applied sciences

MDPI

Article

WaQuPs: A ROS-Integrated Ensemble Learning Model for Precise Water Quality Prediction

Firna Firdiani [1], Satria Mandala [1,*], Adiwijaya [1] and Abdul Hanan Abdullah [2]

[1] Human Centric Engineering & School of Computing, Telkom University, Jl. Telekomunikasi No. 1, Bandung 40257, West Java, Indonesia; vyrnvyrd@gmail.com (F.F.); adiwijaya@telkomuniversity.ac.id (A.)
[2] Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia; hanan@utm.my
* Correspondence: satriamandala@telkomuniversity.ac.id

Abstract: Water presents challenges in swiftly and accurately assessing its quality due to its intricate composition, diverse sources, and the emergence of new pollutants. Current research tends to oversimplify water quality, categorizing it as potable or not, despite its complexity. To address this, we developed a water quality prediction system (WaQuPs), a sophisticated solution tackling the intricacies of water quality assessment. WaQuPs employs advanced machine learning, including an ensemble learning model, categorizing water quality into nuanced levels: potable, lightly polluted, moderately polluted, and heavily polluted. To ensure rapid and precise dissemination of information, WaQuPs integrates an Internet of Things (IoT)-based communication protocol for the efficient delivery of detected water quality results. In its development, we utilized advanced techniques, such as random oversampling (ROS) for dataset balance. We used a correlation coefficient to select relevant features for the ensemble learning algorithm based on the Random Forest algorithm. Further enhancements were made through hyperparameter tuning to improve the prediction accuracy. WaQuPs exhibited impressive metrics, achieving an accuracy of 83%, precision of 82%, recall of 83%, and an F1-score of 82%. Comparative analysis revealed that WaQuPs with the Random Forest model outperformed both the XGBoost and CatBoost models, confirming its superiority in predicting water quality.

Keywords: Internet of Things; machine learning; water quality prediction; Random Forest; random oversampling

Citation: Firdiani, F.; Mandala, S.; Adiwijaya; Abdullah, A.H. WaQuPs: A ROS-Integrated Ensemble Learning Model for Precise Water Quality Prediction. *Appl. Sci.* **2024**, *14*, 262. https://doi.org/10.3390/app14010262

Academic Editor: Douglas O'Shaughnessy

Received: 14 November 2023
Revised: 21 December 2023
Accepted: 22 December 2023
Published: 28 December 2023

1. Introduction

Water is an irreplaceable natural resource, serving as a foundational element for the human body and playing an indispensable role in our survival. Beyond sustaining bodily functions, water is integral to a myriad of daily activities, including cooking, washing, and bathing. Nevertheless, swiftly and accurately assessing water quality poses a significant challenge due to its intricate composition, diverse sources, and the introduction of new pollutants. The task of promptly and precisely evaluating processed water quality remains an enduring challenge yet to be fully resolved. To address this issue, one promising avenue involves the application of machine learning techniques.

Numerous studies have explored the application of machine learning in the classification and prediction of water quality. Iyer et al. [1] conducted research on predicting water quality using machine learning, employing SVM, Random Forest, and Decision Tree models. The findings of their study revealed that the performance of the Random Forest model surpassed the other models, achieving an accuracy rate of 68%. However, it is noteworthy that the accuracy rate obtained by Random Forest in this study remained below the threshold of 70%.

Continuing the exploration of water quality prediction, Sen et al. [2] conducted further research focusing on addressing challenges in aquaculture through the introduction of an

intelligent machine learning and IoT-based biofloc system. This research aimed to enhance efficiency, production, water recycling, and automatic feeding within the aquaculture domain. The system integrates water quality prediction capabilities using advanced machine learning models, including Decision Tree classification and Random Forest. Moreover, real-time monitoring is facilitated through an Android app. The outcomes of this study were highly promising, with Random Forest achieving an impressive accuracy rate of 73.76%.

Xin and Mou [3] conducted research on water quality using a multimodal-based machine learning algorithm. The categories used in this research consisted of only two labels, namely potable or not. The researchers utilized the LGBM, Catboost, and XGBoost algorithms to develop an ensemble learning model from a dataset containing metal elements in water. Employing a 10-fold cross-validation approach and fine-tuning the hyperparameters, XGBoost demonstrated the highest accuracy, with an average of 79%. While this study exhibited relatively robust accuracy, there remains room for potential enhancements in this domain.

Subsequently, Patel et al. [4] conducted an extensive analysis, evaluating the performance of 15 machine learning models for water quality classification. Similar to prior studies, they classified water quality according to a binary distinction, i.e., potable or non-potable. Following their initial assessment, the top five models exhibiting the highest accuracy were selected and further refined through hyperparameter optimization. Notably, among these models, Random Forest emerged as the top performer, achieving an impressive accuracy rate of 81%. The accuracy value in this research was quite good compared to previous research, namely above 80%. While this accuracy value surpassed that of previous research, exceeding 80%, it is important to note that the classification in this study remained limited to only two categories: potable or not.

