Architecture
Image Studies

Vol.6, Issue 4, pp.963-973, 2025
DOI: https://doi.org/10.62754/ais.v6i4.702
© by AP2 on Creative Commons 4.0
International License (CC BY-NC 4.0)
https://journals.ap2.pt/index.php/ais/index

An Ensemble-Based Predictive Learning (EBPl) Model for Optimized Water Quality Analysis in Smart Ecosystems

Manohari P[1], G. Kannan[2]

Abstract

Water quality (WQ) is essential to making sure the sustainability of ecosystems, especially in smart environments where automation and data-driven decision-making play key roles. Monitoring and managing water resources efficiently becomes increasingly important as urbanization and industrial activities intensify. Efficient water quality monitoring is crucial for sustainable water utilization in diverse uses, including drinking, bathing, irrigation, and aquaculture. Water quality is evaluated according to its chemical, biological, and physical constituents, with anthropogenic activities such as industrial waste disposal as a significant influencing element. Machine learning (ML) algorithms have recently emerged as efficient methods for WQ classification. This paper focuses on an Ensemble-based predictive learning (EBPL) model for achieving optimized water quality analysis in smart ecosystems by integrating multiple algorithms to provide more accurate, reliable, and adaptable predictions. This work uses a KaggleWQ dataset to train EBPL models that classify WQ according to the Water Quality Index (WQI). The ensemble-based predictive learning (EBPL) model for WQ analysis is developed using Random Forest (RF), AdaBoost, Support Vector Machine (SVM), and hyper-parameter modification using Randomized Search CV. Ensemble learning was proposed to enhance classification accuracy by combining model outputs using voting, stacking, and boosting techniques. This approach leverages the advantages of each model, producing an extremely accurate and reliable system for water quality monitoring in sustainable smart ecosystems.

Keywords: *Random Forest; Randomized Search CV; SVM; AdaBoost; Machine learning; Ensemble learning; Predictive Learning.*

Introduction

Although water is essential for life, it is frequently polluted by human activities and industrial waste. To address this, organizations like the Central Pollution Control Board (CPCB) and the World Health Organization (WHO) have established standards for potable water, setting limits for various water quality parameters (WQP) based on its intended use, such as drinking, swimming, irrigation, water sports, harbor activities, and aquaculture. Each year, the WHO reports that 485,000 people die from water-borne diseases caused by unsafe water. The chemical, physical, and biological characteristics of water are evaluated to determine its appropriateness for applications like irrigation, shell-fishing, supporting aquatic life, and human consumption. A systematic classification system is developed for WQ to support different uses. The employed dataset features essential parameters including dissolved oxygen, pH, conductivity together with biological oxygen demand (BOD) and fecal coliform to support practical water quality assessment. The traditional WQ categorization process using sample collection with laboratory tests requires substantial time and workforce. There exists a need for better and more precise water classification methods which must accommodate the needs of drinking water alongside irrigation water and aquaculture and bathing water uses.

The implementation of artificial intelligence (AI) presents itself effectively for real-time WQ classification problems during modern years. The subset of Artificial Intelligence known as ML has demonstrated significant potential in enhancing both accuracy and effectiveness of this assignment. The article implements different ML models for its WQ classification approach. The paper introduces Ensemble-based predictive learning for enhancing water quality measurement in smart ecosystems

[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalai Nagar – 608002, India (corresponding author).
[2] Assistant Professor, Department of Computer Science, Govt. Arts and Science College, Manalmedu– 609202. (Deputed from Dept. of CIS, Annamalai University), India.

through multiple algorithm integration to achieve more precise and dependable flexible predictions. The EBPL model applies various machine learning algorithms such as RF, AdaBoost and SVM with RandomizedSearchCV for performing hyper-parameter optimization. Each predictive model demonstrates distinct performance levels according to their success in water quality classification as shown in the results. Through their implementation we gain a successful and precise method for water quality classification which allows us to apply measures properly based on results while improving general water management systems.

