Reliable and Efficient Model for Water Quality Prediction and Forecasting

Azween Abdullah[1], Himakshi Chaturvedi[2], Siddhesh Fuladi[3],
Nandhika Jhansi Ravuri[4], Deepa Natesan[5], M.K Nallakaruppan[6]
Faculty of Applied Science and Technology, Perdana University, Malaysia[1]
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India-632014[2, 3, 4]
Department of Networking and Communication, SRM Institute of Science and Technology, Kattankulathur[5]
School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India-632014[6]

***Abstract*—Water quality is a crucial aspect of environmental and public health. Hence, its assessment is of paramount importance. This research paper aims to leverage machine learning models to classify water quality based on a comprehensive dataset. The dataset contains various water quality indicators, and the primary objective is to predict whether the water is safe or not to consume or use. This research evaluates the performance of diverse machine learning algorithms, such as Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, and more for comparative analysis. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess the models' effectiveness in classifying water quality. The Random Forest algorithm gave the best performance with an accuracy of 95.08%, an F1-Score of 94.69%, a Precision of 90.48%, a Recall of 93.10%, and an AUC score of 0.91. A comparative plot for the ROC AUC curve is also plotted between the various machine learning models used. Feature importance, which can help identify which water quality parameters have the greatest impact on predicting water quality outcomes, is also found in the research work.**

Keywords*—Random forest; logistic regression; feature importance; decision trees; support vector machines

I. Introduction

Access to clean and safe drinking water is a fundamental human right. Waterborne diseases resulting from contaminated water sources have severe consequences on public health. Water quality plays a pivotal role in ensuring the well-being of both ecosystems and human populations. However, despite international efforts to ensure safe water sources for all, the global challenge of providing clean and potable water persists. Hence, monitoring water quality is essential to prevent waterborne diseases and environmental degradation. The World Health Organization (WHO) estimates that millions of people worldwide suffer from waterborne diseases each year due to inadequate water quality. Water quality is essential for the health of ecosystems, wildlife, and human populations. Contaminated water sources pose significant risks to public health, as waterborne diseases are a leading cause of illness and death worldwide. These diseases, often resulting from the consumption of water polluted with pathogens, chemicals, and heavy metals, impose a substantial burden on society, particularly in vulnerable and underserved communities. Moreover, beyond the immediate human health concerns, compromised water quality also leads to environmental degradation, adversely affecting aquatic ecosystems, biodiversity, and the overall sustainability of natural resources. Consequently, the importance of monitoring and maintaining water quality cannot be overstated. In the big picture, water quality analysis and evaluation techniques have substantially improved the efficiency of water pollution control [1]. To date, many methods have been developed to monitor and assess water quality worldwide, such as the multivariate statistical method [2], fuzzy inference [3], and the water quality index (WQI) [4].

Various pollutants and contaminants can compromise the quality of water sources. These include heavy metals like lead, cadmium, and mercury, pathogenic microorganisms such as bacteria and viruses, and chemical compounds like nitrates and arsenic. The presence of these contaminants in drinking water can have dire consequences for public health, causing diseases such as cholera, dysentery, and lead poisoning. Detecting and classifying water as safe or unsafe as a complex and multifaceted challenge. Although highly accurate, traditional laboratory-based methods for water quality assessment are often time-consuming, resource-intensive, and not conducive to real-time monitoring. Therefore, there is a pressing need for innovative approaches that can provide timely and reliable assessments of water quality.

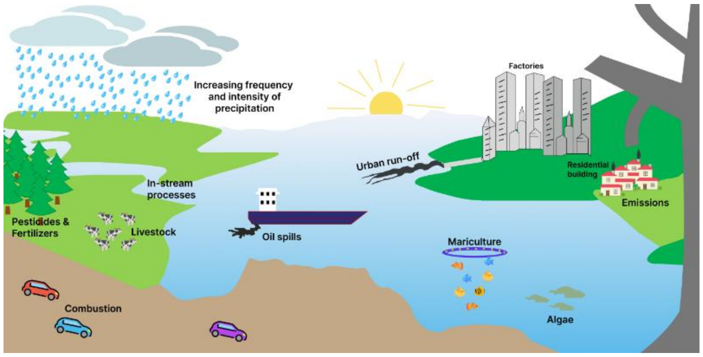

Fig. 1. Contributors towards poor quality of water.

Fig. 1 illustrates the large number of contributors that influence the quality of water. These contributors are entry points for various elements and chemicals that can significantly affect water quality. Just as depicted in the figure, through these various sources and places, a multitude of chemicals are introduced, ultimately influencing the overall quality of water. The objectives for the research are as follows:

- Investigate the utility of machine learning models for water quality classification using a comprehensive set of water quality indicators [5].
- Evaluate and compare the performance of various machine learning algorithms in classifying water sources as safe or unsafe for human consumption.
- Identify critical water quality indicators and features that strongly influence water quality classification, offering insights for targeted monitoring and intervention strategies [6].
- Bridge the gap between traditional water quality assessment methods and emerging technologies to enhance the efficiency and timeliness of water quality monitoring.
- Acknowledge the limitations of considering all water quality parameters due to cost and technical challenges and address the need for more data-driven approaches using machine learning advancements.

The significance of this research extends beyond the confines of academia. It holds practical and societal implications that resonate with the global need for clean and safe drinking water. By developing accurate machine learning models for water quality classification, we contribute to the broader efforts to ensure access to safe and clean drinking water for all. Furthermore, the insights gained from this research have the potential to inform targeted water quality monitoring strategies, enabling more efficient resource allocation and rapid intervention when unsafe water sources are detected. Ultimately, the research aligns with the United Nations Sustainable Development Goals, particularly Goal 6: "Ensure availability and sustainable management of water and sanitation for all", by enhancing our capacity to safeguard water resources and protect human health.

