Educational Administration: Theory and Practice
2023, 29(4), 2374-2381
ISSN: 2148-2403
https://kuey.net/

Research Article

Innovative Water Quality Prediction For Efficient Management Using Ensemble Learning

Madan Mohan Tito Ayyalasomayajula[1*]

[1*]Computer Science, School of Business & Technology Aspen, University, USA, mail2tito@gmail.com

Citation: Madan Mohan Tito Ayyalasomayajula, et.al (2023) Innovative Water Quality Prediction For Efficient Management Using Ensemble Learning, *Educational Administration: Theory and Practice*, 29*(4)* 2374-2381
Doi: 10.53555/kuey.v29i4.7118

ARTICLE INFO

ABSTRACT

Water quality is of utmost importance for the health and welfare of humans, animals, plants, industries, and the whole ecosystem. Contamination and pollution have had a negative impact on water quality in recent decades. Ensuring the water's purity is vital for the well-being of the population and the long-term preservation of the ecosystem. Conventional approaches to monitoring and predicting water quality often lack precision and promptness. This study investigates ensemble machine learning methods to increase water quality forecast accuracy to address these issues. This method integrates Random Forest, K-Nearest Neighbour, and logistic regression to improve prediction accuracy and durability. The ensemble technique is analyzed using a large dataset of pH, turbidity, dissolved oxygen, and chemical contaminants. The results demonstrate significant improvements in predicting accuracy when compared to individual models, offering a more reliable tool for water quality management. This work showcases the efficacy of ensemble machine learning in producing pragmatic insights for optimizing water management techniques. The system attains an exceptional degree of accuracy, with a rate of 99.98%. Additionally, it demonstrates a high level of precision at 99.87%, recall at 99.67%, F1-score at 99.89%, and Matthews correlation coefficient at 97.86%. These findings have significant implications for improving resource management and protecting the environment.

Keywords- water quality, normalization, ensemble learning, water quality Index, and water management

Introduction

Water is an important resource upon which all forms of life rely. Water pollution causes a deterioration in the quality of water, which adversely affects the welfare of marine species and, therefore, the well-being of humans who depend on them. Hence, it is essential to oversee the water's quality and ensure the conservation of marine biodiversity [1]. Mitigating and managing water pollution requires understanding water quality issues. Environmental water management initiatives have been developed by numerous governments globally to comprehend the marine environment. One billion people require drinkable water, and two million die yearly from polluted water and poor sanitation. Thus, freshwater quality must be preserved [2]. Water quality is crucial to the sustainability of the diversion plan. Bad water may be costly since it needs money to restore water distribution infrastructure when problems emerge. The goal of ensuring safe drinking water at affordable prices has led to an increase in the need for improved water management and water quality control. Freshwater, disposal systems, and organizational monitoring difficulties need to be systematically evaluated in order to solve these problems [3]. Forecasting water quality involves making predictions about the changes that will occur in the state of health of a water system at a certain point in time. Planning and regulating water quality depend heavily on the results of water quality assessments. By extrapolating future advancements in water purification under varying pollution levels and formulating pragmatic strategies for preventing and managing water pollution, the efficacy of approaches for mitigating and controlling water pollution may be enhanced. Water diversion plans have to evaluate the general consistency of the water. A significant amount of water is delivered to address daily drinking challenges. As such, research needs to be done on ways to predict the quality of water in today's society [4]. AI and ML are crucial for security. They value knowing system inputs and outcomes above intricate operating processes. In order to effectively manage, regulate, and monitor water,