Literature Survey

Water is a fundamental resource that sustains human existence and animal life together with plant life and supports both personal activities and industrial processes. The research by Ajayi et al. [2022] utilized Logistic Regression, RF and SVM as ML models to perform WQ classification for drinking and irrigation purposes. Logistic Regulation proved most successful in the prediction of drinking water quality yet SVM demonstrated the best results for irrigation water classification. The research relied on information acquired through an IoT data collection system which measured pH together with turbidity and temperature alongside total dissolved solids (TDS). The processed IoT sensor data enabled implementation of ML models after successful pre-processing of the data. The WQI prediction required regression models while classification models determined the WQ Classification. The Multilayer Perceptron (MLP) model provided the best R-squared result of 0.93 according to Rahu et al. [2023] because Support Vector Regression (SVR) showed a lower R-squared of -0.73. RF model achieved an accuracy score of 0.91 in the classification task which exceeded the performance of three other models namely SVM and XGBoost and Decision Tree.

The WQI functions widely to evaluate water quality in management systems. The research by Aslam et al. [2022] applied 12 different data mining algorithms integrated using techniques including bagging and cross-validation while using Random Forest Classifier (RFC) to predict more precisely. Researchers determined that combined techniques of machine learning generate better results than individual implementations because the RT-ANN model reached an R-squared score of 0.951 at peak performance and this makes the system both efficient and economical for WQI prediction. The researchers at Chakravarthy et al. [2023] created an AdCSO-sELM model that achieved outstanding results in potable water quality prediction with accuracy at 96.54% and precision at 98.48%. Cai et al. [2022] determined Random Forest and 1D-CNN as reliable models based on their work which showed that the 1D-CNN model produced an R-squared value of 0.87 but Random Forest delivered reliable predictions while avoiding complex optimization.

The research by Al-Sultani et al. [2021] developed five ensemble ML models to predict biochemical oxygen demand (BOD) through the use of Quantile Regression Forest (QRF), RF, Radial SVM, Stochastic Gradient Boosting (GBM) and (GBM) using H2O (GBM_H2O). The combination of genetic algorithms with principal component analysis developed the best BOD prediction model which PCA-QRF proved to be. Abdullah et al. [2020] improved search space exploration algorithms for water quality prediction through their Dual-KA algorithm yet this method still yielded weaker outcomes than the leading models reported in literature. This method proved to be an effective tool for WQ prediction although it did not match the best-performing techniques in the literature. Kirui et al. [2024] conducted an evaluation of XGBoost alongside SVM, MLP, K-Nearest Neighbors (KNN) and DT in order to establish their capability for water quality classification. The researchers established XGBoost as the most suitable model because it reached 95.12% accuracy indicating the capability of machine learning algorithms to assist in water quality classification procedure.

The research of Hasan et al. [2021] evaluated DT and also included LR, SVM, NB and KNN models for water quality classification in Abu Dhabi drinking water. When using REPTree under these test conditions scientists achieved the best result with 98.47% accuracy. The performance analysis of four ML algorithms KNN, SVM, DT and RF was conducted by Pagadala et al. [2023] to evaluate water potability. The experimentation by the researchers revealed that RF surpassed its competitors at the pre-over-sampling stage and post-over-sampling stage where it accomplished the best possible accuracy. The researchers from Tanega et al. [2020] employed ML models to assess water quality at Taal Lake Philippines by measuring parameters which included temperature, pH, dissolved oxygen (DO), nitrates and phosphates. According to their findings RF and DT reached a 95% accuracy rate and were followed by SVM reaching 93.33%.

Radhakrishnan et al. [2022] employed SVM, DT and NB models together with weighted arithmetic WQI to conduct WQ classification. One dataset delivered 87.10% accuracy through the SVM model while the same model succeeded with 98.50% accuracy using a separate dataset. Both results surpassed the performances of DT and NB models. Khullar et al. [2023] utilized Decision Trees and NB as well as SVM and ensemble KNN and boosted trees in machine learning models which led to 99.65% accuracy from the Decision Tree model. Bearing Abirami et al.'s [2023] study in mind the models RF, DT, SVM, NB and KNN provided results indicating that RF delivered the best accuracy level of 91.97% for water quality classification. According to Yusri et al. [2022] SVM demonstrated lower performance than XGBoost for WQI prediction by achieving 65% cross-validation accuracy yet XGBoost reached 90% prediction accuracy level.

