WATER QUALITY MONITORING OF SUNGAI PUSU USING IOT TECHNOLOGY AND MACHINE LEARNING APPROACHES

BY

TAHSIN FUAD HASAN

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

2025

WATER QUALITY MONITORING OF SUNGAI PUSU USING IOT TECHNOLOGY AND MACHINE LEARNING APPROACHES

BY

TAHSIN FUAD HASAN

A dissertation submitted in fulfilment of the requirement for the degree of Master of Science in Engineering

Kulliyyah of Engineering
International Islamic University Malaysia

DECEMBER 2025

ABSTRACT

Water quality monitoring is crucial for maintaining healthy ecosystems and ensuring the availability of clean water resources. Traditional monitoring methods, relying on laboratory tests, are often insufficient due to their time-consuming nature and lack of real-time data. This research aims to develop a cost-effective, real-time water quality monitoring system using Internet of Things (IoT) technology and machine learning (ML) algorithms. The primary objectives of this study are to characterize water quality in the Sungai Pusu at IIUM Gombak campus, investigate the effectiveness of IoT-enabled sensors and ML algorithms in continuous water quality monitoring, and design an integrated water quality monitoring system. The methodology involved developing an IoT-based device using Arduino UNO microcontroller and multiple sensors to collect water quality data. Two datasets were generated: the first with four parameters (pH, turbidity, temperature, and total dissolved solids) and the second including an additional dissolved oxygen sensor. Data was collected from three distinct water sources: potable water, flowing river water, and stagnant puddle water. Nine different classification algorithms were applied to analyze the collected data. Results showed that the developed system successfully classified water quality conditions with up to 98% accuracy. For the first dataset, Random Forest algorithm performed best with 98.1% accuracy, while for the second dataset, K-Nearest Neighbors (KNN) algorithm achieved 97.2% accuracy. This research demonstrates the potential of combining IoT and ML technologies for efficient, real-time water quality monitoring. The developed system offers a scalable and cost-effective solution for continuous assessment of water quality parameters, which can significantly improve water management in rivers and streams.

المستخلص

مراقبة جودة المياه أمر بالغ الأهمية للحفاظ على النظم البيئية الصحية وضمان توافر موارد المياه النظيفة. الطرق التقليدية لمراقبة جودة المياه، التي تعتمد على الاختبارات المخبرية، غالبًا ما تكون غير كافية بسبب طبيعتها التي تستغرق وقتًا طويلاً وعدم توفر البيانات في الوقت الفعلي. تهدف هذه الدراسة إلى تطوير نظام مراقبة لجودة المياه في الوقت الفعلي وبتكلفة فعالة باستخدام تكنولوجيا إنترنت الأشياء (IoT) وخوارزميات تعلم الآلة (ML). الأهداف الرئيسية لهذه الدراسة تشمل توصيف جودة المياه في نهر سونغاي بوسو بحرم IIUM في غومباك، ودراسة فعالية أجهزة الاستشعار المدعومة بإنترنت الأشياء وخوارزميات تعلم الآلة في المراقبة المستمرة لجودة المياه، وتصميم نظام متكامل لمراقبة جودة المياه.

شملت المنهجية تطوير جهاز يعتمد على إنترنت الأشياء باستخدام المتحكم الدقيق Arduino UNO والعديد من أجهزة الاستشعار لجمع بيانات جودة المياه. تم إنشاء مجموعتين من البيانات: الأولى تحتوي على أربعة معايير (الرقم الهيدروجيني، العكارة، درجة الحرارة، وإجمالي المواد الصلبة الذائبة)، والثانية تضمنت مستشعرًا إضافيًا للأكسجين المذاب. تم جمع البيانات من ثلاثة مصادر مياه متميزة: مياه صالحة للشرب، مياه نهر جار، ومياه راكدة.

تم تطبيق تسعة خوارزميات تصنيف مختلفة لتحليل البيانات التي تم جمعها. أظهرت النتائج أن النظام المطور نجح في تصنيف ظروف جودة المياه بدقة تصل إلى 98% بالنسبة للمجموعة الأولى من البيانات، حقق أفضل أداء بدقة بلغت 98.1%، بينما حققت خوارزمية أقرب خوارزمية الغابات العشوائية (Random Forest) الجيران (K-Nearest Neighbors - KNN) دقة 97.2% للمجموعة الثانية من البيانات.

