Unmanned surface vehicle for intelligent water quality assessment to promote sustainable human health

Qadir, M. I., Mumtaz, R., Manzoor, M., Saleem, M., Khan, M. A. & Charlesworth, S

DOI 10.2166/ws.2024.141
ISSN 1606-9749
ESSN 1607-0798

Publisher: IWA Publishing

Water Supply

© 2024 The Authors

Water Supply Vol 00 No 0, 1 doi: 10.2166/ws.2024.141

Unmanned surface vehicle for intelligent water quality assessment to promote sustainable human health

Muhammad Ibtsaam Qadir[a], Rafia Mumtaz[a,*], Mariam Manzoor[a], Misbah Saleem[a], Muhammad Ajmal Khan[a] and Susanne Charlesworth[b]

[a] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[b] Institute of Sustainability, Equity and Resilience, Centre for Agroecology, Water and Resilience, Coventry University, Coventry, United Kingdom
*Corresponding author. E-mail: rafia.mumtaz@seecs.edu.pk

MIQ, 0009-0002-5521-2855; RM, 0000-0002-0966-3957

ABSTRACT

Deteriorating water quality poses a substantial risk to human health, with billions at risk of waterborne diseases due to contamination. Insufficient water quality data augment risks as conventional monitoring methods lack comprehensive coverage. Technologies like the Internet of Things and machine learning offer real-time water quality monitoring and classification. IoT nodes often provide point data insufficient for monitoring the quality of entire water bodies. Remote sensing, though useful, has limitations such as measuring only optically active parameters and being affected by climate and resolution issues. To address these challenges, an unmanned surface vehicle named `AquaDrone' has been developed. AquaDrone traverses water bodies, collecting data of four key parameters (pH, dissolved oxygen, electrical conductivity, and temperature) along with GPS coordinates. The data is transmitted to a web portal via LoRa communication and Wi-Fi, where visualizations like data tables, trendlines and color-coded heatmaps are generated. A multilayer perceptron classifies water quality into five categories, aiding in real-time classification. A comparative analysis of various oversampling techniques has been conducted in the context of water quality classification. The AquaDrone offers a feasible solution for monitoring quality of small to medium-sized water bodies, crucial for safeguarding public health.

Key words: human health, Internet of Things, machine learning, real-time, unmanned surface vehicle, water quality monitoring

HIGHLIGHTS

- AquaDrone measures four physicochemical parameters of water along with GPS coordinates to monitor water quality in real-time.
- A multilayer perceptron classifies water quality into five categories, with oversampling used to enhance model training.
- AquaDrone wirelessly transmits data to the web portal, where trendlines, data table and a color-coded heatmap are generated to visualize water quality.
- This ensures effective real-time monitoring in small to medium-sized water bodies, protecting both water sources and human health.

INTRODUCTION

Water plays a pivotal role in the development and survival of all living organisms and in maintaining a healthy ecosystem. It is an essential resource for industrial, agricultural, and domestic activities. Despite its importance, water quality is being degraded continuously at an alarming rate due to urbanization, industrialization, and anthropogenic activities. Poor drinking water quality is a major concern all over the world and exposes humans to health risks as it has been linked to the spread of diseases such as cholera, typhoid, dysentery, polio, and hepatitis (WHO 2022). Approximately, 80% of untreated wastewater is released into the environment, which negatively impacts human health (Lin *et al.* 2022). This wastewater contains harmful pollutants and heavy metals such as arsenic, chromium, and cadmium (Chen *et al.* 2019), which are highly carcinogenic (Lin *et al.* 2022). Sewage discharge and agricultural runoff contaminate water with pathogens (Parris 2011), resulting in a harmful impact on human health (Malakar *et al.* 2019) and waterborne diseases like diarrhea (Zhang 2012; Lin *et al.* 2022). Pathogens such as *Escherichia coli* (Weller *et al.* 2021) and *Salmonella* (Buyrukoğlu 2021) increase the likelihood of contamination and pose substantial threats to surface water quality.

In Pakistan, water quality is depleting at a startling speed, owing to overpopulation, rapid urbanization, and the disposal of untreated wastewater, both industrial and domestic, into the rivers (Imran *et al.* 2022). Poor water quality and water shortages have significantly affected the country's agricultural and environmental systems and have made Pakistan a water-stressed country (Zhang *et al.* 2020). Moreover, the lack of clean water availability is the root cause of 40% of all the fatalities in Pakistan (Daud *et al.* 2017). Most of the water in rivers is affected by microbial contamination, with very high levels of total coliforms and fecal coliforms in downstream rivers and tributaries (Imran *et al.* 2022). Notwithstanding the current situation, there is a dearth of water quality monitoring systems and water treatment plants in the country, which further impacts deteriorating water quality. Therefore, there is a need to develop systems that can monitor and predict water quality in real time, further helping relevant organizations and agencies to take timely preventive and remedial measures, as improved water quality results in better health (Zhang 2012).

There are several methods to assess water quality, and the most commonly used ones are laboratory and field testing, Internet of Things (IoT) nodes, and satellite imagery. Laboratory testing for monitoring water quality is costly, cumbersome, and ineffective. Much time is spent on the collection, transportation, testing, and statistical analysis of samples (Ahmed *et al.* 2019); consequently, this method does not provide real-time data. Field testing is also costly in terms of time. Recently, IoT sensors have been widely utilized to assess water quality in real time, saving a great deal of time when compared to manual laboratory techniques. However, IoT nodes provide point data, which is insufficient to represent the water quality of a large water body (Khan *et al.* 2022). In addition, difficult terrain makes certain data sites inaccessible, and the presence of certain bacteria in water may cause health risks to those collecting samples (Koparan *et al.* 2018) or installing sensors. Satellite imagery is considered to be a feasible technique to monitor the water quality of large water bodies, unaffected by the topography, but not without limitations. For example, non-optically active parameters are difficult to measure directly using satellite imagery (Sagan *et al.* 2020), and the acquisition of data is also limited by the revisit time of the satellite over a specific location. With a coarser spatial resolution, monitoring of small and medium inland lakes becomes difficult (Murray *et al.* 2022). Moreover, satellite imagery is also affected by several environmental and climatic effects (Khan *et al.* 2022) such as cloud cover (Murray *et al.* 2022).

