Journal of Science Engineering Technology and Management Science
Volume 02, Issue 08, August 2025 ISSN: 3049-0952
www.jsetms.com DOI: 10.64771/jsetms.2025.v02.i08.590-601

# ML BASED PREDICTING OF DISSOLVED OXYGEN LEVELS FOR SUSTAINABLE RIVER ECOSYSTEM MANAGEMENT

P. Anupama[1],E. Vinay Kumar[2], S. Ajanth[2], Kondaveni Anugna[2], Jeripothula Venkatesh[2]
[1]Assistant Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering
[1,2] Sree Dattha Institute of Engineering and Science, Sheriguda, Ibrahimpatnam, 501510, Telangana.

***To Cite this Article***

*P. Anupama, E. Vinay Kumar, S. Ajanth, Kondaveni Anugna, Jeripothula Venkatesh, "Ml Based Predicting Of Dissolved Oxygen Levels For Sustainable River Ecosystem Management", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 08, August 2025,pp: 590-601, DOI:* http://doi.org/10.64771/jsetms.2025.v02.i08.pp590-601

*Submitted: 15-07-2025 Accepted: 21-08-2025 Published: 28-08-2025*

**ABSTRACT**

Dissolved oxygen (DO) plays a critical role in maintaining the health and stability of river ecosystems, as aquatic organisms rely on it for respiration and survival. Therefore, monitoring and accurately predicting DO levels is essential for effective environmental management, conservation efforts, and water quality assessment. Traditionally, DO prediction involves the physical collection of water samples from various river locations, followed by laboratory analysis. Based on the collected data, researchers build empirical models to estimate DO levels using related physical and chemical factors. However, this conventional approach has significant limitations, including data gaps due to spatial and temporal constraints, delayed results, high costs, and limited coverage, which hinder real-time monitoring and accurate forecasting. Recognizing these challenges and the importance of maintaining appropriate DO levels, there is a growing need for more efficient and precise prediction methods. Machine learning (ML) emerges as a powerful solution, capable of analyzing large volumes of data and uncovering complex relationships that traditional models may overlook. This study aims to develop a machine learning-based predictive model for DO levels in river water, leveraging both historical and real-time environmental data. Such a model would enable continuous, reliable DO predictions, supporting better ecosystem health assessment and ensuring the well-being of aquatic life. Ultimately, the findings of this research will contribute to improved environmental management, informed conservation strategies, and sustainable water resource planning for future generations.

**Keywords:** Dissolved Oxygen Prediction, River Water Quality, Machine Learning, Environmental Monitoring, Aquatic Ecosystems, Real-Time Analysis, Water Resource Management, Conservation, Sustainable Development

## 1. INTRODUCTION

Predicting dissolved oxygen levels in river water using machine learning is a vital and interdisciplinary approach that addresses critical environmental, ecological, and public health challenges. Dissolved oxygen is a key indicator of water quality and aquatic ecosystem health, and its fluctuations, driven by factors such as pollution, climate change, and land use, can have severe consequences for biodiversity and water resource sustainability. Traditional monitoring methods are limited by inefficiency, low spatial coverage, and delayed response times, making real-time and predictive approaches increasingly necessary. Machine learning models, trained on diverse environmental variables like temperature, pH, turbidity, nutrient levels, and weather data, offer accurate and timely forecasts, enabling early detection of water quality issues and supporting proactive environmental management.

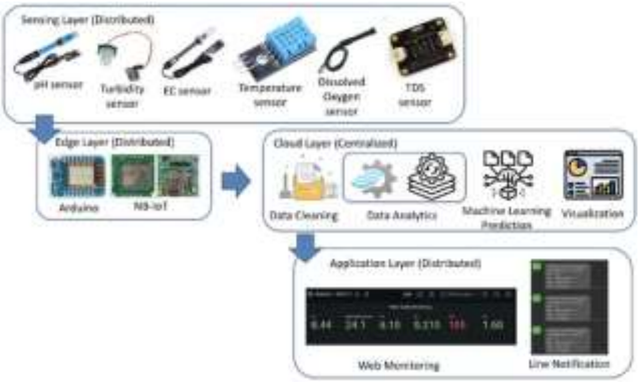


Figure 1: Architecture of the drinking water monitoring system.

These predictive tools find wide-ranging applications including continuous water quality monitoring, aquaculture optimization, regulatory compliance, environmental impact assessments, climate change adaptation, and emergency response. Furthermore, they empower researchers, policymakers, industries, and educators with data-driven insights to enhance conservation strategies, safeguard freshwater ecosystems, and support the sustainable management of water resources, aligning with global sustainability goals such as UN SDG 6.