The analysis of water quality at Korattur Lake through ML models was studied by Danush et al. [2023]. Their experiment conducted evaluations of the Bagging-Decision Tree, Bagging-Support Vector Classifier (SVC), Boosting-Logistic Regression, KNN and XGBoost model models. According to the study results XGBoost yielded highest success with 99.82% accuracy while Bagging-Decision Tree scored 99.67% accuracy and KNN demonstrated 93.23% accuracy. Machine learning algorithms have succeeded remarkably in water quality prediction while providing accelerated results instead of traditional test methods used throughout years. The examined models including Random Forest, Decision Tree, SVM and XGBoost successfully identified and predicted water quality data with high accuracy in diverse experimental and testing conditions.

Xiaotong Zhu et al. [2022] developed a framework which uses ML techniques for WQPs estimation. The authors deployed their ensemble ML model framework through validation using measurements obtained from WQ stations together with monitoring station data. The authors used SHAP (Shapley Additive Explanations) to explain the prediction process of ensemble models. The experimental findings showed that executing ensemble ML models together with feature selection methods produced effective predictions of chlorophyll-a (Chla) and turbidity as well as dissolved oxygen (DO) concentrations.

The author Annie Jose and Srinivas Yasala [2024] developed an ensemble model which combined Hidden Markov Model (HMM) and Artificial Neural Network (ANN) for the prediction of groundwater quality in Kanyakumari District, Tamil Nadu, India. The collected data was pre-processed through cleaning and normalization to ensure it was ready for model training. HMM was accustomed to uncover concealed patterns in the data, while ANN was applied to forecast groundwater quality categories.

Ensemble-Based Predictive Learning (Ebpl) Model

Predictive models based on ensemble strategies use diverse separate models to improve both prediction quality and model reliability as well as strengthen predictive power. Ensemble methods reach better performance by uniting the advantages of AdaBoost, RF, SVM alongside RandomizedSearchCV through their combination of algorithms. Two main ensemble learning strategies include bagging because it generates several models from random subsamples of the data while averaging predictions and boosting through model weight adjustments across successive iterations. Stacking represents an effective technique that trains a meta-model to aggregate base model predictions. The collaborative prediction approach in ensemble learning provides voting as a critical technique which groups multiple model results to achieve enhanced accuracy while maintaining system reliability in classification problems.

AdaBoost

AdaBoost represents a reliable ensemble learning (EL) model which improves weak classifier accuracy through collective output processing into strong classifier outputs. The algorithm achieves its goal through a series of weak classifier training sessions using weighted data sets with extra emphasis on incorrectly handled data points. The structural framework of the final model uses the output from multiple weak classifiers where weights equal their performance rating. The research uses AdaBoost classification to determine water quality fit for drinking purposes and shell-fishing and bathing activities alongside irrigation usage and sports participation and harbor usage. The classification accuracy of water quality improved when AdaBoost worked with standard WQP measures pH, dissolved oxygen (DO), BOD, nitrate and conductivity to provide reliable water quality evaluations. Water quality management receives better support through this decision-making approach which assists monitoring activities.

Adaboost Steps

Initialize Weights:

If there are n training instances, the initial weight for each instance is:

$w_i = 1/n, for\ all\ i = 1, 2, \ldots, n$

For Every Iteration $t = 1, 2, \ldots, T$,

1.Train a Weak Learner

2.Calculate Weighted Error ε_t:

Determine the weighted error of the weak learner $h_t(x)$.

$$\varepsilon_t = \Sigma_{i=1}^{n} w_i * 1(h_t(x_i) \neq y_i) / \Sigma_{i=1}^{n} w_i$$

where $1(h_t(x_i) \neq y_i)$ is an indicator function that is 1 if the prediction $h_t(x_i)$ is incorrect and 0 if it is correct.

3.Compute Weak Learner's Weight:$\alpha_t = 1/2 * log\big((1 - \varepsilon_t)/\varepsilon_t\big)$

4.Update Weights of Training Examples:$w_i^{(t+1)} = w_i^{(t)} * e^{(\alpha_t * 1(h_t(x_i) \neq y_i))}$

Final Hypothesis: Following T iterations, a weighted sum of the final strong classifier's weak learners:

$$H(x) = sign\big(\Sigma_{t=1}^{T} \alpha_t * h_t(x)\big)$$

This last classifier combines all of the weak learners, where each weak learner contributes based on its weight α_t.