The paper is divided into the following sections: Section II delves into an extensive literature survey, identifying existing research gaps and showcasing innovative approaches in the field. Section III intricately details the system model and architecture, comprising a thorough dataset overview, data preprocessing techniques, and model evaluation methods. The presentation of results is encapsulated in Section IV, featuring performance metric graphs, a confusion matrix, and visual representations of feature importance. Section V navigates through in-depth discussions on practical implications and outlines potential avenues for future research. Finally, Section VI encapsulates the conclusions, summarizing key findings.

II. Literature Review

Li, Z., Liu, H., Zhang, C., & Fu, G. [7] introduced a real-time water quality prediction method for distribution networks using Graph Neural Networks. Addressing sparse monitoring data challenges, the approach underscores GNNs' effectiveness in capturing complex relationships.

Garabaghi, F. [8] employed the AdaBoost ensemble method to classify water sources as safe or unsafe for human consumption based on various water quality indicators. By combining these indicators, their model demonstrated the potential of machine learning to ensure access to safe drinking water. This study contributes to public health and environmental protection efforts.

Li, L. [9] tackled the vital issue of model interpretability in water quality assessment. They introduced a method that combined Random Forest with Shapley values to provide insights into the features contributing to water quality predictions. This research emphasized the importance of model transparency and interpretability in building trust in automated water quality assessment systems.

Cruz, R. [10] explored the application of Machine Learning for predicting harmful algal blooms. Their study utilized historical data on water quality parameters and algal bloom occurrences, demonstrating the potential of data-driven approaches in addressing ecological threats and water safety concerns.

Yan, J. [11] introduced a hybrid model that combined Neural Networks and Principal Component Analysis (PCA) for water quality prediction. Their research emphasized the importance of feature reduction and dimensionality reduction techniques in enhancing the efficiency and effectiveness of predictive models.

Ighalo, J. O. [12] proposed a novel approach that integrated Internet of Things (IoT) technology with machine learning for real-time water quality monitoring. Their study focused on sensor networks and data analytics, enabling proactive responses to water quality deviations.

Barzegar, R. [13] employed Extreme Learning Machines (ELM) for water quality prediction. Their research highlighted the speed and efficiency of ELM in handling large datasets, making it a valuable tool for real-time monitoring and forecasting.

Mosavi, A. [14] researched water quality assessment using ensemble models. By combining Random Forest, Gradient Boosting, and AdaBoost, their approach improved the robustness and accuracy of predictions, addressing the need for reliable water safety assessments.

Li, L. [15] explored the application of Recurrent Neural Networks (RNNs) for predicting temporal water quality variations. Their study emphasized the importance of considering historical data and dynamic patterns in water quality assessment.

Kadinski, L. [16] proposed a data-driven approach to identify contamination sources in water distribution systems. By integrating machine learning and network analysis, their research contributed to the early detection and management of waterborne risks.

Haghiabi, A. H. [17] introduced a framework for water quality prediction using multiple machine learning models. Their approach combined Support Vector Machines, Decision Trees, and K-nearest neighbors to improve predictive accuracy and model robustness.

Chakravarthy [18] introduced a method focusing on water quality prediction, employing SoftMax-ELM optimized with the Adaptive Crow-Search Algorithm. This innovative technique aims to enhance accuracy in water quality predictions by optimizing the SoftMax-ELM model.

Dogo, E. M. [19] explored using unsupervised learning for water quality anomaly detection. Their research applied Self-Organizing Maps (SOM) to identify deviations from normal water quality conditions, enhancing the early detection of water contamination incidents.

Solanki, A. [20] laid the foundation for machine learning models in water quality assessment. Their early work paved the way for subsequent research by highlighting the potential of data-driven approaches in this domain. This study marks the inception of applying machine learning to water quality analysis.

Chang, N. B. [21] explored integrating remote sensing data with machine learning techniques to assess water quality in large water bodies. Their approach showcased the scalability of machine learning in monitoring vast aquatic environments, with implications for environmental conservation and management.

Wu, J. [22] delved into using Long Short-Term Memory (LSTM) neural networks for time-series-based water quality prediction. Their study focused on predicting temporal variations in water quality indicators, aiding in forecasting changes over time. This research contributes to a better understanding of the dynamic nature of water quality.

Moayedi, H. [23] developed a hybrid model that combined machine learning and physical models for water quality assessment. This integrated approach improved prediction accuracy by considering both data-driven and mechanistic aspects, bridging the gap between empirical and theoretical approaches in water quality research.

Yan, K. [24] introduced a novel feature engineering technique called Recursive Feature Extraction (RFE) for water quality data. This method improved model performance by selecting the most relevant features for prediction, enhancing the efficiency and effectiveness of water quality assessment models.

Ahmed, M. [25] explored the application of Deep Belief Networks (DBNs) for water quality monitoring. Their study highlighted the potential of deep learning techniques in capturing complex patterns in water quality data, paving the way for advanced modeling approaches in the field.

Liu, S. [26] investigated the use of evolutionary algorithms in optimizing machine learning models for water quality prediction. Their research emphasized the importance of model tuning and parameter optimization for improved prediction accuracy, contributing to more reliable water quality assessments.

Iqbal, K. [27] applied clustering techniques to segment water quality data into distinct groups. This unsupervised learning approach facilitated.

Identifying common patterns and anomalies in water quality profiles offers insights for targeted monitoring and intervention strategies.

Table I provides the research gaps in the previous approaches for tackling the issues with water quality.

