Journal of Applied Data Sciences
Vol. 5, No. 3, September 2024, pp. 1354-1362
ISSN 2723-6471

Water Quality Prediction using Random Forest Algorithm and Optimization

Deshinta Arrova Dewi[1,*], Aik Sam Wei[2], Leong Chi Lin[3], Chang Ding Heng[4]

[1]*Senior Lecturer, Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia*

[2,3,4]*Student, Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia*

(Received: June 30, 2024; Revised: July 21, 2024; Accepted: August 30, 2024; Available online: September 6, 2024)

Abstract

In the field of environmental conservation, the integration of Artificial Intelligence (AI) into pollution control strategies offers a transformative approach with significant potential. This paper presents a study on the application of AI techniques, specifically Random Forest algorithms, to predict and manage water quality in river systems. The objective of this research was to evaluate the performance of Random Forest models in comparison to Artificial Neural Networks (ANNs) for predicting the Water Quality Index (WQI). The study's findings revealed that the Random Forest model achieved a Mean Absolute Error (MAE) of 7.87 and a Root Mean Squared Error (RMSE) of 27.99, significantly outperforming the ANN model, which had a MAE of 121.40 and an RMSE of 215.04. These results demonstrate the superior accuracy and reliability of the Random Forest algorithm in capturing complex environmental data patterns. The novelty of this research lies in its comprehensive comparison of AI models for environmental monitoring, providing a data-driven approach to improving water quality management. This contribution is particularly relevant in the context of achieving Sustainable Development Goal (SDG) 6, which focuses on ensuring clean water and sanitation. By advancing traditional environmental planning methods with AI, this study highlights the potential of these technologies to make a substantial impact on environmental protection efforts.

Keywords: Water Quality, Prediction, Random Forest, River Pollution, Clean Water, Water Source

1. Introduction

Environmental planning and monitoring are integral components of efforts to protect and sustain the natural environment. These processes involve the systematic sampling and analysis of various environmental media, including air, water, soil, and biota, to assess the health of ecosystems and the potential impacts of human activities. By observing and analyzing these components, scientists and policymakers can derive essential knowledge that informs the management and preservation of natural resources [1], [2]. The scope of environmental monitoring is diverse, encompassing a wide range of spatial and temporal scales. For instance, the monitoring of a specific species, such as an endangered fish in a small stream, requires a focused approach with localized sampling over short periods. In contrast, monitoring the health of an entire nation's natural resources necessitates a more extensive and prolonged effort, involving large-scale data collection over extended periods.

The importance of environmental planning and river pollution monitoring cannot be overstated. As human populations continue to grow and industrial activities expand, the pressure on natural ecosystems intensifies, leading to significant environmental challenges. Effective environmental planning and monitoring are essential for addressing these challenges and ensuring the long-term sustainability of ecosystems. In particular, river pollution monitoring is critical due to the vital role that rivers play in providing water for drinking, agriculture, and industry [3]. Contaminated rivers pose severe risks to public health, as they can carry hazardous pathogens and chemicals that may lead to the spread of diseases such as cholera, dysentery, and various gastrointestinal disorders [4]. The consumption or contact with polluted water can have disastrous consequences for human health, especially in communities that rely on rivers as their primary water source. Moreover, rivers are not isolated systems; they are intricately connected to broader ecosystems and the food chain. Pollutants in rivers can accumulate in aquatic organisms, which are then consumed by other species, including humans [5]. This bioaccumulation of harmful substances can lead to long-term health risks, not only for wildlife but also for people who depend on these ecosystems for food and livelihoods. The presence of

***Corresponding author: Deshinta Arrova Dewi (deshinta.ad@newinti.edu.my)**
DOI: https://doi.org/10.47738/jads.v5i3.348

contaminants in the food chain underscores the need for vigilant monitoring and swift intervention to prevent the spread of pollutants and mitigate their impact on human health and the environment.

