International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Monitoring Water Quality in Pusu River Using Internet of Things (IoT) and Machine Learning (ML)

Nassereldeen Kabbashi [1*], Tahsin Fuad Hasan [2], M.d Zahangir Alam [3], T. Saleh [4], Aisha Hassan [5]

Submitted: 13/05/2024 **Revised:** 28/06/2024 **Accepted:** 12/07/2024

Abstract: The availability of clean water, a vital natural resource that supports diverse ecosystems, is increasingly threatened by sediment accumulation which impacts rivers, oceans, and coastal life, which is in line with sustainable development number goal 6 clean water and sanitation. Rapid industrialization and urbanization have intensified these challenges, leading to the degradation of natural water ecosystems and placing an undue strain on water resources. Pollution from sediments and human activities carries harmful contaminants, reduces visibility, disrupts aquatic life, and impairs ecosystem function. Maintaining the health of rivers and other water bodies requires the timely detection of changing conditions and deterioration, which is crucial for implementing effective countermeasures. However, current water quality monitoring methods primarily rely on laboratory tests, which require specialized staff, chemicals, and expertise. These traditional methods are often insufficient for addressing the complex and dynamic issues of water quality. Fortunately, the advent of the Internet of Things (IoT) technology has enabled real-time collection of water quality data. In addition, the application of soft computing technology for water quality assessment offers a more efficient, faster, and environmentally friendly alternative to conventional laboratory-based techniques. In this dissertation, we propose the use of an IoT device to monitor the performance of a water treatment system and collect data on key water quality indicators. Machine learning (ML) tools will be employed to analyze and simulate these data, enabling the prediction of future water quality parameters. The water quality dataset was collected in two stages. During the first iteration, data were gathered using sensors that measured four parameters: pH, turbidity, temperature, and total dissolved solids (TDS). In the subsequent iteration, the dataset was expanded to include a dissolved oxygen sensor in addition to the initial four sensors. The data collection process for turbidity and other water quality parameters involved more than just 879 data points, the data collection process was comprehensive, and the dataset was validated and analyzed with seasonal changes in mind, systematic approach ensured that the water quality parameters data collected were reliable, accurate, and actionable for monitoring water quality in the river. The dataset encompasses samples from three distinct potability classes: potable water sources, free-flowing river water from the Pusu River, and stagnant water from the puddles, and potholes. Nine proven classification algorithms were applied to the datasets, successfully classifying the water quality conditions with up to 98% accuracy. The best-performing model was then deployed and integrated into a graphical user interface (GUI) for rapid water condition testing, thereby facilitating the instantaneous assessment of water quality.

Keywords: *Water management; IoT, ML, GUI.*

Highlights

i. IoT and ML enable real-time water quality monitoring and prediction.
ii. The dataset was collected using sensors that measured pH, turbidity, temperature, TDS, and dissolved oxygen.
iii. Samples from potable water, the Pusu River, and stagnant water sources were analyzed.
iv. Nine classification algorithms were applied, achieving an accuracy of up to 98 %.
v. The best-performing model was integrated into the GUI for rapid water condition testing.
vi. Traditional water quality monitoring methods are inadequate for addressing these complex challenges.
vii. Innovative approaches that leverage ML and the IoT are required for robust monitoring systems.

Abbreviations

BOD	Biochemical oxygen demand
COD	Chemical oxygen demand
GUI	Graphical user interface
IoT	Internet of Things
KNN	K-Nearest Neighbors
ML	Machine-learning

[1] Department of Chemical Engineering and Sustainability, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia. Email: nasreldin@iium.edu.my
[2] Department of Chemical Engineering and Sustainability, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia
[3] Department of Chemical Engineering and Sustainability, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia
[4] Department of Mechatronics Engineering, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia
[5] Department of Electrical Engineering, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia

MLP	Multi-Layer Perceptron
TDS	Total dissolved solids

These abbreviations are used but we could not find a definition

CHES	Chemical Engineering and Sustainability Department
IIUM	International Islamic University Malaysia
KOE	Kulliyyah of Engineering
WHO	World Health Organization

Summary (for editorial board)

This study proposes the use of the IoT and ML techniques to monitor water quality in the Pusu River, Malaysia. Water quality data were collected using sensors measuring pH, turbidity, temperature, TDS, and dissolved oxygen from potable, river, and stagnant water sources. Nine classification algorithms were applied to the datasets, achieving up to 98% accuracy in classifying the water quality conditions. The best-performing model was integrated into a GUI for rapid water condition testing. This study highlights the limitations of traditional water quality monitoring methods and emphasizes the need for innovative approaches that leverage IoT and ML for robust, real-time monitoring systems to address the complex challenges posed by water pollution.

1. Introduction

Water pollution remains a pervasive issue, with both organic and inorganic contaminants from agricultural, industrial, and domestic sewage compromising water sources worldwide. This contamination poses serious threats to human health and agriculture, leading to the bioaccumulation of toxic metals within the food chain. Pollution, in general, is a significant problem for modern society, with many water sources being tainted by harmful substances from various human activities. In 2015, water and soil pollution accounted for 16% of global deaths, with around 92% of these fatalities occurring in developing economies (Landrigan et al., 2018). Moreover, river pollution not only harms ecosystems but also diminishes the utility of rivers for agriculture, urban and industrial water supply, and irrigation.