Random Forest (RF)

The RF algorithm operates as an EL method to execute classification and regression assignments. RF uses several DT and integrates them via merging techniques to achieve better predictive accuracy. The training data gets split into many random data subsets through bootstrap sampling that involves duplicated samples. Each training subset trains its own different DT distribution. Instead of evaluating all potential features at tree nodes the algorithm randomly selects a subset of features which results in finding the best split among them. The implementation of diverse trees prevents the model from fitting the noise in the data by reducing overfitting. The algorithm allows each DT to achieve its maximum depth without any form of pruning to produce deep trees that exhibit low bias with high variability.

Voting Mechanism: For classification tasks, the final prediction of the Random Forest for a particular target variable (e.g., 'Drinking', 'Bathing', etc.) is determined by a majority voting mechanism among the individual decision trees.

yRF =model ($\{y_i(x)\}_{i=1}^{N_{trees}}$)

where:yRF(x) is the predicted class by the RF for input x, $y_i(x)$ is the predicted class by the i-thDT and N_{trees} is the total number of DTs in the RF.

Support Vector Machine (SVM)

Supervised learning with SVM provides a dependable approach to examine the boundaries between different classes while working with regression tasks. The SVM system determines the maximum marginal hyperplane which supports every nearest data point from different classes. Through kernel transformations SVM extends dimensional values of input data to achieve versatility in linear and non-linear data classification. SVM demonstrates exceptional effectiveness when processing complex datasets since it provides flexible operation capabilities.

Step 1: Calculate Euclidian distance between data points in different classes and identify the support vectors (s1,s2,..,sn)

Step 2: Use a bias to modify the support vectors

Step 3: Apply support vector equations and solve to find the α values.

α1 Φ(s1) · Φ(s1) + α2 Φ(s2) · Φ(s1) + …..+αn Φ(sn) · Φ(s1)= classlabel

where Φ is the kernel function

Step 4: Calculate Weight Vector by Apply α values in weight vector w=$\sum_i \alpha_i . s_i$

Step 5: y=wx+b produce maximum margin hyperplane.

RandomizedSearchcv

Randomized Search Cross-Validation (RandomizedSearchCV) represents an efficient approach to perform machine learning model hyperparameter tuning. The RandomizedSearchCV method selects random parameter combinations from specified spaces instead of conducting complete searches like grid search. Using this technique provides substantial reductions in required computing power as well as corresponding resource demands mainly because of its ability to handle large datasets along with complex model structures. RandomizedSearchCV operates through distribution definition of hyperparameters followed by random combination sampling and cross-validation testing to find optimal model performance. Performance evaluations of model accuracy or F1 score determine the selection of the best hyperparameter combination. RandomizedSearchCV achieves a practical equilibrium between extensive parameter evaluation and computational efficiency because it normally generates results which compete against exhaustive search approaches.

Ensemble Model

The ensemble model establishes voting algorithms to merge the classifiers through hard voting which realizes predictions by majority agreements and soft voting which calculates class average probabilities. The real-time water quality prediction phase concentrates on various use cases which involve drinking water and irrigation and sports activities with attention to reliability and accuracy. The ensemble model gets evaluated using performance metrics which include precision and accuracy and recall and these metrics demonstrate its success against leading contemporary methods for smart water quality surveillance. Figure 1 demonstrates the suggested EBPL model's architecture.

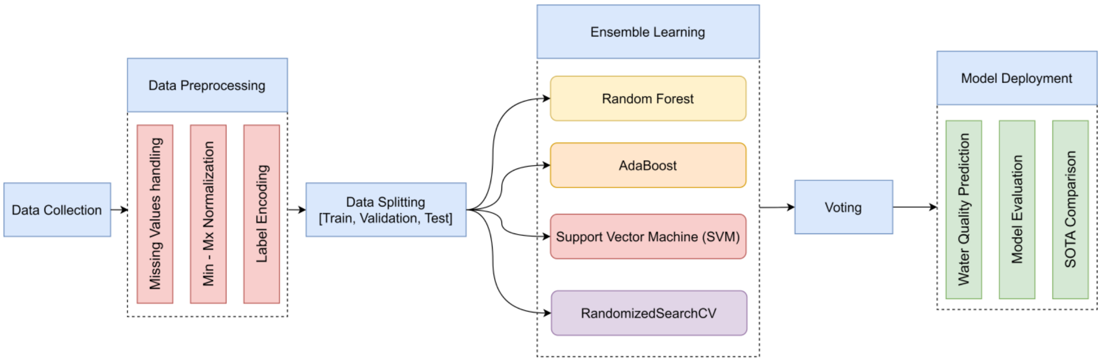

Figure 1. Architecture of the proposed EBPL model.