it is essential to have a good understanding of its quality. Preserving water ecosystems is crucial for doing research on water contamination. Therefore, a practical and efficient water quality forecasting method is needed. Preventing rapid water quality changes and giving solutions requires water quality forecasting. Thus, precise water quality projections can maintain biodiversity, control fishing productivity, and assure drinking water safety [7]. The standard water quality forecast method ignores biology, physics, hydraulics, chemistry, and weather. Researchers prioritize groundwater forecasting functionality and reliability. ANN, stochastic mathematics, fuzzy mathematics, 3S technology, and other techniques have improved water quality forecasts and broadened their applications [8]. Real-time or near-real-time predictive algorithms estimate the Water Quality Index (WQI) and Water Quality Classification (WQC) more efficiently and cheaply than laboratory analysis. This allows continuous water quality monitoring, early deterioration detection, and quick response to threats or pollutants. Predictive models may effectively handle such conditions by using available data and estimating missing values, ensuring that the WQI is generated even when the complete dataset cannot be immediately accessed [9]. By enhancing WQI and WQC projections, resource distribution becomes more efficient. Based on projected water quality categories, decision-makers may prioritize sampling, focus monitoring on places of interest, and enhance treatment techniques [10]. Early water quality warning systems may use predictive algorithms. Continuous WQI and WQC monitoring and forecasting allow for possible water quality issues to be identified in advance. This allows for proactive actions to minimize the effects and safeguard water resources. Considering this, the accomplishments of this effort may be summarised as follows:

- Data normalization was conducted to standardize the data and facilitate further processing.
- Utilizing ensemble models like Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression to enhance performance in various water quality situations by selecting or weighting models depending on data attributes.

In the following parts, Section 2 examines previous research investigations, Section 3 demonstrates the proposed object detection model, Section 4 shows the findings and discussion of the experiments, and Section 5 concludes with the conclusion and prospective future work.

Related works

Machine learning is very suitable for forecasting water quality since it can detect the elements responsible for changes and uncover intricate correlations between variables and their anticipated results. ML models have been widely used in several domains. It attains a precision rate of 99.50% in forecasting WQC values. The study in [12] examines the feasibility of constructing an advanced monitoring system that enables water operators to conduct real-time quality control. ANN and other advanced pattern recognition systems, such as SVM, are modern sensor technologies used to detect abnormalities and assess their severity. A nine-layer MLP and KNN imputer are used [13] to handle missing data. The nine-layer MLP model may predict water quality with 0.99 accuracy using the KNN imputer. The research modeled and forecasted water quality indicators using BPNN, RBFNN, SVM, and LSSVM [14]. The SVM predicted 99% of publicly accessible and industrial aquaculture system data. [15] compares G-Naive Bayes, B-Naive Bayes, SVM, KNN, X Gradient Boosting, and RF. The SVM was 78.96% accurate. The study conducted by IN [16] revealed that Bayesian-optimized machine learning achieved the best levels of accuracy of 0.992 and the Kappa coefficient of 0.987. The performance and PREI findings of the DT, ExT, and GXB models suggest they can efficiently and consistently anticipate WQIs with little model uncertainty.

Machine learning models for predicting water quality exhibit high accuracy but face limitations such as data quality issues, model complexity, and interpretability challenges. Models like Random Forest and SVM achieve impressive results but often require extensive computational resources and can be challenging to interpret. Additionally, variability in environmental factors and the need for robust real-time implementation pose significant challenges. Ensuring accurate and reliable predictions necessitates addressing these limitations through improved data handling, model optimization, and comprehensive validation. Ensemble machine learning methods are adopted to overcome these issues, which enhance water quality prediction by increasing accuracy and robustness through the combination of multiple models, which mitigates overfitting and captures complex relationships between variables. These methods provide stable and consistent predictions, making them resilient to anomalies and variations in data. Additionally, ensembles can handle large datasets efficiently and offer valuable insights into feature importance, contributing to more reliable and precise water quality monitoring systems.

System model

Initial loading includes 30 stations and 10 features from the water sample dataset. We divide the dataset into X and Y for training and testing. First, the X train and X test datasets are subjected to data preparation and normalization. As shows that figure-1 an ensemble machine learning model uses normalization findings to enhance water quality prediction.