Keeping in view the drawbacks and limitations of the above techniques, this research combines IoT sensors with an unmanned surface vehicle (USV) designed and developed to monitor and classify the water quality of an entire water body in real time. The USV, named 'AquaDrone', is equipped with sensors that can monitor pH, dissolved oxygen (DO), electrical conductivity (EC), and temperature. The AquaDrone travels on the surface of the water body to send the values of water quality parameters to the onsite receiver using Long-Range (LoRa) communication and then to the database using Wi-Fi. A machine learning model (multilayer perceptron, MLP) was trained using historical data from rivers in Hong Kong (1986–2020) to classify the water quality ranging from Very Bad to Excellent. For comprehensive data visualizations, color-coded heat maps and data tables were generated on the web portal. Typically, water quality data are gathered from the edges of the water body, leading to the presence of class imbalance, which in turn can result in the underperformance of the machine learning model. Therefore, a comparative analysis has also been conducted between several oversampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic (ADASYN), SMOTE + Tomek Links, and SMOTE + ENN (Edited Nearest Neighbor). For intelligent classification of water quality, a MLP was trained on the original dataset as well as all the oversampled datasets for comparison. The development of the AquaDrone, the use of sensors, wireless communication, and machine learning model together with the development of a web portal for water quality assessment differentiates the proposed system from existing systems.

The paper is organized as follows. The following section highlights the related work and research that has been conducted in the domain of water quality monitoring thus far, using IoT, unmanned vehicles, remote sensing, and machine learning. In the subsequent section, the proposed methodology for the design and development of the AquaDrone, including the development and training of the deep learning model will be discussed. The findings of the research, including the accuracy metrics for the deep learning model, the comparative analysis of the oversampling techniques, and the development of a web portal for detailed and comprehensive data visualizations are given in the Results and discussion section. The last section encapsulates the conclusion of the research.

RELATED WORK

Researchers are exploring methods and techniques to monitor water quality; typically this involves laboratory analysis, field monitoring, and manual calculation methods. In recent years, the focus has turned to IoT-based solutions, remote sensing, and machine learning methodologies to find automated and optimized solutions to the problem.

Shafi *et al.* (2018) proposed an embedded prototype for monitoring of water quality parameters in real time. The research further compared the performance of several machine learning models in binary classification of water quality. They found that the deep neural network outperformed other algorithms with an accuracy of 93%. Koparan *et al.* (2018) developed an autonomous water sampling method. They deployed an unmanned aerial vehicle (UAV) equipped with floating attachments to navigate to a specific location. In addition, they designed a water-capturing mechanism to collect water samples from this location. The sample required further laboratory testing for measuring water quality parameters, which is the limitation of this approach. Another study (Pasika & Gandla 2020) proposed a smart water quality monitoring system using a microcontroller and sensors to measure pH, turbidity, temperature, water level, and humidity; and developed a mobile application to visualize the data.

Harmful algal blooms are known to cause damage to the environment as well as human health (Berdalet *et al.* 2015); to predict water quality and the presence of algal blooms, Lee & Lee (2018) employed MLP, recurrent neural networks, and long short-term memory (LSTM) using chlorophyll-a as an indicator. They found that the LSTM model showed the best performance among all the models used. Xu *et al.* (2020) compared four machine learning algorithms to monitor water quality by predicting the presence of fecal indicator bacteria in water and used ADASYN oversampling, finding that MLP performed most effectively. Weller *et al.* (2021) also compared several machine learning models to predict the levels of *E. coli* in agricultural water finding that models incorporating turbidity and weather factors outperformed all other models, irrespective of the algorithm.

Dunbabin *et al.* (2009) designed a 16 ft solar-powered catamaran for gathering water quality data, integrated with a floating sensor network enabling remote mission control and data collection. The designed Autonomous Surface Vehicle (ASV) can operate in diverse weather conditions and improves current manual monitoring by offering extensive and frequent water storage monitoring across large distances. Madeo *et al.* (2020) developed a USV for water quality monitoring, but their system lacks an intelligent, machine learning-based, water quality classification and a web portal for data visualization. Cheng *et al.* (2021) proposed another solution for water quality monitoring by designing a UAV with a floating structure and sensor array to measure the pH, DO, ammonia, nitrogen, and temperature at the landing point only. The study also investigated and predicted trends in water quality using time series analysis. Bayusari *et al.* (2021) designed an autonomous underwater vehicle (AUV) equipped with a camera, sonar, and depth sensor for self-navigation and IoT-based sensing of temperature, pH, and DO; this includes non-real-time monitoring as data were retrieved when the AUV returned to the surface.

In another study, Khurshid *et al.* (2022) employed IoT and machine learning to make real-time bacterial predictions by measuring five water quality parameters: pH, temperature, turbidity, total dissolved solids, and DO. Several classical machine learning and deep learning models were trained, out of which Support Vector Machine (SVM) and Bayesian Regression outperformed the others in bacterial predictions with mean squared errors of 0.356 and 0.396 and mean absolute errors of 30.76 and 31.25, respectively. Ahmed *et al.* (2019) conducted a comparative analysis of various supervised machine learning models for efficient water quality classification and found that MLP performed the best in classifying water quality using four parameters with an accuracy of 85%.

Researchers have also opted for a multisource approach to monitor and predict water quality. A recent study (Zubair *et al.* 2022) integrated three data modalities (Geographical Information System (GIS) satellite imagery, and IoT nodes) to predict the water quality holistically using time series analysis. Khan *et al.* (2022) worked along the same lines and singled out a multimodal approach to classify the water quality, acquiring data from IoT nodes and satellite imagery. The artificial neural network (ANN) outperformed SVM and random forest with an accuracy of 97% in classifying water quality.

The latest research has mostly focused on the use of the IoT and satellite imagery for water quality monitoring along with laboratory methods as conventional laboratory methods are expensive, time-consuming, and non-real-time. Both IoT nodes and manual sampling provide point data, which are insufficient to represent the entire water body, and satellite imagery has its own drawbacks other than its revisit time, as it can only measure optically active parameters (Sagan *et al.* 2020) and is affected by climatic effects (Khan *et al.* 2022). The proposed methodology addresses the limitations of previous methods and techniques and provides a holistic method to monitor water quality.