## 2. LITERATURE SURVEY

Traditional water quality monitoring methods, which rely on manual sampling and laboratory analysis, are inherently limited by low timeliness and insufficient spatiotemporal resolution [**12**]. With recent technological advancements, IoT-based sensors now facilitate continuous monitoring of various physical, chemical, and biological water quality parameters, transmitting data in real time to support more responsive and data-driven water management However, sensor data remains vulnerable to environmental disturbances, fluctuating hydrological conditions, and technical issues such as drift, malfunction, and data loss, particularly in complex environments (e.g., underwater or outdoor settings), which poses challenges to the accuracy and reliability of monitoring results Integrating AI—particularly ML algorithms—offers a promising solution to these challenges AI algorithms can intelligently process raw sensor data by performing real-time calibration, error correction, anomaly detection, and denoising, thereby significantly improving data quality and monitoring accuracy [**8**]. Additionally, AI's powerful pattern recognition capabilities enable automatic diagnosis of sensor faults, self-calibration, and data compensation, thereby enhancing the stability and robustness of monitoring systems. The deployment of AI models at sensor terminals (i.e., edge computing) allows for localized intelligent analysis of data, reducing transmission delays, improving response speed, and granting sensors the ability to adapt to environmental changes and optimize autonomously. This collaborative integration of AI, sensors, and IoT not only greatly improves the accuracy, real-time performance, and intelligence of water quality monitoring but also effectively compensates for the inherent limitations of sensors under complex conditions [**16**]. summarizes current applications of AI algorithms in water quality monitoring, from traditional ML algorithms such as SVM and K-Means to DL models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). AI technology is widely applied in key tasks such as water quality classification, pollutant concentration prediction, image-based pollution detection, and equipment fault diagnosis. These systems typically integrate IoT platforms with various sensors and combine RS or image data to enable low-cost, real-time remote data collection. Their significant advantages include high accuracy, real-time monitoring, low deployment costs, and strong environmental adaptability. The application fields cover various water environments such as drinking water, wastewater treatment, lakes, and aquaculture, significantly promoting the intelligent development of water quality monitoring and laying the foundation for building efficient water environment monitoring systems.

RS technology is particularly valuable for providing large-scale, near-synchronous observations of surface water bodies, which is crucial for assessing the overall condition of vast water areas . AI algorithms enable efficient analysis of RS imagery, allowing for precise extraction of surface water information, and enabling rapid identification of water quality issues such as algal blooms and pollution

## 3. PROPOSED METHODOLOGY

The research begins with acquiring a high-quality dataset containing key river water quality parameters such as temperature, pH, BOD, COD, and total suspended solids, aimed at predicting Dissolved Oxygen (DO) concentration. This dataset, sourced from real-time sensors or trusted repositories, undergoes thorough preprocessing involving missing value treatment, outlier detection, label encoding, feature correlation analysis, and normalization. Initially, a Linear Regression (LR) model is implemented as a benchmark to capture linear relationships between features and DO levels, though it shows limitations in modeling complex nonlinear dependencies. To address this, a novel hybrid Random Forest Regressor (RFR) is proposed, incorporating a two-level stacked ensemble approach enhanced by k-means-based feature clustering and a dynamic feature importance reweighting mechanism. This advanced model significantly improves prediction accuracy, achieving an $R^2$ score of 0.97 compared to 0.71 for the LR model. Performance is evaluated using metrics like MAE, MSE, RMSE, and $R^2$, confirming the superiority of the RFR. Finally, the trained RFR model is tested on new unseen data using the same preprocessing pipeline, demonstrating robust generalization and high predictive accuracy, making it suitable for real-world deployment in river water quality monitoring systems.