According to the Department of Environment Malaysia (DoE) (2017), the number of rivers in Malaysia decreased from 579 in 2008 to 477 in 2019. Alongside this decline, the quality of river water has worsened, making it increasingly challenging to utilize for various purposes. Malaysian rivers are threatened by both point and non-point sources of pollution, including sewage treatment plants, agro-industrial activities, manufacturing, commercial and residential wastewater, and pig farms (Che Mahmud, 2021). As a result, river water management in Malaysia remains a critical issue, especially with high turbidity and sediment contributing significantly to pollution. Suspended solids (SS), originating from sources such as soil erosion, runoff, and algal blooms, further exacerbate these challenges. Excess sediment in rivers not only affects biodiversity and coastal aquatic life but also leads to increased concentrations of suspended matter, particularly laterite clay particles, during the rainy season. These particles can absorb pollutants and give the water a brownish hue, further compromising water quality.

Traditional methods of analyzing water quality in treatment facilities typically involve sample collection, laboratory testing, and examination—processes that are labor-intensive, costly, and ultimately fall short in providing real-time feedback on water conditions (Das and Jain, 2017). However, the advancement and widespread availability of Internet of Things (IoT) technology now enables the real-time collection of water quality data. By incorporating soft computing technology, water quality assessment becomes more efficient, faster, and environmentally friendly compared to traditional laboratory-based methods. Additionally, the integration of machine learning (ML) algorithms allows for the analysis and classification of the acquired datasets, providing valuable insights into water quality conditions.

The Pusu River, situated on the campus of the International Islamic University Malaysia, has been significantly impacted by pollution and waste disposal issues stemming from the rapid population and industrial growth in the surrounding area. Urbanization has led to the river's water becoming increasingly cloudy, while waste dumping near its tributaries raises concerns about leachate seeping into the river. The river's location within a densely populated university environment further exacerbates its vulnerability to daily pollution (Hamid, 2020).

The widespread impacts of industrialization, combined with the accumulation of excess sediment in bodies of water such as rivers and oceans, pose significant risks to ecosystems, reduce the usability of water resources, and increase treatment costs. Traditional water quality monitoring methods, like laboratory tests, are often inadequate for effectively addressing these challenges due to their limited spatial and temporal coverage, high costs, lack of scalability, and delayed detection of water quality issues. To overcome these limitations, there is an urgent need to develop innovative approaches that leverage machine learning and IoT technologies to create more robust monitoring systems capable of continuously assessing water quality parameters in real-time. Research into the implementation of advanced, data-driven solutions offers actionable insights for maintaining and improving water quality, ultimately safeguarding ecosystems and enhancing water usability. By optimizing proven classification algorithms with IoT-based connectivity and infrastructure, we can address pressing environmental concerns and establish a more resilient water management framework (Zhu *et al*., 2022).

2. Water Management and IoT

Water quality encompasses the chemical, physical, and biological characteristics of water, determining its suitability for specific purposes, such as recreation, drinking, fisheries, agriculture, or industry (Rishika, 2019). Water can be broadly classified into two categories based on its origin: groundwater and surface water. Both types are vulnerable to contamination from various sources, including agriculture, industry, and domestic activities. Pollutants such as heavy metals, pesticides, fertilizers, hazardous chemicals, and oils can compromise the quality of both groundwater and surface water (Omer, 2019). Physical water quality parameters include turbidity, temperature, color, and odor, among others. Turbidity specifically measures the cloudiness of water, reflecting how much light can pass through it. High turbidity indicates the presence of suspended materials such as clay, silt, organic matter, plankton, and other particles (Kothari et al. 2021). Chemical parameters of water quality include pH, acidity, alkalinity, and the presence of ions such as chloride and sulfate, as well as metals and dissolved substances like oxygen, biochemical oxygen demand (BOD), chemical oxygen demand (COD), and any toxic substances. Pollution can alter the pH of water, potentially causing harm to aquatic life and disrupting ecosystems. Regarding biological parameters, the American Public Health Association (APHA, 2005) emphasizes that the presence or absence of living organisms is a crucial indicator of water quality. While most microorganisms in wastewater are benign, the presence of harmful microorganisms can signal the presence of disease-causing agents, indicating potential health risks.

The Internet of Things (IoT) enables the connection of devices through the internet, offering significant advantages for automating water distribution and monitoring for leaks. IoT is structured into three main layers: the physical layer, where sensors collect environmental data; the network layer, where data is converted into digital streams for processing; and the application layer, which delivers specific services to users as shown in Figure 1. To avoid network congestion, it is crucial to process or store data immediately upon collection or in the cloud (Yasin *et al*., 2021).

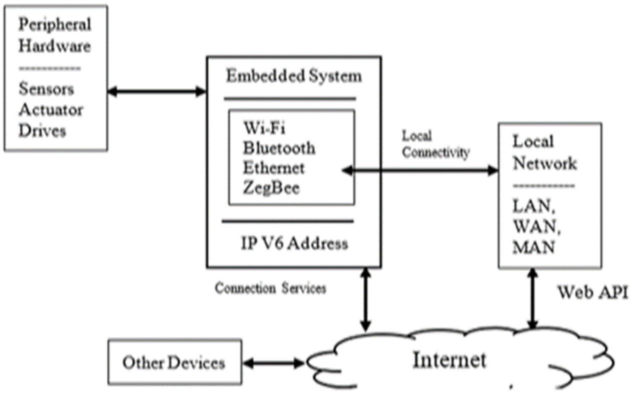


Fig. 1. Basic architecture of IoT [3]

In 2016, S. Geetha *et al*. introduced an innovative IoT-based solution for monitoring water quality within pipes. Their model tests water samples and analyzes the data online to enhance measurement accuracy and detect deviations from predefined standards. The system features a controller equipped with a built-in WiFi module and a cost-effective, intelligent water quality monitoring device capable of measuring pH, turbidity, and conductivity. It also incorporates a warning system to alert users to any fluctuations in water quality parameters. The experimental setup included five key parameters: conductivity, pH, turbidity, temperature, and water level. This setup was connected to the Ubidots network, and the collected data were compared against WHO guidelines for drinking water quality. K. Gupta *et al*. (2018) developed a device that can be continuously monitored via a mobile app from any location. This device offers full automation and intelligent management capabilities. It is designed to be stable, easy to install, and compact, making it highly efficient and user-friendly. V. Ranjan *et al.* (2020) introduced a smart rainwater harvesting system utilizing IoT technology. This model features a segregation mechanism that allocates rainwater into two tanks in a 60-40 percent ratio. A rainfall detection sensor is positioned at the top of the system to accurately monitor and determine rainfall events. Smart water management through IoT technology focuses on collecting and analyzing data related to a city's water supply, pressure, and distribution to enhance the efficiency of water transportation and usage. As economic development, climate change, and population growth increasingly impact water resource availability, the adoption of IoT solutions becomes crucial for optimizing water management practices and ensuring sustainable water use.

