International Journal of Innovation and Industrial Revolution
EISSN: 2637-0972
IJIREV
Volume 8 Issue 25 (June 2026) PP. 241-252

INTERNATIONAL JOURNAL OF INNOVATION AND INDUSTRIAL REVOLUTION (IJIREV)
www.gaexcellence.com/ijirev

A HYBRID MACHINE LEARNING APPROACH FOR ENHANCED WATER QUALITY PREDICTION: INTEGRATING RANDOM FOREST AND GRADIENT BOOSTING

Syaimak Abdul Shukor[1]*, Cheng Peng[2]

[1] Center for Artificial Intelligence Technology, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
syaimak@ukm.edu.my https://orcid.org/0000-0002-7694-154X
[2]Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
p129807@siswa.ukm.edu.my https://orcid.org/0009-0001-3516-2195
*Corresponding Author

Article Info:

Article history:
Received date: 07.04.2026
Revised date: 29.04.2026
Accepted date: 20.05.2026
Published date: 22.06.2026

To cite this document:

Shukor, S. A., & Cheng, P. (2026). A Hybrid Machine Learning Approach for Enhanced Water Quality Prediction: Integrating Random Forest and Gradient Boosting. *International Journal of Innovation and Industrial Revolution, 8*(25), 241-252.

Abstract:

The escalating contamination of global water resources poses significant challenges to human health and environmental sustainability, necessitating the development of rapid, high-precision monitoring technologies. While traditional chemical and biological assessments are reliable, they are often hindered by high costs, complexity, and significant analytical latency. In response to the Industry 4.0 paradigm, this research proposes a hybrid machine learning framework that integrates Random Forest (RF) and Gradient Boosting (GB) through two fusion strategies: error-based boosting and weighted model averaging to automate and optimise water quality prediction. It implements a structured methodology including data preprocessing, feature extraction, and hyperparameter tuning. By leveraging the complementary strengths of RF (overfitting resistance) and GB (sequential error reduction), the fusion model is designed to capture nonlinear relationships in the datasets. Experimental results demonstrate that the optimised hybrid approach outperforms standalone models, achieving higher prediction accuracy and better generalisation. Furthermore, it provides a comparative analysis of feature configurations, identifying the optimal parameters for reliable detection. The findings suggest that integrating such intelligent algorithms into environmental management systems can improve automation and decision-making in water monitoring. The study concludes by outlining future trajectories, emphasising the integration of Deep Learning and Big Data analytics to further refine the practicality of water quality assessment. Ultimately, it provides a basis

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 25 (June 2026) PP. 241-252

for further development and application for sustainable water resource management in the era of digital transformation.

DOI: 10.35631/IJIREV.825014

Keyword:

Gradient Boosting, Machine Learning Fusion, Random Forest, Water Quality Prediction

Introduction

Ensuring the availability of safe drinking water is a paramount global concern, as water plays a fundamental role in ecosystems, economies, and public health. However, water quality is increasingly threatened by diverse anthropogenic activities, which compromise the sustainability of this critical resource (Ahmed et al., 2020). Traditional assessment techniques often fall short of modern requirements, as they are labour-intensive, costly, and frequently lack the real-time accuracy necessary for proactive management (Kristensen et al., 1996). In the context of the Fourth Industrial Revolution, computational techniques have introduced transformative possibilities for predictive analytics. Machine learning (ML) technologies, in particular, offer an alternative approach with precise, cost-effective solutions that can discern intricate patterns in multidimensional datasets (Mishra & Silakari, 2022). Among these, the Random Forest (RF) algorithm is a widely used ensemble learning method capable of processing complex environmental variables such as pH, sulphates, and turbidity to determine water potability with superior speed and efficiency (Liu et al., 2022). By prioritising impactful features and resisting overfitting, these technologies are enabling more responsive, data-driven operations.