تُظهر هذه الدراسة إمكانيات الجمع بين تقنيات إنترنت الأشياء وتعلم الآلة لمراقبة جودة المياه بكفاءة وفي الوقت الفعلي. يوفر النظام المطور حلاً قابلًا للتطوير وبتكلفة فعالة للتقييم المستمر لمعايير جودة المياه، مما يمكن أن يُحسن إدارة المياه بشكل كبير في الأنهار والجداول.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Engineering.

…………………………………

Nassereldeen Ahmed Kabbashi
Supervisor

…………………………………

Tanveer Saleh
Co-Supervisor

…………………………………

Md Zahangir Alam
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Engineering.

Ma'an Fahmi Rashid Al Khatib
Examiner

Abdurahman Hamid Nour
External Examiner

This dissertation was submitted to the Department of Mechatronics Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science in Engineering.

…………………………………
Ali Sophian
Head, Department of Mechatronics Engineering

This thesis was submitted to the Kulliyyah of Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science in Engineering.

…………………………………
Khairul Azami Bin Sidek
Dean, Kulliyyah of Engineering

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Tahsin Fuad Hasan

15/09/2024

Signature…………………………… Date……………………

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

WATER QUALITY MONITORING OF SUNGAI PUSU USING IOT TECHNOLOGY AND MACHINE LEARNING APPROACHES

I declare that the copyright holder of this thesis are jointly owned by the student and IIUM.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Tahsin Fuad Hasan

05/12/2025

…………………………… ……………………..

Signature Date

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 OVERVIEW

Rivers are considered major natural infrastructures due to their role in providing water for agricultural, urban, and industrial uses, and they are essential for the sustainable development of countries. It is also a crucial ingredient within the sustainable development of countries. The waterways serve as habitats for all sorts of organisms and is part of varying ecological biodiversity. However, with the exponential growth of economic development and urbanization, river pollution has risen significantly, causing severe ecological damage and deterioration of the water environment. Large amounts of domestic and industrial wastewater are discharged into rivers, leading to the loss of their function as resources and harming human health and agriculture through the bioaccumulation of toxic materials in the food chain. By the early 20th century, few natural rivers remained globally, and water pollution continues to be a widespread problem, with organic and inorganic contaminants affecting ecosystems and reducing the benefits rivers provide for agriculture, water supply, and irrigation. In 2015, water and soil pollution were responsible for 16% of global deaths, with 92% of these occurring in developing economies (Landrigan et al., 2018).

According to data from the Department of Environment (DoE) (2017), the number of rivers in Malaysia decreased from 579 in 2008 to 477 in 2019. The quality of river water has also declined, making it more difficult to use for various purposes. The rivers are threatened by both point and non-point sources of pollution, including sewage treatment plants, agro-industry, manufacturing, commercial and residential wastewater, and pig farms (Che Mahmud, 2021). Consequently, river water management in Malaysia remains a significant issue, particularly with high turbidity and sediment contributing to river pollution. These suspended solids (SS) can come from various sources, including soil erosion, runoff, and algal blooms. Excess sediment in rivers can also affect biodiversity and coastal aquatic life. During the rainy season, the concentration of suspended matter in rivers, particularly laterite clay particles, is high, which can adsorb pollutants, and give the water a brown color.

When considering water treatment facilities and mechanisms, traditional means of analyzing water quality required the procurement of sample collection, laboratory testing and examination which involved multiple processes, labor and cost and ultimately lacked in providing real-time feedback of the water conditions (Das and Jain, 2017). Concurrently, the advancement and availability of internet of things (IoT) technology has made it possible to collect real-time data on the quality of water in bodies of water. Using soft computing technology for water quality assessment can be a more efficient, faster, and environmentally friendly alternative to laboratory-based techniques. Furthermore, inclusion of machine learning (ML) algorithms will allow for the analysis of acquired dataset, in order to classify the water quality conditions.

The case study's subject, Sungai Pusu—located on the campus of the International Islamic University Malaysia—has faced pollution and waste disposal issues due to the rapid growth of population and industrialization in the area. The river has been affected by urbanization, leading to cloudy water, and waste dumping near its tributary may result in leachate seeping into the river. The location of the river within a university area with many residents makes it vulnerable to daily pollution. To address these issues, implementing a real-time water quality monitoring system using IoT and machine learning could provide valuable insights for effective pollution control and river management strategies.

1.2 PROBLEM STATEMENT

The rampant industrialization and urbanization have led to increased water pollution and sediment accumulation in rivers and oceans, posing significant risks to ecosystems and water resources (Landrigan et al., 2018; Wang et al., 2012). Conventional water quality monitoring methods, such as laboratory tests, have proven inadequate in addressing these issues due to limitations in spatial and temporal coverage, high costs, lack of scalability, and delayed detection of water quality problems (Das and Jain, 2017; Banna et al., 2014). While recent studies have explored the use of IoT and machine learning for water quality monitoring (Geetha & Gouthami, 2016; Pappu et al., 2017), there remains a gap in developing comprehensive, real-time monitoring systems that integrate multiple sensors and tested through the plethora of machine learning algorithms.