Machine learning is a robust technique for analyzing large datasets, uncovering patterns, and making predictions. The process involves several key steps: data acquisition, algorithm selection, model training, and model validation. Choosing the right algorithm is critical, and machine learning technologies are broadly categorized into supervised and unsupervised learning (Sagan, 2020).

Supervised learning utilizes labeled training datasets to develop predictive functions, making it suitable for tasks such as data classification and regression. Unsupervised learning, on the other hand, deals with unlabeled data and is primarily used for discovering hidden patterns and relationships through techniques like clustering and association mining. Data collection is a fundamental step in developing machine learning models as seen in figure 2. Both continuous and intermittent water quality monitoring results can serve as essential benchmarks for effective water system management. Traditional environmental monitoring methods, commonly employed by government agencies, often face challenges with in-situ monitoring due to practical constraints. In contrast, remote sensing technologies offer real-time, large-scale water quality monitoring capabilities. They can reveal pollutant movement and distribution patterns that are difficult to detect using conventional methods (Zhu *et al*., 2022).

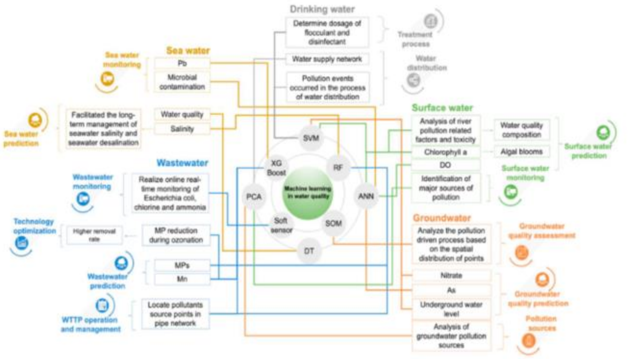

Drinking water
Determine dosage of flocculant and disinfectant
Treatment process
Water supply network
Pollution events occurred in the process of water distribution
Water distribution
Sea water
Sea water monitoring
Pb
Microbial contamination
Sea water prediction
Facilitated the long-term management of seawater salinity and seawater desalination
Water quality
Salinity
Wastewater
Wastewater monitoring
Realize online real-time monitoring of Escherichia coli, chlorine and ammonia
Technology optimization
Higher removal rate
MP reduction during ozonation
Wastewater prediction
MPs
Mn
WTTP operation and management
Locate pollutants source points in pipe network
SVM
XG Boost
RF
PCA
Machine learning in water quality
ANN
Soft sensor
SOM
DT
Surface water
Analysis of river pollution related factors and toxicity
Chlorophyll a
DO
Identification of major sources of pollution
Water quality composition
Algal blooms
Surface water prediction
Surface water monitoring
Groundwater
Analyze the pollution driven process based on the spatial distribution of points
Groundwater quality assessment
Nitrate
As
Underground water level
Groundwater quality prediction
Analysis of groundwater pollution sources
Pollution sources

Figure 2: Applications of different machine learning algorithms in different water treatment and management systems. (Zhou *et al.*, 2022)

Pappu et al. (2017) developed an IoT-based water quality monitoring system utilizing machine-to-machine (M2M) communication. This system integrates pH and TDS sensors to collect data, which is processed by an edge processor running a machine learning algorithm. The algorithm predicts water quality based on a trained dataset, and both predicted and historical data are stored in the cloud for easy access via mobile phones. This system automates water quality monitoring in residential areas, eliminating the need for human intervention.

Sagan et al. (2020) demonstrated that integrating machine learning with real-time sensor data and satellite observations enables advanced optimization. Their experiments revealed that models such as partial least squares regression, support vector regression, and deep neural networks achieved higher accuracy compared to traditional methods.

Shen *et al*. (2020) developed a geo-dataset to estimate and map nitrogen and phosphorus concentrations in rivers and streams across the contiguous United States. Their approach provided detailed spatial resolution, approximately 1 km (30 arc-seconds), capturing various chemical forms of these nutrients.

Wu *et al*. (2020) focused on classifying water images into distinct categories of clean and polluted water. This approach aims to provide immediate feedback for a water pollution monitoring system that utilizes IoT technology to capture and analyze water images. Overall, machine learning has proven to be a powerful tool for addressing water-related challenges, such as predicting water quality, optimizing resource allocation, and managing water shortages. Despite its effectiveness, several challenges remain in fully leveraging machine learning for water quality assessment. One significant hurdle is the reliance on large volumes of high-quality data, which can be difficult to acquire due to cost constraints or technological limitations in water treatment and management systems. Another challenge is the complexity of real-world water treatment and management systems, which often limits the applicability of current algorithms to contexts or conditions.

3. Research Method

After careful selection of the sensors for water quality assessment, the next steps involve establishing and configuring an IoT framework. Utilizing technologies like NodeRed and Arduino components, the IoT infrastructure will be precisely set up. A preliminary data collection trial will be strategically conducted across various river profile points at specific intervals. Concurrently, an exhaustive search for the most suitable machine learning model will be conducted, using the pre-existing classified dataset as a benchmark for testing and selection.