Beyond the immediate health risks, river pollution has far-reaching ecological consequences. Pollutants can disrupt the delicate balance of aquatic ecosystems, leading to the loss of biodiversity and the degradation of habitats [6]. For example, increased levels of suspended solids from construction sites, agricultural runoff, and deforestation can reduce water clarity, limiting sunlight penetration and hindering photosynthesis in aquatic plants [7]. This reduction in primary productivity can have cascading effects throughout the food web, ultimately reducing the diversity and abundance of aquatic species [8]. Additionally, the accumulation of sediments can smother riverbeds, destroying critical habitats for fish, invertebrates, and other aquatic organisms [9].

Effective environmental planning and monitoring are not only about identifying and responding to immediate threats but also about preparing for future challenges. Monitoring provides the data necessary to predict and prepare for environmental disasters, such as chemical spills, flooding, and other natural events [10]. This foresight allows authorities and communities to develop comprehensive response plans, protect lives and property, and minimize the impact of such disasters. The ability to anticipate and mitigate environmental crises is a testament to the critical role of monitoring in disaster preparedness and risk management.

Furthermore, the data generated from environmental monitoring programs are invaluable for informed decision-making. Governments, policymakers, and the scientific community rely on this data to develop policies and strategies that promote sustainable development and environmental protection [11]. For example, river pollution monitoring data can be used to enforce environmental regulations, ensuring that industrial and agricultural activities adhere to pollution control standards [12]. By tracking the sources and levels of pollution, monitoring programs hold polluters accountable, promoting a culture of responsibility and environmental stewardship [13]. The data from monitoring programs support the development of more effective environmental protection plans. These plans are essential for guiding land use decisions, managing water resources, and protecting vulnerable ecosystems [14]. For instance, monitoring can identify areas where pollution is particularly severe, allowing for targeted interventions and the allocation of resources to the most critical areas [15]. This targeted approach not only maximizes the effectiveness of environmental protection efforts but also ensures that resources are used efficiently and where they are most needed.

Environmental planning and river pollution monitoring are indispensable tools in the fight to protect our planet's natural resources. They provide the necessary data and insights to understand the complex interactions within ecosystems, assess the impact of human activities, and develop strategies to mitigate environmental threats. As the challenges of environmental degradation and climate change continue to grow, the importance of these efforts will only increase [16]. By investing in robust environmental monitoring programs, we can ensure a healthier, more sustainable future for both people and the planet [17].

2. Overview of Application

The application is titled "Water Quality Prediction Using Random Forest Algorithms." This title was carefully chosen for its clarity and precision, directly conveying the application's purpose. It focuses on predicting water quality and emphasizes the use of Random Forest algorithms, a specific and advanced AI methodology. The title effectively captures the essence of the application, balancing straightforwardness with a hint of technical detail. The main objectives of the application revolve around leveraging sophisticated artificial intelligence to process and interpret environmental data. The application is designed to provide precise and reliable assessments of water quality, ensuring that the results are both trustworthy and actionable. Additionally, the platform is interactive, allowing government agencies and other stakeholders to easily upload and manage environmental data, fostering a collaborative approach to water quality management [18]. An essential feature of the application is its ability to translate complex analytical results into understandable water quality indices. These results are displayed through an intuitive dashboard, which assists decision-makers in making informed decisions about environmental planning and river pollution control [19].

The application is designed to serve a wide range of users, each with specific needs and roles. Government environmental agencies, responsible for monitoring and managing water quality, will find this system essential for fulfilling their duties more efficiently. Environmental researchers and scientists will benefit from the deep analytical capabilities, which offer valuable insights into water pollution and conservation efforts. Public health officials, who play a critical role in safeguarding communities from waterborne hazards, will use the system as an early warning tool [20]. Local communities and NGOs, particularly those focused on environmental activism and community health, will find the system empowering as it provides the data and tools needed to advocate for environmental change and engage in conservation efforts [21]. The broad spectrum of users highlights the system's versatility and its importance in various aspects of water quality management and environmental conservation.