This research aims to address this gap by developing an innovative approach that leverages machine learning and IoT capabilities to create a more robust monitoring system capable of continuously assessing water quality parameters in real-time. Unlike previous studies that focused on limited parameters or specific water bodies (Sagan et al., 2020; Wu et al., 2020), our research integrates a wider range of sensors and applies multiple machine learning algorithms to improve classification accuracy and adaptability across diverse water environments. By optimizing proven classification algorithms with IoT-based connectivity and infrastructure, this study aims to provide a more comprehensive and scalable solution for water quality monitoring. This approach offers a more holistic integration of sensor data, real-time analysis, and adaptive machine learning techniques, potentially paving the way for a more resilient and responsive water management framework.

1.3 RESEARCH OBJECTIVES

The primary objective of this research is to develop a real-time water quality platform using the Internet of Things and Machine learning on rivers. The following issues need to be addressed to achieve the main objectives:

1) To investigate the effectiveness of IoT-enabled sensors and machine learning algorithms in continuously monitoring water quality indicators, such as turbidity, pH and total dissolved solids to provide real-time and accurate data.
2) To design and develop an integrated water quality monitoring system that utilizes IoT-enabled sensors and leverages classification algorithms based on machine learning techniques.

1.4 RESEARCH SCOPE

This study will focus on conducting analysis and experiment of water quality measurement along the Sungai Pusu River in IIUM. The experiment will be done using IoT and machine learning tools to collect parameter readings including pH, temperature, turbidities, total dissolved solids and dissolved oxygen level. Data will be collected and analyzed using 9 different ML algorithms to complete assessment for identification of optimal means of water quality monitoring system.

1.5 THESIS ORGANIZATION.

This thesis consists of five chapters, each of which delves into different aspects of a proposed water quality monitoring system. Chapter Two presents the literature review, which covers research on the topics of water quality analysis, principles of IoT and ML, and existing systems that are currently utilized for water monitoring. Chapter Three details the methodology of the proposed system, explaining how data was collected using Arduino Uno Microcontroller aided with compatible sensors for parameter data collection. The chapter also describes the system setup and data collection process. Chapter Four presents the evaluation and analysis of the system on varying datasets, using various analytical tools. Finally, Chapter Five presents the conclusion and future works that can be made to the system to further increase its accuracy and capability. Overall, this thesis provides a comprehensive overview of a proposed water quality monitoring system and its potential to improve water quality management.

1.6 CHAPTER SUMMARY.

The chapter details the very general overview of the thesis, emphasizing on developing a water quality monitoring system using IoT and ML. The chapter includes the background and motivation behind the study, highlighting the need for a real time water monitoring system. The chapter also describes the research objectives that we hope to meet in order to achieve a conducive and sustainable system.

CHAPTER TWO:

LITERATURE REVIEW

2.1 OVERVIEW

Rivers are regarded as the primary source of irrigation and water supply for agriculture, cities, and industries, both of which are crucial for the long-term growth and development of any country. Numerous creatures that are necessary for the right balance of biodiversity can be found in rivers and other bodies of water as habitats. Rivers' beneficial uses are diminished, and biodiversity is seriously threatened by increased silt in the water. Meanwhile, the increase in industrialization and human activities has led to a significant increase in the release of pollutants, including organic and inorganic substances, into the environment through waste and wastewater. These substances, particularly heavy metals, can be harmful to public health and can damage physiological functions by altering enzymes and proteins. Overall, clean water is becoming increasingly scarce on a global scale. The World Water Assessment Programme (WWAP) and the Water Resources Group (WRG) have estimated that water consumption will increase by 2030 and that 40% of the global population will suffer from water scarcity (Almuktar et al., 2018). In addition, it is predicted that 34% of countries will experience water scarcity by 2025, with many of these being developing nations (Stikker, 1998).