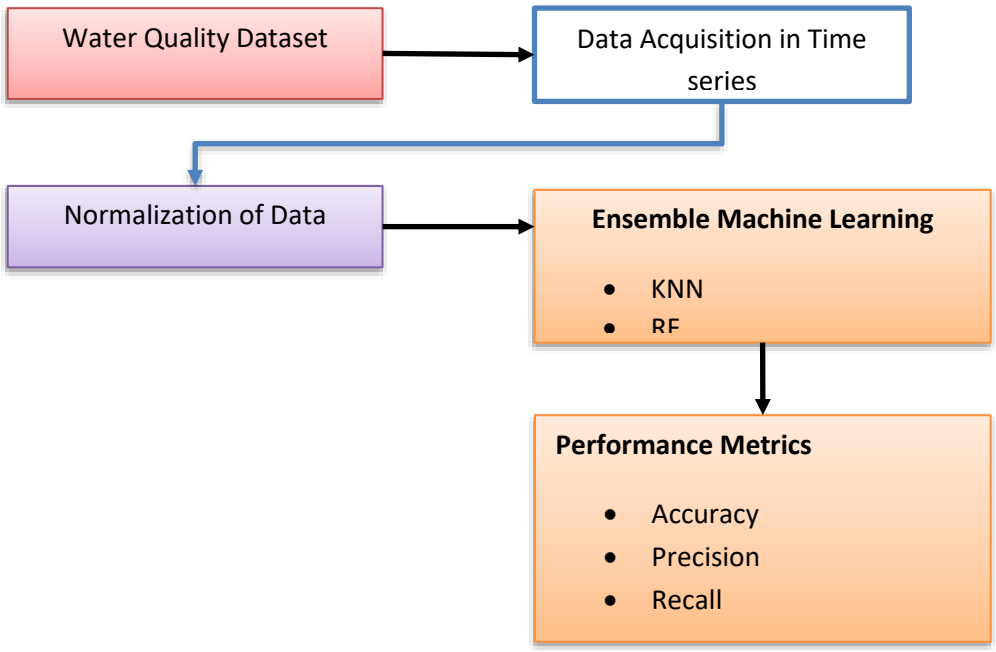

Figure-1 Overall block diagram for water quality prediction

3.1 Dataset details

This spatiotemporal "power of hydrogen (pH)" forecast for the next day is based on data from several water-measurement indices. USAGS releases this data [18]. The Savannah River basin and Atlanta water system are given high-level geographical information. It comprises December 1, 2014–January 28, 2016 training data. It includes 282 days of testing data from March 25, 2017, to January 1, 2018. The pH values were estimated from 37 Georgia water station daily samples. Spatial connectedness, or the exact linkages between locations via water streams, is ambiguous due to the complicated water system design. Eleven standard criteria determine each water station's pH, including temperature, specific conductance, and dissolved oxygen. A matrix with 37 rows and 11 columns holds the input data. The rows are water stations, and the columns are pH-altering materials. The training data comprises 423 instances of spatial matrices, each with dimensions of 37 × 11. The dimensions of the test input are 282 units in length, 37 units in width, and 11 units in height. This provides the daily pH readings for 37 water stations. The training and test output data dimensions are (423 × 37) and (282 × 37), respectively.

3.2 Data preparation

Monitoring station data includes observed values $\{x_t\}$ that are recorded at certain intervals t. Here, hourly time series data is collected. After data imputation, we normalize all observations inside [0,1] using the following approach.

$$X_t = \frac{x_t - \min(x_t)}{max(x_t) - \min(x_t)} \quad (1)$$

In addition, the time series is decomposed into its trend, seasonality, and irregular components using an additive model. Nevertheless, the research does not include the cyclic component:

$$X_t = trend_t + cyclic_t + seasonal_t + irregular_t \quad (2)$$

The trend component $trend_t$ at time t shows the series' long-term development, which might be linear or non-linear. The seasonal component $seasonal_t$, indicates seasonal fluctuations at time t. The "irregular" component (also known as "noise") at time t exhibits stochastic and unpredictable impacts. Occasionally, time series may show cyclic variations $cyclic_t$, which are not exactly periodic.