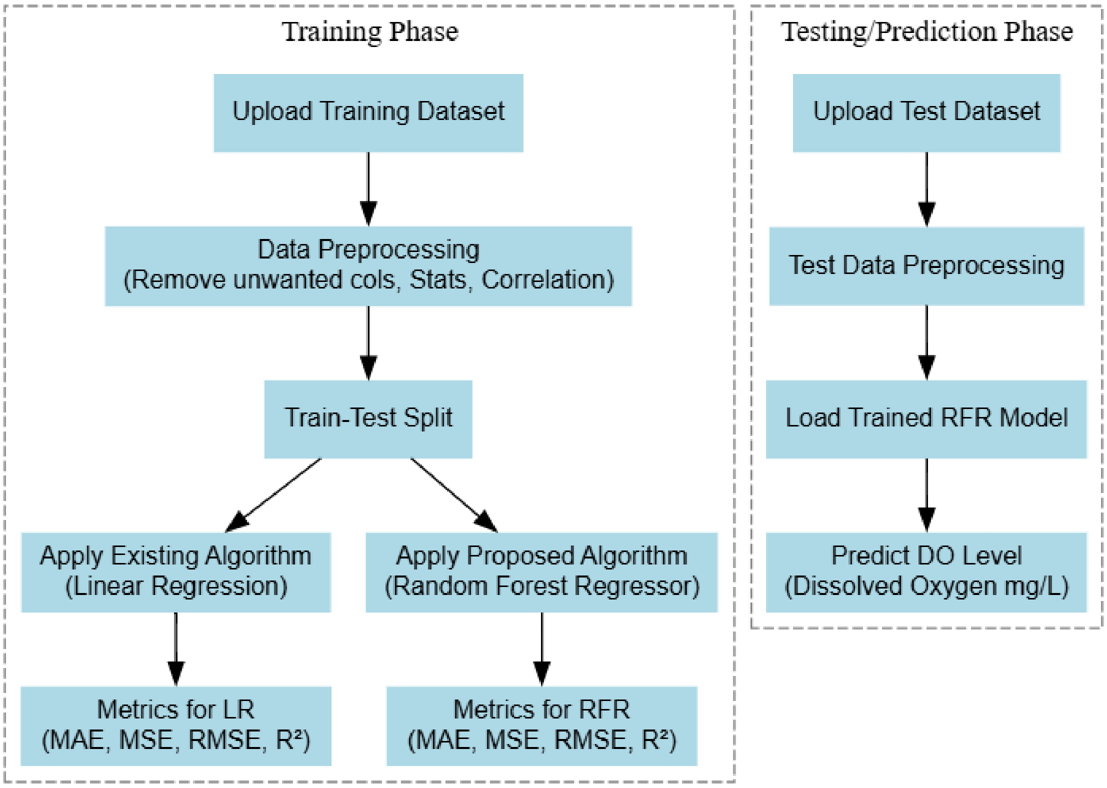


Figure 2(a): Proposed Methodology

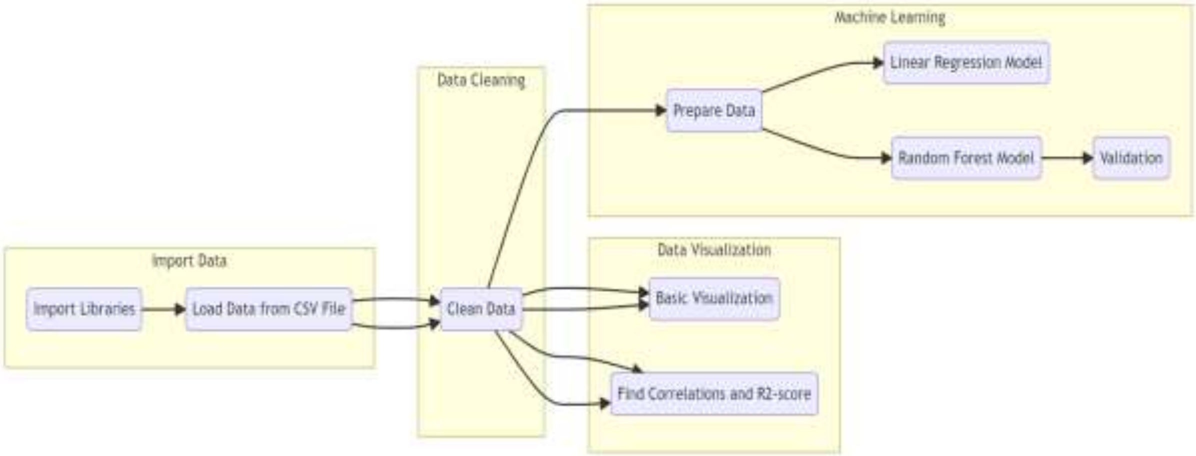


Figure 2(b) : Proposed System Model.

Data preprocessing is a critical step in preparing raw data for machine learning models, as real-world datasets often contain noise, missing values, and inconsistent formats that can hinder model performance. This process involves cleaning and transforming the data to improve model accuracy and efficiency. Key tasks include acquiring the dataset, importing necessary libraries, identifying and handling missing data (often through imputation techniques), encoding categorical variables into numerical formats, and applying feature scaling to maintain consistency across features. One essential part of preprocessing is splitting the dataset into a training set and a test set. This division ensures that the model learns patterns from the training data and is then evaluated on unseen test data to validate its generalization ability. The training set is used to train the model with known outputs, while the test set is used to assess its predictive performance on unknown data. Proper dataset splitting prevents overfitting and ensures the model performs well not only on training data but also on new, real-world inputs.

**Random Forest Regression**

Random Forest is a widely used supervised learning algorithm suitable for both classification and regression tasks, based on the ensemble learning principle which combines multiple decision trees to enhance model performance and accuracy. It constructs several decision trees from different subsets of the dataset and aggregates their results—via majority voting for classification or averaging for regression—to deliver more reliable predictions. This ensemble approach reduces overfitting and improves model robustness. The algorithm begins by randomly sampling data, building individual decision trees on each sample, generating independent outputs, and then combining them for the final prediction. Key features of Random Forest include its diversity (as not all features are used in each tree), resistance to high dimensionality, ability to train in parallel, and built-in validation due to out-of-bag data. It assumes that trees have low correlation and that meaningful data exists in feature variables for accurate prediction. Random Forest employs **bagging**, which involves bootstrap sampling and aggregation, in contrast to **boosting**, which builds sequential models to enhance performance. The proposed system leverages these strengths and introduces additional enhancements such as adaptive feature clustering and dynamic reweighting, making it highly accurate, stable, and efficient even with missing values or high-dimensional data. Its parallel nature and reduced risk of overfitting make it a powerful solution for real-world environmental prediction tasks.

.

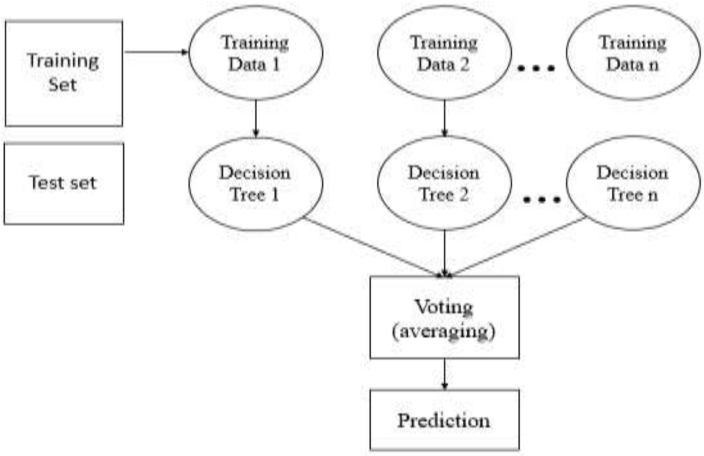


Figure 3: Random Forest algorithm.

**Database**

A database is a structured collection of data stored electronically, enabling efficient storage, retrieval, and management through a Database Management System (DBMS) such as SQLite, MySQL, or PostgreSQL. In the *SmartRiver: AI-Powered Dissolved Oxygen Monitoring and Prediction* project, SQLite—a lightweight, file-based relational DBMS—is used due to its simplicity, no-server requirement, and seamless integration with Python via the sqlite3 module. The database handles user authentication and role-based access control with a primary table named login, which stores user ID, username, password, and role (admin or user). The table is created at runtime if not already present, and new records are inserted during user registration. Upon login, the system verifies credentials by querying the table and redirects users based on their role. Though passwords are stored in plain text for demonstration, in real-world applications, hashing would be applied for security.