Henceforth, there is an increasing need for water quality monitoring across several domains, from drinking water, and municipal and industrial wastewaters, to environmental waterways, as water quality and availability are emerging as a crucial concern throughout the world (rivers, lakes, groundwater, and oceans). Since 2009, there have been numerous discussions about the creation of a smart water grid (SWG or smart water network) to achieve an effective, dependable, and high-quality water supply. The term "Smart Water Grid" (SWG) has been often used, particularly since the "Water Innovation Alliance" in 2009, when the leading water-relevant companies from all over the world took part in the launch of the Smart Water Grid Initiative. The next-generation water quality monitoring paradigm is thought to be represented by the smart water quality monitoring system

(SWQMS) (Baanu et al, 2021), which is seen as a crucial component of SWG and refers to the implementation of intelligent water information systems by IT convergence. Online water quality monitoring (OWQM) systems with a collection of online automatic monitoring devices for data capture, transmission networks, and business software for data analysis formed the basis of the original SWQMS concept. Although OWQM is designed to measure physicochemical parameters in real-time in waters (including groundwater, industrial wastewater, urban drainage, rivers, streams, lakes, and oceans), there are currently efforts being made to enhance how monitoring is carried out, how information is shared, and how decisions are made based on monitoring (Dong et al., 2014).

2.2 WATER QUALITY ANALYSIS:

Water quality refers to the chemical, physical, and biological properties of water and its suitability for a specific purpose such as recreation, drinking, fisheries, agriculture, or industry. Each purpose has different standards for chemical, physical and biological properties. Water quality standards are established through research to ensure the efficient use of water for a particular purpose. Water quality analysis is conducted to measure essential water quality parameters and assess whether they meet the required standards for human use or consumption. This analysis is mainly utilized for monitoring and ensuring compliance with standards, regulations and maintaining the efficiency of systems. It is crucial in the public health and industrial sectors, particularly for drinking water (Ritabara, 2019).

Water can be first separated into two categories based on its origin, ground water and surface water. Both types of water are at risk of contamination from sources like agriculture, industry, and domestic activities which may include pollutants like heavy metals, pesticides, fertilizers, hazardous chemicals and oils. Water quality itself can be classified into four categories: water that is safe to drink, water that is safe to consume but may have an unpleasant taste, water that is contaminated or polluted and water that is infected with harmful organisms. Furthermore, according to Omer (2019), water quality parameters can be subdivided into three categories: physical, chemical, and biological. The focus of this study will comprise mainly of physical and chemical parameters.

Within the physical parameters of water quality, there is turbidity, temperature, color, odor, etc. Turbidity refers to the cloudiness of water, which is a measure of how well light can pass through it. This indicates the presence of suspended materials such as clay, silt, organic matter, plankton, and other particles in the water. Turbidity in drinking water can make it appear unappealing and has several negative impacts. It can increase the cost of water treatment, provide shelter for harmful microorganisms, making them resistant to disinfection processes, it can also clog or damage fish gills, reducing growth, affecting maturity of eggs and larvae, and lowering the efficiency of fishing methods. Kothari et al.'s (2021) correlation study of water quality parameters and water quality index of districts of Uttarakhand, the study found that turbidity significantly correlates with the presence of nitrate in drinking water. Temperature influences various characteristics of water such as taste, thickness, ability to dissolve, smell and chemical reactions. This means that processes like sedimentation, chlorination, and biological oxygen demand (BOD) are affected by temperature. Additionally, the ability of water to remove dissolved heavy metals through the process of biosorption is also influenced by temperature. According to Wisewell (n.d.)l, water between 10°C and 22°C (50°F and 72°F) is generally preferred and optimal for allowing the body to absorb the water and rehydrate.

Another key physical parameter is Total Dissolved Solids (TDS). is a measure of all inorganic and organic substances dissolved in water, including minerals, salts, and ions. TDS levels can affect water taste, smell, and overall quality. High TDS concentrations could signal the presence of harmful contaminants, although TDS itself is considered a secondary standard by the EPA (Sensorex, 2021). TDS can come from natural sources like springs and soil, as well as human activities such as agricultural runoff and water treatment processes. Elevated TDS levels can impact aquatic ecosystems, affect the efficiency of water treatment systems, and potentially lead to scale buildup in pipes and appliances (Fresh Water Systems, 2024).