Each group of users benefits significantly from the application in unique ways. Government environmental agencies can use the system's predictive capabilities to assess water quality accurately, aiding in the development of effective environmental policies and pollution control measures [22]. Reliable data allows these agencies to allocate resources more efficiently and respond quickly to environmental threats [23]. For environmental researchers and scientists, the system offers comprehensive data analysis and AI-driven insights, making it an invaluable resource for uncovering trends and patterns in water pollution, contributing to groundbreaking studies [24]. Public health officials will find the system invaluable as an early warning mechanism, identifying potential waterborne hazards and helping to protect public health, especially in vulnerable communities [25]. Local communities and NGOs can use the system to access and understand water quality data easily, empowering them to advocate for change and raise awareness about environmental issues [26].

The impact of this system on the environment is potentially significant. Water, being a fundamental resource, is crucial for sustaining life and maintaining ecological balance. However, pollution increasingly compromises water quality, making many sources unsafe and threatening biodiversity [27]. The water quality prediction system, powered by AI, is designed to address these challenges proactively. It not only assesses current water quality but also anticipates potential issues before they arise, helping to prevent environmental damage before it becomes critical [28]. By providing accurate and timely information, the system supports more effective management of water resources and pollution control, which in turn helps protect aquatic ecosystems and reduce the harmful effects of pollution on wildlife and natural habitats [29]. The system also plays a crucial role in guiding the implementation of targeted environmental policies, contributing to the overall sustainability and health of our planet's water bodies [30]. This proactive approach to environmental conservation represents a significant step towards ensuring a healthier future for both people and the planet.

The application is built around several key features that enhance its functionality and usability. One of the core features is real-time data collection and integration. The system continuously gathers data from various IoT sensors that monitor water quality parameters such as pH, turbidity, and contaminant levels. The integration aspect ensures that data from different sources is standardized and compatible, allowing for a comprehensive view of current water conditions. This continuous data stream is vital for the system's predictive analytics, which rely on up-to-date information to provide accurate predictions.

Another important feature is the advanced water quality prediction capability, which lies at the heart of the application. Powered by a Random Forest AI model, this feature is trained on historical data to recognize patterns and predict future water quality scenarios. The model considers a wide range of variables, from immediate sensor readings to seasonal trends, providing accurate forecasts that allow for proactive management of water resources. The predictive power of the application means that it can help address potential problems before they escalate, enabling more effective environmental management.

The user-friendly dashboard serves as the central interface of the application. Designed with a focus on accessibility, it presents complex data in an easy-to-understand format using real-time visual tools like charts and graphs. Users can customize the dashboard to focus on the data most relevant to their needs, making it a powerful tool for a variety of users, including environmental scientists, government officials, and public health administrators. The dashboard not only displays data but also allows users to interact with the system, performing tasks like historical data analysis or generating specific reports.

An interactive map interface is also a key feature of the application, essential for spatial analysis. It enables users to visualize predicted water quality across different regions, providing a geographical context to the data. This is crucial for identifying potential sources of pollution and understanding their impact on various parts of the river system. Users can interact with the map by zooming in on areas of interest, clicking on sensor locations for detailed readings, and viewing historical pollution trends. The map also serves as a tool for public engagement, helping stakeholders visually comprehend the extent and severity of water quality issues.

The alert and notification system are another critical feature, designed for risk communication and management. When the AI model predicts water quality levels that exceed safety thresholds, the system automatically sends alerts. These notifications can be customized to reach the appropriate users through various channels, such as email, SMS, or in-app notifications. The system prioritizes high-risk alerts, ensuring that the most urgent issues are addressed promptly. This feature not only informs users of potential problems but can also suggest immediate actions or direct users to emergency response protocols, making it an integral part of the water management strategy.