3.3 Normalization of data

Standardizing the original data guarantees consistent magnitude, therefore allowing comparison and evaluation. A linear transformation that equally rescales data from 0 to 1 is minimum-maximum normalization, often referred to as outlier normalization. Z-score standardization is yet another approach to data normalization. However, Z-scores carry potential dangers. At first, the Z-score requires knowing the average and spread of the dataset, which might be difficult in real-world research and data exploration. Usually, it is replaced by the average and standard deviation of a sample. A normal distribution is the most suitable data distribution for obtaining Z-scores. Therefore, we opt for min-max normalization. It is more effective when the data contains values that are in close proximity to one another. In this investigation, the transformation function employed for min-max normalization is as follows:

$$x' = \frac{x - x(min)}{x(max) - x(min)} \quad (3)$$

Where $x(max)$ and $x(min)$ is the maximum and minimum value in the sample data.

3.4 Water quality prediction process

The dataset has two parts: training and testing. The classification model will be trained on 80% of the data. The testing data set, 20% of the data, will assess the model's accuracy on untested data. The first phase involves constructing a model by thoroughly evaluating and analyzing the data tuples from the training data, which consists of a set of characteristics. The need for a class label attribute is recognized for each tuple in the training data. Figure 2 displays the three ensemble classifiers along with a quality indicator.

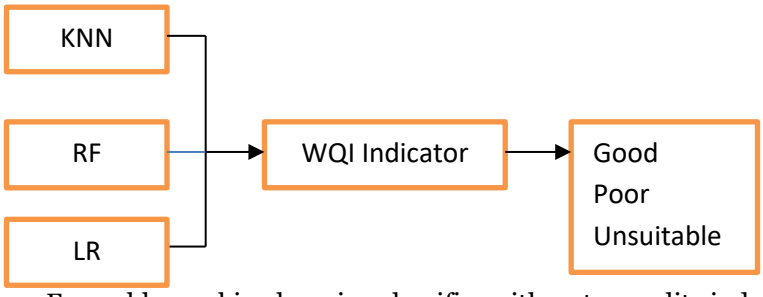

Figure-2 Ensemble machine learning classifier with water quality index block

3.4.1 K-means nearest neighbor

The KNN classifier is a popular pattern recognition algorithm due to its simplicity and power. It is also non-parametric, meaning it makes no data distribution assumptions. It involves determining the Euclidean distance between test data and training data samples. Test data is then categorized by the majority of its k-closest training data neighbors. Usually, "k" is a tiny positive integer. Differentiating across classes becomes harder as K increases. The best K value is chosen via cross-validation. KNN classifiers compare test data points to all training data points. The classifier guesses the test data point's label based on the d_1 distance to the nearest training class.

$$d_1(I_1, I_2) = (\sum_p |I_1^p - I_2^p|) \qquad (4)$$

where I_1 and I_2 represent the vectors for points 1 and 2, respectively. Let d_1 signify the distance and Σ be the sum calculated overall points. Typically, in the KNN algorithm, the number of neighbors is represented by N, N samples are selected and evaluated using a certain distance metric,

$$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|.\, 1/p) \qquad (5)$$

where p = 1 represents Manhattan distance, p = 2 Euclidean distance, and $p = \infty$ Chebyshev distance.
This is the K-NN algorithm:

- Step 1: Input and split the dataset into training and testing sets.
- Step 2: Select an instance from the testing sets and compute its distance from the training set.
- Step 3: Sort distances ascendingly.
- Step.4: The instance's class dominates the three initial training examples (k=3).

3.4.2 Random forest

RF is also a widely used supervised ML algorithm. This method applies to both classification and regression tasks; however, it is generally more effective in classification tasks. Additionally, it functions effectively with datasets of considerable size and high dimensionality. The goal is to integrate several weak learners into one strong learner. Ensemble learning approach RF builds numerous decision trees during training. Individual tree predictions define the modal class. RF consists of tree predictors that rely on a random vector selected separately with the same distribution for all forest trees,

$$entrophy\ (s) = \sum_{i=1}^{c} -kn[gf] \qquad (6)$$

The RF algorithm is:

- Step.1: Randomly selected K data points from the training set.
- Step.2: The K data points create the decision trees.
- Step.3: Select the desired N-tree from the created trees and proceed to repeat the first two steps.
- Step.4: Develop an N-tree that can forecast a new data point's categorization and assign it to the most likely category.