Chemical parameters range from pH, acidity, alkalinity to presence of chloride, sulfate and metals as well as dissolved oxygen, BOD, chemical oxygen demand (COD), and any toxic substances. pH is a critical parameter of water quality, representing the measure of a solution's acidity or basicity. An excessive pH can be detrimental to the use of water, high pH can make the taste bitter and decrease the effectiveness of chlorine disinfection while low pH can corrode or dissolve metals and other substances. Pollution can also change the pH of water, which can harm aquatic animals and plants. Most aquatic animals and plants have adapted to a specific pH, and even a slight change can cause harm. The effects of pH on animals and plants are significant, low pH can decrease the number of hatched fish eggs, irritate fish and aquatic insect gills, and damage membranes, whereas water with a very low or high pH is fatal. Additionally, pH can also affect other chemicals in water, for example heavy metals such as cadmium, lead and chromium dissolve more easily in highly acidic water, and a change in pH can change the forms of some chemicals in the water, making them more toxic to aquatic organisms. Dissolved Oxygen (DO) is a crucial chemical parameter to be considered. It refers to the amount of oxygen dissolved in water and is essential for the survival of aquatic organisms. DO levels are influenced by factors such as water temperature, atmospheric pressure, and biological activity. In drinking water, healthy DO levels typically range from 6.5-8.0 mg/L (Sensorex, 2022). Low DO levels can lead to hypoxic conditions, stressing aquatic life and potentially causing fish kills. On the other hand, while high DO levels can improve water taste, they may also increase corrosion in water distribution systems. DO is considered a key indicator of water quality and ecosystem health (U.S. Environmental Protection Agency [EPA], 2025).

2.3 IoT AND WATER MANAGEMENT

The IoT allows for connecting devices through the internet, which can be beneficial for automating the distribution of water and monitoring for leaks. The main structure of the IoT includes three layers: the physical layer, where sensors gather data from the environment; the network layer, where the data is converted into digital streams for processing; and the application layer, which provides specific services to the user. Figure 2.1 displays a basic architecture of what an IoT system can comprise of. The data needs to be processed immediately when it is collected or stored in the cloud to prevent network congestion (Yasin, et al, 2021). Many sensors are available in electronic stores for water monitoring including fabricated sensors, capacitive sensors, Turbidity Sensor, Soil Moisture sensor, etc.

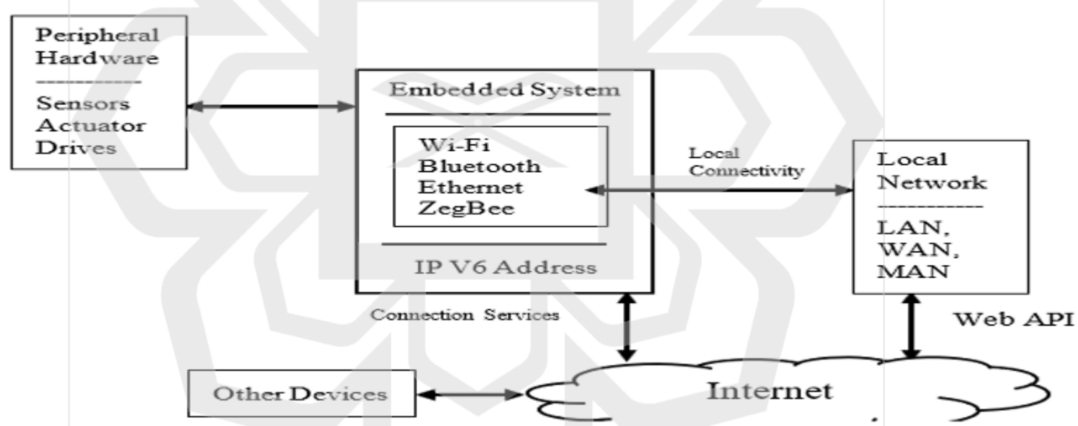


Figure 2.1: Basic Architecture of IoT (Yasin, et al. 2021)

In 2016, Geetha et al. proposed a solution using IoT technologies to monitor in-pipe water quality. The model tests water samples and analyzes data posted on the web to increase the accuracy of water measurements and identify deviations from predefined values. The system also includes a controller with a built-in WiFi module and a low-cost, simple intelligent water quality monitoring device that can measure pH, turbidity, and conductivity. The system also includes a warning system to alert users of any variations in water quality parameters. The experimental setup includes 5 parameters: conductivity, pH, turbidity, temperature, and water level, it was connected to the Ubidots network, and the results were compared to the WHO's guidelines for drinking water quality. While innovative in its approach to real-time monitoring, the system was limited to controlled environments and lacked scalability for open water bodies.

Gupta et al. (2018) proposed a device that can be constantly monitored by a mobile app from anywhere. The device can be automatically managed and provides complete intelligent automation. It is a stable, easy to install device with minimal dimensions. The project allows residents to track and control the water management system through their smartphones remotely. It also enables company managers to take appropriate actions if the water quantity drops below a certain threshold. The project uses an ultrasonic water depth indicator and a turbidity sensor. The sensor continuously monitors the water level and sends data to the company's inhabitants via the cloud, allowing the secretary to call for a water tanker if the water level drops. The turbidity sensor also tracks basic water quality features, which can be helpful in providing historical data on the relationship between quality and quantity. The system advanced the field by introducing user accessibility and automated management. However, their focus on water quantity over quality parameters revealed a significant gap in comprehensive water quality assessment.