## 4 RESULTS

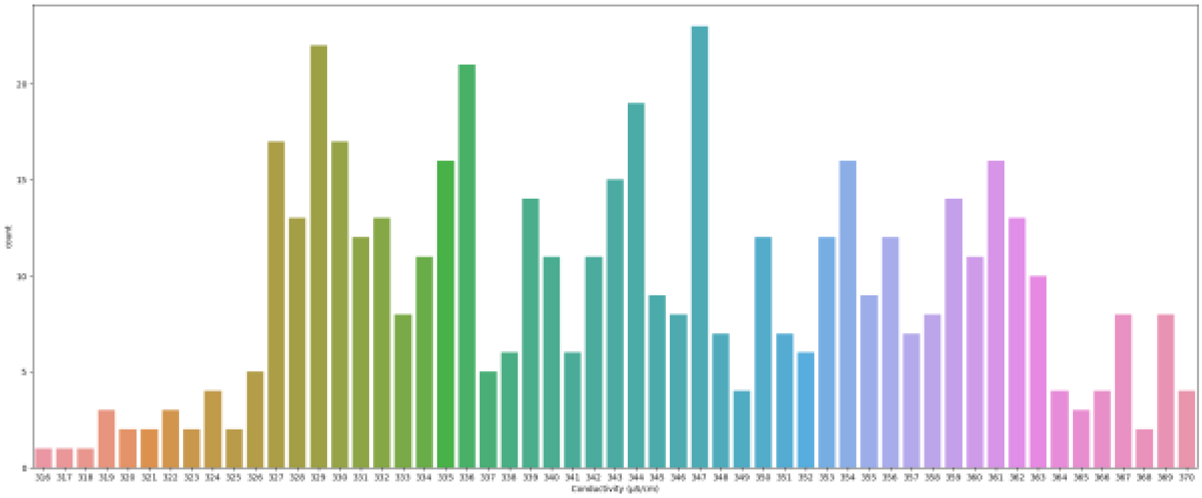


Figure 4: Count plot to visualize the distribution of values in the "Conductivity (µS/cm)"

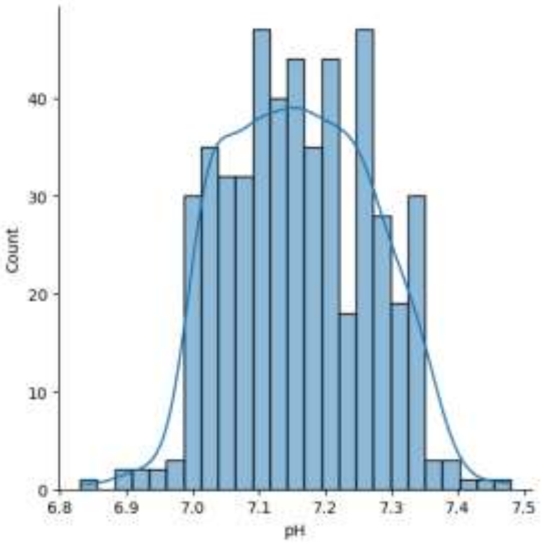


Figure 5: Distribution plot for distribution of values in the "pH" column

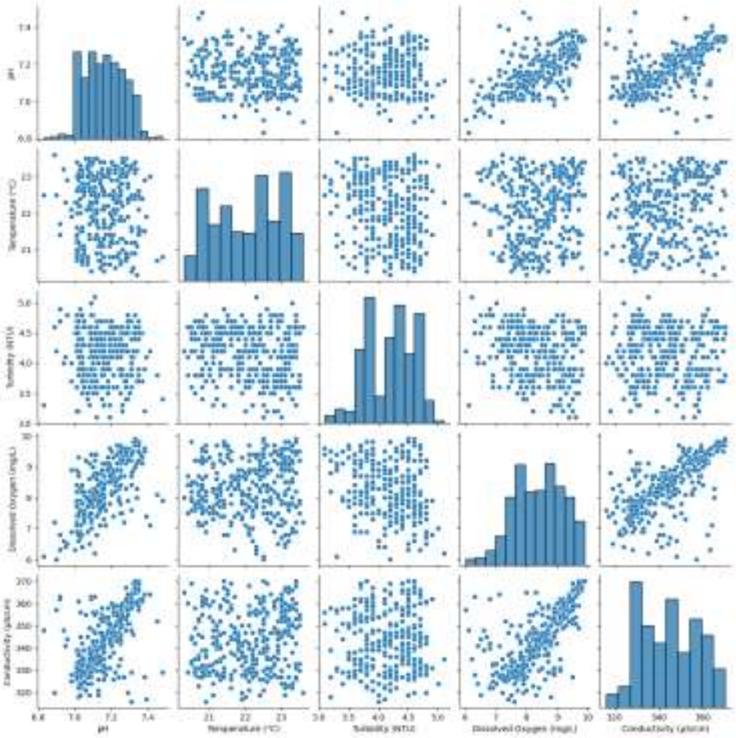


Figure 6: illustrating pairwise relationships and distributions of numerical variables in the Dataset