The critical phase focuses on data refinement, ensuring the removal of any extraneous or irrelevant information gathered during the collection process. This meticulous data cleaning is essential to eliminate potential inaccuracies and discrepancies in the final analysis. The ultimate objective is to enable precise feature extraction, laying the foundation for a robust and reliable water quality analysis.

The foundation of this system was established through the careful selection of a microcontroller, with the Arduino UNO ATmega328 emerging as the cornerstone. This choice was primarily motivated by its cost-effectiveness and, importantly, Arduino's inherent versatility in seamlessly integrating a variety of sensor systems—a crucial factor in optimizing our data collection capabilities. Sensors connect to the Arduino UNO via analog or digital pins, where the Arduino processes the data, converts it into meaningful readings, and then transmits it through serial communication or wireless modules for real-time monitoring and analysis.

Sensors interfacing with the Arduino UNO typically involve connecting the sensor's output to the appropriate input pins on the Arduino, which acts as a central processing unit. Here’s how the process generally works, Sensor Interface: sensors are connected to the Arduino UNO via its analog or digital input pins, depending on the type of sensor. Analog sensors, such as temperature or pH sensors, typically connect to the analog input pins (A0-A5), while digital sensors, like flow sensors, connect to the digital pins (D2-D13). Power Supply: The Arduino provides power to the sensors through its 5V or 3.3V pins. Ground (GND) connections are also made to ensure a complete circuit. Data Acquisition: the Arduino’s analog-to-digital converter (ADC) reads the voltage signal from the analog sensors. This signal is then converted into a digital value ranging from 0 to 1023, which represents the sensor’s reading within its specific range. Digital Sensors: Digital sensors provide a high or low signal (1 or 0), which the Arduino reads directly. Some digital sensors may also communicate using protocols like I2C or SPI, requiring additional code for data interpretation. Data Processing: The Arduino is programmed using the Arduino IDE to process the sensor data. This could involve calibrating raw sensor data, converting it into meaningful units (e.g., converting voltage to temperature), and performing initial data analysis or filtering. Thresholds and Alarms: The Arduino can be programmed to trigger alerts if sensor readings exceed certain thresholds, providing immediate feedback or action. Data Transmission: the processed data can be sent to a computer or another device via the Arduino’s USB port using serial communication. This allows real-time monitoring and data logging on a connected computer. Wireless Transmission: If the system includes wireless modules (like Wi-Fi, Bluetooth, or Zigbee), the Arduino can transmit data wirelessly to a remote server or cloud platform for further analysis, storage, and real-time monitoring. Storage and Display: data can also be displayed on an LCD screen connected to the Arduino or stored on an SD card for later retrieval. This setup allows for real-time data collection, processing, and transmission, enabling efficient monitoring and management of water quality or any other environmental parameters the sensors are designed to measure. The Smart Water Quality Monitoring System (SWQMS) enhances pool management by providing real-time monitoring of water parameters such as pH, chlorine levels, and temperature. This allows for immediate adjustments, ensuring optimal water quality and safety. The system also automates routine checks, reducing the need for manual testing, and can predict maintenance needs, thereby improving efficiency and reducing operational costs. By maintaining consistent water quality, the SWQMS helps prevent health issues and prolongs the life of pool infrastructure.

Our initial data collection and experimentation phase utilized four key water parameter sensors, each playing a critical role in capturing specific aspects of water quality:

Temperature Sensor: This sensor provides crucial insights into water temperature variations, a key factor influencing aquatic life and biochemical processes.

pH Sensor: Essential for measuring the acidity or alkalinity of the water, the pH sensor offered valuable data for assessing water suitability for various applications.

Turbidity Sensor: Focused on water clarity, the turbidity sensor served as a vital indicator of particulate matter and sediment levels, which are critical for ecosystem health.

TDS (Total Dissolved Solids) Sensor: This sensor was instrumental in measuring the concentration of dissolved substances in the water, helping to determine its overall purity and suitability for various uses.

In the second phase of data collection and testing, we incorporated an additional sensor:

Dissolved Oxygen Sensor: A crucial component for evaluating water quality, this sensor provided insights into the amount of oxygen available in the water, which is vital for sustaining aquatic life.

The error rate for each sensor varies depending on several factors, including the sensor's quality, environmental conditions, and how well the sensor is maintained. Common sources of errors include pH Sensor: Errors might arise due to electrode fouling, temperature fluctuations, or improper calibration. Typical error rates could be around 2-5%, with higher rates in extreme environments. Turbidity Sensor: Errors might occur due to sediment build-up on the sensor lens or interference from large particulate matter. Error rates might range from 3-7%. Dissolved Oxygen Sensor: Errors could be due to membrane fouling or temperature effects. Error rates might be in the range of 4-6%. These error rates are identified through calibration tests, comparison with known standards, and during the initial setup phase. Ongoing monitoring and regular calibration are essential to minimize errors and maintain data integrity.

IoT SET-UP

Node-RED stands out as an exceptional tool for building a server dedicated to water quality monitoring within IoT systems, thanks to its versatile and user-friendly visual programming interface. Its node-based architecture streamlines the integration of various sensors, IoT devices, and data streams, enabling seamless communication and data collection from multiple sources critical to water quality analysis. The extensive library of pre-built nodes in Node-RED accelerates the development of data collection workflows, allowing users to easily configure tasks for data processing, aggregation, and analysis.

Its flexibility and compatibility with numerous protocols make Node-RED adaptable to a wide range of sensor technologies used in water quality monitoring, ensuring smooth integration and interoperability as shown in Figure 3. Additionally, Node-RED's web-based dashboard creation tools empower users to visualize data in real-time, facilitating efficient monitoring and informed decision-making in water quality management systems.