Ranjan et al. (2020) proposed a concept for smart rainwater harvesting using IoT. The model includes a segregation system that divides the water into two tanks in a 60-40 percent ratio. The rainfall detection sensor is placed on top of the system to detect whether it is raining. The model was tested using two bottles of water, one with regular drinking water and the other with lemon added to make it mildly acidic. The system calculates the pH value of the rainwater and checks if it is higher than 5. If the pH is greater than 5, the servo motor on the hinge rotates in a clockwise direction, filling the tank on the right-hand side. If the pH is less than 5, the servo motor rotates in the opposite direction, filling the tank on the left-hand side. This separates the drinking water from the acidic water, isolating water with a pH of 5 or higher. All of this is achievable using the NodeMCU WiFi module. Overall, the smart rainwater harvesting system demonstrated the evolution of IoT applications in water management. Their pH-based segregation system showed innovation in automated water quality control, though the binary classification (pH > 5 or pH < 5) may be oversimplified for complex water quality assessment.

Hamid et al. (2020) proposed a smart water quality monitoring system (SWQMS) design for monitoring and evaluation water quality within swimming pool systems whererby factors influencing pH and temperature is analyzed . The system is using NodeMCU V3 processing unit with ESP8266 Wi-Fi module attached with pH sensor and temperature sensor (DS18B20) enabling constant monitoring of pH and temperature at real time. Monitoring is done via IoT cloud (Ubidots apps). The investigation was done as well to find the significant factor influencing pH and temperature particularly time of day which had no influence on pH but did on temperature. DOE and ANOVA were the statistical tools used however the scope relied on only pH and temperature for quality evaluation.

Overall, smart water management using IoT is a technology that aims to collect and analyze data on a city's water supply, pressure, and delivery to improve efficiency in water transportation and usage. It is becoming increasingly important as economic growth, climate change, and population growth affect the availability of water resources. Information and communication technology plays a significant role in addressing these challenges through various technologies that help reduce water waste, improve water quality, and manage water resources. There are multiple smart technologies available for water services, including exploration, technical methods, filtering and processing, and many others.

2.4 USE OF MACHINE LEARNING TECHNIQUES

Historically, water quality has been monitored through a three-step process which involves collecting water samples, conducting tests, and investigating the results. This process tends to be typically carried out by scientists manually, and it can be unreliable and doesn't provide advance warning about water quality. However, with the development of wireless sensor technology, some research has been done on using sensors placed in water to monitor its quality and send updates to farmers. Additionally, machine learning algorithms have also been used to analyze water quality (Pappu et al, 2017). The use of machine learning (ML) techniques has been a major development in the analysis of hydrological parameters. Recent research suggests that these methods can effectively model and predict water quality parameters.

Machine learning is a powerful approach for analyzing large data sets and identifying patterns or making predictions. The process of applying machine learning includes data acquisition, appropriate algorithm selection, model training, and model validation. The main area of interest tends to be the choice of algorithm is crucial, and the two main classes of machine learning technologies are supervised and unsupervised learning. Supervised learning uses labeled training data sets to deduce predictive functions and can be used for data classification and regression. Unsupervised learning is used to handle data without labels and is mainly used for classification and association mining. Reinforcement learning is another class of machine learning algorithms that focuses on a machine's ability to correctly answer unlearned problems but is not widely used in the field of water environment. Collecting data is a crucial step in creating machine learning models. Both continuous and intermittent water quality monitoring results can serve as benchmarks for managing water systems. Traditional environmental monitoring techniques are commonly used by government agencies, but they face practical challenges when it comes to in-situ monitoring. Remote sensing technologies, on the other hand, can provide real-time and large-scale water quality monitoring and can reveal the movement and distribution patterns of pollutants that are hard to detect with traditional methods. (Zhu et al, 2022). Figure 2.2 displays all the different algorithms and water quality and management-based application that have been studied prior.