METHODOLOGY

Design of the AquaDrone

A USV, AquaDrone, was designed and developed using a powerful brushless DC (BLDC) motor and an electronic speed controller (ESC) to provide the required thrust combined with a high-power servo motor and a rudder for steering. The mechanical system runs on electric power supplied by a rechargeable Li-Po battery. A 30A ESC has been used to control the speed of the BLDC motor. The speed of the propeller and the position of the rudder are regulated by a remote control (2.4 GHz ISM, 6 Channels). The design of the AquaDrone is focused on reducing drag and resistance to enhance its maneuverability in the water. The streamlined shape from the front makes AquaDrone more hydrodynamically efficient and able to maintain its speed for longer distances, as depicted in Figure 1. The AquaDrone is a lightweight USV that is easily transportable. However, in wavy and turbulent water conditions, its small form factor and light weight influence the movement making it unstable.

Water quality parameters

Four water quality parameters were chosen for this research: temperature, pH, DO, and EC. Temperature is an important water quality parameter, which influences other water quality parameters and aquatic life. DO is another important parameter and is significantly affected by temperature (Zhi *et al.* 2023) since it is the amount of free oxygen available in the water that supports aquatic life; low DO can cause aquatic life to suffocate (Bayusari *et al.* 2021). pH is a measure of acidity or alkalinity with a safe environmental range of 6.5–8.5 (Cheng *et al.* 2021). EC is the ability of water to conduct an electric current and is an indication of the concentration of dissolved salts. Human activities raise dissolved solids entering the waters, elevating conductivity. Higher conductivity may correlate with other indicators of alteration in water quality.

Development of water quality monitoring sensor network

The AquaDrone is housed with a microcontroller (Arduino UNO), global positioning system (GPS) sensor, and sensors to measure four physicochemical parameters of water (pH, temperature, EC, and DO). Vernier water quality sensor probes are used, which provide 0–5 V output and are connected to the analog pins of the Arduino UNO board using Vernier Arduino Interface Shields. For wireless communication, AquaDrone is equipped with a LoRa communication module and acts as a transmitter node, enabling long-range transmission with low power consumption. The microcontroller is programmed to send data packets after an interval of 5 s using LoRa WAN, and each data packet contains information about AquaDrone's

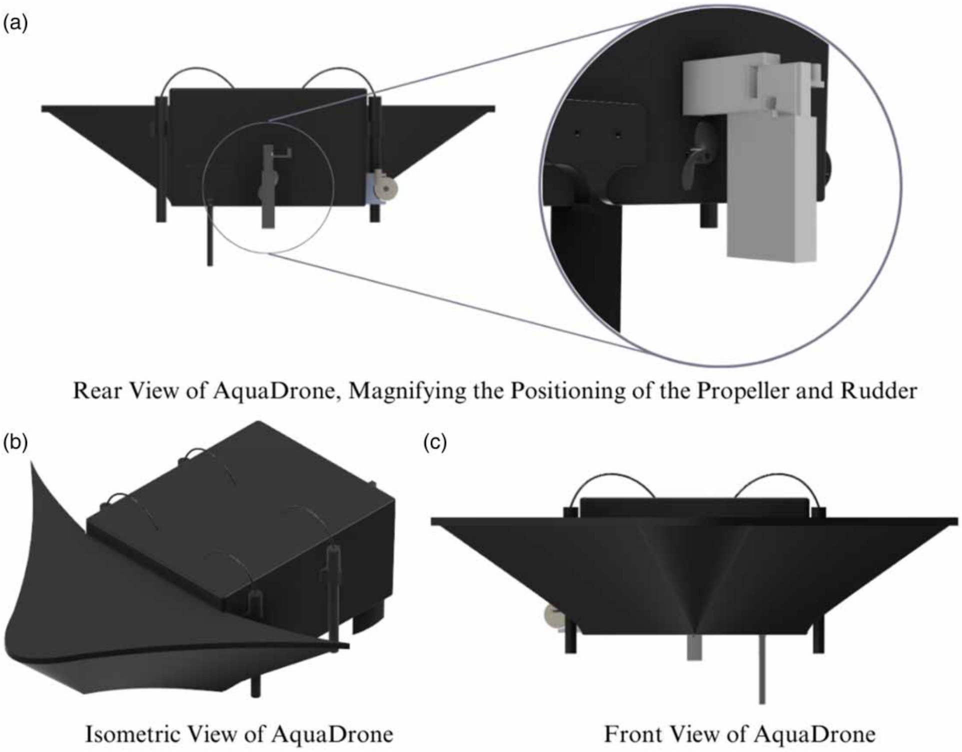

(a)
Rear View of AquaDrone, Magnifying the Positioning of the Propeller and Rudder
(b)
(c)
Isometric View of AquaDrone
Front View of AquaDrone

Figure 1 | Three-dimensional renders of the waterproofed AquaDrone illustrating the design of the hull, positioning of the water quality sensor probes, propeller, and rudder.

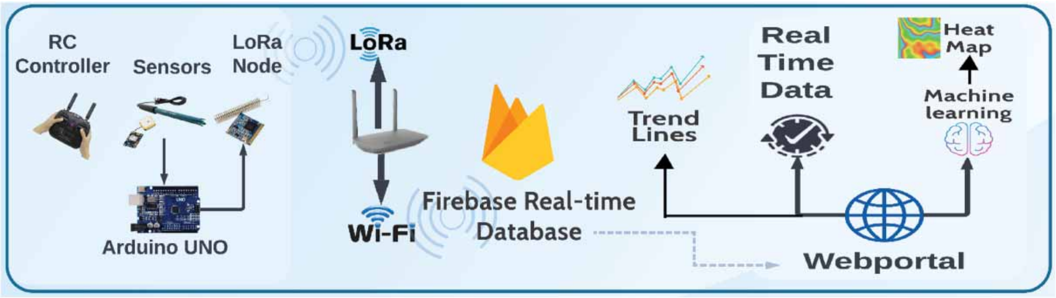


Figure 2 | System architecture showing the main components of the proposed system.

location (GPS coordinates) and the four water quality parameters. The onsite receiver node consists of a Node-MCU ESP32 connected to the LoRa module. The ESP32 and LoRa are programmed to receive data packets based on receiving interrupts (Figure 2).

Firebase Realtime Database (RTDB) is a cloud-hosted platform that stores data received from the client and allows the server to fetch it in real-time. The ESP32 connects to the Firebase RTDB using Wi-Fi and sends the received data packets from AquaDrone to the RTDB.