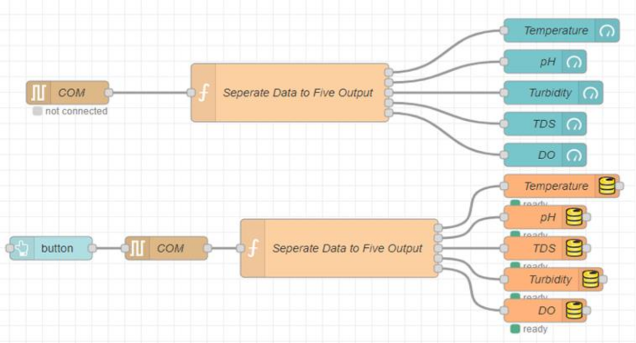


Figure 3: Basic Node-RED node structure for the 5-feature data collection

Firebase was utilized as the data server host, offering a robust platform for managing data due to its comprehensive features and user-friendly interface as seen in Figure 4. One of its standout capabilities is its real-time database, which enables instant updates and synchronization across multiple devices—an essential feature for IoT-based water quality monitoring systems that rely on live data feeds. This real-time functionality ensures that any changes detected by sensors or devices are immediately reflected throughout the system, providing up-to-the-minute information for analysis and decision-making.

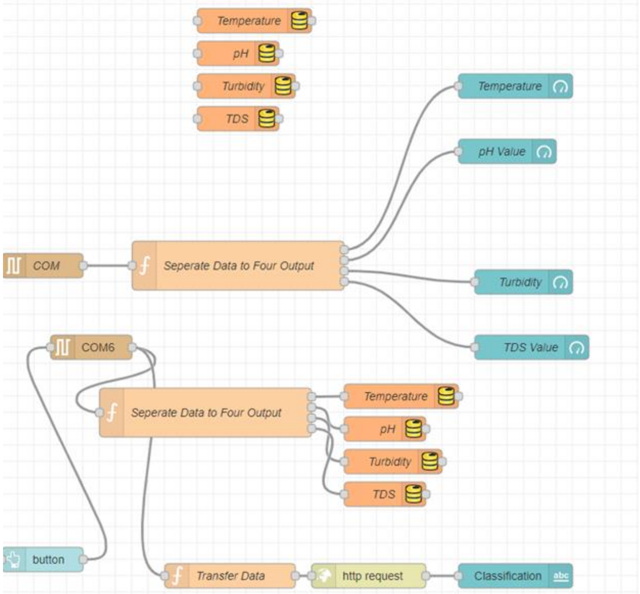
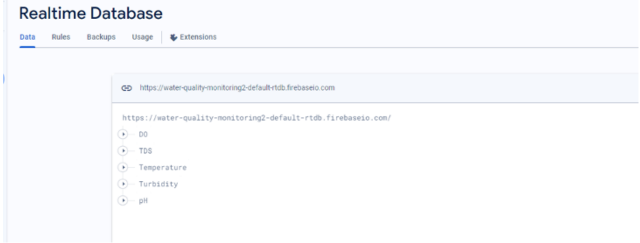


Figure 4: Set-up with the Machine Learning deployment and classification added for the 4-feature data collection

To enhance data handling, we integrated an app script that facilitates the transfer of data to Google Sheets. This integration (Figure 5) allows for easier data extraction and processing, particularly for machine learning applications, streamlining the workflow from data collection to analysis.



Realtime Database

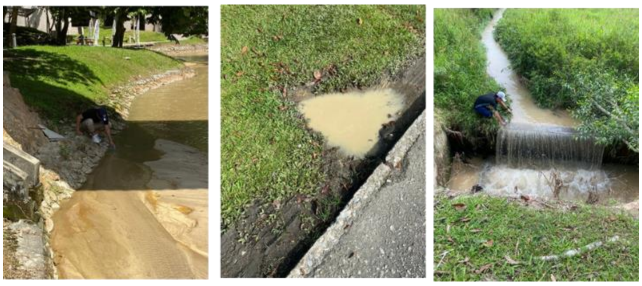
Data Rules Backups Usage Extensions

https://water-quality-monitoring2-default-rtdb.firebaseio.com

https://water-quality-monitoring2-default-rtdb.firebaseio.com/
- DO
- TDS
- Temperature
- Turbidity
- pH

Figure 5: Firebase Server

In our initial experimental attempt, we focused on collecting data for four key parameters: temperature, pH, total dissolved solids (TDS), and turbidity. The collected data was categorized into three distinct classifications (Figure 6). The first category, "potable water," included filtered water, bottled drinking water, and tap water. The second category, "flowing river water," consisted of samples taken from bodies of water in constant motion, primarily centered around the Pusu River and its streams. The third and final category, "still puddle water," encompassed water samples collected from stationary sources, such as puddles and potholes.



(a) (b) (c)

Figure 6: Data collection from: (a) a point from Pusu River lower stream (b) a puddle (c) a point from Pusu River upper stream

4. Machine Learning

The selected algorithms underwent thorough testing and evaluation to determine their effectiveness in addressing this specific classification task. The algorithms included:

Random Forest Classifier: Known for its ensemble learning technique, this algorithm combines the output of multiple decision trees to enhance predictive accuracy. Its robustness and ability to handle complex classification tasks make it a strong contender in data analysis.

K-Nearest Neighbors (KNN): Utilizing a proximity-based approach, KNN classifies data points based on their similarity to neighboring data points. Its simplicity and effectiveness often make it a reliable baseline for classification tasks.

Naïve Bayes Classifier: This algorithm operates on the principles of Bayes' theorem and assumes independence among features. It is efficient in processing large datasets while maintaining reasonable accuracy levels, making it a practical choice for many classification problems.