3.4.3 Logistic regression

This approach uses the Ridge estimator multinomial logistic regression to construct classifiers. The matrix may be constructed as a $n * (c - 1)$ matrix when a dataset has c classes for m data cases with n variables. The probability for class j, excluding the final class, is computed as

$$pj(Xi) = exp(XiBj)/((sum[j = 1,2, \ldots . (k - 1)]exp(Xi * Bj)) + 1) \quad (7)$$

Logistic regression analyses data to find relationships between independent factors and dependent variables. Usually, when using this approach for predictive tasks, the input dataset consists of two potential values for the dependent variable (target class). Comprehending how uncertainty and model structure interact is essential to conduct accurate water quality evaluations. The combination of Random Forest and KNN, together with the iterative optimization process, enhances their robust performance but also increases unpredictability. The WQI is a prominent measure that influences the quality. The WQI is calculated by including many criteria,

$$WQI = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i=1}^{N} w_i} \quad (8)$$

where N, q_i, and w_i represent the number of parameters, quality rating scale, and unit weight for parameter i , respectively. Equation (9) calculatesq_i,

$$q_i = 100 \times \left(\frac{v_i - v_{id}}{s_i - v_{id}}\right) \quad (9)$$

where v_i is the projected value for the parameter i, v_{id} is the ideal value for pure water and s_i is the standard value. From equation (10), w_i is the unit weight,

$$w_i = \frac{k}{s_i} \quad (10)$$

The constant of proportionality, denoted as k, is calculated using equation (11),

$$k = \frac{1}{\sum_{i=1}^{N} s_i} \quad (11)$$

Consequently, the quality of water is categorized as shown in Table 1.

Table-1 Classification of Water Quality Level

WQI rate	**Classified result**
0–50	Good
51–100	Poor
More than 100	Unsuitable

4. Performance analysis

Experimental setup—We trained the model using an 8GB Nvidia Geforce 2080 GPU. The training method used 100 epochs and 8 batches. This research used a personal computer equipped with an Intel Core i3 central processing unit and 8 gigabytes of random-access memory. Python 3.9.7 is used for classification and optimization.

Performance matrix- Many performance metrics may assess water quality prediction. Performance evaluation methods, including accuracy, precision, recall, and Matthew's correlation coefficient, tested the proposed model's WQ prediction capacity. The statistical techniques are well-defined, and Equation (12) calculates accuracy,

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (12)$$

Precision is, Eq. (13):

$$precision = \frac{TP}{TP+FP} \quad (13)$$

Recall is, Eq (14)

$$recall = \frac{TP}{TP+FN} \quad (14)$$

F1 score is, Eq. (15)