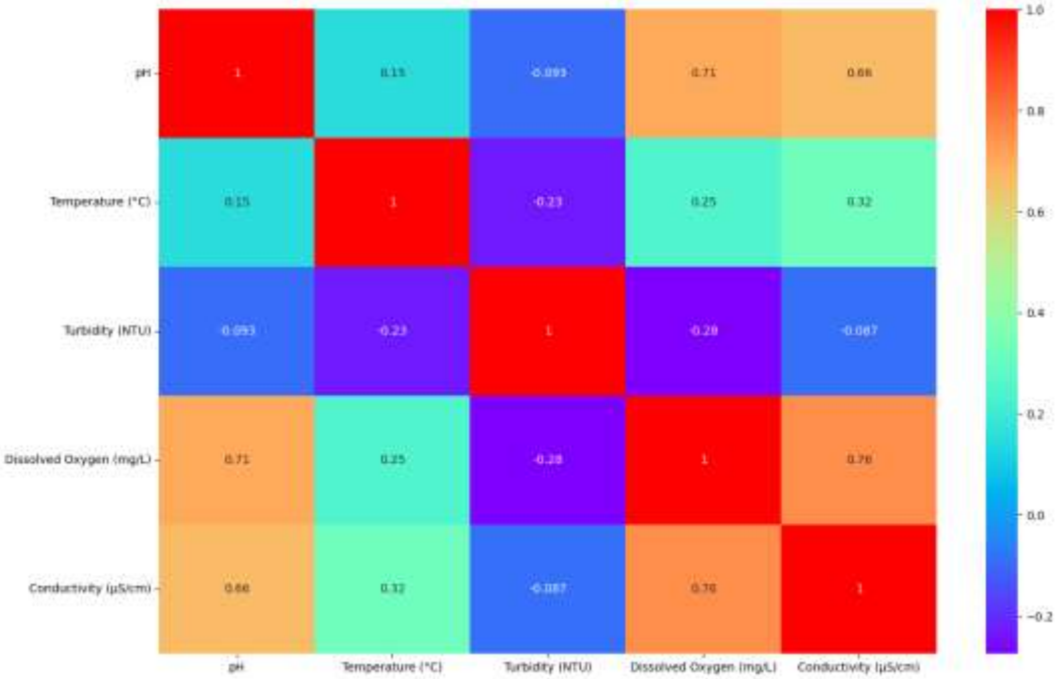


Figure 7: Visualizing the correlation matrix of numerical columns in the DataFrame

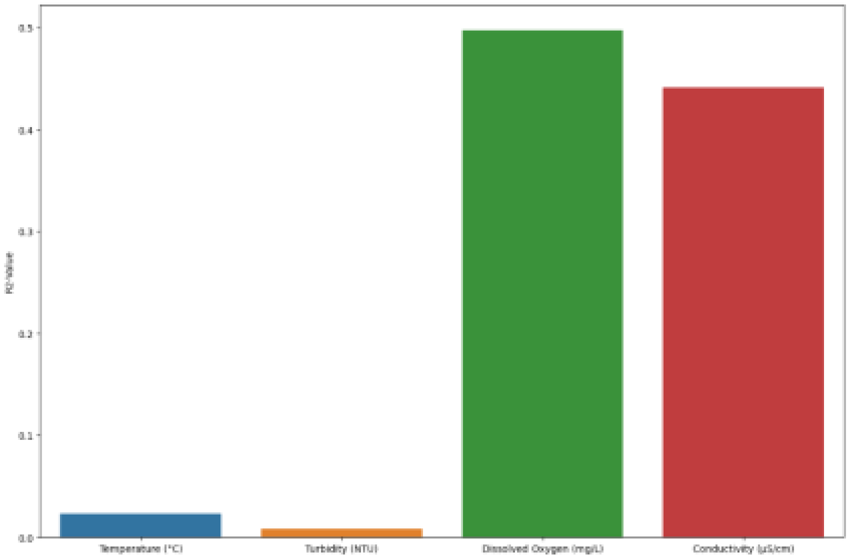
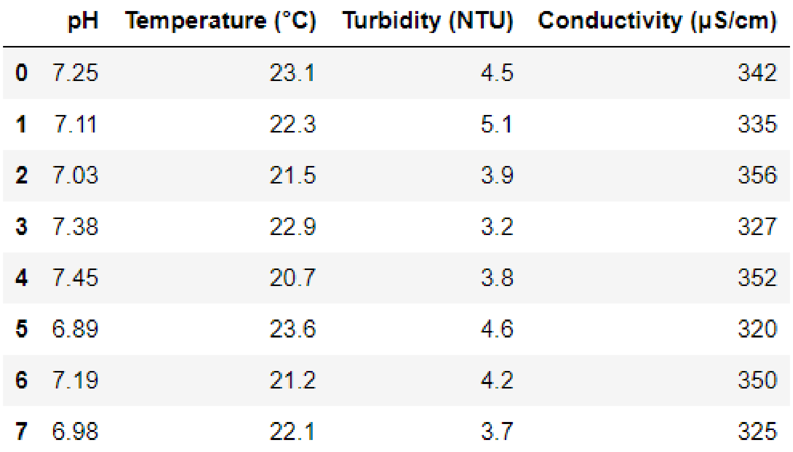


Figure 8: showcase the R2-Values ("R-squared values") from the Data Frame .



| | pH | Temperature (°C) | Turbidity (NTU) | Conductivity (µS/cm) |
|---|---|---|---|---|
| **0** | 7.25 | 23.1 | 4.5 | 342 |
| **1** | 7.11 | 22.3 | 5.1 | 335 |
| **2** | 7.03 | 21.5 | 3.9 | 356 |
| **3** | 7.38 | 22.9 | 3.2 | 327 |
| **4** | 7.45 | 20.7 | 3.8 | 352 |
| **5** | 6.89 | 23.6 | 4.6 | 320 |
| **6** | 7.19 | 21.2 | 4.2 | 350 |
| **7** | 6.98 | 22.1 | 3.7 | 325 |

Figure 9: features from the Data Frame

Table 1: Performance comparison of quality metrics obtained using linear regressor (LR) model and random forest regressor (RFR) model.