A. Research Gaps

- Data-Driven Insights for Targeted Monitoring: The paper fills a research gap by providing insights into data-driven approaches that help identify critical water quality indicators and features.
- Integration of Various Machine Learning Models: The paper addresses the gap in research related to integrating and comparing multiple machine learning models for water quality assessment, providing insights into the performance of different algorithms.
- Interpretability and Transparency: The paper bridges the gap by addressing the need for model interpretability and transparency in the context of water quality assessment

In Fig. 2, we visually represent the diverse contributions from various fields that have shaped and influenced our research. This figure illustrates the multidisciplinary nature of the research endeavor and the collaborative efforts of experts from different domains.

TABLE I. RESEARCH GAPS

Ref No	Author	Proposed Method	Limitations
3	Garabaghi, F.	AdaBoost for classifying water sources as safe or unsafe	Limited discussion of model performance and real-world application challenges.
17	Chang, N. B.	Integration of remote sensing with ML for assessing water quality	Challenges in remote sensing data quality and applicability to different ecosystems.
8	Ighalo, J. O.	IoT and ML for real-time water quality monitoring	Potential issues with sensor network deployment, data quality, and security.
10	Mosavi, A.	Ensemble models combining RF, GB, and AdaBoost for prediction	Complexity in model interpretation and computational resources required.
23	Iqbal, K.	Clustering to segment water quality data into distinct groups	Dependency on the quality of input data and the choice of clustering algorithm.

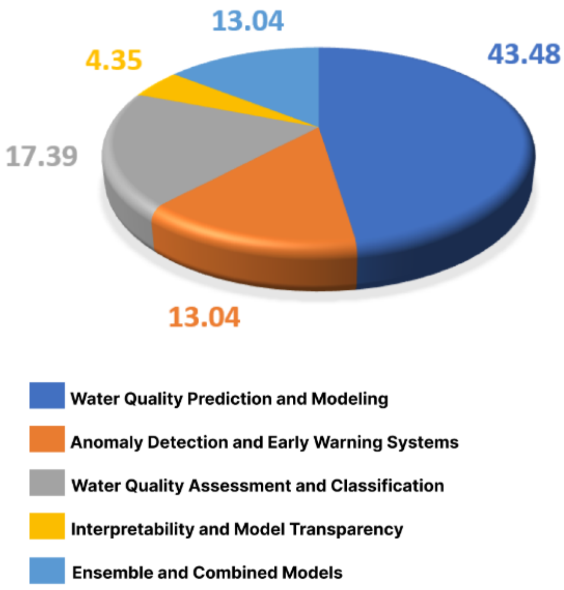

Fig. 2. Contributions for the research.

III. System Model and Architecture

A. Dataset Overview

The dataset was obtained from Kaggle, a well-known platform for sharing and exploring datasets. The dataset contains information related to water quality parameters and attributes for classifying water sources as safe or unsafe for human consumption. The dataset comprises 21 columns and 8000 rows, providing substantial data for robust analysis and modeling.

Below are the dataset values and their significance in deteriorating or enhancing water quality.

1) *Aluminum:* Measures the concentration of aluminum in the water.

2) *Ammonia:* Indicates the ammonia level in the water.

3) *Arsenic:* Reflects the concentration of arsenic in the water.

4) *Barium:* Represents the amount of barium in the water.

5) *Cadmium:* Measures the cadmium content in the water.

6) *Chloramine:* Indicates the chloramine level in the water.

7) *Chromium:* Reflects the concentration of chromium in the water.

8) *Copper:* Measures the copper content in the water.

9) *Fluoride:* Represents the fluoride level in the water.

10) *Bacteria:* Reflects the presence or absence of bacteria in the water.

11) *Viruses:* Indicates the presence or absence of viruses in the water.

12) *Lead:* Measures the lead content in the water.

13) *Nitrates:* Reflects the nitrate level in the water.

14) *Nitrites:* Indicates the nitrite level in the water.

15) *Mercury:* Measures the mercury content in the water.

16) *Perchlorate:* Represents the amount of perchlorate in the water.

17) *Radium:* Reflects the concentration of radium in the water.

18) *Selenium:* Measures the selenium content in the water.

19) *Silver:* Indicates the silver level in the water.

20) *Uranium:* Reflects the concentration of uranium in the water.

21) *Is_safe:* The class attribute, where '0' represents not safe and '1' represents safe water sources.

B. Data Preprocessing and Feature Engineering

Given the diverse range of water quality indicators, such as aluminum, ammonia, arsenic, and others, the data preprocessing phase involved several key steps. We addressed missing values by mean imputation to prevent gaps in the dataset, normalized the values of these indicators, and encoded categorical attributes. Eq. (1) is used to handle missing values.

$$x_i = \frac{1}{N}\sum_{j=1}^{N} x_j \qquad (1)$$

Moreover, the most critical aspect of feature engineering was the transformation of raw indicator values into binary representations. This transformation enabled us to create the "is_safe" feature, serving as the class attribute for classification and ultimately facilitating accurate water quality assessment. To facilitate the modeling process, we employed the StandardScaler function to scale the features into a common range. This normalization was vital for ensuring that each indicator contributed to the classification process on an equal footing. For a given feature X, the standardization using StandardScaler transforms it into a new feature X', given in Eq. (2). This transformation ensures that the standardized feature has a mean of 0 and a standard deviation of 1, which is important for many machine learning algorithms, especially those sensitive to the features' scale.

$$X' = \frac{X-\mu}{\sigma} \qquad (2)$$

where:

X' is the standardized feature.

X is the original feature.

μ is the mean(average) of the feature X.

σ is the standard deviation of the feature X.

