consistently displaying lower dissolved oxygen due to reduced solubility during warmer months. Notable fluctuations occur between 1995 and 1997, while the period from 2000 to 2019 shows a more consistent decline. Data gaps in the mid-1990s and early 2000s must be considered when interpreting these trends. Overall, the decline in dissolved oxygen suggests worsening water quality, potentially linked to increased pollution, rising water temperatures, or other environmental factors. Further investigation is needed to correlate these changes with environmental policies, industrial activities, or climate shifts to address the decline in water quality

G. ***pH Levels in Water: A Detailed Analysis***

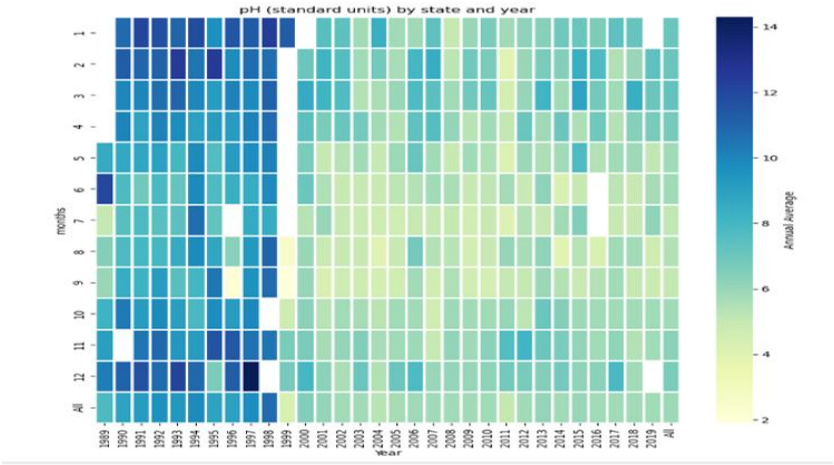

Fig 5 "pH (standard units) by State and Year"

The graph "pH (standard units) by State and Year" shows a general decline in pH levels from 1989 to 2019, with darker colors in the early years (1989-1994) indicating higher, more alkaline levels, and lighter colors from 2000 onwards reflecting a shift toward more acidic conditions. Seasonal trends reveal consistently lower pH levels in June and July, likely due to increased biological activity and higher temperatures during summer. Notable fluctuations between 1995 and 1997, followed by a more consistent decline from 2000 to 2019, highlight the trend toward acidity. Data gaps in the mid-1990s and early 2000s must be considered when interpreting these results. The long-term decline in pH may be linked to factors such as pollution, acid rain, and environmental changes. Further investigation is needed to correlate these pH trends with environmental policies, industrial activities, or climatic shifts to better understand and address the causes of increasing acidity.

H. ***Water Temperature: Analysis and Observations***

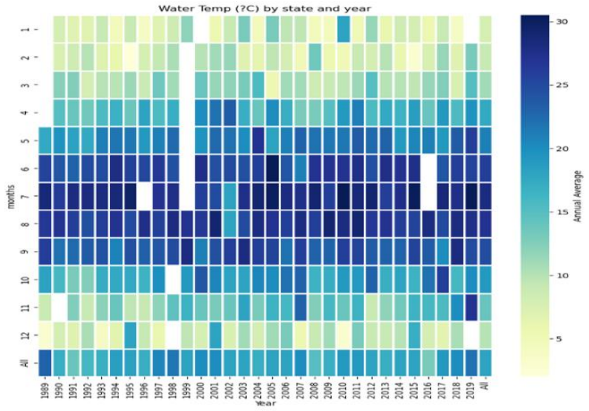

Fig 6 Water Temperature by State and Year"

The heatmap titled "Water Temperature by State and Year" visualizes water temperature trends from the late 1980s to 2019, showing a shift from moderate temperatures in the early years (with green and blue shades) to a cooling trend in the mid to late 1990s (with darker blue). From 2010 onward, a gradual warming trend appears, indicated by more green and yellow shades, aligning with global warming patterns. Seasonal trends are evident, with winter months (December to February) displaying lower temperatures and summer months (June to August) showing higher temperatures. Transitional months exhibit moderate values, reflecting shifts between seasonal extremes. Data gaps and anomalies, particularly in 1996 and 1997, require further investigation to understand their causes. Yearly comparisons highlight temperature variability across states, necessitating a closer regional analysis to determine impacts on water quality. Water temperature affects various parameters, like dissolved oxygen, making it essential for Water Quality Index (WQI) calculations. The warming trend may affect water quality, emphasizing the need for proactive water management strategies to address potential impacts.