Conventional water monitoring methods are often too slow and resource-heavy to facilitate timely, proactive decision-making. While standard machine learning models offer improvements over traditional techniques, they often lack the precision needed to capture the non-linear interactions inherent in complex environmental data fully. Current models often struggle with data gaps and modelling complexity, limiting their effectiveness in large-scale computing environments. There is a need for an improved predictive framework that integrates high-level modelling complexity with fast execution. To address these limitations, a hybrid RF-GB framework is developed that provides more cost-effective, continuous monitoring and enhanced predictive accuracy compared to existing standalone methodologies.

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 25 (June 2026) PP. 241-252

The primary objective is to improve the precision and efficiency of water quality forecasting by implementing a hybrid framework that bridges the gap between traditional environmental monitoring and intelligent industrial analytics, thereby enhancing prediction performance. The approach focuses on developing a hybrid model by integrating the ensemble-learning capabilities of Random Forest (RF) with the sequential optimisation of Gradient Boosting (GB) to identify optimal predictive outcomes. Furthermore, the evaluation of the improved algorithm's efficacy is based on systematic weight allocation and benchmarking against a comprehensive suite of statistical metrics, specifically Recall, F1-score, ROC-AUC, and Mean Squared Error (MSE), to assess predictive performance and generalisation across complex environmental datasets.

The analysis focuses on optimising the RF algorithm for predicting water potability using data sourced from the Kaggle repository, which includes statistically verified information from the Environmental Protection Agency (EPA). The investigation targets key indicators, including pH, dissolved oxygen, and turbidity, utilising several decades of historical data. The scope encompasses the full data lifecycle, including automated preprocessing for missing values and outliers, feature selection, and hyperparameter tuning, benchmarked against standard models such as Support Vector Machines (SVM) and Gradient Boosting (GB). Furthermore, it examines the model's generalizability across different water bodies to ensure broader geographic applicability.

A quantitative approach is adopted to validate the effectiveness of three distinct hybrid algorithm configurations experimentally. Quantitative modelling is selected for its ability to provide objective, quantifiable results that clarify the specific advantages of algorithm performance. Using a robust sample from the Kaggle database, the data undergoes rigorous cleaning, normalisation, and consistency checks. The hybrid RF and GB methods are analysed using multiple performance metrics, including Mean Squared Error (MSE), F1-scores, and ROC curves, to provide a comprehensive picture of predictive stability and accuracy.

This study provides a faster, more cost-effective alternative to traditional laboratory methods, enabling continuous and proactive water resource protection. By accurately identifying pollution sources and predicting quality deterioration, this hybrid model supports the establishment of agile, data-oriented management systems. From an industrial perspective, optimising these algorithms enhances the operation of water treatment facilities, ensuring the effective removal of pollutants and a reliable water supply. Ultimately, the integration of advanced feature selection and optimised hybrid strategies leads to more balanced generalisation, providing managers with a robust tool for environmental protection and public health.

The primary goal of this research is to revolutionise water quality detection by optimising the Random Forest algorithm and integrating it into a comprehensive hybrid predictive framework. By addressing raw data challenges such as incompleteness and outliers through robust preprocessing and fine-tuning, this study demonstrates promising results in both accuracy and efficiency. The resulting system serves as a vital tool for environmental science, advocating for more intelligent resource management and ensuring universal access to clean, safe water.

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 25 (June 2026) PP. 241-252

Literature Review

The transition toward automated water quality monitoring represents a critical innovation in the context of the Fourth Industrial Revolution, moving beyond traditional manual techniques (Meng et al., 2024). Traditional methods, while reliable, are increasingly seen as inadequate for the speed and scale required by modern industrial and environmental standards (Sawyer et al., 1978). Recent advancements in machine learning (ML) have enabled the development of intelligent systems that can predict potential risks and improve the accuracy of water assessments, thereby facilitating timely intervention (Meng et al., 2024). This technological leap enables the identification of intricate patterns within multidimensional datasets that traditional manual sampling might overlook, establishing ML as a standardised data processing method in the field of artificial intelligence (Mishra & Silakari, 2022; Meng et al., 2024).