Gradient Boosting Classifier: This algorithm excels in sequentially training weak learners, gradually improving predictive performance with each iteration. It is particularly well-suited for capturing complex relationships within the data, making it a powerful tool for nuanced classification tasks.

Logistic Regression: Despite its name, Logistic Regression is a robust classification algorithm. It is particularly effective for binary classification tasks but is also adaptable for handling multinomial classifications.

Support Vector Machine (Polynomial): Utilizing a kernel-based approach, the Support Vector Machine with a polynomial kernel excels at identifying optimal decision boundaries, making it effective in capturing complex relationships among features.

Decision Tree: With its intuitive, tree-like structure, Decision Trees are adept at interpreting and representing data relationships, offering clear decision paths based on feature attributes.

Multi-Layer Perceptron Classifier: As a type of artificial neural network, the Multi-Layer Perceptron is well-suited for learning complex patterns in data, providing adaptability and versatility in managing intricate classification tasks.

Each of these algorithms was rigorously tested, compared, and evaluated to assess their effectiveness in classifying the collected water quality data into the specified categories. The aim was to identify the algorithm or combination of algorithms best suited to address the specific complexities and nuances of this water quality monitoring application.

5. Results and Discussion

Training Dataset 1 was collected using four sensors as outlined in the previous chapter. The dataset included over 879 entries and upon initial data cleaning and processing which included removal of duplicates, null values and faulty data set moved the final dataset to 526 entries from the 3 dataset classes (Figure 7). The "faulty data" removed during the data cleaning process typically included sensor readings that were inconsistent, out-of-range, or clearly erroneous due to sensor malfunctions, environmental interference, or transmission errors. This data could also include incomplete records or anomalies caused by temporary system glitches. Removing this faulty data was crucial to ensure the accuracy and reliability of the final analysis.

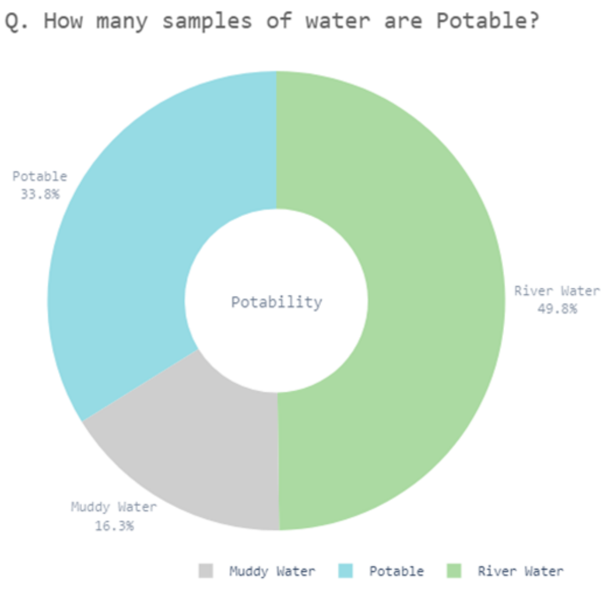


Figure 7: The distribution of the dataset

Figure 7 illustrates the distribution of the dataset across the potability classes, highlighting a degree of class imbalance. River water accounts for over 45% of the data, while potable water represents 33.8%, indicating a mild imbalance. Muddy water, at 16.3%, shows a moderate imbalance. These class imbalances present challenges, particularly due to the overrepresentation of the river water class, which could potentially bias the predictive models.

Figure 8 presents the correlation matrix, providing insights into the impact of each parameter on water quality. A key observation is the weak correlation between temperature sensor readings and the other features, as well as water potability. This lack of correlation can be attributed to the multifactorial nature of temperature, which varies significantly based on factors such as the collection point and time of day, independent of water potability. As a result, the temperature feature was excluded from model training and testing due to its limited predictive value.

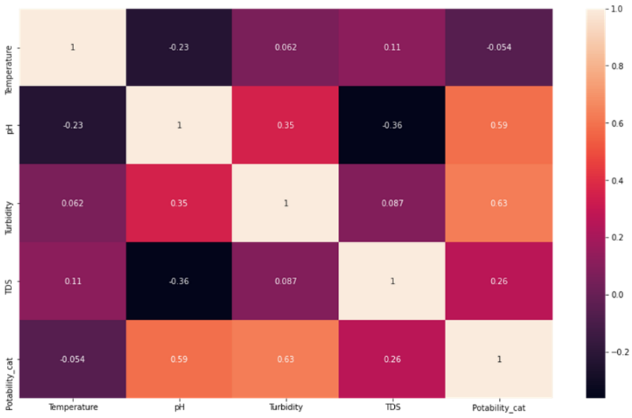


Figure 8: Correlation Matrix for dataset 3

In this project, six sensors were installed along different strategic points of the river to monitor various water quality parameters. These sensors were deployed at locations identified as critical for understanding the overall health of the river, such as upstream, midstream, and downstream areas, as well as near potential pollution sources. The sensors were configured to collect data at regular intervals, typically every 15 minutes. This frequency allowed for continuous monitoring and capturing of real-time changes in water quality. The sensors were connected to an IoT framework, which allowed for real-time data transmission to a central server. This setup enabled continuous, remote monitoring of the water quality, with data being accessible in real-time via a web-based dashboard or mobile application. In addition to continuous data collection, the system was programmed to send real-time alerts if any of the measured parameters exceeded predefined thresholds, enabling prompt intervention if necessary.

Both turbidity and pH demonstrate a robust correlation of approximately 50% with water potability, underscoring their importance in classification tasks. In contrast, total dissolved solids (TDS) exhibit a lower correlation of 23% with potability. Additionally, the correlation between individual features and other features in the dataset is minimal. These results highlight the distinct role of each parameter in evaluating water quality and emphasize the critical importance of feature selection in model development, as illustrated in Figure 9.