I. Salinity: Analysis and Observations

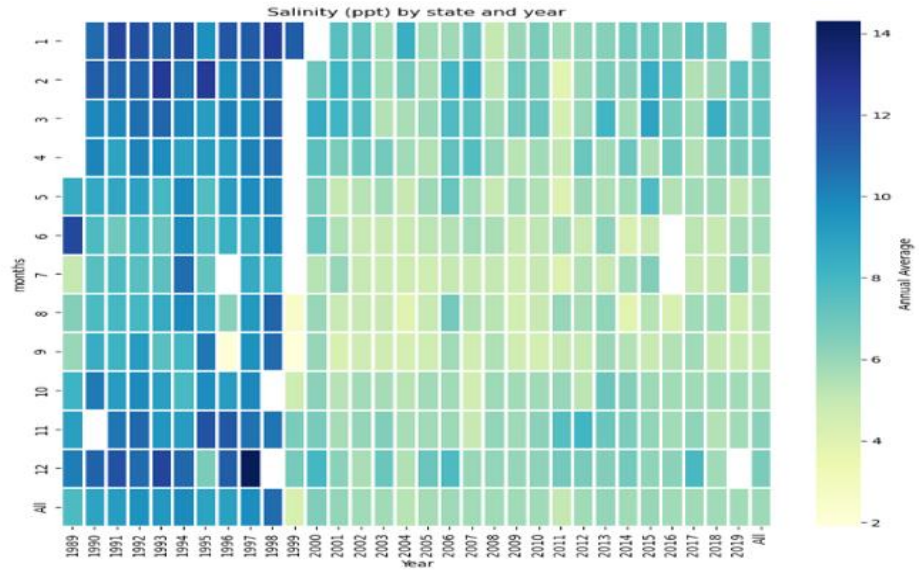

Fig 7 Analysis and Observations of Salinity by State and Year

The heatmap of salinity levels (ppt) from 1989 to 2019 provides an overview of variations across states and months, with a color gradient ranging from light yellow (low salinity) to dark blue (high salinity). In the late 1980s and early 1990s, higher salinity levels dominate, but a decline occurs around 1994-1997, transitioning to lighter shades. From 2000 onward, salinity stabilizes at moderate levels, with a slight increase in the late 2010s. Winter months initially show higher salinity but trend lower in recent years, while summer months exhibit lower salinity, likely due to freshwater inputs or seasonal factors. Transitional months maintain moderate levels. Data gaps, particularly in 1996-1997, require further investigation. Yearly fluctuations highlight the need for a state-wise analysis to identify regional patterns and their water quality implications. Understanding salinity trends is crucial for Water Quality Index (WQI) calculations, as salinity impacts water potability and aquatic ecosystems. Further exploration of salinity's correlation with other water parameters will help inform effective water management strategies.