Historically, water quality assessment has relied on physical and chemical analysis, often referred to as classical methods, including spectral analysis, chromatography, and electrochemical methods (Noh et al., 2022; Jannath et al., 2022).

- Electrochemical and Spectral Analysis: Electrochemical methods, proposed as early as 1896, were designed to detect harmful pollutants in water through physical-chemical interactions (Guth et al., 2013). Spectral methods identify organic and inorganic substances by measuring the absorbance of the sample at different wavelengths (Guo et al., 2020).
- Chromatographic Techniques: Mature technologies such as gas, liquid, and ion chromatography provide high accuracy and repeatability for quantitative and qualitative analysis but are limited by long processing times and the need for specialised operators (Starek et al., 2023; Kowalska et al., 2022).

The primary industrial bottleneck of these classical approaches is that they typically require expensive instrumentation and specialised labour, which limits their application to rapid, large-scale monitoring (Sawyer et al., 1978; Brett, 2022). Consequently, the introduction of numerical methods is viewed as a more economical and scientifically valuable approach to meet modern safety standards (Costa et al., 2019).

Predictive analytics transforms raw data into meaningful insights by harnessing statistical models to forecast future outcomes from historical trends (Shmueli et al., 2011).

- Analytics Hierarchy: Descriptive analytics identifies past patterns, while diagnostic analytics utilises data mining to answer the "why" question about specific changes, such as a decline in water quality (Patel et al., 2020; Smith et al., 2020).
- The Predictive Framework: Modern predictive analytics encompasses modelling and ML to analyse current and historical data to predict future development trends (Shmueli et al., 2011).

Implementing these models requires a systematic pipeline: defining objectives, data collection (facilitated by IoT technology), and rigorous data preparation, including cleaning and normalisation to ensure data reliability (Patel et al., 2020; Zheng et al., 2018).
Supervised learning utilises labelled data to perform classification and regression tasks essential for environmental monitoring (Bishop et al., 2006).

- Random Forest (RF) is a paradigm-shifting ensemble method that combines multiple decision trees to provide improved predictive performance than any single model alone (Breiman, 2001, Abdul Latif, L.I. et al., 2025). It reduces overfitting by combining predictions from multiple trees trained on random subsets of the original data a process known as Bootstrap Aggregating (Bagging) (Breiman, 1996). RF is renowned for its ability to handle large datasets and automatically prioritise relevant features, such as pH, dissolved oxygen, and organic matter concentration (Liaw et al., 2001). This approach significantly reduces model variance without increasing bias, allowing it to be effectively generalised to unseen data (Dietterich, 2000; Louppe, 2014).
- Gradient Boosting (GB) involves the sequential addition of "weak" models to correct errors from previous iterations, with each model minimising error at each step (Friedman et al., 2011). It has become a staple in predictive analytics for its flexibility in solving complex supervised problems such as regression and classification (Friedman et al., 2011).
- Support Vector Machines (SVMs) are valued for their ability to construct optimal hyperplanes in multidimensional space to separate different classes (Hastie et al., 2009). While effective for high-dimensional environmental datasets, SVMs can be computationally intensive and sensitive to kernel selection compared to tree-based models (Ruospo et al., 2021).

The introduction of ML algorithms has revolutionised water quality monitoring by reducing reliance on manual labour and providing policymakers with medium- to long-term analytical support (Meng et al., 2024; Han et al., 2011). Compared to Deep Learning Neural Networks (DLNNs), traditional ML models like RF and GB are often more computationally efficient and effective when data is limited (Shallue et al., 2019; Salman & Liu, 2019). Ultimately, water testing technology is moving toward greater intelligence and automation, supporting the development of stringent regulations and more proactive environmental management (Tyagi et al., 2013).