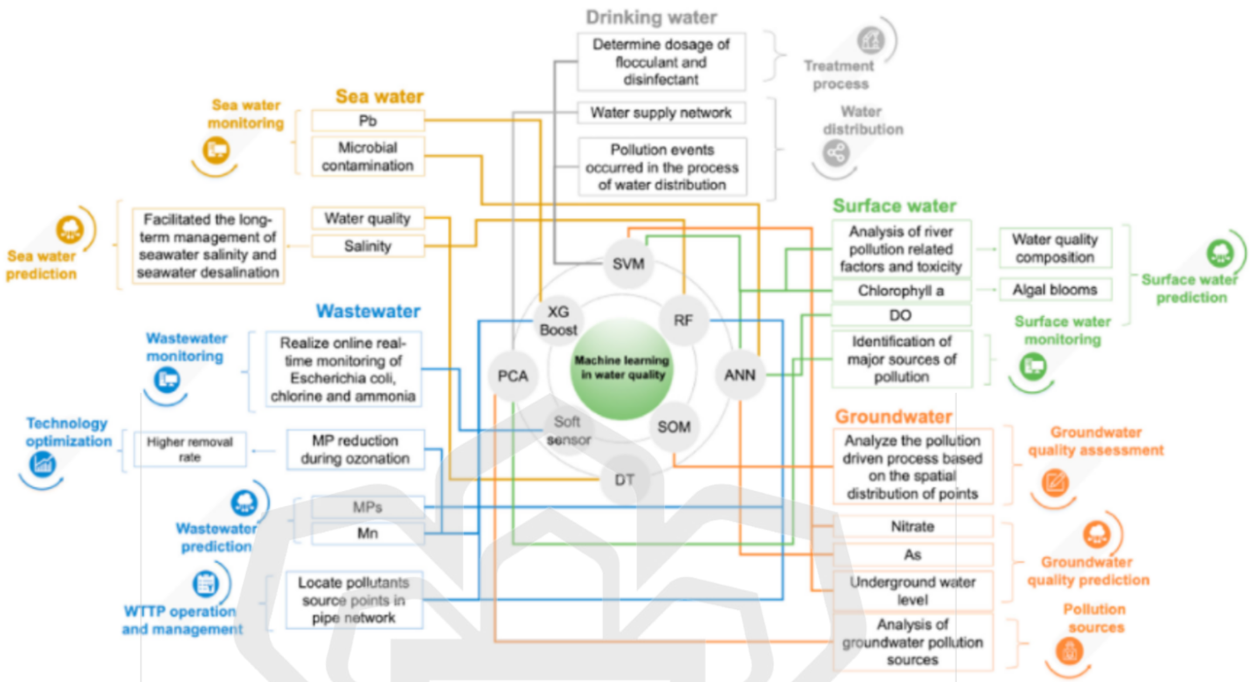


Figure 2.2: Applications of different machine learning algorithms in different water treatment and management systems. (Zhou et al., 2022)

Pappu et al. (2017), developed an IoT-based Water Quality Monitoring system that uses machine to machine communication. The system receives input from pH and TDS sensors, which is processed by an edge processor with a machine learning algorithm that predicts water quality based on a trained dataset. The predicted and trained data are stored in the cloud and can be accessed from a mobile phone. This system automates water quality monitoring for residential areas and does not require human involvement. The Intelligent IoT-based Water Quality Monitoring System used Arduino and Raspberry Pi3 as a microcontroller and processing unit, respectively. It also utilizes pH and TDS sensors to collect data on hydrogen ion concentration and total dissolved solvents in the water. The Arduino unit communicates with the Pi3 to transmit data for analysis using the K-Means Clustering machine learning algorithm. The paper however was limited in terms of only relying on 2 parameters (pH and TDS) and achieved only basic binary classification and was restricted to storage tank applications.

The K-Means Clustering algorithm was used to train a dataset on different types of water, including mud, lemon, salt, tap, and drinking water. The algorithm clusters the data based on similar PH and TDS values, so each cluster represents a specific type of water. When a new dataset is tested, the machine will classify it into the closest cluster, which represents the predicted type of water. The K-Means algorithm was used in machine learning to monitor and predict water quality. There are several advantages to using the K-Means algorithm over others:

It was computationally faster than hierarchical clustering when there are many variables and k is small:

- It produces tighter clusters than hierarchical clustering, especially for globular clusters.
- It was fast, robust, and easy to understand.
- It was relatively efficient with a complexity of O(tknd), where n is the number of objects, k is the number of clusters, d is the number of dimensions of each object, and t is the number of iterations. Typically, k, t, and d are much smaller than n.
- The K-Means Clustering algorithm is also iterative.

The IoT-based Water Quality Monitoring system, initially designed for storage tanks, demonstrated the potential for broader applications, including monitoring bodies of water such as ponds and rivers, as well as water pipes. It can also be extended to consider other water parameters besides pH and TDS, and to control the flow of water based on water quality. It is important to ensure data security and integrity when transmitting data for analysis, prediction, and controlling the valve of the water tank and storage area. The system's data flowchart is display in figure 2.3.