$$F1 - score = \frac{2*recall*precision}{recall+precision} \quad (15)$$

MCC is, eq.(16)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (16)$$

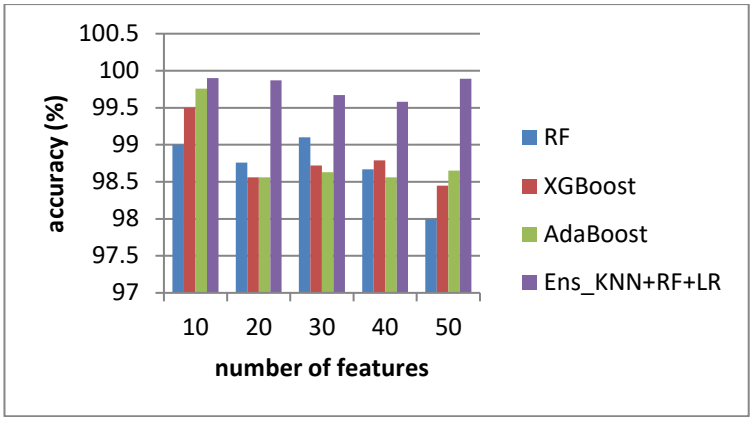

Figure-3 Calculation of accuracy

Figure 3 presents the accuracy results of various models when using 50 features. The existing RF method has an accuracy of 98%, while XGBoost performed with an accuracy of 98.45%, and AdaBoost showed further improvement, reaching an accuracy of 98.65%. However, the proposed ensemble method (Ens_KNN+RF+LR), significantly outperformed the other models with an impressive accuracy of 99.89%.

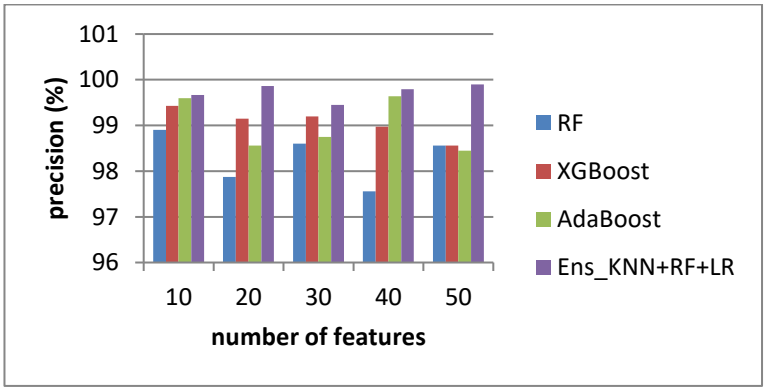

Figure-4 Calculation of precision

Figure 4 shows the precision scores of various models using 50 features. The Existing Random Forest (RF) and XGBoost both achieved a precision of 98.56%. AdaBoost was slightly behind, with a precision of 98.45%. The proposed ensemble method (Ens_KNN+RF+LR), outperformed the other models, achieving a precision of 99.9%.

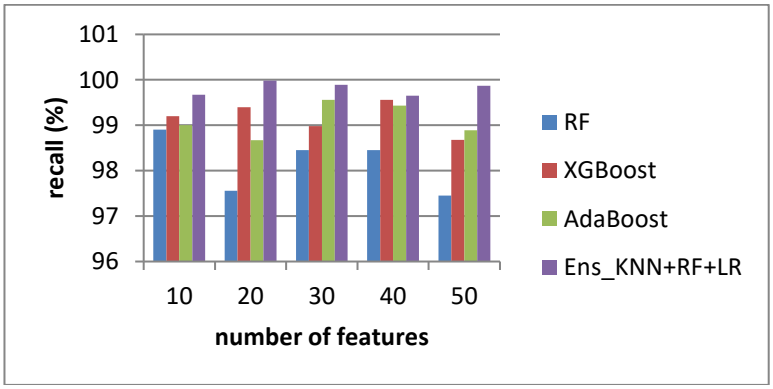

Figure-5 Calculation of recall

Figure 5 shows the recall scores for different models using 50 features. The existing Random Forest (RF) achieved a recall of 97.45%, while XGBoost recorded a recall of 98.68%, and AdaBoost demonstrated a slightly higher recall at 98.89%. The ensemble method (Ens_KNN+RF+LR), achieved the highest recall score of 99.87%.