The transition to automated water quality monitoring marks a critical shift in the Fourth Industrial Revolution, moving from traditional manual assessments to intelligent machine learning (ML) systems. While classical methods such as spectral analysis, chromatography, and electrochemical techniques offer high reliability, they are often hindered by long processing times, high costs, and the need for specialized labour. Consequently, numerical methods and predictive analytics have emerged as more economical solutions to meet modern industrial standards. By utilising a systematic pipeline from data collection via IoT technology to rigorous preparation predictive frameworks transform raw data into actionable insights. Supervised learning algorithms, particularly ensemble methods like Random Forest (RF), significantly enhance accuracy by combining multiple decision trees to reduce variance and handle large datasets effectively. Additionally, Gradient Boosting (GB) and Support Vector Machines (SVM) offer flexible solutions for complex classification tasks, though they may be more computationally intensive. Ultimately, the integration of these ML algorithms revolutionizes monitoring by providing policymakers with reliable long-term analytical support and fostering proactive, automated environmental management.

While previous studies have applied Random Forest (RF), Gradient Boosting (GB), or their combinations for water quality prediction, most approaches rely on single integration strategies or standard ensemble implementations, with limited comparison between alternative

hybridisation mechanisms. In addition, feature selection is often treated as a preprocessing step without systematic evaluation of its impact on hybrid model performance.

This study addresses these gaps by (i) implementing and comparing two distinct RF–GB integration strategies; a sequential error-correction approach and a weighted averaging approach, (ii) introducing a correlation-based feature selection threshold to reduce model complexity, and (iii) evaluating model performance across both classification and regression metrics to provide a more comprehensive assessment.

Methodology

A quantitative experimental design is employed to develop an optimised hybrid machine learning framework for water quality prediction. The methodology is structured into three distinct phases: comprehensive data preprocessing, model construction via ensemble techniques, and comparative performance evaluation. To evaluate the predictive system's reliability, the methodology examines the synergy between Random Forest (RF) and Gradient Boosting (GB) to bridge the gap between traditional laboratory assessments and Industry 4.0 automated environmental monitoring.

A high-dimensional water potability dataset comprising 3,277 samples is used, initially characterised by nine physical and chemical features, including pH, Sulfate, hardness, and turbidity. Real-world environmental data often contains inherent noise and inconsistencies that can degrade model performance. Consequently, a preprocessing procedure was applied to convert raw, unrefined data into a structured format suitable for modelling.

- Outlier Management: To maintain dataset integrity, 366 harmful data points were identified using standard deviation boundaries and removed. This step ensures the model is not distorted by extreme values inconsistent with underlying natural patterns.
- Missing Value Imputation: The dataset contained 1,434 missing records, which were imputed using the mean to preserve the data distribution and minimise statistical bias.
- Feature Engineering and Scaling: Non-numeric categorical variables were transformed using One-Hot Encoding to ensure compatibility with ML algorithms. Furthermore, Z-score normalisation was applied to scale variables to a mean of 0 and a standard deviation of 1, preventing features with larger scales from disproportionately influencing the model's decision-making.

Feature selection was conducted to enhance the model's interpretability and transparency. Analysis revealed that Sulfate, pH, and Hardness are the dominant influence factors, with Sulfate exhibiting the highest relative importance at approximately 0.12. By focusing on features with a correlation index greater than 0.1 with the target variable, the research identified seven primary classification elements, thereby reducing computational complexity while improving prediction precision.

The core of the methodology involves the architectural synthesis of RF and GB models. The dataset was partitioned into a training set (80%) and a testing set (20%).

- Random Forest Implementation: RF was employed as the foundational ensemble method to reduce variance through Bootstrap Aggregating (Bagging). Hyperparameters

International Journal of Innovation and Industrial Revolution
EISSN: 2637-0972 IJIREV
Volume 8 Issue 25 (June 2026) PP. 241-252

were set to 100 trees with a maximum depth of 10 to balance accuracy against computational cost.