In another study, Ahmed et al. [5] addressed the critical issue of declining water quality attributed to rapid urbanization and industrialization, posing considerable health risks. The study considered the application of supervised machine learning algorithms to predict a water quality index (WQI) and a water quality class (WQC) based on four input parameters: temperature, turbidity, pH, and total dissolved solids. Among the classification algorithms employed in this research, Random Forest was included. However, it is noteworthy that Random Forest achieved an accuracy value of only 76% in this particular study.

Furthermore, Wong et al. [6] conducted a comprehensive exploration and analysis of 17 novel input features. Their goal was to formulate an enhanced water quality index (WQI) capable of adapting to the land use activities surrounding the river. For model selection, the researchers employed five regression algorithms—specifically, Random Forest, AdaBoost, Support Vector Regression, Decision Tree Regression, and Multilayer Perceptron. Among these algorithms, Random Forest exhibited superior prediction performance. This study introduced a modified Random Forest method that incorporated the synthetic minority oversampling technique, yielding accuracy values reaching 77.68%. While Ahmed et al. [5] and Wong et al. [6] classified water into five distinct classes in their research, it is noteworthy that the accuracy rate fell slightly below the 80% benchmark.

To address these challenges, we developed a water quality prediction system (WaQuPs), a sophisticated solution tackling the intricacies of water quality assessment employing advanced machine learning, including an ensemble learning model. WaQuPs categorizes water quality into nuanced levels: potable, lightly polluted, moderately polluted, and heavily polluted, aligning with the guidelines set by Government Regulation 82 of 2001. To ensure rapid and precise dissemination of information, WaQuPs integrates an Internet of Things (IoT)-based communication protocol for the efficient delivery of the detected water quality results.

In its development, WaQups leverages ensemble machine learning techniques, specifically combining multiclass classification with random oversampling (ROS) rather than using the synthetic minority oversampling technique (SMOTE) [7], to enhance performance. The main classification algorithm employed in our model is Random Forest, serving as the primary method of analysis. Additionally, we conduct comparative assessments with other ensemble learning algorithms, such as XGBoost and Catboost, to evaluate their effective-

ness in this context. To optimize our classification model, we implement cross-validation techniques and carry out hyperparameter tuning.

The adoption of ensemble machine learning in our study is driven by its acknowledged superiority over classical machine learning methodologies. The primary advantage of ensemble learning lies in its utilization of multiple algorithms simultaneously, enhancing overall proficiency and robustness [8,9]. Moreover, ensemble learning, marked by the fusion of predictions from multiple models, leads to an elevated level of predictive accuracy [10]. This strategic choice is underscored by the findings presented in Ajayi's paper [10], where the ensemble learning model demonstrated superior accuracy compared to classical machine learning. In our pursuit of water quality prediction, we embraced ensemble learning and specifically chose Random Forest as the primary model to handle multi-class classification cases. This decision was informed by insights extracted from prior research, as referenced in research by Patel et al. [4].

Previous studies have underscored Random Forest's superiority within ensemble learning, exhibiting higher accuracy values in determining water quality into two classes [4]. Moreover, Random Forest has demonstrated suitability for handling multi-class classification challenges [11]. It is noteworthy that while Random Forest has proven effective for multi-class classification, it has been underutilized in water quality studies. We found only two studies [5,6] which to date have explored its application in solving multi-class cases within the realm of water quality.

For comparative analysis, we selected the XGBoost and CatBoost models, both falling within the ensemble learning paradigm and chosen based on insights from prior research [3]. CatBoost is renowned for its adeptness in handling categorical and heterogeneous data [12], aligning well with our dataset, which involves categorizing water quality into specific levels: potable, lightly polluted, moderately polluted, and heavily polluted. On the other hand, XGBoost excels in terms of operating speed, scalability, and effectiveness with large datasets [13], attributes that resonate with the size and complexity of our dataset. Considering these factors, we decided to compare the XGBoost and CatBoost models with our primary model, Random Forest.

2. Related Works

Numerous prior studies have investigated the realm of water quality prediction. In a notable contribution, Torky and colleagues [14] conducted an extensive investigation aimed at ensuring the delivery of safe drinking water and estimating water quality indices through the implementation of machine learning techniques. Their research resulted in the development of a dual-component system: the first component's role is to assess the potability of water, while the second component is focused on predicting water quality index (WQI) values through regression analysis. The initial component employs a range of machine learning algorithms to facilitate water classification. The study strongly emphasizes the essential role of machine learning in improving the precision of water quality prediction models. As a result, it provides valuable insights that can advance the application of machine learning in the field of water quality assessment. The research is of significant importance by highlighting the critical need to ensure safe drinking water for the well-being of communities and the preservation of environmental sustainability.

In 2022, Krtolica et al. [15] conducted a study concerning a crucial aspect of water quality assessment through the presence of macrophytes. Their research was grounded in existing literature pertaining to water quality assessment, highlighting the significance of considering macrophytes as valuable indicators of water quality. They employed a variety of machine learning algorithms to construct a robust water quality assessment model. Their experimental results revealed that the SVM algorithm achieved the highest accuracy rate, an impressive 88%, in evaluating water quality based on the presence of macrophytes. This achievement underscores the substantial potential of machine learning in enhancing the precision and reliability of water quality detection methodologies, particularly in ecosystems where macrophytes play a vital role.