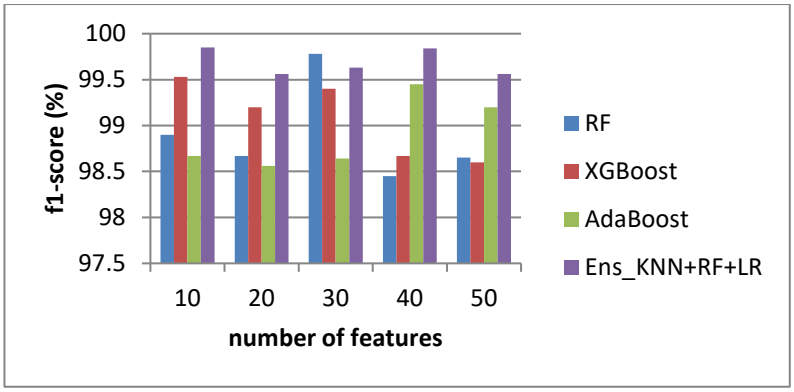

Figure-6 Calculation of f1-score

Figure 6 shows the F1 Scores for various models using 50 features. Random Forest (RF) achieved an F1 Score of 98.65%, while XGBoost followed closely with an F1 Score of 98.6%. AdaBoost performed slightly better, with an F1 Score of 99.2%. The ensemble method (Ens_KNN+RF+LR) achieved the highest F1 Score of 99.56%.

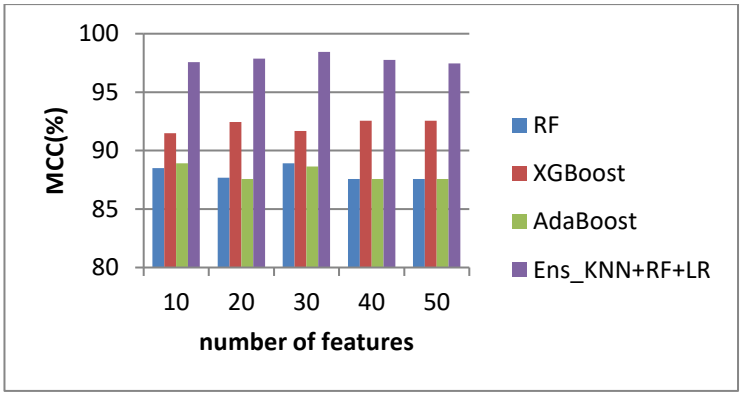

Figure-6 Calculation of MCC

Figure 6 shows the Matthews Correlation Coefficient (MCC) scores for various models using 50 features. Random Forest (RF) and AdaBoost both achieved an MCC of 87.56%, indicating a moderate level of performance. XGBoost demonstrated a higher MCC of 92.56%, reflecting improved correlation between the predicted and actual classifications. The proposed ensemble method (Ens_KNN+RF+LR) achieved the highest MCC of 97.45%.
Table 2 displays the existing RF, XGBoost, AdaBoost, and proffered Ens_KNN+RF+LR classification models by performance.

Table 2. Performance of the existing and Proffered method

Parameters	Accuracy	Precision	Recall	F1-score	MCC
RF	98.01	97.92	97.92	97.92	87.52
XGBoost	98.31	98.21	98.24	98.25	90.55
AdaBoost	98.12	98.03	98.05	98.03	87.93
Proposed Ens_KNN+RF+LR	**99.98**	**99.87**	**99.67**	**99.85**	**97.56**

Conclusion

Recent environmental factors have resulted in water pollution, leading to detrimental impacts on human health and the emergence of several complex diseases. Therefore, the ability to forecast water quality is crucial for ensuring the survival of the human species. Ensemble machine learning approaches have shown significant promise in improving water management via water quality forecasting. By incorporating the strengths of multiple models, the proposed Ens_KNN+RF+LR has demonstrated enhanced accuracy and robustness in predicting complex water quality parameters. This approach not only reduces the risk of overfitting but also ensures better generalization across different water bodies and environmental conditions. Furthermore, timely and accurate forecasts are supported by the integration of several data sources, such as historical records and real-time sensor data, which helps with proactive decision-making for pollution control, water treatment optimization, and guaranteeing a safe water supply. Future work should focus on enhancing model scalability, incorporating more diverse datasets, improving real-time data processing capabilities, and exploring hybrid models that combine machine learning with domain-specific knowledge to further refine prediction accuracy and operational effectiveness.