The methodology for classifying water quality using ensemble learning starts with collecting data, utilizing a water quality dataset from Kaggle. This dataset encompasses various features that represent water quality's chemical, biological, and physical attributes, consisting of pH, DO, TDS, and turbidity. Following data collection, pre-processing is conducted to address missing values through imputation techniques and to apply Min-Max normalization, adjusting the characteristics to a consistent range of 0 to 1. Additionally, label encoding is implemented to convert categorical labels (such as drinking and irrigation) into numerical values suitable for ML models. The preprocessing operations enable the division of the dataset into training, validation and test sets that support reliable model training along with hyperparameter optimization and assessment steps. The next phase involves training various models within the ensemble learning framework, which includes RF, AdaBoost, SVM, and RandomizedSearchCV for hyper-parameter optimization. These models classify water quality into distinct categories, with AdaBoost achieving the highest accuracy at 99.5%, Random Forest at 98%, SVM at 91%, and RandomizedSearchCV at 99.3%.

Experiment and Result Analysis

This paper presents an ensemble-based predictive learning model designed to optimize water quality analysis in smart ecosystems by combining multiple algorithms to deliver more accurate, reliable, and adaptable predictions. To train the model, a Kaggle water quality dataset with the objective of classifying water quality according to the WQI. The dataset, which includes WQ metrics from 620 different water bodies, was sourced specifically for this study. The objective of this research aims to determine the most appropriate ML algorithm for classifying water quality across various categories, including drinking, bathing, shellfishing, irrigation, water sports, harbour activities, and aquaculture.

Each category of water quality is assessed against established thresholds based on factors like pH, BOD, DO, nitrate levels, and conductivity. The classification of water usage is guided by different parameter values presented in the study. Table 1-7 outlines the water quality standards for drinking, bathing, irrigation, shell-fishing, water sports, harbour, and Aquaculture.

Table 1. Water Quality Metrics for Drinking Water

Parameter	Range
PH values	6.5 –8.5
BOD	1 mg/L - 3mg/L
DO	6.5 mg/L - 12 mg/L
Conductivity	>500 microS/cm
Nitrate	<4.5 mg/L

Table 2. Water Quality Metrics for Bathing Water

Parameter	Range
PH values	6.5 –8.5
BOD	<3 mg/L
DO	4 mg/L - 18 mg/L
Conductivity	100>micro S/cm –<1000 microS/cm
Nitrate	<10 mg/L

Table 3. Water Quality Metrics for Irrigation

Parameter	Range
PH values	5.5 –8.5
BOD	2 mg/L - 10 mg/L
DO	5 mg/L - 18 mg/L
Conductivity	100>micro-S/cm –< (800-1000) microS/cm
Nitrate	< 10 mg/L

Table 4. Water Quality Metrics for Shell fishing

Parameter	Range
PH values	7.0-8.5
BOD	4 mg/L - 10 mg/L
DO	5 mg/L - 12 mg/L
Conductivity	2>micro-S/cm – <30 microS/cm
Nitrate	< 10 mg/L

Table 5. Water Quality Metrics for Water Sports

Parameter	Range
PH values	(6.5-7.0) –8.5
BOD	2 mg/L –5 mg/L
DO	5 mg/L - 12 mg/L
Conductivity	50>micro-S/cm – <1000 microS/cm
Nitrate	< 10 mg/L

Table 6. Water Quality Metrics for harbour

Parameter	**Range**
PH values	(6.5-7.0) -(8.5-9.0)
BOD	2 mg/L - (10mg/L)
DO	5 mg/L - 12 mg/L
Conductivity	2000>micro-S/cm - <50000 microS/cm
Nitrate	< 10 mg/L

Table 7. Water Quality Metrics for Aquaculture

Parameter	**Range**
PH values	6.5 - 9.0
BOD	2 mg/L - (5-6 mg/L)
DO	5 mg/L - 15 mg/L
Conductivity	100>micro-S/cm - <2000 microS/cm
Nitrate	< 10 mg/L

This study employed AdaBoost, SVM, RF, RandomizedSearchCV, and ensemble models utilizing both hard and soft voting to assess the performance of these algorithms. Table 8 presents the accuracy of the predictive learning models in classifying water quality for various applications. Among the individual models, AdaBoost attained the highest accuracy at 96%, while RandomizedSearchCV achieved an accuracy of 95%, outperforming both Random Forest and SVM.