C. Model Evaluation

Evaluating the performance of machine learning models for water quality classification is a critical aspect of our research. To gauge the effectiveness of these models, we considered multiple performance metrics. Accuracy was an essential metric that measures the proportion of correctly classified instances. We also examined the F1 score, which balances precision and recall, ensuring that our models can efficiently identify both safe and unsafe water sources. Furthermore, precision and recall were critical for understanding the model's ability to minimize false positives and false negatives, respectively. The ROC AUC score assessed the models' ability to distinguish between safe and unsafe sources. By employing these metrics, our research ensures rigorous evaluation and validation of the water quality classification models.

Algorithm 1 shows how to calculate the accuracy of a machine-learning model. As more accurate model outcomes result in better decisions, it is important to identify which model would work best for a given dataset.

Algorithm 1: To Calculate Accuracy for Machine Learning Model

```
Input:
- Trained machine learning model (ML_model)
- Test data (X_test) with corresponding true labels (y_true)
Output:
- Accuracy of the machine learning model
1. Initialize a variable 'correct_predictions' to 0.
2. Initialize a variable 'total_predictions' to 0.
3. For each data point (x) and true label (y_true) in the test data (X_test, y_true):
  a. Use the trained machine learning model (ML_model) to make a prediction (y_pred) for x.
  b. Increment 'total_predictions' by 1.
  c. If the model's prediction (y_pred) matches the true label (y_true):
     - Increment 'correct_predictions' by 1.
4. Calculate the accuracy as follows:
  - Accuracy = (correct_predictions / total_predictions) * 100
5. Output the accuracy value as the result.
End
```

Algorithm 2 shows how the calculation for ROC AUC score is performed and how the plot for the ROC curve is designed. The ROC AUC score tells us how efficient the model is. The higher the AUC, the better the model's performance at distinguishing between the positive and negative classes. An AUC score of 1 means the classifier can perfectly distinguish between all the Positive and the Negative class points.

Algorithm 2: To Calculate ROC AUC Score and Plot ROC Curve

```
Input:
- Trained machine learning model (ML_model)
- Test data (X_test) with corresponding true binary labels (y_true)
Output:
- ROC AUC Score
- ROC Curve Plot
1. Use the trained machine learning model (ML_model) to predict probabilities for each data point in the test data (X_test).
2. Calculate the ROC AUC score using the predicted probabilities and true labels:
  - ROC_AUC_Score = roc_auc_score(y_true, predicted_probabilities)
3. Compute the ROC curve by varying the decision threshold:
  - FPR (False Positive Rate), TPR (True Positive Rate), thresholds = roc_curve(y_true, predicted_probabilities)
4. Plot the ROC curve:
  - Plot FPR is on the x-axis, and TPR is on the y-axis.
  - Add a diagonal line representing a random classifier (FPR = TPR) for reference.
  - Label the curve and the diagonal line accordingly.
  - Add a legend to the plot.
5. Output the ROC AUC Score and the ROC Curve Plot.
End
```

C. Model Evaluation Equations for Machine Learning Algorithms

The selection of appropriate models plays a crucial role in the success of any research endeavor, as different models employ distinct algorithms and mathematical techniques to model the underlying data. This subsection presents an overview of the machine learning models employed in this research, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Each model's Equation is discussed in Eq. (3) to Eq. (7), elucidating the mathematical foundation upon which they operate, allowing for a comprehensive understanding of their implementation in the context of this research.

Logistic Regression:

$$P(y = 1|x) \;=\; 1 \,/\, (1 \;+\; exp(-z) \tag{3}$$

where:

P(y=1|x) is the likelihood that the positive class will exist.

z is the linear combination of the input features and their corresponding coefficients.

Decision Tree:

$$Prediction = Tree(x) \tag{4}$$

where:

Tree(x) represents the traversal of the decision tree to assign the class label to the instance x based on the

Decision Tree learned rules.is a tree-based classifier that splits the feature space based on a set of rules.

Random Forest:

$$Prediction = Average(Tree1(x), Tree2(x), \ldots, TreeN(x) \tag{5}$$

where:

Tree(x) represents the traversal of the decision tree.

An ensemble technique called Random Forest blends various decision trees to produce estimations.

The prediction in a Random Forest is obtained by averaging the predictions of individual decision trees.

Support Vector Machines (SVM):

$$Prediction \;=\; sign(w\text{^}T \ast c \;+\; e) \tag{6}$$

where:

Prediction is the predicted class label.

w is the weight vector.

c are the input features.

e is the bias term.

The sign function assigns the class label based on the sign of the linear combination.

Support Vector Machines are binary classifiers that aim to find the hyperplane that maximizes the margin between two classes.

KNeighborsClassifier:

dist(*x*, *z*) = (*Σ*_{*r* = 1}^*d* |*x*_*r* − *z*_*r*|^*p*)^(1/*p*) (7)

where:

dist(x,z) represents the distance between the two points x and z.

Σ_{r=1}^d sums all the features of the data.

|x_r - z_r|^p calculates the absolute difference in each dimension.

D. Architecture

Fig. 3 provides an overview of how data processing works in our paper by showing each step involved in the process from input to output as well as illustrating how information flows between these steps.