Deep learning model for water quality classification

The dataset that the deep learning model has been trained on was obtained from the Environment Protection Department (EPD), Hong Kong, containing more than 31,000 data points, and spanning more than 30 years from 1986 to 2020. Six water quality parameters (temperature, pH, DO, EC, nitrates, and turbidity) were chosen for labeling the dataset. The dataset of the six above-mentioned parameters was normalized to bring the data to a common range (0–100). The Water Quality Index (WQI) was calculated using the weighted averages method in which each parameter was assigned a weight. The WQI is derived by the multiplication of normalized parameters with their respective weights and subsequent aggregation of the weighted values. This resultant sum was divided by the cumulative sum of weights assigned (Ahmed *et al.* 2019). The calculated WQI was used to classify the data points into five classes based on the ranges as given in Table 1.

Having labeled the dataset based on six parameters and keeping only the four parameters which AquaDrone can measure (temperature, DO, EC, pH), the data was split into training and test sets (80 and 20% respectively). The visualization in Figure 3 shows that there is a class imbalance in the dataset. Class imbalance typically leads to misclassification of instances from the minority class more often than instances from the majority class, even with high accuracy (Johnson & Khoshgoftaar 2019). One approach to address an imbalanced dataset is to oversample the minority class, where synthetic data instances are synthesized from the existing data instances. Two oversampling techniques SMOTE and ADASYN, and two hybridization approaches SMOTE + Tomek Links and SMOTE + ENN were used. Hybridization combined both undersampling and oversampling techniques.

Synthetic minority oversampling technique

In SMOTE, synthetic samples are generated for the minority class (Chawla *et al.* 2002). It focuses on the feature space to create new instances by using interpolation between the instances that are near each other.

Table 1 | WQI ranges

WQI range	Class
0–25	Very Bad
25–50	Bad
50–70	Medium
70–90	Good
90–100	Excellent

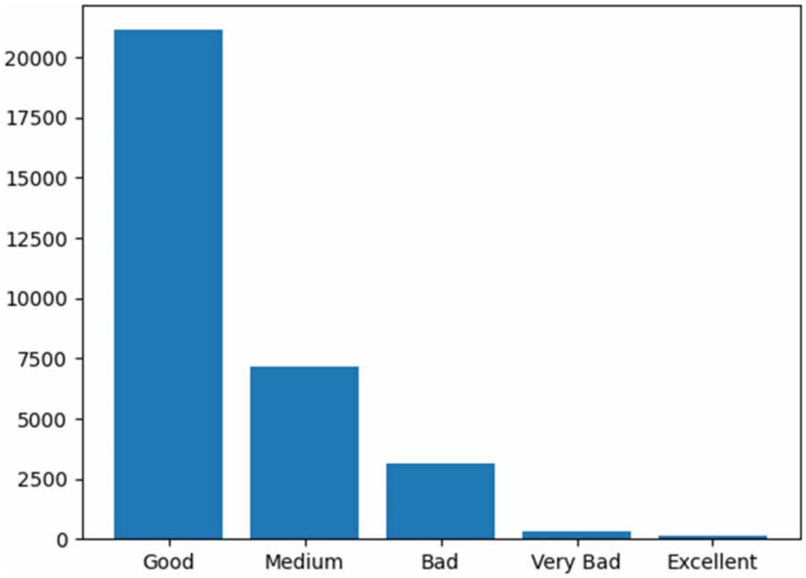


Figure 3 | Class imbalance in the dataset.

ADASYN sampling

ADASYN employs a weighted distribution for various instances of the minority class, depending on their learning difficulty. This approach generates more synthetic data for minority class instances that are difficult to learn while producing relatively fewer synthetic data for those minority instances that are easier to learn (He *et al.* 2008).

SMOTE + Tomek links

This hybridization technique synergistically integrates the capabilities of SMOTE, which generates synthetic data for the minority class, and Tomek links, which eliminates data from the majority class recognized as Tomek links (Batista *et al.* 2003). Tomek links identify data samples from the majority class that are in close proximity to the minority class data.

SMOTE + ENN

SMOTE + ENN combines SMOTE's ability of synthetic data generation for the minority class together with the ENN's ability to omit some observations from both classes that are identified as having different class between the observation's class and its *K*-nearest neighbor majority class (Batista *et al.* 2004).

After applying these techniques, the resulting dataset distributions were substantially adjusted, allowing for a more equitable representation of all five water quality classes. Table 2 presents a detailed comparison of class distribution in the water quality dataset before and after the application of oversampling techniques. The values describe the instance counts within each class, revealing the notable changes accomplished through the implementation of oversampling techniques.

For the purposes of water quality classification, a MLP is employed. MLP is a classical ANN, predominantly employed for classification problems. The model used for the research comprises of three hidden layers. An excessively large learning rate

Table 2 | Class distribution comparison before and after oversampling techniques

		After oversampling			
Classes	**Before oversampling**	**SMOTE**	**ADASYN**	**SMOTE + Tomek links**	**SMOTE + ENN**
Very Bad	329	16,911	16,924	16,902	16,728
Bad	3,140	16,911	16,834	16,746	15,265
Medium	7,151	16,911	16,989	16,016	11,406
Good	21,117	16,911	16,911	16,121	11,344
Excellent	128	16,911	16,920	16,890	16,719

might result in performance that explodes or oscillates throughout training epochs and lowers overall performance, whereas an extremely small learning rate will lead to very slow learning or possibly the inability to learn at all. For our model, the learning rate of 0.001 is an optimum value facilitating stable and efficient convergence. Moreover, the Adam optimizer has been used due to its faster convergence and robustness. The model has been trained on both the original dataset and four oversampled datasets. To reduce the risk of overfitting, which is prevalent when training deep learning models, we employed the early stopping technique, setting the patience parameter to 5. Early stopping is a well-known technique used in neural network training to track the model's performance on a different validation dataset. To avoid the model from overfitting to the training data, this strategy stops the training process when the model's performance on the validation set does not improve for a certain number of consecutive epochs.

RESULTS AND DISCUSSION

In this section, a detailed analysis is presented discussing the performance of AquaDrone, a thorough comparison of machine learning model performances on both the original and oversampled datasets, and the data visualizations depicting the findings.