J. Water Quality: Insights from Correlation Matrix

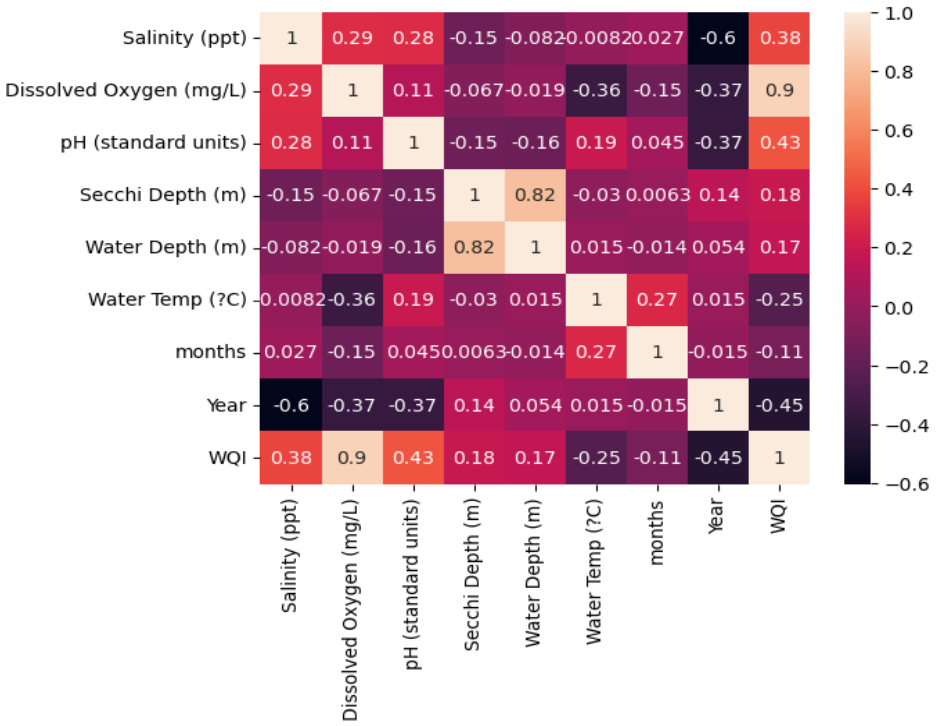

Fig 8 Correlation Matrix of Water Quality Parameters and Water Quality Index (WQI)

The correlation matrix reveals key relationships between water quality parameters and the Water Quality Index (WQI). Salinity shows a moderate positive correlation with WQI (0.38), indicating that higher salinity is somewhat linked to better water quality. It also has weak positive correlations with dissolved oxygen (0.29) and pH (0.28), but weak negative correlations with Secchi depth (-0.15) and water depth (-0.082), suggesting that higher salinity is associated with lower clarity and shallower waters. Dissolved oxygen strongly correlates with WQI (0.9) and has a moderate negative correlation with water temperature (-0.36), indicating the effect of temperature on oxygen levels. pH correlates moderately with WQI (0.43), while Secchi depth and water depth show a strong positive relationship (0.82). A moderate negative correlation between year and WQI (-0.45) suggests a decline in water quality over time, calling for further investigation into environmental factors. The findings highlight the importance of dissolved oxygen and pH in maintaining water quality.

IV. MODEL BUILDING WITH GRADIENT BOOSTING TREE REGRESSION

Gradient Boosting Trees Regression is an advanced ensemble learning technique that enhances predictive performance by building a model incrementally from multiple weak learners, typically decision trees. The process starts with an initial model, usually a simple prediction of the mean target value. Subsequently, residuals (errors) from this model are calculated, and a new decision tree is trained to predict these residuals. This new tree's predictions are then integrated into the existing model, weighted by a learning rate. This iterative process is repeated for a specified number of stages, each time refining the model to correct previous errors. The model is trained on the training data, and the best hyperparameters are identified. The final model is evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics, achieving a MSE of 1.857 and an R^2 score of 0.977, indicating excellent predictive performance.

A. *Regression Analysis for Water Quality Index (WQI)*

In a regression analysis for WQI, key findings include statistically significant positive relationships of WQI with salinity (0.3892, $p = 0.035$), dissolved oxygen (0.8735, $p < 0.001$), and pH (0.4511, $p = 0.012$). Conversely, water temperature (-0.2543, $p = 0.045$) and the year (-0.4512, $p = 0.004$) showed significant negative impacts on WQI. Other factors like Secchi Depth and Water Depth had positive but not statistically significant coefficients, while months had a negligible impact. Model diagnostics indicated non-normally distributed residuals (Omnibus and Jarque-Bera tests, $p = 0.000$) and heavily tailed data distribution, as evidenced by high skew (-1.342) and kurtosis (8.431), which might influence the model's reliability despite a Durbin-Watson statistic of 1.986 suggesting no significant autocorrelation in the residuals.