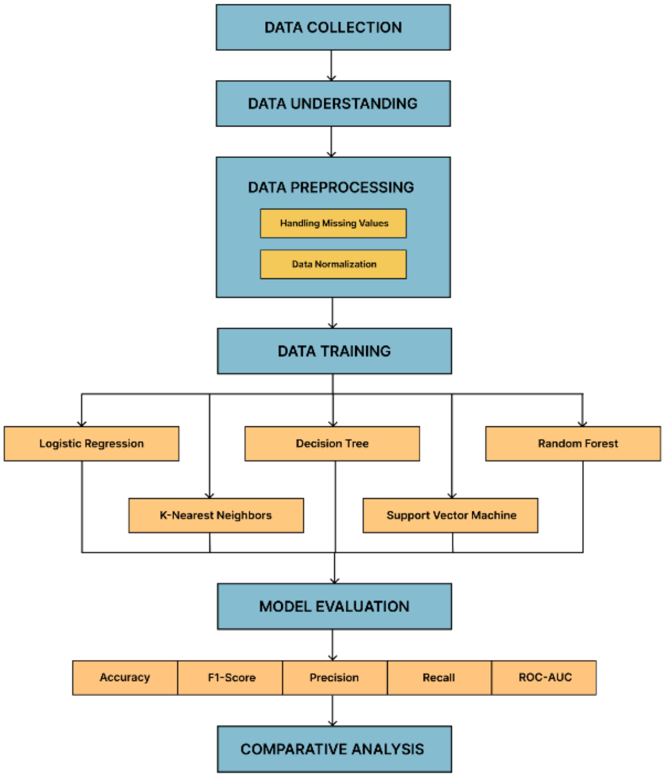

Fig. 3. Architecture diagram.

IV. Results

Our research, produced significant findings that hold practical implications for ensuring safe drinking water. Fig. 4 highlights the Accuracy of our machine learning models on the test set. Notably, the Random Forest model achieved an accuracy of 95.08% reflecting its robustness in classifying water sources as safe or unsafe. The Decision Tree model closely followed, demonstrating an accuracy of 94.62%. Logistic Regression, K-Nearest Neighbors, and Support Vector Machine also provided competitive results. Fig. 5 highlights the F1-Score for the various models. The Random Forest model achieved a score of 94.69%, followed by Decision Tree model with a score of 94.63%. Fig. 6 highlights the Precision for the models. Random Forest model achieved a score of 90.48%, which is followed by Support Vector Machine model with a score of 86.58 %. Fig. 7 highlights the Recall of the various models used. Random Forest model gave the best score of 93.10%, which is followed by Support Vector Machine model with a score of 87.60%.

Fig. 8 illustrates the ROC-AUC curve, which depicts the performance of different machine learning models in classifying water quality ranging from a score of 0 to 1. Random Forest outperforms other models with the highest AUC score of 0.91, showcasing its superior ability to discriminate between various water quality levels.

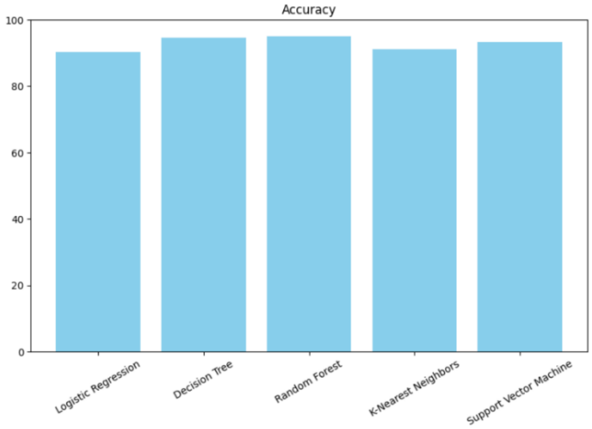

Fig. 4. Model Comparison in terms of accuracy.

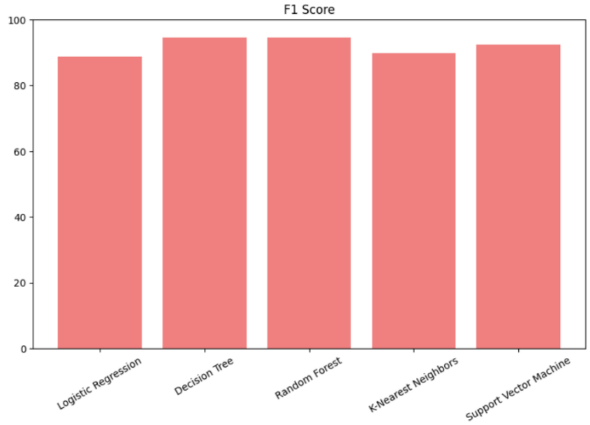

Fig. 5. Model comparison in terms of F1-score.

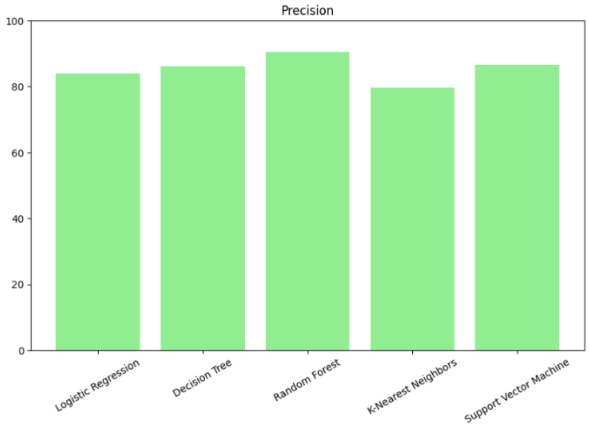

Fig. 6. Model Comparison in terms of precision.

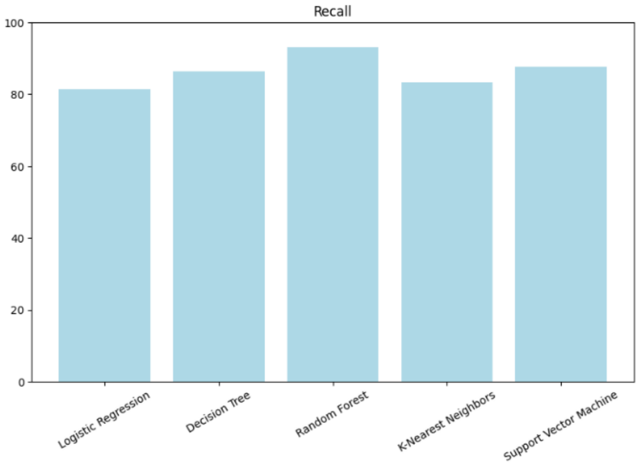

Fig. 7. Model comparison in terms of recall.