The performance comparison between the Linear Regression (LR) model and the proposed Random Forest Regressor (RFR) model clearly demonstrates the superiority of the RFR approach. The LR model, which serves as a baseline, yielded a Mean Absolute Error (MAE) of 0.308565, Mean Squared Error (MSE) of 0.187902, Root Mean Squared Error (RMSE) of 0.433477, and an R² Score of 0.714892, indicating a moderate fit to the data. In contrast, the RFR model significantly outperforms LR, achieving a much lower MAE of 0.076527, MSE of 0.020768, RMSE of 0.144113, and a remarkably high R² Score of 0.968488. This high R² suggests that the RFR model captures the underlying data patterns much more accurately, making it a highly reliable choice for predicting Dissolved Oxygen levels in water quality assessments. The results validate the effectiveness of the proposed ensemble learning-based approach over traditional linear modeling.

Table 1

| Model | MAE | MSE | RMSE | R2SCORE |
|---|---|---|---|---|
| LR model | 0.308565 | 0.187902 | 0.433477 | 0.714892 |
| RFR model | 0.076527 | 0.020768 | 0.144113 | 0.968488 |

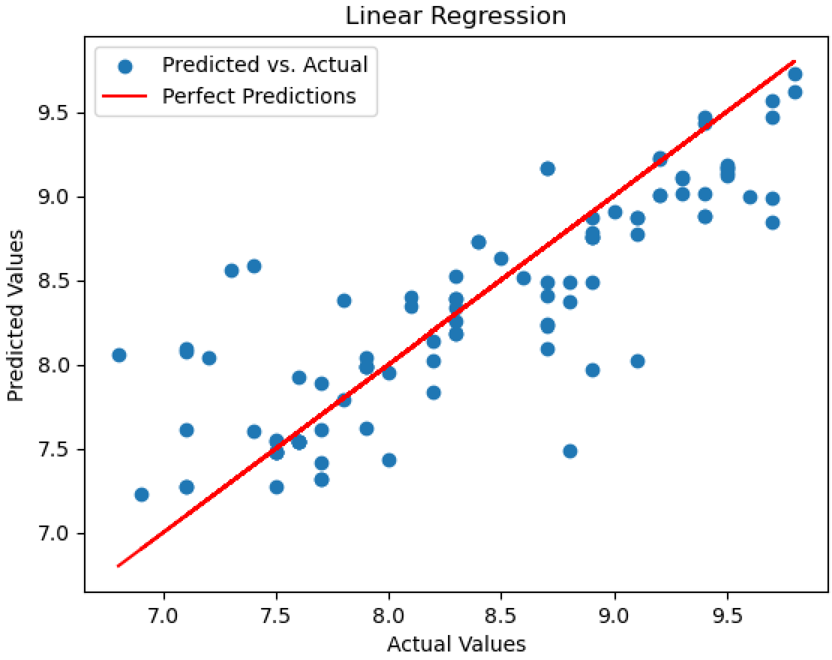


Figure 10: compare predicted values with actual values for Existing Linear Regression

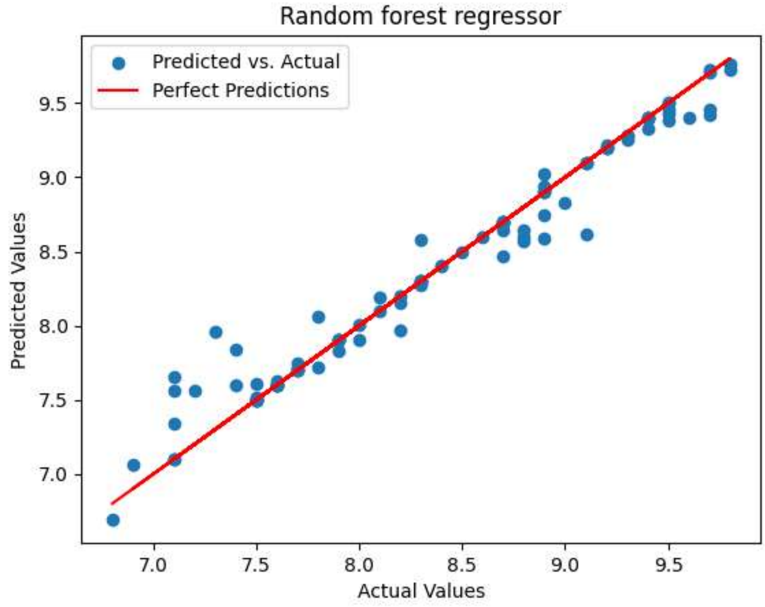
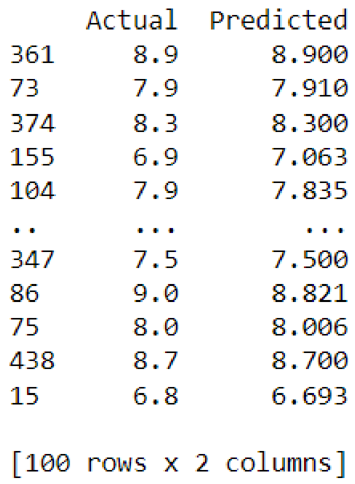