Performance of AquaDrone

The AquaDrone was controlled remotely and had a sail time of approximately 10–15 min at full speed and maximum battery capacity. With a tested operational range of approximately 500 m (0.5 km), the AquaDrone offers a significant coverage area for water bodies. The AquaDrone was designed to have optimal efficiency in still water bodies (see Figure 4(a)) since in flowing water (see Figure 4(b)), its stability and movement were affected by the movement of water.

Performance of the deep learning model

The performance of a machine/deep learning model is typically measured through various accuracy metrics that assess its predictive accuracy. The metrics utilized for evaluating the performance of the implemented algorithm include accuracy, precision, recall, and F1 score.

i. Accuracy:

Accuracy is the count of correct classifications made by the model across all observed values (Ahmed *et al.* 2019) and measured using Equation (1) (Deng *et al.* 2021).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{1}$$

ii. Precision:

Precision is the ratio of accurately classified instances belonging to a specific positive class, relative to the total instances classified as that class (Ahmed *et al.* 2019), as shown in Equation (2) (Deng *et al.* 2021).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

iii. Recall:

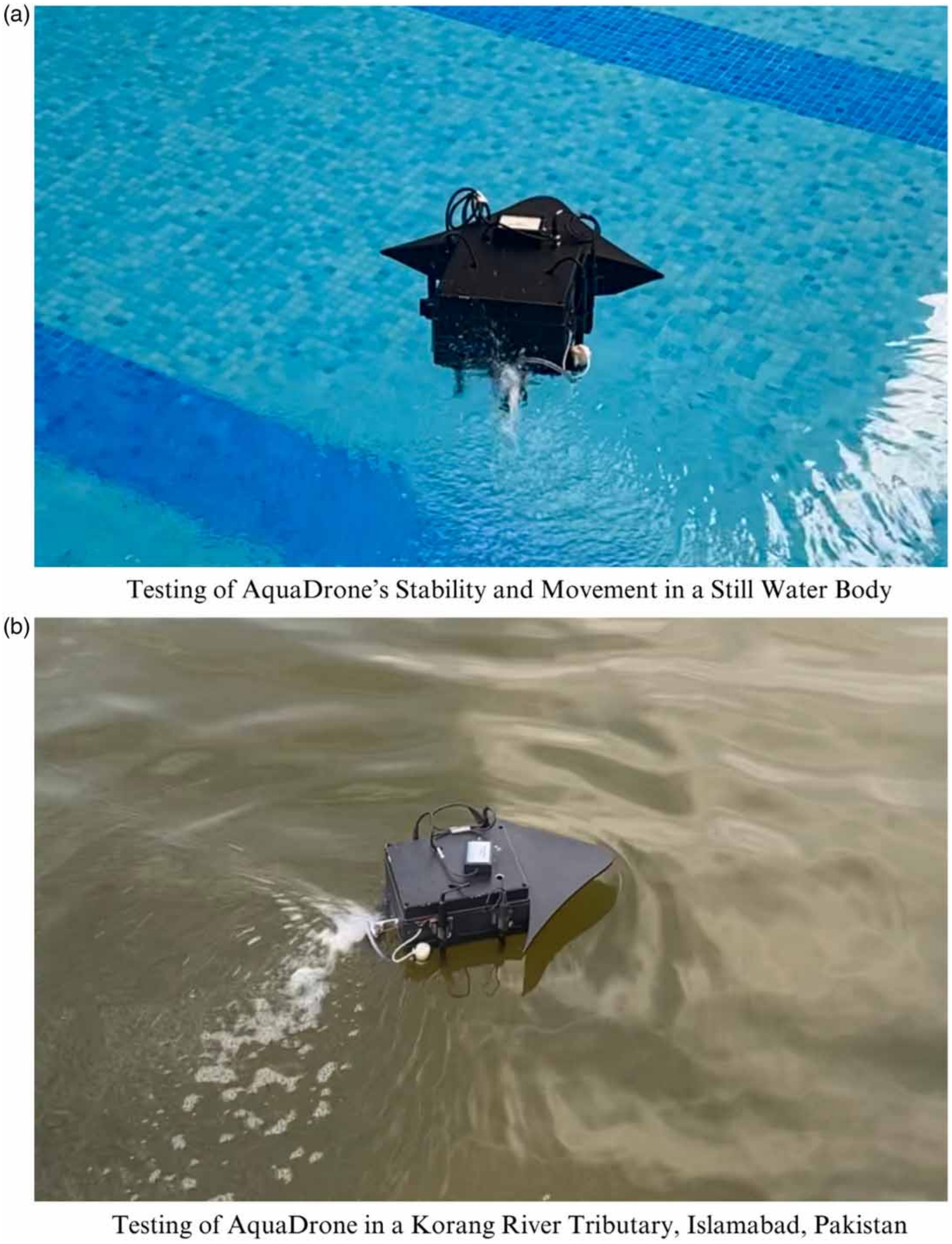
Recall quantifies the model's ability to correctly identify all relevant instances of a specific class (Ahmed *et al.* 2019) and measured using Equation (3) (Deng *et al.* 2021).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

iv. F1 score:

The F1 score is the harmonic mean of precision and recall (Ahmed *et al.* 2019), as shown in Equation (4) (Deng *et al.* 2021), providing a balanced measure of a model's precision and recall.

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$



Testing of AquaDrone's Stability and Movement in a Still Water Body

Testing of AquaDrone in a Korang River Tributary, Islamabad, Pakistan

Figure 4 | Testing of AquaDrone.

True positive (TP) and true negative (TN) refer to the count of accurately classified samples. False positive (FP) and false negative (FN) represent the count of samples that are incorrectly categorized into different water quality classes.