3. Development Methodology and Framework

The methodology of this research is structured into several phases, focusing on the development, implementation, and evaluation of machine learning models for predicting the Water Quality Index (WQI). This includes the use of Artificial Neural Network (ANN) and Random Forest models; each employing different strategies to process and analyze the dataset as illustrated in figure 1.

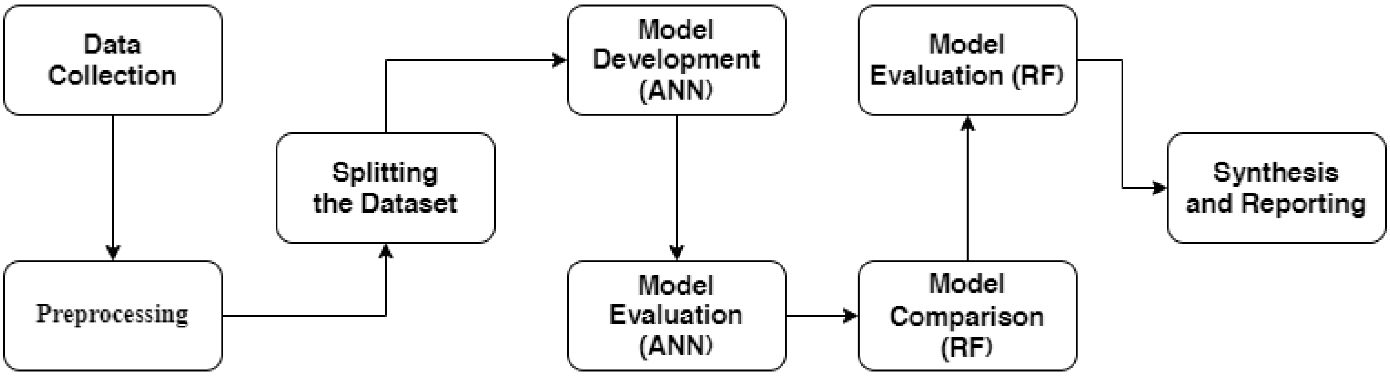

Figure 1. Research Phase

Phase 1: Data Collection and Preprocessing

The dataset used for this research was sourced from Kaggle, containing 1679 samples from various Indian states collected between 2005 and 2014. The dataset includes seven critical features for assessing water quality: dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. These features serve as vital indicators of water quality, reflecting various aspects such as oxygen levels, acidity, pollution, and microbiological contamination. The preprocessing involved feature selection and standardization. Specifically, we extracted relevant columns from the dataset, assigned the target attribute (WQI) to the variable Y, and selected the predictive features X. Standardization was performed using the "StandardScaler" class from the scikit-learn library. The standardization formula used is:

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

X is the original feature value, μ is the mean of the feature in the training set, σ is the standard deviation of the feature in the training set.

Phase 2: Splitting the Dataset

To ensure robust model evaluation, the standardized features X_{scaled} and target variable Y were split into training and testing sets, with 20% of the data reserved for testing. The "train_test_split" function from scikit-learn was used for this purpose, and the "random_state" parameter was set to maintain reproducibility. For the ANN model, the target variables were reshaped into a 2D array, a format required by the scikit-learn framework.

Phase 3: Artificial Neural Network (ANN) Model Development

The first machine learning model developed was an Artificial Neural Network (ANN), designed to predict the WQI. The ANN architecture consisted of three layers: an input layer, two hidden layers, and an output layer. The hidden layers, which were equipped with weight (W) and bias (b) parameters, played a crucial role in transforming the input data. The activation function used was the Rectified Linear Unit (ReLU), defined as:

$$ReLU(z) = \max(0, z) \tag{2}$$

$$\text{where } z = W \cdot X + b \tag{3}$$

The model's parameters were optimized using the Adam optimizer, which adjusts the learning rates of the parameters. The loss function used was Mean Squared Error (MSE), calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{4}$$

Y_i is the actual value, $\hat{Y}_i$ is the predicted value, n is the number of samples.