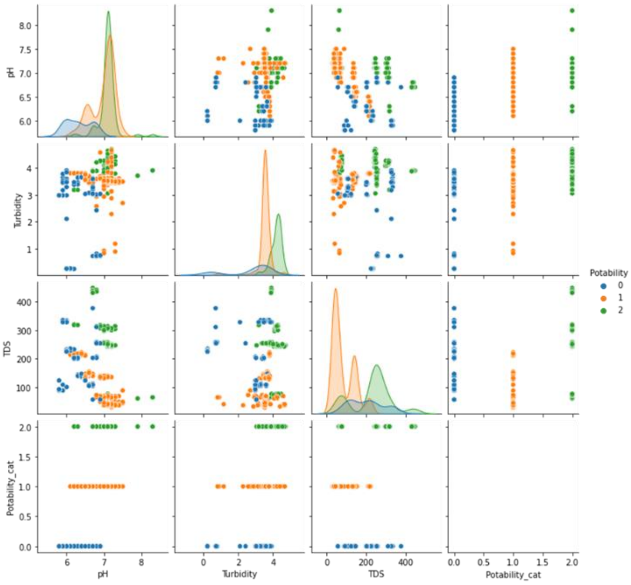


Figure 9: Scatter Plot of each feature against another

Figure 9 illustrates the scatter plot distribution of the classes across each feature, aiming to reveal class overlaps and improve model differentiation. The plots show distinct clusters for each class, though some regions exhibit overlap. Muddy water (0) displays the most dispersed distribution, with noticeable clustering primarily in the turbidity vs. pH plot. River water (1) forms a more compact cluster, which may be influenced by its higher representation in the dataset. Clear distinctions are evident, such as potable water (0), which consistently falls within the pH range of 6.5 to 7.5 and shows the highest turbidity values, reflecting its clarity.

6. Model Performance

To evaluate the performance of the selected algorithms, the following metrics are utilized:

Classification Accuracy: This metric measures the proportion of correctly predicted instances among all input samples, offering a straightforward assessment of model effectiveness. A higher accuracy score indicates better overall performance of the model.

Confusion Matrix: The confusion matrix provides a detailed breakdown of the model’s performance by categorizing predictions into true positives, true negatives, false positives, and false negatives. This detailed view allows for the derivation of accuracy and offers deeper insights into the model’s predictive capabilities.

Precision, Recall, and F1 Score:

Precision quantifies the ratio of true positives to all instances predicted as positive, reflecting the accuracy of positive identifications. It is computed by dividing true positives by the sum of true positives and false positives.

Recall measures the proportion of correctly identified positive instances out of all actual positive instances, indicating the model’s ability to capture true positives. It is calculated by dividing true positives by the sum of true positives and false negatives.

F1 Score combines precision and recall into a single metric, providing a balance between the two. It is the harmonic mean of precision and recall, particularly useful in scenarios with class imbalance, offering a comprehensive evaluation of model performance.

Decision Boundary: The decision boundary represents the line or surface that separates different classes in a classification problem. It provides insights into the model’s classification accuracy and helps in assessing the model’s robustness. A clear, well-defined boundary suggests good model performance, while a fuzzy or overly complex boundary may indicate issues such as suboptimal modeling or overfitting.

Overall, as shown in Table 1, all datasets achieved accuracy rates above 85% across various models. This high performance can be attributed to the large size of the datasets and the well-defined clusters among the classes. Gradient Boosting emerged as the most effective model, achieving an impressive accuracy of 99.3%. Similarly, other ensemble methods, such as Random Forests, as well as non-linear models like Decision Trees and K-Nearest Neighbors (KNN), demonstrated exceptional performance in classifying water potability across multiple classes. In contrast, linear models, such as Logistic Regression and Support Vector Machines, faced challenges with the non-linear relationships present in the data, which likely affected their classification accuracy in this multi-class scenario.

Table 1: Accuracy Scores Across Models for dataset 1

Model	Accuracy score	Score
6	Gradient Boosting	0.993671
0	Random Forest	0.981013
3	Decision Tree	0.974684
1	KNN	0.962025
5	Multilayer Perceptron	0.962025
2	Logistic Regression	0.898734
4	Support Vector Machines	0.860759
7	Naive Bayes	0.841772

As illustrated in Table 2, all models demonstrate high precision, recall, and F1 scores, showcasing their effectiveness in accurately classifying instances across multiple classes. Gradient Boosting achieves the highest scores, with Decision Tree, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) following closely behind. Although Random Forest exhibits slightly lower performance, it still shows strong results. These metrics offer valuable insights into each model's classification capabilities, helping to identify the most suitable model for the specific classification task.

Table 2: Confusion Matrix for Top 5 models in Dataset 1

	Precision	Recall	F1 Score
Gradient Boosting	0.9866666667	0.9933333333	0.99
Random Forest	0.9666666667	0.9766666667	0.97
Decision Tree	0.9766666667	0.985	0.98
KNN	0.9766666667	0.985	0.98
MLP	0.9733333333	0.9822222222	0.9766666667

Dataset 1, despite exhibiting some class imbalance, served as the initial benchmark for evaluating model performance. The results across various metrics highlighted crucial areas for improvement in the experimentation process. Correlation analysis underscored the significance of turbidity and pH in determining water potability, while the temperature sensor showed negligible correlation and was subsequently excluded from further analysis. The scatter plots revealed substantial overlap among the three classes, particularly between puddle and river water, and occasionally between river water and potable water.