The MLP was trained on the original as well as the oversampled datasets. To assess the performance of the model, several metrics were observed such as accuracy, precision, recall, and F1 score. The model was used to classify the water quality based on measured pH, temperature, DO, and EC. The accuracy metrics displayed in Table 3 demonstrate performance of the model across all the training datasets. The model performance is nearly the same when trained on the original dataset and the oversampled dataset using SMOTE, both with an accuracy of approximately 81%. However, the latter has the best F1 score among all the models. Overall, the model performed the best when trained on the oversampled dataset using SMOTE. As this technique oversamples the minority class by generating synthetic examples, this in consequence leads to better representation of the minority class. Furthermore, neural networks are data hungry. Therefore, more data are likely to improve the performance of our MLP model. However, the performance of the oversampling techniques is

Table 3 | Accuracy metrics of model for All training datasets

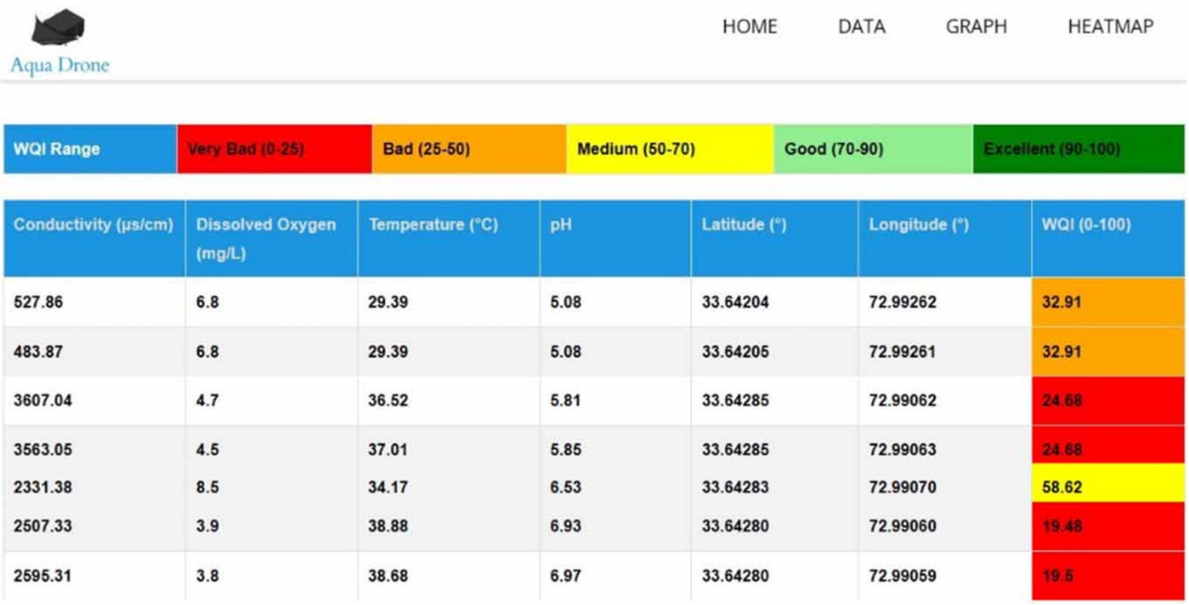
Dataset	Training dataset accuracy	Validation dataset accuracy	Test dataset			
			Accuracy	Precision	Recall	F1 score
Original	0.8124	0.8133	0.8146	0.6101	0.4373	0.4535
SMOTE	0.8404	0.8409	0.8145	0.6025	0.8109	0.6554
ADASYN	0.7936	0.7877	0.7393	0.4933	0.7241	0.5441
SMOTE + Tomek links	0.8429	0.8406	0.7908	0.5690	0.8122	0.6157
SMOTE + ENN	0.8934	0.8960	0.7814	0.5778	0.7994	0.6328

contingent upon the classifiers used and the performance metrics employed to evaluate the respective models (Chakravarthy *et al.* 2019). Because oversampling can result in overlapping ranges for the input variables for various classes, oversampling may not always produce the best results. This means that the values of the variables for one class could be similar to those of another. This makes it difficult for the model to determine which class a specific input instance belongs to based only on the input variables, which results in confusion and misclassification. In these instances, the model might have challenges to correctly classify data, which would lead to poorer overall performance metrics.

The weighted average method is a widely used technique for water quality classification. This method was used to label the dataset based on six water quality parameters (temperature, DO, pH, EC, nitrates, and turbidity). However, the model was trained only using the parameters for which the sensors have been installed on the AquaDrone, which are temperature, EC, DO, and pH. The parameters selected for training are the fundamental parameters of water quality. While training based on four parameters is restrictive, the model can learn the underlying relationships among the water quality parameters. This approach reduces the risk of overfitting while improving the generalizability of the developed model.

Data visualization

The web portal, with an interactive user interface, displays the values of water quality parameters, GPS coordinates, and WQI in the form of various visualizations. A color scheme is used to represent different levels of water quality. Red is used to indicate 'Very Bad' water quality, while orange represents 'Bad' water quality. Yellow is designated to indicate a 'Medium' level of water quality, while light green and dark green are employed to depict 'Good' and 'Excellent' levels of water quality, respectively. The data fetched from the Firebase RTDB are presented in the form of a data table (see Figure 5). The deep learning



WQI Range	Very Bad (0-25)	Bad (25-50)	Medium (50-70)	Good (70-90)	Excellent (90-100)

Conductivity (µs/cm)	Dissolved Oxygen (mg/L)	Temperature (°C)	pH	Latitude (°)	Longitude (°)	WQI (0-100)
527.86	6.8	29.39	5.08	33.64204	72.99262	32.91
483.87	6.8	29.39	5.08	33.64205	72.99261	32.91
3607.04	4.7	36.52	5.81	33.64285	72.99062	24.68
3563.05	4.5	37.01	5.85	33.64285	72.99063	24.68
2331.38	8.5	34.17	6.53	33.64283	72.99070	58.62
2507.33	3.9	38.88	6.93	33.64280	72.99060	19.48
2595.31	3.8	38.68	6.97	33.64280	72.99059	19.5

Figure 5 | Data table showing the data fetched from Firebase RTDB, which includes values of EC, DO, temperature, pH, GPS coordinates, and the calculated WQI.