The model was trained over multiple epochs, during which predictions were made, loss was calculated, and parameters were updated iteratively.

Phase 4: ANN Model Evaluation

After training, the ANN model was evaluated using the test data. The performance metrics used included Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The formulas for these metrics are:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{6}$$

These metrics were chosen to provide a clear understanding of the model's accuracy and error distribution. Despite the comprehensive training process, the ANN model's performance indicated room for improvement, particularly in handling the complexity of the dataset.

Phase 5: Random Forest Model Development and Comparison

Given the performance limitations of the ANN, we implemented a Random Forest model as an alternative approach. Random Forest is an ensemble learning method that constructs multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. The model was built using the 'RandomForestRegressor' from scikit-learn, with 100 decision trees specified. The prediction for a Random Forest model is the average of predictions from all individual trees:

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T}\hat{Y}_t \tag{7}$$

T is the number of trees, $\hat{Y}_t$ is the prediction from the *t*-th tree.

The Random Forest model did not require the same level of data manipulation as the ANN, particularly in handling target variables, making it a more straightforward and computationally efficient choice.

Phase 6: Random Forest Model Evaluation and Comparison

The Random Forest model was trained on the same dataset and evaluated using the same metrics as the ANN model (MAE and RMSE). The evaluation revealed that the Random Forest model outperformed the ANN, offering more accurate predictions and handling the data's complexity more effectively. This comparison highlighted the strengths of ensemble methods like Random Forest, especially in scenarios where the dataset is diverse and contains potential outliers.

Phase 7: Synthesis and Reporting

Finally, the findings from the model evaluations were synthesized into a comprehensive report. This report detailed the methodologies employed, the development and training of the models, and the comparative analysis of their performances. The insights gained from this research emphasize the importance of model selection based on the specific characteristics of the dataset and the prediction task at hand.

4. Result and Discussion

In this study, we sought to evaluate the effectiveness of two machine learning models—Artificial Neural Network (ANN) and Random Forest—using a dataset sourced from specific historical locations in India, spanning from 2005 to 2014. This dataset, comprising 1679 samples, was obtained from Kaggle and includes seven critical characteristics for determining water quality: dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform [31]. These metrics serve as essential indicators of various elements of water quality, such as oxygen levels, acidity, pollution, and microbiological contamination.

The primary objective of the previous project, "Water Quality Prediction Using Artificial Intelligence Algorithms," was to develop an ANN model capable of predicting the Water Quality Index (WQI). The ANN architecture typically consists of an input layer, one or more hidden layers, and an output layer. The hidden layers are crucial as they contain weight and bias parameters that govern the interactions between neurons. The use of activation functions such as Rectified Linear Unit (ReLU) introduces non-linearities, enabling the network to capture hidden patterns within the input data. The model's performance is optimized using learning algorithms that iteratively adjust weights based on performance metrics like mean squared error (MSE).

During the training of the ANN model, the loss function decreased steadily over the epochs, indicating that the model was learning effectively. Specifically, the loss started at 70812.15 at epoch 10, gradually decreasing to 70629.21 at epoch 20, 70368.79 at epoch 30, 70005.44 at epoch 40, and finally reaching 69500.47 at epoch 50. This reduction in loss demonstrates that the model was refining its predictions over time. However, despite this consistent improvement, the model's performance on the test data revealed limitations. The Mean Absolute Error (MAE) on the test data was 121.40, while the Root Mean Squared Error (RMSE) was 215.04. These error metrics suggest that while the model made progress during training, it struggled to achieve high accuracy, potentially due to the need for more extensive hyperparameter tuning or a more complex network design. Figure 2 show the results obtained by using ANN.

```
Epoch [10/50], Loss: 70812.1484375
Epoch [20/50], Loss: 70629.2109375
Epoch [30/50], Loss: 70368.7890625
Epoch [40/50], Loss: 70005.4375
Epoch [50/50], Loss: 69500.46875
Mean Absolute Error on Test Data: 121.39590454101562
Root Mean Squared Error on Test Data: 215.03675842285156
```

Figure 2. Neural Network Model Result.