During model training and testing, ensemble methods such as Gradient Boosting and Random Forest, along with non-linear models like K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP), exhibited exceptional performance, achieving accuracies exceeding 95%. Among these, Random Forest emerged as the most effective model for deployment, owing to its robust performance across multiple evaluation metrics and its adept handling of non-linear relationships in the data. The evaluation of both datasets and machine learning models highlights the effectiveness of integrating IoT-enabled water quality monitoring systems with machine learning for analysis and classification. The results validate the proposed framework's ability to accurately assess and classify water quality. They emphasize the crucial role of feature selection, model evaluation, and iterative refinement in achieving precise and dependable assessments. Additionally, the choice of Random Forest for Dataset 1 and K-Nearest Neighbors (KNN) for Dataset 2 underscores the importance of tailoring model selection to the specific characteristics of each dataset and classification task. Integrating IoT devices into water treatment systems enables real-time monitoring, automated control, and predictive maintenance, leading to improved efficiency, cost savings, and proactive water quality management.

The long-term benefits of using IoT and machine learning for water quality monitoring include significant cost savings and efficiency improvements. These technologies enable real-time monitoring and analysis, reducing the need for manual sampling and laboratory testing, which lowers labor and operational costs. Early detection and automated intervention help prevent costly repairs and environmental damage by addressing issues before they escalate. Additionally, IoT systems reduce the need for extensive physical infrastructure, minimizing infrastructure costs. The scalability and adaptability of these systems allow for efficient resource allocation and continuous optimization, further enhancing the overall effectiveness and sustainability of water quality management.

7. Acknowledgement

The author wishes to acknowledge the support received from the KOE and the CHES department, as well as express gratitude to all the technicians and colleagues involved. Special thanks are extended to RMC IIUM for enabling this grant under Project ID: IUMP-SRCG22-005-0005 and Project Title: "Investigation of Water Quality Monitoring Using IoT and Machine Learning Techniques: Pusu River IIUM Gombak Case Study."

References:

[1] Landrigan, P. J., Fuller, R., Acosta, N. J. R., Adeyi, O., Arnold, R., Basu, N. (Nil), Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J. I., Breysse, P. N., Chiles, T., Mahidol, C., Coll-Seck, A. M., Cropper, M. L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Zhong, M., 2018. The Lancet Commission on pollution and health. The Lancet Commissions, 391(10119), 462–512. https://doi.org/10.1016/S0140- 6736(17)32345-0.

[2] Department of Environment Malaysia (DoE)., 2017. Environmental Essentials for Siting of Industries in Malaysia (EESIM).

[3] Che Mahmud, N. A., 2021. River water quality issues in Malaysia. UMP News. Retrieved January 8, 2023, from https://news.ump. edu.my/experts/river-water-quality-issues-malaysia.

[4] Das B and Jain, P. C., 2017. Real-time water quality monitoring system using Internet of Things. International Conference on Computer, Communications and Electronics (Comptelix), Jaipur, India, 2017, pp. 78-82, doi: 10.1109/COMPTELIX.2017.8003942.

[5] Hamid, S.A., Rahim, A.M., Fadhlullah, S.Y., Abdullah, S.B., Muhammad, Z., & Leh, N.A., 2020. IoT based Water Quality Monitoring System and Evaluation. 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 102-106.

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

[7] Rishika Chakraborty, Khalid M Khan, Daniel T Dibaba, Md Alfazal Khan, Ali Ahmed, Mohammad Zahirul Islam., 2019. Health Implications of Drinking Water Salinity in Coastal Areas of Bangladesh. Int J Environ Res Public Health. 2019 Oct 4;16(19):3746. doi: 10.3390/ijerph16193746.

[8] Omer, N.H., 2019. Water Quality Parameters. Water Quality - Science, Assessments and Policy.

[9] Kothari, V., Vij, S., Sharma, S., & Gupta, N., 2021. Correlation of various water quality parameters and water quality index of districts of Uttarakhand. Environmental and Sustainability Indicators, 9, 100093. doi:10.1016/j.indic.2020.100093.

[10] APHA., 2005. Standard Methods for the Examination of Water and Wastewater, Washington DC: American Public Health Association.

[11] Yasin, H. M., Zeebaree, S. R., Sadeeq, M. A., Ameen, S. Y., Ibrahim, I. M., Zebari, R. R., & Sallow, A. B., 2021. IoT and ICT based smart water management, monitoring and controlling system: A review. Asian Journal of Research in Computer Science, 8(2), 42-56.

[12] Geetha S, Gouthami S., 2016. Internet of things enabled real time water quality monitoring system. Smart Water.; 2:1-19.

[13] Gupta K, Kulkarni M, Magdum M, Baldawa Y, Patil S., 2018. Smart water management in housing societies using IoT.

Proceedings of the International Conference on Inventive Communication Computational Technologies; 1609-1613.

[14] Ranjan V, Reddy MV, Irshad M, Joshi N., 2020. The Internet of Things (IOT) based smart rain water harvesting system. IEEE, 6th International Conference on Signal Processing and Communication, ICSC 2020;302-305.

[15] Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., 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, 103187. doi:10.1016/j.earscirev.2020. 1031.

[16] Pappu, S., Vudatha, P., Niharika, A. V., Karthick, T., & Sankaranarayanan, S., 2017. Intelligent IoT based water quality monitoring system. International Journal of Applied Engineering Research, 12(16), 5447-5454.

[17] Shen, L.Q., Amatulli, G., Sethi, T. et al., 2020. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Sci Data 7, 161. https://doi.org/10. 1038/s41597-020-0478-7.

[18] Wu, Y., Zhang, X., Xiao, Y., & Feng, J., 2020. Attention Neural Network for Water Image Classification under IoT Environment. Applied Sciences, 10(3), 909. doi:10.3390/app10030909.