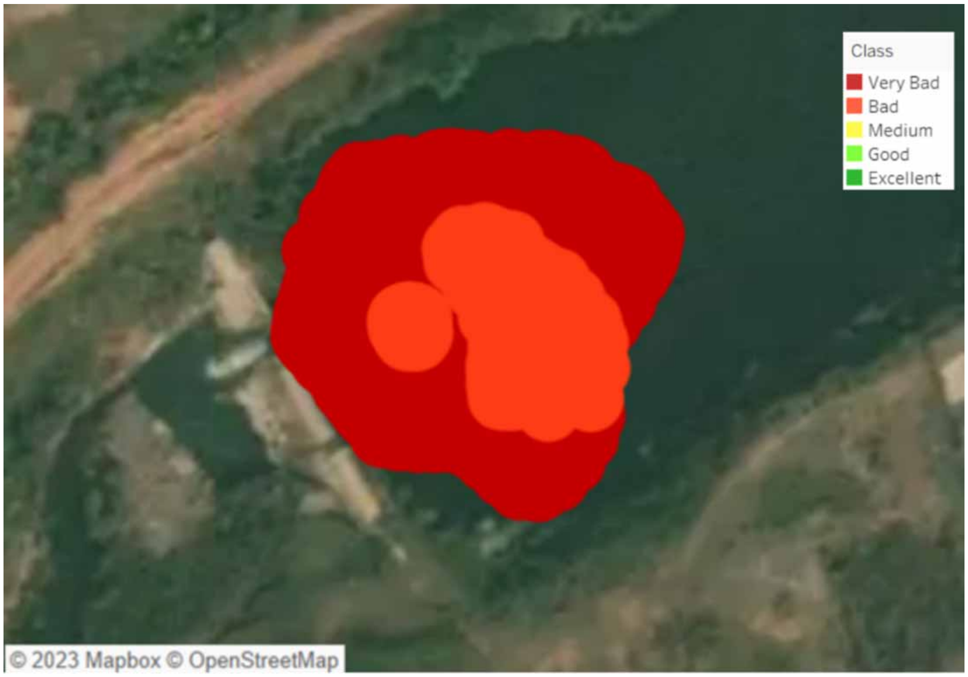


Figure 6 | Generated heatmap from Korang River tributary testing.

model was trained and deployed to classify water quality, which is further used together with GPS coordinates to generate a color-coded heatmap layer of the tested water body overlaid on the satellite view map as shown in Figure 6. The study site was a tributary of the Korang River located in Islamabad. The AquaDrone was controlled to traverse a specific area within this water body, with the resulting heatmap displayed in Figure 6.

The research introduces an alternative approach for real-time water quality monitoring that overcomes the constraints associated with the previous water quality monitoring methods, namely, laboratory techniques, IoT nodes, and satellite imagery. It addresses the issue of accessing inaccessible data sites, which has posed a major challenge when dealing with manual laboratory methods and IoT-based solutions. The proposed solution, AquaDrone, is capable of accessing unreachable data sites providing an ample number of data instances to effectively represent the complete water body. The solution can help determine the regions afflicted by water pollution by means of detailed visualizations presented on the web portal.

CONCLUSION

The availability of clean drinking water is decreasing across the globe due to deteriorating water quality in our water bodies. Consequently, a significant portion of the population lacks access to potable water and is subject to increased health risks and outbreaks of waterborne diseases. Conventional laboratory techniques are commonly used to monitor water quality but are time-consuming and expensive. IoT nodes are helpful in monitoring water quality, but they provide point data, which is insufficient to adequately represent the whole water body. Satellite imagery is also used but the acquired data is not real-time, and varying atmospheric conditions may affect the accuracy of obtained data. The proposed solution, AquaDrone, is a USV with sensors that can perform real-time measurements of various water quality parameters of the entire water body. Real-time data are sent to the web portal along with the GPS coordinates, where the color-coded heatmap is generated to visualize and represent the water quality of the entire site, classified using a deep learning model. In addition, the dataset has been oversampled using SMOTE, ADASYN, SMOTE + Tomek Links, and SMOTE + ENN techniques. Subsequently, the model has been trained on all five datasets including the original and four oversampled variations. The model trained on the oversampled dataset using SMOTE achieved the highest performance, with an accuracy of approximately 81%.

The future scope of this research will focus on refining the design to ensure that the AquaDrone's stability and movement remain minimally influenced by the flow of water. Due to limited resources, the system has been developed with just four fundamental parameters. However, expanding the sensor array to include measurements for additional water quality parameters like turbidity, total dissolved solids, phosphates, and nitrates could enhance the AquaDrone's capabilities and accuracy of water quality analysis significantly. Moreover, the AquaDrone can also be made autonomous instead of being controlled remotely.

The project is associated with the United Nations Sustainable Development Goal 6 – 'Clean Water and Sanitation'.

ACKNOWLEDGEMENTS

This research work is the joint output of the British Council researcher exchange project under British Council-HEC Pak-UK Faculty Mobility Grant (F137 – 2023).

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R. & García-Nieto, J. 2019 Efficient water quality prediction using supervised machine learning. *Water* **11** (11), 2210. https://doi.org/10.3390/w11112210.

Batista, G. E., Bazzan, A. L. & Monard, M. C. 2003 Balancing training data for automated annotation of keywords: A case study. *Wob* **3**, 10–18.

Batista, G. E., Prati, R. C. & Monard, M. C. 2004 A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* **6** (1), 20–29.

Bayusari, I., Adawiyyah, N. A., Dwijayanti, S., Hikmarika, H., Husin, Z. & Suprapto, B. Y. 2021 Water quality monitoring system in autonomous underwater vehicle based on Internet of Things (IoT). In *2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. https://doi.org/10.23919/eecsi53397.2021.9624211.

Berdalet, E., Fleming, L. E., Gowen, R., Davidson, K., Hess, P., Backer, L. C., Moore, S. K., Hoagland, P. & Enevoldsen, H. 2015 Marine harmful algal blooms, human health and wellbeing: Challenges and opportunities in the 21st century. *Journal of the Marine Biological Association of the United Kingdom* **96** (1), 61–91. https://doi.org/10.1017/s0025315415001733.

Buyrukoğlu, S. 2021 New hybrid data mining model for prediction of *Salmonella* presence in agricultural waters based on ensemble feature selection and machine learning algorithms. *Journal of Food Safety* **41** (4). https://doi.org/10.1111/jfs.12903.

Chakravarthy, A. D., Bonthu, S., Chen, Z. & Zhu, Q. 2019 Predictive models with resampling: A comparative study of machine learning algorithms and their performances on handling Imbalanced datasets. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. https://doi.org/10.1109/icmla.2019.00245.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002 SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357. https://doi.org/10.1613/jair.953.

Chen, B., Wang, M., Duan, M., Ma, X., Hong, J., Xie, F., Zhang, R. & Li, X. 2019 In search of key: Protecting human health and the ecosystem from water pollution in China. *Journal of Cleaner Production* **228**, 101–111. https://doi.org/10.1016/j.jclepro.2019.04.228.