Table 8. Accuracy of the Predictive Learning Model Based on Water Classification Type

	Drinking	Bathing	Irrigaton	Shell fishing	Water sports	Harbour & Aquaculture	Total model accuracy
Adaboost	0.97	0.97	0.95	0.96	0.97	0.96	0.96
RF	0.95	0.96	0.94	0.95	0.96	0.93	0.95
Randomizedsearchcv	0.96	0.97	0.95	0.96	0.97	0.94	0.96
SVM	0.94	0.91	0.91	0.91	0.90	0.93	0.92
EBPL_Hard Voting	0.97	0.97	0.95	0.96	0.97	0.96	0.96
EBPL-Soft Voting	0.99	1.00	0.99	0.99	1.00	1.00	0.99

All predictions from four separate models were combined through the hard and soft voting ensemble methods whose results are displayed in Table 8. The hard voting approach reaches its decision through the majority vote from participating models whereas soft voting determines output from weighted average probability score calculations. The ensemble model with hard voting obtained 97% accuracy but the ensemble model with soft voting scored 99% accuracy. This superior performance can be attributed to the integration of multiple classifiers, which enhances robustness and overall accuracy. This is because the more accurate models, AdaBoost and RandomizedSearchCV, dominate the ensemble, ensuring robust classification.

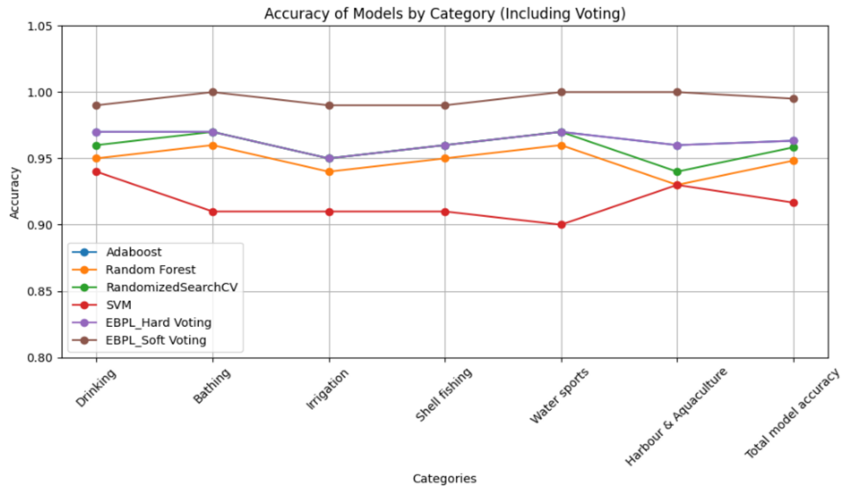

Figure 2. Performance of EBPL Model Vs Other ML model

Figure 2 illustrates that the proposed ensemble method, the EBPL -soft voting model, consistently achieves high performance with an accuracy of 0.99 across all water categories. AdaBoost also demonstrates strong performance in most categories, although it experiences a slight decline in the "Irrigation" category, maintaining a score of around 0.95 in the others. In contrast, the SVM exhibits the lowest performance among the models, particularly struggling in the "Shellfishing" and "Water Sports" categories, where its accuracy falls close to 90%. This analysis indicates that the proposed EBPL-soft voting model is the most reliable for these categories, while other machine learning models may require further tuning to achieve competitive performance in specific areas.

Table 9. Classification Performance metrics of EBPL Model Vs Other ML model

	Accuracy	**Precision**	**Recall**	**F1-Score**
Adaboost	0.96	0.97	0.98	0.96
RF	0.95	0.93	0.91	0.92
Randomizedsearchcv	0.96	0.95	0.96	0.96
SVM	0.92	0.86	0.81	0.83
EBPL_ Hard Voting	0.96	0.97	0.97	0.97
EBPL_Soft Voting	0.99	0.98	1.00	0.99