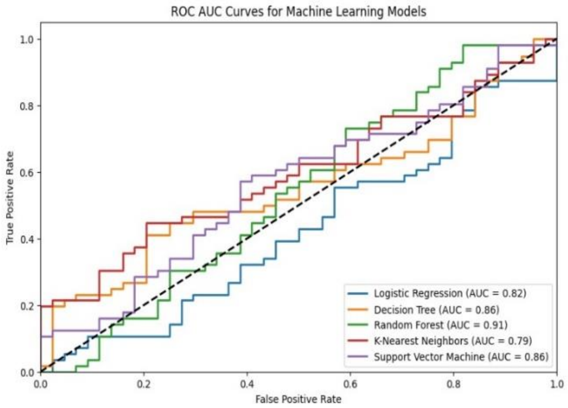

Fig. 8. ROC AUC score.

Feature importance is typically calculated based on how often a feature is selected to split nodes in decision trees within the ensemble and how much the feature contributes to reducing impurity in those splits. Feature importance using a Random Forest model can help identify which water quality parameters have the greatest impact on predicting water quality outcomes. This knowledge is essential for understanding the most influential factors affecting water quality, enabling better decision-making. Fig. 9 illustrates the results.

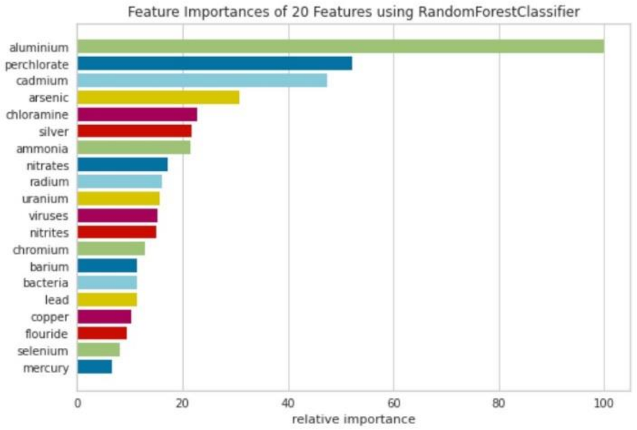

Fig. 9. Feature importance.

Table II presents the importance of the feature obtained from a random forest model. The "Feature" column lists the input variables, while the "Importance" column quantifies the significance of each feature in the model's predictions. These importance scores, ranging from 0 to 1, reveal the relative influence of each variable in detecting the water quality.

TABLE II. IMPORTANCE OF A SPECIFIC FEATURE

Feature	Importance
aluminum	0.214372
perchlorate	0.119331
cadmium	0.113347
arsenic	0.065622
ammonia	0.046725
chloramine	0.046548
Silver	0.045136
nitrates	0.037736
nitrites	0.033658
uranium	0.033620
radium	0.032812
viruses	0.031122
chromium	0.029673
barium	0.028580
bacteria	0.026482
lead	0.023878
copper	0.023279
fluoride	0.019532
selenium	0.016568
Mercury	0.011981

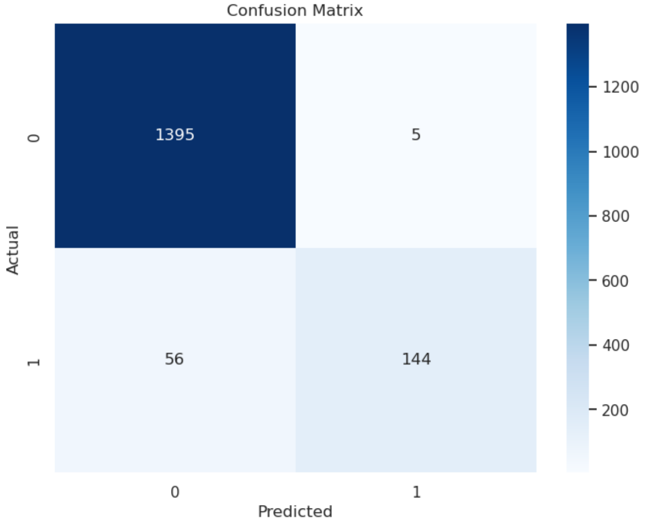

Fig. 10. Confusion matrix.

The confusion matrix illustrated in Fig. 10 visually represents the classification performance of the random forest model. It offers a detailed breakdown of true positives, true negatives, false positives, and false negatives, providing insights into the model's accuracy and error patterns.

The heatmap in Fig. 11 represents the relationships between the different water quality features in the dataset, to help identify which parameters are strongly correlated (positively or negatively) and which are not significantly related.