- Gradient Boosting Integration: To minimise bias, GB was integrated to sequentially correct errors from the RF base. While RF processes trees in parallel, GB builds them step by step, focusing on difficult-to-classify instances.
- Hybridisation Strategy: Two strategies were investigated: a boosting phase applying iterative weight adjustments to RF errors, and a mean-value weighting approach after cross-validation. Experimental results indicate the effectiveness of the weighted hybrid approach in achieving superior generalisation on unseen data.

To ensure the model's reliability and prevent overfitting, a 5-fold cross-validation was used. The original dataset was divided into five folds, with each fold serving as a validation set across five iterations. The final model performance was quantified using a multidimensional metric suite:

- F1-Score: Calculated as the harmonic mean of precision and recall providing a balanced assessment of classification coverage.
- AUC-ROC: Employed to measure the model's capacity to discriminate between safe and unsafe water categories across different thresholds.
- Mean Squared Error (MSE): Used to quantify the discrepancy between predicted and actual values, serving as a measure of the model's stability and fit.

The resulting framework provides a systematic approach to water quality prediction by leveraging the complementary strengths of RF and GB. Through the preprocessing and hybrid architectural design, an approach for environmental monitoring that maintains high accuracy under the resource constraints typical of industrial applications is proposed.

Results and Discussion

Performance Metric Evaluation

The efficacy of the water quality classification models was assessed using a multifaceted experimental framework, utilising comparative indicators to measure predictive accuracy across the 3,277 samples in the water dataset. The models were evaluated for their ability to accurately predict water quality levels, using key metrics such as pH, turbidity, and contaminant concentrations, which are widely used indicators in environmental monitoring (Tyagi et al., 2013; Kristensen et al., 1996).

The study benchmarked four distinct algorithmic configurations: Simple Weight, which calculates the mean value of different models; Random Forest (RF) in its original ensemble form proposed by Leo Breiman (2001); Hybrid Weight; and Hybrid 2, a model utilising cross-validation and a specific hybrid weighting method.

Evaluation was conducted through a comprehensive suite of metrics, including F1-score, Mean Squared Error (MSE), Accuracy, and the Area Under the Receiver Operating Characteristic (AUC-ROC) curve, which are standard performance indicators in predictive analytics and machine learning (Han et al., 2011; Shmueli & Koppius, 2011; Hastie et al., 2009).

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 25 (June 2026) PP. 241-252

Comparative Results Analysis

The experimental data reveal disparate performance profiles across the tested algorithms, as summarised in the integrated performance table (Table 1). Table 1 summarises the comparative performance of all evaluated models, highlighting that no single configuration dominates across all metrics, reinforcing the importance of metric-specific model selection.

Table 1: Consolidated Water Quality Dataset Test Scores Example

Algorithms	F1-score	MSE	AUC-ROC score	Accuracy
Simple weight	0.6521	0.3140	0.6873	0.686
Random Forest	0.4658	0.3216	0.6871	0.6784
Hybrid 2	0.3558	0.2668	0.6481	0.6798
Hybrid weight	0.4111	0.3231	0.6841	0.6768

The Simple Weight algorithm showed the best performance among the evaluated models for the classification task, achieving the highest F1-score of 0.6521 and an AUC-ROC of 0.6873. These results indicate improved performance in delivering balanced precision and recall while effectively differentiating between safe and unsafe water classes, consistent with findings in classification evaluation literature (Hastie et al., 2009; Han et al., 2011).

In contrast, the Hybrid 2 algorithm exhibited the lowest predictive error, with an MSE of 0.2668, suggesting it has the highest predictive power for regression-aligned tasks in this test environment. This aligns with prior studies highlighting the strength of hybrid and ensemble approaches in minimising prediction error (Mishra & Silakari, 2022; Liu et al., 2022).