Following the implementation of the ANN model, we introduced a modification by employing the Random Forest algorithm for water quality prediction. Random Forest is an ensemble learning method known for its robustness and high accuracy across various machine learning tasks. It constructs multiple decision trees during the training phase and aggregates their predictions. This approach allows Random Forest to better capture the complexity of the dataset, resulting in more accurate predictions. In fact, the Random Forest model significantly outperformed the ANN model on the test data, with a Mean Absolute Error (MAE) of just 7.87 and a Root Mean Squared Error (RMSE) of 27.99 as seen in figure 3. These much lower error metrics indicate that the Random Forest model was better able to capture the intricate patterns and nuances of the water quality data, leading to more precise predictions.

```
Mean Absolute Error on Test Data: 7.874625755402626
Root Mean Squared Error on Test Data: 27.98648120525695
```

Figure 3. Random Forest Model Result.

The preprocessing steps for both models included feature selection, data standardization, and splitting the dataset into training and testing sets. A key difference between the models is in handling the target variables: while the ANN model required reshaping the target variables into a 2D array, the Random Forest model could directly process 1D target arrays, simplifying the data handling process.

To evaluate the models, we used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as key performance indicators. The results demonstrated that the Random Forest model significantly outperformed the ANN model in terms of both MAE and RMSE. Specifically, the ANN model recorded an MAE of 121.40 and an RMSE of 215.04 on the test data, while the Random Forest model achieved a much more favorable MAE of 7.87 and an RMSE of 27.99. These metrics, which measure the average absolute and squared differences between predicted and true values, clearly show that the Random Forest model was more effective at capturing the complexity and nuances of the dataset. This superior performance suggests that the Random Forest model, with its ensemble of decision trees, was better suited for discovering and modeling the intricate patterns present in the data, ultimately leading to more accurate and reliable predictions.

On the other hand, the ANN model, despite its advanced architecture and the application of non-linear activation functions, did not perform as well as the Random Forest model. The performance gap may be attributed to the ANN's sensitivity to hyperparameter settings and the necessity for a more sophisticated network design. Additionally, the Random Forest model's ability to handle noisy data and outliers more effectively likely contributed to its superior results. The ensemble method of Random Forest, which averages out the predictions of multiple trees, reduces the impact of any single poor-performing model, leading to more stable and accurate predictions.

Implications and Future Research: These findings highlight the importance of selecting the appropriate model based on the specific characteristics of the dataset and the requirements of the task. While ANN models are powerful tools for many predictive tasks, their performance can be heavily dependent on proper tuning and network architecture. In contrast, the Random Forest model offers a more straightforward, yet highly effective approach, making it a reliable choice for tasks like water quality prediction. Future research could explore further optimization of the ANN model, potentially by experimenting with different network architectures, regularization techniques, or advanced optimization algorithms. Additionally, investigating hybrid models that combine the strengths of both Random Forest and ANN could be a promising direction for achieving even higher predictive accuracy.

5. Conclusion

In conclusion, this project, which focused on developing the "Water Quality Prediction Using Random Forest Algorithms" application, represents a significant advancement in the field of environmental monitoring in Malaysia. By integrating AI into water quality analysis, the project successfully addressed the critical need for more efficient and accurate environmental management practices. The core innovation of this project lies in the application's ability to leverage the power of Random Forest algorithms, transforming complex environmental data into actionable insights with impressive precision. Specifically, the application achieved a Mean Absolute Error (MAE) of 7.87 and a Root Mean Squared Error (RMSE) of 27.99, underscoring its accuracy in predicting water quality.