References

1. Jain D, Shah S, Mehta H et al (2021) A Machine Learning Approach to Analyze Marine Life Sustainability. In: Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Springer, pp 619–632
2. Clark RM, Hakim S, Ostfeld A (2011) Handbook of water and wastewater systems protection. In: Protecting Critical Infrastructure. Springer, pp 1–29. https://doi.org/10.1007/978-1-4614-0189-6
3. Hu Z, Zhang Y, Zhao Y et al (2019) A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. Sensors 19:1420
4. Zhou J, Wang Y, Xiao F et al (2018) Water quality prediction method based on IGRA and LSTM. Water 10:1148
5. Waqas M, Tu S, Halim Z et al (2022) The role of artifcial intelligence and machine learning in wireless networks security: principle, practice and challenges. Artif Intell Rev 55:5215–5261. https://doi.org/10.1007/s10462-022-10143-2
6. Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text features for the identifcation of dominant emotion in an email. Knowl Based Syst 208:106443. https://doi.org/10.1016/j.knosys.2020.106443
7. Wu J, Wang Z (2022) A Hybrid Model for Water Quality Prediction Based on an Artifcial Neural Network, Wavelet Transform, and Long Short-Term Memory. Water 14:610
8. Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. Int J Environ Res Public Health 15:1322
9. Liu P, Wang J, Sangaiah AK et al (2019) Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability 11:2058
10. Hmoud Al-Adhaileh M, Waselallah Alsaade F (2021) Modelling and prediction of water quality by using artifcial intelligence. Sustainability 13:4259
11. Azrour, Mourade, et al. "Machine learning algorithms for efficient water quality prediction." *Modeling Earth Systems and Environment* 8.2 (2022): 2793-2801.
12. Abba, Sani Isah, et al. "Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index." *Environmental Science and Pollution Research* 27 (2020): 41524-41539..
13. Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water quality prediction using KNN imputer and multilayer perceptron. *Water*, *14*(17), 2592.
14. Li, T., Lu, J., Wu, J., Zhang, Z., & Chen, L. (2022). Predicting aquaculture water quality using machine learning approaches. *Water*, *14*(18), 2836.
15. Ahmed, Umair, et al. "Efficient water quality prediction using supervised machine learning." *Water* 11.11 (2019): 2210.
16. Li, Shicheng, and James Yang. "Modelling of suspended sediment load by Bayesian optimized machine learning methods with seasonal adjustment." *Engineering Applications of Computational Fluid Mechanics* 16.1 (2022): 1883-1901.
17. Uddin, M. G., Nash, S., Diganta, M. T. M., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*, *321*, 115923.
18. Zhao, L.; Gkountouna, O.; Pfoser, D. Spatial Auto-regressive Dependency Interpretable Learning Based on Spatial Topological Constraints. ACM Trans. Spat. Algorithms Syst. 2019, 5, 1–28.
19. Ahmed, Ali Najah, et al. "Machine learning methods for better water quality prediction." *Journal of Hydrology* 578 (2019): 124084.
20. Ayyalasomayajula, Madan Mohan Tito, and Sailaja Ayyalasomayajula. "Proactive Scaling Strategies for Cost-Efficient Hyperparameter Optimization in Cloud-Based Machine Learning Models: A Comprehensive Review."
21. Ayyalasomayajula, Madan Mohan Tito, Sathishkumar Chintala, and Sandeep Reddy Narani. "Optimizing Textile Manufacturing With Neural Network Decision Support: An Ornstein-Uhlenbeck Reinforcement Learning Approach." *Journal of Namibian Studies: History Politics Culture* 35 (2023): 335-358.
22. Ayyalasomayajula, Madan Mohan Tito, Sathish Kumar Chintala, and Sailaja Ayyalasomayajula. "A Cost-Effective Analysis of Machine Learning Workloads in Public Clouds: Is AutoML Always Worth Using?."
23. Ayyalasomayajula, Madan Mohan Tito, Sathishkumar, Chintala. "Fast Parallelizable Cassava Plant Disease Detection using Ensemble Learning with Fine Tuned AmoebaNet and ResNeXt-101". *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 11. 3(2020): 3013–3023.