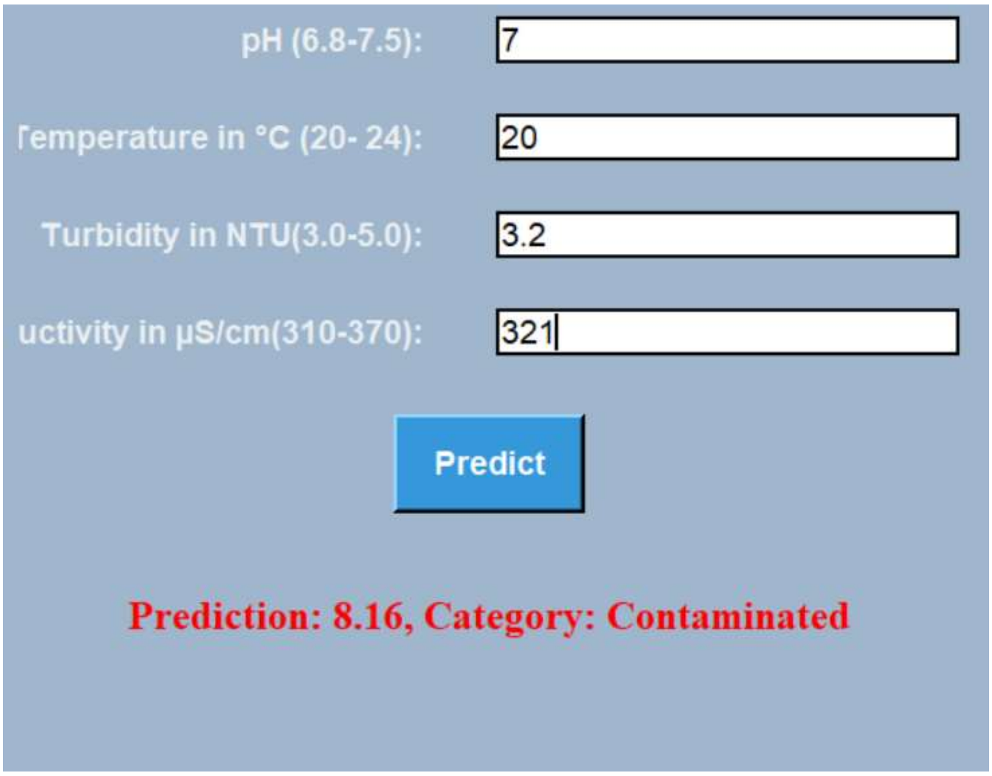
Figure 11: compare predicted values with actual values for Proposed random forest Regression



```
     Actual  Predicted
361     8.9      8.900
73      7.9      7.910
374     8.3      8.300
155     6.9      7.063
104     7.9      7.835
..      ...        ...
347     7.5      7.500
86      9.0      8.821
75      8.0      8.006
438     8.7      8.700
15      6.8      6.693

[100 rows x 2 columns]
```

Figure 12: Data Frame showing the actual and predicted values side by side.



pH (6.8-7.5): 7

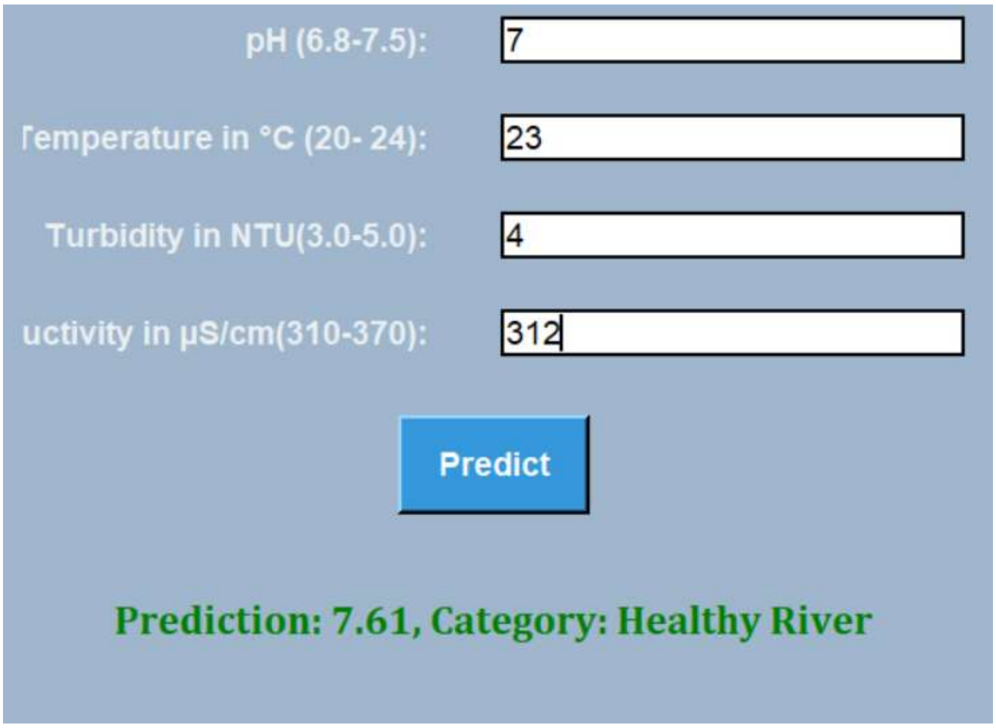
Temperature in °C (20- 24): 20

Turbidity in NTU(3.0-5.0): 3.2

uctivity in µS/cm(310-370): 321

Prediction: 8.16, Category: Contaminated

Figure 13: Model Prediction on Test Case 1.



pH (6.8-7.5): 7

Temperature in °C (20- 24): 23

Turbidity in NTU(3.0-5.0): 4

uctivity in µS/cm(310-370): 312

Prediction: 7.61, Category: Healthy River

Figure 14: Model Prediction on Test Case 2.