Discussion of Findings and Algorithmic Trade-offs

A critical analysis highlights a potential conflict between accuracy-based metrics (MSE) and classification-centric metrics (F1/AUC-ROC). While the Hybrid 2 method achieved the lowest MSE, signifying that its predictions are, on average, closer to the true values, the Simple Weight method's superior F1 and AUC-ROC scores make it more reliable for categorical potability detection.

In classification problems, metrics such as F1-score and AUC-ROC are generally considered more relevant than MSE, as they account for the trade-off between precision and recall, especially in datasets with potentially imbalanced class distributions (Hastie et al., 2009; Han et al., 2011).

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 25 (June 2026) PP. 241-252

The Simple Weight approach's dominance in these areas is a suitable model for classification tasks in this dataset. However, the Hybrid Weight method may benefit from further optimisation with larger datasets, as ensemble techniques often do; this model may eventually surpass current benchmarks (Ahmed et al., 2020; Patel et al., 2020). Furthermore, ensemble and hybrid learning approaches are widely recognised for their adaptability and improved generalisation capability in complex environmental datasets (Mishra & Silakari, 2022; Coastal et al., 2022).

Ultimately, the choice of algorithm must be guided by the specific industrial requirements and the relative importance of different evaluation metrics. For scenarios where distinguishing between classes with balanced precision is paramount, the Simple Weight algorithm is preferred. If the primary goal is minimising the raw magnitude of error in continuous value prediction, the hybrid approach providing the lowest MSE may be prioritised (Shmueli & Koppius, 2011).

The experimental results confirm that the enhanced Simple Weight algorithm provides the best balance of accuracy and differentiation for the water quality dataset. By integrating the strengths of Random Forest and Gradient Boosting, the research has produced a model capable of addressing the non-linear interactions inherent in environmental data, as widely supported in machine learning literature (Breiman, 2001; Hastie et al., 2009).

Conclusion

This research successfully contributes to water quality prediction by developing an optimised hybrid machine learning framework. It evaluated two RF-GB hybridisation strategies: sequential error correction and weighted averaging, within a unified experimental framework. In addition, a correlation-based feature selection method was applied to examine its impact on model performance. By integrating ensemble learning from Random Forests (RF) with the error-correction capabilities of Gradient Boosting (GB), some limitations of traditional methods and labour-intensive laboratory methods are being addressed. The hybrid approach was rigorously validated using 5-fold cross-validation on 3,277 environmental samples, providing a model evaluated on complex datasets and non-linear interactions in water-quality data.

The experimental results demonstrate that the proposed hybrid configurations provide improved predictive accuracy compared to baseline models and greater stability than standalone models. While the Hybrid 2 model achieved the lowest Mean Squared Error (0.2668), the Simple Weight hybrid algorithm emerged as the most effective for classification, achieving an F1-score of 0.6521 and an AUC-ROC of 0.6873. These findings highlight the importance of meticulous data preprocessing, including outlier removal and missing-value interpolation, for refining raw environmental data to achieve high-precision industrial applications. The results show that no single hybrid strategy consistently outperforms others across all metrics, highlighting the importance of aligning model design with specific prediction objectives.

Ultimately, it aligns with ongoing developments in Industry 4.0 by providing a technical foundation for automated, real-time water-monitoring systems. The hybrid model’s ability to accurately predict key metrics, such as pH, dissolved oxygen, and turbidity, offers a potential approach to supporting environmental and resource management. Future research will focus

on integrating these models with IoT sensors and big data analytics to enhance further the intelligence and responsiveness of global water safety infrastructure.

The results highlight the importance of a comprehensive metric evaluation, demonstrating that while standalone models such as Random Forests perform well, optimised weight-based hybridisation can support the further development of automated water monitoring systems in an Industry 4.0 context. These findings highlight that the effectiveness of hybridisation depends on the evaluation objective: weighting-based strategies favour classification performance, while boosting-based integration favours error minimisation.