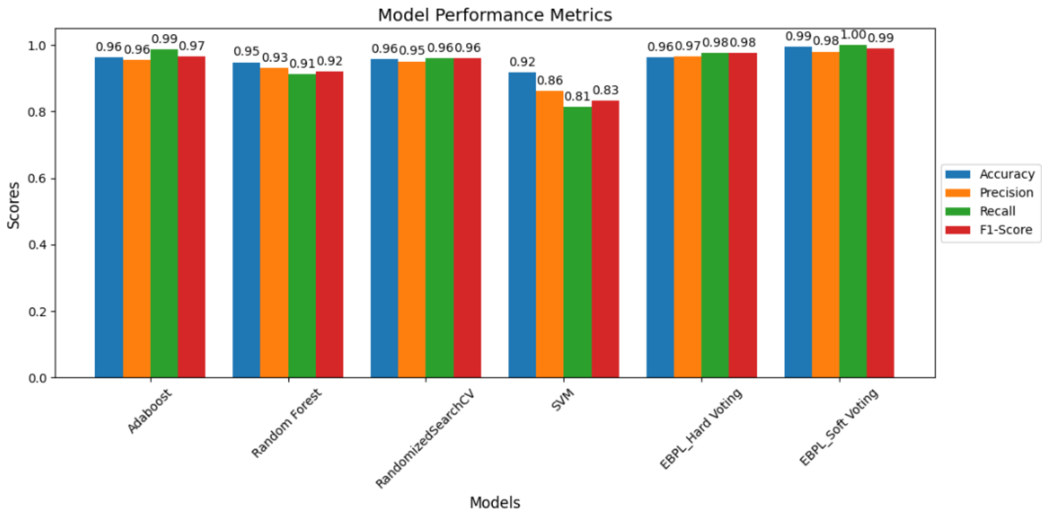

Figure 3. Classification Performance metrics of EBPL Model Vs Other ML model

A comparison of six approaches—AdaBoost, RF, RandomizedSearchCV, SVM, EBPL_Hard Voting, and EBPL_Soft Voting (as shown in Figure 3 and Table 9)—reveals their performance across various metrics. The EBPL_Soft Voting system produced the most optimal results, achieving an accuracy of 0.99, precision of 0.98, recall of 1, and an F1-Score of 0.99. This model is particularly effective in scenarios requiring high precision, ensuring that nearly all positive predictions are accurate, while also maintaining excellent recall and F1-Score. It also outperformed other machine learning models, achieving an accuracy of 0.96, precision of 0.97, recall of 0.97, and an F1-Score of 0.97.

AdaBoost and RandomizedSearchCV both recorded an accuracy of 0.96, with slight differences in precision, recall, and F1-Scores, reflecting their effectiveness in hyperparameter tuning. In contrast, SVM had the lowest performance across all metrics, accuracy of 0.92, precision of 0.86, recall of 0.81, and F1-Score of 0.83, indicating that it may require further tuning for this specific dataset or problem.

Conclusion

The paper shows how ML approaches succeed at WQ prediction tasks. The proposed ensemble model, EBPL_soft voting, achieves remarkable accuracy and robustness, reaching an impressive accuracy of 99%, which outperforms other ML approaches such as RF, SVM, and RandomizedSearchCV. Key parameters influencing water quality predictions include pH, DO, BOD, nitrate levels, and conductivity. This model effectively classifies water into six distinct usage categories: drinking water, bathing, shellfishing, irrigation, water sports, and harbour and aquaculture. These findings indicate promising opportunities for leveraging machine learning to enhance WQ management and decision-making processes, Health advantages and environmental sustainability develop as a result of combined heat and power systems.

Declarations

Funding

On Behalf of all authors the corresponding author states that they did not receive any funds for this project.

Conflicts Of Interest

The authors declare that we have no conflict of interest.

Competing Interests

The authors declare that we have no competing interest.

Data Availability Statement

All the data is collected from the simulation reports of the software and tools used by the authors. Authors are working on implementing the same using real world data with appropriate permissions.

Ethics Approval

No ethics approval is required.

Consent To Participate

Not Applicable

Consent For Publication

Not Applicable

Human And Animal Ethics

Not Applicable.

Code Availability

Not Applicable.

Author contributions

All authors significantly contributed to the development of the described tool, and are currently actively involved in it. The first draft of the manuscript was written by Corresponding Author and co-author improved on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

There is no acknowledgement involved in this work.

References

A. N. Hasan and K. M. Alhammadi, 2021 “Quality Monitoring of Abu Dhabi Drinking Water Using Machine Learning Classifiers," 14th International Conference on Developments in eSystems Engineering (DeSE), Sharjah, United Arab Emirates, pp. 1-6, doi: 10.1109/DeSE54285.2021.9719373.