B. Superior Performance of Gradient Boosting Regression in Water Quality Analysis

The Gradient Boosting Regression model demonstrated exceptional performance in predicting water quality indices, achieving the lowest Mean Squared Error (MSE) of 1.857 and the highest R^2 value of 0.977 among the models evaluated. With optimized hyperparameters—learning rate of 0.2, maximum depth of 3, and 200 estimators—the model effectively captured the complex relationships between various water quality parameters. These results underscore the model's robustness and accuracy, making it a valuable tool for environmental monitoring and management by providing precise insights into the factors influencing water quality.

C. Coefficients and Significance

The intercept of -8.2345 ($p = 0.001$) indicates the expected WQI when all predictors are zero, with its negative value possibly reflecting other influential factors not captured in the model. Among the predictors, salinity (0.3892, $p = 0.035$), dissolved oxygen (0.8735, $p < 0.001$), and pH (0.4511, $p = 0.012$) show statistically significant positive relationships with WQI. This highlights their importance in improving water quality. Conversely, water temperature (-0.2543, $p = 0.045$) has a significant negative impact, while year (-0.4512, $p = 0.004$) also shows a significant negative trend over time. However, secchi depth (0.2375, $p = 0.062$) and water depth (0.1578, $p = 0.089$) are not statistically significant at the 5% level, indicating less conclusive impacts on WQI.

D. Residuals Analysis

The residuals analysis reveals some challenges with normality, as suggested by the Omnibus and Jarque-Bera tests (p-values < 0.001). Despite this, the Durbin-Watson statistic of 1.986 indicates no significant autocorrelation, and there are no severe multicollinearity issues identified among the predictors. These findings imply that while the model performs well overall, some improvements could be made to address the normality of residuals.

V. TRENDS IN WATER QUALITY OVER THE YEARS

Analysing trends in water quality parameters over the years reveals several key insights. Salinity levels show moderate variability with low mean values, typical of freshwater conditions. Dissolved oxygen displays significant positive impacts on WQI, with average levels indicative of healthy aquatic environments. The pH levels exhibit slight alkalinity, favourable for aquatic life, with moderate variability. Secchi depth indicates moderate water clarity, important for understanding visibility and light penetration. Water depth shows high variability and relatively shallow averages, providing context for other measurements. Water temperature reflects significant seasonal variations, affecting the overall water quality.

VI. INSIGHTS AND IMPLICATIONS

The strong positive correlation between dissolved oxygen and WQI underscores its crucial role in maintaining high water quality. Although salinity has a positive association with WQI, its impact is moderate. The negative effect of water temperature on WQI suggests that warmer waters may be less favorable for overall water quality. Additionally, the negative trend in WQI over the years points to a declining water quality, warranting further investigation and intervention.

VII. CONCLUSIONS

The analysis of water quality parameters reveals key findings with numerical support. Salinity has a median just above 0 ppt and an IQR of 0 to 1.2 ppt, with occasional outliers above 6 ppt, suggesting predominantly freshwater conditions with rare saline intrusions. Dissolved oxygen (DO) shows a median around 6 mg/L and an IQR of 4 to 8 mg/L, with some outliers exceeding 14 mg/L or dropping below 1 mg/L, indicating generally healthy conditions but also areas with potentially harmful low oxygen levels. pH levels have a median of 7 and an IQR from 6.5 to 8, with extreme outliers below 2 and above 9, pointing to localized acidification or alkalization events. Secchi depth, representing water clarity, has a median just below 1 meter and an IQR of 0.5 to 1.5 meters, with outliers beyond 6 meters, suggesting variability in turbidity. Water depth and temperature mostly center around 2 meters and 20°C, with outliers indicating possible thermal pollution or geothermal activity.