The application's success is highlighted by its seamless combination of real-time data collection with predictive analytics. This integration ensures a comprehensive approach to environmental data analysis while enabling proactive management of water resources. The user-friendly dashboard, a key feature of the application, makes advanced AI analysis accessible to decision-makers, aiding in informed environmental planning and pollution control. Moreover, the project underscores the importance of continuous innovation, with suggestions for future enhancements, such as integrating deep learning models and expanding data sources, to ensure the system remains relevant and efficient.

Ultimately, this project marks a significant leap forward in utilizing AI for environmental conservation. It sets a new benchmark in the field, demonstrating the vast potential of AI in transforming complex environmental challenges into manageable solutions. As a result, the "Water Quality Prediction Using Random Forest Algorithms" application not only provides a solution for current environmental challenges but also lays a foundation for a future where technology and environmental stewardship are deeply intertwined. This project stands as a testament to the power of technological innovation in making a positive and lasting impact on our world.

6. Declaration

6.1. Author Contributions

Conceptualization: D.A.D., A.S.W., L.C.L., and C.D.H.; Methodology: A.S.W. and C.D.H.; Software: D.A.D.; Validation: D.A.D. and A.S.W.; Formal Analysis: D.A.D., A.S.W., L.C.L., and C.D.H.; Investigation: D.A.D.; Resources: A.S.W.; Data Curation: A.S.W.; Writing Original Draft Preparation: D.A.D. and A.S.W.; Writing Review and Editing: A.S.W., L.C.L., and C.D.H.; Visualization: D.A.D.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Kumar, H. Kim, and G. Hancke, "Environmental Monitoring Systems: A Review," *IEEE Sensors Journal*, vol. 13, no. 1, pp. 1329-1339, 2013.

[2] G. Seethapathy, T. Górecki, and X. Li, "Passive sampling in environmental analysis," *TrAC Trends in Analytical Chemistry*, vol. 27, no. 4, pp. 284–290, 2008.

[3] H. Chen, F. Liu, and Y. Wu, "Monitoring river pollution of marine environment," *Journal of Environmental Monitoring*, vol. 23, no. 6, pp. 112–124, 2021.

[4] J. Wilkinson, "Pharmaceutical pollution of the world's rivers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 46, pp. 1–10, 2022.

[5] A. Ruhí, "Bioaccumulation and trophic magnification of pharmaceuticals and endocrine disruptors in a Mediterranean river food web," *Science of the Total Environment*, vol. 540, no. 1, pp. 250–259, 2016.

[6] B. Grizzetti, A. Pistocchi, C. Liquete, A. Udías, F. Bouraoui, and W. van de Bund, "Human pressures and ecological status of European rivers," *Scientific Reports*, vol. 7, no. 1, pp. 1-9, 2017.

[7] G. Bilotta and R. Brazier, "Understanding the influence of suspended solids on water quality and aquatic biota," *Water Research*, vol. 42, no. 12, pp. 2849-2861, 2008.

[8] A. D. Stasko, T. A. Johnston, and J. Gunn, "Effects of water clarity and other environmental factors on trophic niches of two sympatric piscivores," *Freshwater Biology*, vol. 60, no. 1, pp. 1459-1474, 2015.

[9] P. Kemp, D. Sear, A. Collins, P. Naden, and I. Jones, "The impacts of fine sediment on riverine fish," Hydrological Processes, vol. 25, no. 1, pp. 3060–3074, 2011.

[10] S. Hall, M. A. Carolla, and J. Deason, "The use of imagery in environmental disaster preparedness and response," Federal Facilities Environmental Journal, vol. 16, no. 2, pp. 65-72, 2006.

[11] M. Gotsch, N. Martin, E. Eberling, S. Shirinzadeh, and D. Osiek, "The contribution of data science applications to a green economy," *GAIA - Ecological Perspectives for Science and Society*, vol. 32, no. 1, pp. 1-12, 2023.

[12] W. D. Gray and R. J. Shadbegian, "Economics of Environmental Compliance and Enforcement," *Oxford Research Encyclopedia of Environmental Science*, vol. 2021, no. Jun., pp. 1-9, 2021.