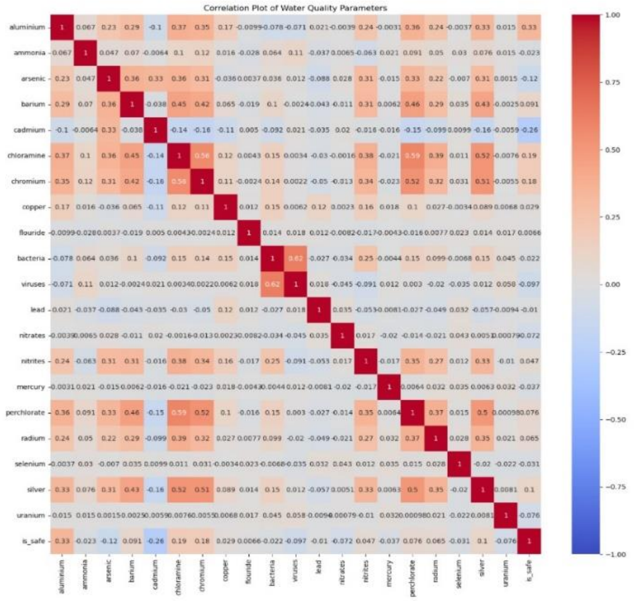

Fig. 11. Heatmap between water quality features.

TABLE III. SUMMARY OF RESULTS

Model	Accuracy	F1-Score	Precision	Recall	ROC-AUC
Logistic Regression	90.25	88.73	83.91	81.37	0.82
Decision Tree	94.62	94.63	86.10	86.38	0.86
Random Forest	95.08	94.69	90.48	93.10	0.91
K-Nearest Neighbors	91.08	89.86	79.66	83.27	0.79
Support Vector Machine	93.25	92.36	86.58	87.60	0.86

Table III provides a comparative analysis of all the models used and their performance measures.

V. DISCUSSIONS

A. *Interpretation of Results*

The interpretation of our research results reveals a comprehensive understanding of their implications for water quality assessment. Our machine learning models, especially the Random Forest and Decision Tree, have proven their capability to effectively classify water sources as safe or unsafe. This has significant practical implications, particularly in the context of providing safe drinking water. The high accuracy and F1 scores signify the models' ability to minimize false positives and negatives, an essential characteristic when dealing with public health issues related to water quality. These results demonstrate the potential for real-time water quality monitoring, allowing for the swift detection of anomalies and timely intervention.

B. *Feature Importance*

Analyzing the importance of features, as highlighted in our feature importance plots, provides valuable insights into the significant indicators influencing water quality classification. This knowledge empowers decision-makers and water quality management authorities to prioritize interventions. By identifying which indicators have the most substantial impact, targeted actions can be taken to ensure water safety. The binary representation of these features makes it easy to understand and act upon the results. Furthermore, feature importance analysis complements traditional laboratory-based methods by providing a data-driven approach to identifying critical water quality indicators.

C. *Practical Implications*

The practical implications of our research are substantial and extend far beyond the scope of our findings. Our models have showcased their potential for real-time water quality monitoring, which can be a game-changer in ensuring the safety of drinking water. The ability to rapidly detect water sources that pose health risks can transform public health management. This research is especially relevant in regions where water quality can fluctuate significantly, potentially impacting the health of communities. By harnessing machine learning models, authorities and stakeholders can efficiently monitor and manage water quality, taking timely actions to address concerns.

D. Limitations

Our research, while promising, relies on historical data, which may not fully capture evolving water quality dynamics. Additionally, our approach assumes fixed threshold values for safety, which may not be universally applicable across different regions and water sources. It is essential to recognize that water quality can vary significantly due to geographical and environmental factors. Future research should aim to address these limitations by considering real-time data and accounting for regional variations. In conclusion, our research has provided valuable insights into the potential of machine learning models for water quality assessment. The interpretability of results, the identification of significant features, and the practical implications of our findings underscore the significance of automated systems in ensuring safe drinking water. The limitations highlighted here serve as a roadmap for future research endeavors to continually improve water quality management and public health worldwide.

E. Future Scope

The success of our current research opens up several promising avenues for future investigations in the field of water quality assessment and management. Some areas where further research and development can make significant contributions are:

1) Real-time data integration: Future research should focus on integrating real-time data sources into the machine learning models. By leveraging the power of continuous data streams from various sensors and sources, we can create more adaptive and responsive models that can detect water quality anomalies as they happen. This would enhance the timeliness and accuracy of intervention strategies.

2) Advanced sensor technology: Researchers can explore the development of cutting-edge sensor technologies that can provide more granular data on water quality parameters. This could involve using nano-sensors, microfluidic devices, and remote sensing technologies to measure a wide range of chemical, biological, and physical indicators in real time.

3) Deep learning and neural networks: Investigate the application of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for analyzing complex, high-dimensional water quality data.

4) Explainable AI (XAI): Develop explainable AI techniques to enhance the interpretability of machine learning models. This is crucial for gaining the trust of stakeholders and decision-makers in the water quality management process. Methods such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) can be employed to provide insights into model predictions.

5) Edge computing and real-time processing: Leverage edge computing and real-time data processing to reduce anomaly detection and response latency. Edge devices equipped with machine learning capabilities can make rapid decisions on data streams, enabling timely intervention.

6) Blockchain for data verification: Use blockchain technology to enhance data integrity and trustworthiness. Blockchain can be employed to securely record and verify water quality data, ensuring its accuracy and preventing tampering.

VI. Conclusion

The research highlights the efficacy of machine learning models in classifying water quality, offering significant practical implications for ensuring safe drinking water. The Random Forest model stood out as the top performer, achieving an accuracy of 95.08% and an F1-score of 94.69. Its precision of 90.48% and recall of 93.10% underscore its ability to identify safe water sources while minimizing false alarms accurately. The ROC-AUC curve further emphasizes the Random Forest's superiority, with the highest AUC of 0.91, signifying its reliability in discriminating between water quality levels. Feature importance analysis using the Random Forest model unveiled crucial insights into the most influential factors affecting water quality outcomes, providing valuable knowledge for decision-making in water quality management.