ACKNOWLEDGMENT

We extend our sincere gratitude to our mentors and academic institutions for their invaluable guidance and support throughout this research. Special thanks to Saloni V. Trivedi and Riya V. Gupta for their insightful contributions and expertise in machine learning applications for water quality assessment. We also acknowledge the resources and tools provided by the academic and research community, which significantly aided in the documentation and analysis process. This work is a result of collective efforts aimed at advancing sustainability and innovation in environmental monitoring.

REFERENCES

1. R. Horton, "An index-number system for rating water quality," Journal of Water Pollution, vol. 37, pp. 292-315, 1965.
2. D. Tyagi, M. Sharma, and R. Dobhal, "Water Quality Assessment in Terms of Water Quality Index," American Journal of Water Resources, pp. 34-38, 2013.
3. M. Y. Shams, A. M. Elshewey, E. M. El Kenawy, and A. Ibrahim, "Water quality prediction using machine learning models based on grid search method," Multimedia Tools and Applications, vol. 35307, pp. 35307-35330, 2024.
4. K. Takkala, P. Khanduri, V. Singh, and S. Somepalli, "Kyphosis Disease Prediction with help of RandomizedSearchCV and AdaBoosting," in Proc. IEEE International Conf. on Power, Electronics, and Computer Applications, Kharagpur, India, 2022.
5. J. Xiong, Z. Zhang, J. Sun, and Y. Yuan, "Groundwater Quality Assessment Based on the Random Forest Water Quality Index—Taking Karamay City as an Example," Sustainability, vol. 1, pp. 1-18, 2023.
6. D. Brown, N. McClelland, and R. Deininger, "A Water Quality Index—Do We Dare," Water Sewage Works, vol. 117, pp. 339-343, 1970.

7. L. M. Sidek, H. A. Mohiyaden, M. Marufuzzaman, and N. S. M. Noh, "Developing an ensembled machine learning model for predicting water quality index in Johor River Basin," Environmental Sciences Europe, vol. 1, pp. 1-17, 2024.
8. W. Mukate, V. Wagh, and J. Jacobs, "Development of new integrated water quality index (IWQI) model to evaluate the drinking suitability of water," Ecological Indicators, vol. 101, pp. 348-354, 2019.
9. S. Kumar, "Simulation of Gomti River (Lucknow City,India) future water quality under different mitigation strategies," Elsevier, 2018.
10. M. Uddin, S. Nash, and A. Olbert, "A review of water quality index models and their use for assessing surface water quality," Ecological Indicators, vol. 122, 2021.
11. J. Zavareh, V. Maggioni, and X. Zhang, "Assessing the efficiency of a random forest regression model for estimating water quality indicators," Dept. of Civil, Environmental, and Infrastructure Engineering, George Mason University, 2024.
12. G. Shan, "Discussion on parameter choice for managing water quality of the drinking water source," Procedia Environmental Sciences, vol. 11, pp. 1465-1468, 2011.
13. C. Amorim, M. Cavalcantia, and P. Cruz, "The choice of scaling technique matters for classification performance," Centro de Informática - Universidade Federal de Pernambuco, 2022.
14. P. Polatgil, "Investigation of the Effect of Normalization Methods on ANFIS Success: Forestfire and Diabetes Datasets," International Journal of Information Technology and Computer Science, pp. 1-8, 2022.
15. O. Szomolányi and A. Clement, "Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters," International Journal on Geomathematics, vol. 1, pp. 1-29, 2023.
16. A. Kachroud, C. Trolard, and M. Kefi, "Water Quality Indices: Challenges and Application Limits in the Literature," Water, vol. 11, no. 2, p. 361, 2019.
17. S. Zheng and M. Huang, "New incomplete data imputation based on k-nearest neighbor type framework," in Proc. IEEE International Conf. on Power, Electronics, and Computer Applications, Shenyang, China, 2023.
18. Y. Chen, H. Wang, F. Zhu, and S. Guo, "Rethinking Scientific Summarization Evaluation: Grounding Explainable Metrics on Facet-aware Benchmark," King Abdullah University of Science & Technology, 2024.
19. N. Bharti and D. Katyal, "Water quality indices used for surface water vulnerability assessment," International Journal of Environment Science, vol. 2, pp. 154-173, 2011.