## 5.CONCLUSION

In conclusion, the comprehensive data analysis and machine learning work performed on the water quality testing dataset represent a crucial endeavor for understanding and predicting critical aspects of water quality, specifically the levels of dissolved oxygen. The work begins with data preparation, encompassing data loading and cleaning, where the identification and rectification of missing values and duplicates are essential for data integrity. Subsequently, basic visualizations provide an initial exploration of the data's distribution and interrelationships among variables. A notable highlight is the correlation analysis, visualized through a heatmap, enabling the identification of significant associations between various water quality parameters. The subsequent linear regression analysis, focusing on the relationship between pH and other variables, offers valuable insights into the dataset. In the machine learning phase, the work splits the data into features and the target variable, followed by the training and evaluation of two regression models—linear regression and random forest regression. The scatter plots of predicted versus actual values serve as a visual gauge of model performance. Furthermore, the creation of the predictions DataFrame facilitates in-depth analysis and comparison of model outcomes. Altogether, this work serves as a foundational step in leveraging data-driven insights to monitor and manage water quality effectively, which is vital for environmental preservation and ensuring the availability of clean and safe water resources.

## REFERENCES

1. Jones, E.R.; Bierkens, M.F.P.; Wanders, N.; Sutanudjaja, E.H.; van Beek, L.P.H.; van Vliet, M.T.H. Current Wastewater Treatment Targets Are Insufficient to Protect Surface Water Quality. *Commun. Earth Environ.* **2022**, *3*, 221.
2. Shi, X.; Mao, D.; Song, K.; Xiang, H.; Li, S.; Wang, Z. Effects of Landscape Changes on Water Quality: A Global Meta-Analysis. *Water Res.* **2024**, *260*, 121946.
3. Salerno, F.; Gaetano, V.; Gianni, T. Urbanization and Climate Change Impacts on Surface Water Quality: Enhancing the Resilience by Reducing Impervious Surfaces. *Water Res.* **2018**, *144*, 491–502.
4. Su, J.; Xu, W.; Lin, Z. Algorithm for Monitoring Water Quality Parameters in Optical Systems Based on Artificial Intelligence Data Mining. *Sci. Rep.* **2024**, *14*, 28142.
5. Essamlali, I.; Nhaila, H.; Khaili, M.E. Advances in Machine Learning and IoT for Water Quality Monitoring: A Comprehensive Review. *Heliyon* **2024**, *10*, e27920.
6. Sela, L.; Sowby, R.B.; Salomons, E.; Housh, M. Making Waves: The Potential of Generative AI in Water Utility Operations. *Water Res.* **2025**, *272*, 122935.
7. AI in Water Management Market Growth Rate, Industry Analysis with Key Companies 2025–2032. Available online: **https://www.datamintelligence.com/research-report/ai-in-water-management-market** (accessed on 13 May 2025).
8. Olawade, D.B.; Wada, O.Z.; Ige, A.O.; Egbewole, B.I.; Olojo, A.; Oladapo, B.I. Artificial Intelligence in Environmental Monitoring: Advancements, Challenges, and Future Directions. *Hyg. Environ. Health Adv.* **2024**, *12*, 100114.
9. Ponnuru, A.; Madhuri, J.V.; Saravanan, S.; Vijayakumar, T.; Manimegalai, V.; Das, A. Data-Driven Approaches to Water Quality Monitoring: Leveraging AI, Machine Learning, and Management Strategies for Environmental Protection. *J. Neonatal Surg.* **2025**, *14*, 664–675.
10. Frincu, R.M. Artificial Intelligence in Water Quality Monitoring: A Review of Water Quality Assessment Applications. *Water Qual. Res. J.* **2024**, *60*, 164–176.
11. Bagheri, M.; Farshforoush, N.; Bagheri, K.; Shemirani, A.I. Applications of Artificial Intelligence Technologies in Water Environments: From Basic Techniques to Novel Tiny Machine Learning Systems. *Process Saf. Environ. Prot.* **2023**, *180*, 10–22.
12. Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. *Sensors* **2016**, *16*, 1298.
13. Nallakaruppan, M.K.; Gangadevi, E.; Shri, M.L.; Balusamy, B.; Bhattacharya, S.; Selvarajan, S. Reliable Water Quality Prediction and Parametric Analysis Using Explainable AI Models. *Sci. Rep.* **2024**, *14*, 7520.
14. Jiang, Y.; Li, C.; Sun, L.; Guo, D.; Zhang, Y.; Wang, W. A Deep Learning Algorithm for Multi-Source Data Fusion to Predict Water Quality of Urban Sewer Networks. *J. Clean. Prod.* **2021**, *318*, 128533.

15. Wai, K.P.; Chia, M.Y.; Koo, C.H.; Huang, Y.F.; Chong, W.C. Applications of Deep Learning in Water Quality Management: A State-of-the-Art Review. *J. Hydrol.* **2022**, *613*, 128332.
16. Durgun, Y. Real-Time Water Quality Monitoring Using AI-Enabled Sensors: Detection of Contaminants and UV Disinfection Analysis in Smart Urban Water Systems. *J. King Saud Univ.-Sci.* **2024**, *36*, 103409.