Acknowledgements:	The authors would like to express their sincere gratitude to Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
Funding Statement:	This research received financial support from Universiti Kebangsaan Malaysia under Grant Number [TAP-K010342].
Conflict of Interest Statement:	The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the International Journal of Innovation and Industrial Revolution (IJIREV).
Ethics Statement:	This study did not involve any human participants, animals, or sensitive data requiring ethical approval. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.
Author Contribution Statement:	All authors contributed significantly to the development of this manuscript. Syaimak Abdul Shukor was responsible for the conceptualization, methodology, critical revision of the manuscript and overall supervision of the study. Cheng Peng handled literature review, drafting, data collection, analysis, and interpretation of results. All authors read and approved the final version of the manuscript prior to submission.

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 25 (June 2026) PP. 241-252

References

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2020). Efficient water quality prediction utilizing memory-based deep learning. *IEEE Access*, *8*, 106967–106981.

Abdul Latif, L. I., Abu Bakar, A., Ali Othman, Z., Abdul Rais, M. S., & Berahim, M. (2025). The Random Forest algorithm for modelling the overspending behaviour of Malaysian households' income class. *Asia-Pacific Journal of Information Technology and Multimedia*, *14*(1), 40–60. https://doi.org/10.17576/apjitm-2025-1401-03

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brett, C. M. A. (2022). Professional requirements for electrochemical environmental monitoring. *Sensors*.

Coastal, A., et al. (2022). Evaluation of machine learning models for predicting microbial contamination in coastal waters. *Journal of Environmental Management*.

Costa, C., et al. (2019). The economic value of numerical methods in water quality monitoring. *Environmental Economics*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*.

Friedman, J. H., et al. (2011). The elements of statistical learning. *Springer*.

Guo, Y., et al. (2020). Infrared spectroscopy for organic pollution detection. *Analytical Methods*.

Guth, U., et al. (2013). History of electrochemical sensors. *Journal of Sensors*.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Jannath, K. A., et al. (2022). Electrochemical sensing of pollutants. *Chemical Engineering Journal*.

Kowalska, T., et al. (2022). *Chromatography in environmental analysis*. CRC Press.

Kristensen, P., Bogestrand, J., & Houzig, R. (1996). *Water quality assessment: A guide to the use of biota, sediments and water in environmental monitoring*. World Health Organization.

Liaw, A., & Wiener, M. (2001). Classification and regression by randomForest. *R News*.

Liu, Y. W., et al. (2022). Analysis of water quality parameters using ensemble learning techniques. *Environmental Monitoring and Assessment*.

Louppe, G. (2014). *Understanding random forests: From theory to practice*. University of Liege.

Meng, Q., et al. (2024). Innovative approaches in machine learning for water quality detection. *Journal of Innovation in Water Science*.

Mishra, N., & Silakari, S. (2022). Predictive analytics in water quality management using machine learning. *International Journal of Environmental Research and Public Health*.

Noh, S., et al. (2022). Spectral analysis of water quality. *Spectrochimica Acta*.

Patel, S. P., et al. (2020). Data preprocessing and feature selection in environmental data analytics. *Journal of Big Data*.

Ruospo, A., et al. (2021). Efficiency of SVMs on large-scale datasets. *Machine Learning Journal*.

Salman, S., & Liu, X. (2019). Overcoming limitations of deep learning with traditional ML. *Deep Learning Review*.

Sawyer, C., et al. (1978). *Chemistry for environmental engineering*. McGraw-Hill.
Shallue, C. J., et al. (2019). Computational efficiency in traditional machine learning. *Artificial Intelligence*.
Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.
Smith, J. P., et al. (2020). Diagnostic analytics in environmental science. *Journal of Environmental Management*.
Starek, M., et al. (2023). Chromatographic methods for water quality testing. *Chromatography Review*.
Tyagi, S., Sharma, B., Singh, P., & Dobhal, R. (2013). Water quality index: A review of trends and applications. *American Journal of Environmental Protection*, *1*(3), 34–38.
Zheng, A., et al. (2018). *Feature engineering for machine learning*. O'Reilly Media