Cheng, L., Tan, X., Yao, D., Xu, W., Wu, H. & Chen, Y. 2021 A fishery water quality monitoring and prediction evaluation system for floating UAV based on time series. *Sensors* **21** (13), 4451. https://doi.org/10.3390/s21134451.

Daud, M. K., Nafees, M., Ali, S., Rizwan, M., Bajwa, R. A., Shakoor, M. B., Arshad, M. U., Chatha, S. A., Deeba, F., Murad, W., Malook, I. & Zhu, S. J. 2017 Drinking water quality status and contamination in Pakistan. *BioMed Research International*. https://doi.org/10.1155/2017/7908183.

Deng, F., Huang, J., Yuan, X., Cheng, C. & Zhang, L. 2021 Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. *Laboratory Investigation* **101** (4), 430–441. https://doi.org/10.1038/s41374-020-00525-x.

Dunbabin, M., Grinham, A. & Udy, J. 2009 An autonomous surface vehicle for water quality monitoring. In *Australasian Conference on Robotics and Automation (ACRA)*. Citeseer, pp. 2–4.

He, H., Bai, Y., Garcia, E. A. & Li, S. 2008 ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. https://doi.org/10.1109/ijcnn.2008.4633969.

Imran, S., Rasheed, H. & Ashraf, M. 2022 *Water Quality Profile of Surface Water Bodies in Pakistan: Situation Analysis and Future Management Strategies*. Pakistan Council of Research in Water Resources (PCRWR). Available from: https://pcrwr.gov.pk/wp-content/uploads/2023/02/Water-Quality-Profile-of-Surface-Water-Bodies-in-Pakistan-2022.pdf.

Johnson, J. M. & Khoshgoftaar, T. M. 2019 Survey on deep learning with class imbalance. *Journal of Big Data* **6** (1). https://doi.org/10.1186/s40537-019-0192-5.

Khan, A. F., Mumtaz, R., Usama, M. & Mahsud, T. K. 2022 Enhanced water quality monitoring and estimation using a multi-modal approach. In: *Empowering Sustainable Industrial 4.0 Systems with Machine Intelligence*, pp. 113–131. https://doi.org/10.4018/978-1-7998-9201-4.ch006.

Khurshid, H., Mumtaz, R., Alvi, N., Haque, A., Mumtaz, S., Shafait, F., Ahmed, S., Malik, M. I. & Dengel, A. 2022 Bacterial prediction using Internet of Things (IoT) and machine learning. *Environmental Monitoring and Assessment* **194** (2). https://doi.org/10.1007/s10661-021-09698-4.

Koparan, C., Koc, A., Privette, C., Sawyer, C. & Sharp, J. 2018 Evaluation of a UAV-assisted autonomous water sampling. *Water* **10** (5), 655. https://doi.org/10.3390/w10050655.

Lee, S. & Lee, D. 2018 Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. *International Journal of Environmental Research and Public Health* **15** (7), 1322. https://doi.org/10.3390/ijerph15071322.

Lin, L., Yang, H. & Xu, X. 2022 Effects of water pollution on human health and disease heterogeneity: A review. *Frontiers in Environmental Science* **10**. https://doi.org/10.3389/fenvs.2022.880246.

Madeo, D., Pozzebon, A., Mocenni, C. & Bertoni, D. 2020 A low-cost unmanned surface vehicle for pervasive water quality monitoring. *IEEE Transactions on Instrumentation and Measurement* **69** (4), 1433–1444. https://doi.org/10.1109/tim.2019.2963515.

Malakar, A., Snow, D. D. & Ray, C. 2019 Irrigation water quality – A contemporary perspective. *Water* **11** (7), 1482. https://doi.org/10.3390/w11071482.

Murray, C., Larson, A., Goodwill, J., Wang, Y., Cardace, D. & Akanda, A. S. 2022 Water quality observations from space: A review of critical issues and challenges. *Environments* **9** (10), 125. https://doi.org/10.3390/environments9100125.

Parris, K. 2011 Impact of agriculture on water pollution in OECD countries: Recent trends and future prospects. *International Journal of Water Resources Development* **27** (1), 33–52. https://doi.org/10.1080/07900627.2010.531898.

Pasika, S. & Gandla, S. T. 2020 Smart water quality monitoring system with cost-effective using IoT. *Heliyon* **6** (7), e04096. https://doi.org/10.1016/j.heliyon.2020.e04096.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., Maalouf, S. & Adams, C. 2020 Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews* **205**, 103187. https://doi.org/10.1016/j.earscirev.2020.103187.

Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M. & Khurshid, H. 2018 Surface water pollution detection using Internet of Things. In *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*. https://doi.org/10.1109/honet.2018.8551341.

Weller, D. L., Love, T. M. & Wiedmann, M. 2021 Interpretability versus accuracy: A comparison of machine learning models built using different algorithms, performance measures, and features to predict *E. coli* levels in agricultural water. *Frontiers in Artificial Intelligence* **4**. https://doi.org/10.3389/frai.2021.628441.

WHO. 2022 *Drinking-water*. World Health Organization (WHO). Available from: https://www.who.int/news-room/fact-sheets/detail/drinking-water.

Xu, T., Coco, G. & Neale, M. 2020 A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Research* **177**, 115788. https://doi.org/10.1016/j.watres.2020.115788.

Zhang, J. 2012 The impact of water quality on health: Evidence from the drinking water infrastructure program in rural China. *Journal of Health Economics* **31** (1), 122–134. https://doi.org/10.1016/j.jhealeco.2011.08.008.

Zhang, D., Sial, M. S., Ahmad, N., Filipe, A. J., Thu, P. A., Zia-Ud-Din, M. & Caleiro, A. B. 2020 Water scarcity and sustainability in an emerging economy: A management perspective for future. *Sustainability* **13** (1), 144. https://doi.org/10.3390/su13010144.

Zhi, W., Ouyang, W., Shen, C. & Li, L. 2023 Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. *Nature Water* **1** (3), 249–260. https://doi.org/10.1038/s44221-023-00038-z.

Zubair, H., Mumtaz, R., Ali, H. K. & Nasir, A. 2022 Time-series analysis and prediction of water quality through multisource data. In: *Empowering Sustainable Industrial 4.0 Systems with Machine Intelligence*, pp. 1–24. https://doi.org/10.4018/978-1-7998-9201-4.ch001.

First received 30 November 2023; accepted in revised form 2 June 2024. Available online 14 June 2024