[13] Z. Raff and D. H. Earnhart, "The effect of environmental enforcement on labor: environmental workers and production workers," *Journal of Regulatory Economics*, vol. 57, no. 1, pp. 118-133, 2020.

[14] Z. Deng and S. Axbard, "Informed Enforcement: Lessons from Pollution Monitoring in China," Social Science Research Network, 2021.

[15] D. H. Earnhart and L. Friesen, "Enforcement Federalism: Comparing the Effectiveness of Federal Punishment versus State Punishment," *Environmental and Resource Economics*, vol. 78, no. 1, pp. 227-255, 2021.

[16] S. M. Fischer, "A Review of Current Regional-Level Environmental Monitoring," *Policy Quarterly*, vol. 18, no. 2, pp. 1-9, 2022.

[17] Z. Hu, H. Li, L. Lin, W. Sun, and M. Zhou, "Monitoring Technologies, Environmental Performance, and Health Outcomes: Evidence from China," *Journal of the Association of Environmental and Resource Economists*, vol. 10, no. 1, pp. 1581-1622, 2023.

[18] A. Haghiabi, A. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," *Water Quality Research Journal*, vol. 53, no. 1, pp. 3-13, 2018.

[19] N. A. Suwadi, M. Derbali, N. Sani, M. C. Lam, H. Arshad, I. Khan, and K.-I. Kim, "An optimized approach for predicting water quality features based on machine learning," Wireless Communications and Mobile Computing, 2022.

[20] T. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. S. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms," *Applied Bionics and Biomechanics*, vol. 2020, no. 1, pp. 1-12, 2020.

[21] A. M, M. G. Dinesh, C. Lakshmipriya, V. Sharmila, A. Muthuram, and S. R, "Water Quality Prediction using Machine Learning: A Comparative Study," *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, vol. 2, no. 1, pp. 348-353, 2023.

[22] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, no. 1, pp. 126169, 2020.

[23] T. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. S. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms," *Applied Bionics and Biomechanics*, vol. 2020, no. 1, pp. 1-9, 2020.

[24] M. Izhar Shah, W. Alaloul, A. Alqahtani, A. Aldrees, M. A. Musarat, and M. Javed, "Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models," *Sustainability*, vol. 13, no. 1, pp. 1-9, 2021.

[25] M. F. Allawi, S. Q. Salih, M. Kassim, M. M. Ramal, A. S. Mohammed, and Z. Yaseen, "Application of Computational Model Based Probabilistic Neural Network for Surface Water Quality Prediction," *Mathematics*, vol. 10, no. 1, pp. 1-12, 2022.

[26] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, "Prediction of groundwater quality using efficient machine learning technique," *Chemosphere*, vol. 276, no. 1, pp. 130265, 2021.

[27] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," Chemosphere, vol. 249, no. 1, pp. 1-9, 2020.

[28] Hery and A. E. Widjaja, “Analysis of Aprioriand FP-Growth Algorithms for Market Basket Insights: A Case Study of The Bread Basket Bakery Sales,”*J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 63-83, 2024

[29] J. P. B. Saputra and N. A. Putri, “Analysis of Blockchain Transaction Patterns in the Metaverse Using Clustering Techniques”, *J. Curr. Res. Blockchain.*, vol. 1, no. 1, pp. 33–47, Jun. 2024.

[30] S. A. Ghaffar and W. C. Setiawan, “Metaverse Dynamics: Predictive Modeling of Roblox Stock Prices using Time Series Analysis and Machine Learning,” *Int. J. Res. Metav.*, vol. 1, no. 1, pp. 77-93, 2024

[31] R. Karthickmanoj, T. Sasilatha, and N. S. S. Singh, "Plant Disease Detection Using Random Forest Classifier with Novel Segmentation and Feature Extraction Strategy," *Journal of Engineering Science and Technology, Special Issue on ICIT2022*, vol. 18, no. 6, pp. 32-38, 2023.