A. O. Al-Sulttani, M. Al-Mukhtar, A. B. Roomi, A. A. Farooque, K. M. Khedher and Z. M. Yaseen, 2021, "Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction," in IEEE Access, vol. 9, pp. 108527-108541, doi: 10.1109/ACCESS.2021.3100490.

1. Annie Jose, Srinivas Yasala 2024, Machine learning-based ensemble model for groundwater quality prediction: A case study. Water Practice and Technology 1 June 2024; 19 (6): 2364–2375. doi: https://doi.org/10.2166/wpt.2024.139

B. Aslam, A. Maqsoom, A. H. Cheema, F. Ullah, A. Alharbi and M. Imran, 2022, "Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach," in IEEE Access, vol. 10, pp. 119692-119705, doi: 10.1109/ACCESS.2022.3221430.

2. G. Danush, A. Pradeeshwar, D. Tharun Kumar, N. Vamsi Krishna and D. V. V. Prasad, 2023,"Machine Learning and Ensemble Techniques for Water Quality Analysis of Korattur Lake," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, pp. 1-9, doi: 10.1109/GCAT59970.2023.10353317.
3. H. I. H. Yusri, A. A. Ab Rahim, S. L. M. Hassan, I. S. A. Halim and N. E. Abdullah, 2022,"Water Quality Classification Using SVM And XGBoost Method," 2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, pp. 231-236, doi: 10.1109/ICSGRC55096.2022.9845143.
4. J. Cai, J. Chen, X. Dou and Q. Xing, 2022, "Using Machine Learning Algorithms With In Situ Hyperspectral Reflectance Data to Assess Comprehensive Water Quality of Urban Rivers," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, Art no. 5523113, doi: 10.1109/TGRS.2022.3147695.
5. J. Kirui, 2024, "Machine Learning Models for Drinking Water Quality Classification," International Conference on Control, Automation and Diagnosis (ICCAD), Paris, France, pp. 1-5, doi: 10.1109/ICCAD60883.2024.10553712.
6. K. Abirami, P. C. Radhakrishna and M. A. Venkatesan, 2023,"Water Quality Analysis and Prediction using Machine Learning," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, pp. 241-245, doi: 10.1109/CSNT57126.2023.10134661.
7. M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein and M. S. Anwar, 2023,"Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality," in IEEE Access, vol. 11, pp. 101055-101086, doi: 10.1109/ACCESS.2023.3315649.
8. N. Radhakrishnan and A. S. Pillai, 2020, "Comparison of Water Quality Classification Models using Machine Learning," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 1183-1188, doi: 10.1109/ICCES48766.2020.9137903.

9. N. S. Pagadala, M. Marri, A. Myla, B. Abburi and K. S. Ramtej, 2023, "Water Quality Prediction Using Machine Learning Techniques," 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, pp. 358-362, doi: 10.1109/SPIN57001.2023.10117415.
10. O. O. Ajayi, A. B. Bagula, H. C. Maluleke, Z. Gaffoor, N. Jovanovic and K. C. Pietersen, 2022,"WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes," in IEEE Access, vol. 10, pp. 48318-48337, doi: 10.1109/ACCESS.2022.3172274.
11. S. Abdullah and N. S. Jaddi, 2020, "Dual Kidney-Inspired Algorithm for Water Quality Prediction and Cancer Detection," in IEEE Access, vol. 8, pp. 109807-109820, doi: 10.1109/ACCESS.2020.3001685.
12. S. Khullar and N. Singh, 2022, "River Water Quality Classification using a Hybrid Machine Learning Technique," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 808-813, doi: 10.23919/INDIACom54597.2022.9763301.
13. S. R. S. Chakravarthy et al., 2023, "Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm," in IEEE Access, vol. 11, pp. 140900-140913, doi: 10.1109/ACCESS.2023.3339564.
14. Xiaotong Zhu, Hongwei Guo, Jinhui Jeanne Huang, Shang Tian, Wang Xu, YouquanMai, 2022, An ensemble machine learning model for water quality estimation in coastal area based on remote sensing imagery, Journal of Environmental Management, Volume323, 116187, ISSN03014797, https://doi.org/10.1016/j.jenvman.2022.116187