In summary, this study demonstrates that machine learning, particularly the Random Forest algorithm, is a powerful tool for classifying water quality with high accuracy. As far as the practical implications are considered, this research can be applied to the regions where water quality can fluctuate significantly. By harnessing machine learning models, authorities and stakeholders can efficiently monitor and manage water quality, taking timely actions to address concerns. These findings can inform policies and strategies to ensure clean and safe water sources, ultimately enhancing environmental and public health.

References

[1] Alam, R., Ahmeahd, Z., Seefat, S. M., & Nahin, K. T. K. (2021). Assessment of surface water quality around a landfill using multivariate statistical method, Sylhet, Bangladesh. Environmental Nanotechnology, Monitoring & Management, 15, 100422.

[2] Oladipo, J. O., Akinwumiju, A. S., Aboyeji, O. S., & Adelodun, A. A. (2021). Comparison between fuzzy logic and water quality index methods: A case of water quality assessment in Ikare community, Southwestern Nigeria. Environmental Challenges, 3, 100038.

[3] Wang, J., Fu, Z., Qiao, H., & Liu, F. (2019). Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China. Science of the Total Environment, 650, 1392-1402.

[4] dos Santos Simoes, F., Moreira, A. B., Bisinoti, M. C., Gimenez, S. M. N., & Yabe, M. J. S. (2008). Water quality index as a simple indicator of aquaculture effects on aquatic bodies. Ecological indicators, 8(5), 476-484.

[5] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., ... & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment & Health.

[6] Jalal, D., & Ezzedine, T. (2020, June). Decision tree and support vector machine for anomaly detection in water distribution networks. In 2020 International Wireless Communications and Mobile Computing (IWCMC) (pp. 1320-1323). IEEE.

[7] Li, Z., Liu, H., Zhang, C., & Fu, G. (2023). Real-time water quality prediction in water distribution networks using graph neural networks with sparse monitoring data. Water Research, 121018.

[8] Garabaghi, F. H., Benzer, S., & Benzer, R. (2022). Performance evaluation of machine learning models with ensemble learning approach in classification of water quality indices based on different subset of features.

[9] Li, L., Qiao, J., Yu, G., Wang, L., Li, H. Y., Liao, C., & Zhu, Z. (2022). Interpretable tree-based ensemble model for predicting beach water quality. Water Research, 211, 118078.

[10] Cruz, R. C., Reis Costa, P., Vinga, S., Krippahl, L., & Lopes, M. B. (2021). A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. Journal of Marine Science and Engineering, 9(3), 283.

[11] Yan, J., Liu, J., Yu, Y., & Xu, H. (2021). Water quality prediction in the luan river based on 1-drcnn and bigru hybrid neural network model. Water, 13(9), 1273.

[12] Ighalo, J. O., Adeniyi, A. G., & Marques, G. (2021). Internet of things for water quality monitoring and assessment: a comprehensive review. Artificial intelligence for sustainable development: theory, practice and future applications, 245-259.

[13] Barzegar, R., Asghari Moghaddam, A., Adamowski, J., & Ozga-Zielinski, B. (2018). Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. Stochastic environmental research and risk assessment, 32, 799-813.

[14] Mosavi, A., Hosseini, F. S., Choubin, B., Abdolshahnejad, M., Gharechaee, H., Lahijanzadeh, A., & Dineva, A. A. (2020). Susceptibility prediction of groundwater hardness using ensemble machine learning models. Water, 12(10), 2770.

[15] Li, L., Jiang, P., Xu, H., Lin, G., Guo, D., & Wu, H. (2019). Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China. Environmental Science and Pollution Research, 26, 19879-19896.

[16] Kadinski, L., Salcedo, C., Boccelli, D. L., Berglund, E., & Ostfeld, A. (2022). A hybrid data-driven-agent-based modelling framework for water distribution systems contamination response during COVID-19. Water, 14(7), 1088.

[17] Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1), 3-13.

[18] Chakravarthy, S. S., Bharanidharan, N., kumar Venkatesan, V., Abbas, M., Rajaguru, H., Mahesh, T. R., & Venkatesan, K. (2023). Prediction of Water Quality using SoftMax-ELM optimized using Adaptive Crow-Search Algorithm. IEEE Access.

[19] Dogo, E. M., Nwulu, N. I., Twala, B., & Aigbavboa, C. (2019). A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water Journal, 16(3), 235-248.

[20] Solanki, A., Agrawal, H., & Khare, K. (2015). Predictive analysis of water quality parameters using deep learning. International Journal of Computer Applications, 125(9), 0975-8887.

[21] Chang, N. B., Bai, K., & Chen, C. F. (2017). Integrating multisensor satellite data merging and image reconstruction in support of machine learning for better water quality management. Journal of environmental management, 201, 227-240.

[22] Wu, J., & Wang, Z. (2022). A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. Water, 14(4), 610.

[23] Moayedi, H., Salari, M., Dehrashid, A. A., & Le, B. N. (2023). Groundwater quality evaluation using hybrid model of the multi-layer perceptron combined with neural-evolutionary regression techniques: case study of Shiraz plain. Stochastic Environmental Research and Risk Assessment, 1-16.

[24] Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B: Chemical, 212, 353-363.

[25] Ahmed, M., Mumtaz, R., Anwar, Z., Shaukat, A., Arif, O., & Shafait, F. (2022). A multi–step approach for optically active and inactive water quality parameter estimation using deep learning and remote sensing. Water, 14(13), 2112.

[26] Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., & Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Mathematical and Computer Modelling, 58(3-4), 458-465.

[27] Iqbal, K., Ahmad, S., & Dutta, V. (2019). Pollution mapping in the urban segment of a tropical river: is water quality index (WQI) enough for a nutrient-polluted river?. Applied Water Science, 9(8), 1-16.