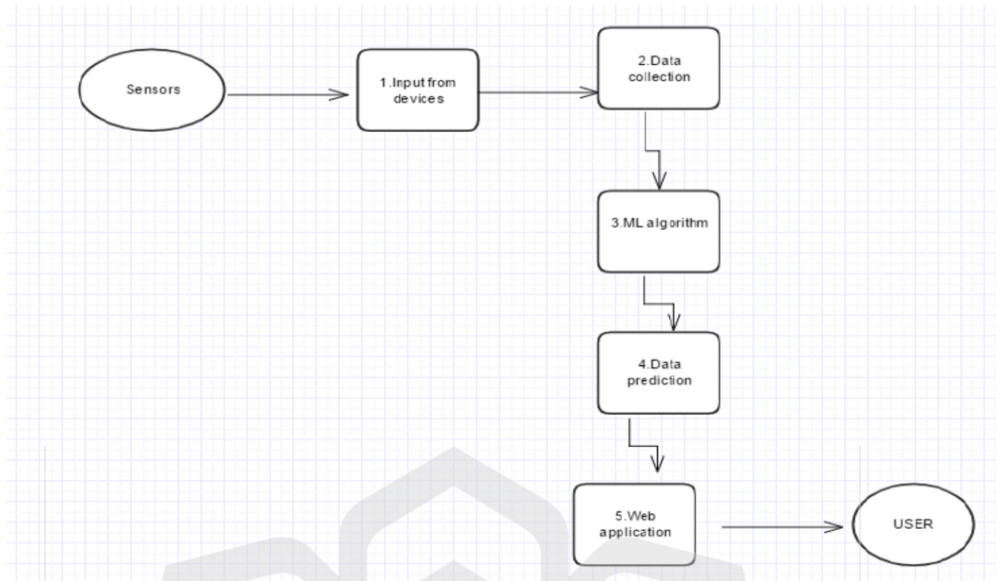


Figure 2.3: Data Flow Chart (Pappu et al, 2019)

Sagan et al. (2020) found that using machine learning based on experiments allowed for advanced optimization by combining sensor data from real-time monitoring and satellite data. The accuracy of models such as partial least squares regression, support vector regression, and deep neural network were higher than traditional models. However, certain water quality variables, like pathogen concentration, cannot be directly measured by remote sensing because they are not optically active or lack high-resolution hyperspectral data, but can be inferred using other measurable data. The key highlights are in the use of integrated satellite data with sensor reading. However, the system couldn't measure certain parameters directly and limited by optical data availability.

Shen et al. (2020) estimated the nitrogen and phosphorous concentration in rivers and streams by creating a geo-dataset estimate and map the concentrations of Nitrogen and Phosphorous in their various chemical forms at a spatial resolution of 30 arc-seconds (approximately 1 km) for the contiguous United States. The models were then built using Random Forest (RF), a machine learning algorithm that regresses the seasonally measured N and P concentrations collected at 62,495 stations across the United States streams for the period of 1994-2018 onto a set of 47 in-house built environmental variables that are available at a near-global extent. The seasonal models were validated through internal and external validation procedures and the predictive powers are measured by Pearson Coefficients, which on average, reach approximately 0.66. This study presented the elaboration and analysis possible by focusing on the facets of limited parameters, in this case being N and P concentration.

In 2015, Guo et al. developed and tested two machine learning models, Artificial Neural Network (ANN) and Support Vector Machines (SVM), for predicting the early 1-day interval total nitrogen (T-N) concentration of effluent in a wastewater treatment plant. The purpose of this study was to avoid the impact of high FW leachate T-N concentration loading to the wastewater treatment. Both daily water quality data and meteorological data were used as input parameters and a pattern search algorithm was utilized for model parameter optimization. Additionally, sensitivity analysis was conducted to determine the effectiveness of each input parameter using the Latin-Hypercube one-factor-at-a-time (LH-OAT) method. The study found that the SVM model performed better than the ANN model, but the ANN model provided more physical-related cause-and-effect relationships between the T-N concentration of effluent and other input parameters. The study concluded that machine learning models can be a reliable method for water quality prediction as an early warning water quality control of wastewater treatment. The paper suggested that long-term modeling for the input value sampling could be suggested to improve the accuracy of the ANN and SVM models. While, the system achieved reliable predictions for wastewater treatment, it was limited to single parameter (T-N) prediction and suggested need for long-term modeling.

Wu et al. (2020) focused on classifying water images into sub-categories of clean and polluted water, in order to provide instant feedback for a water pollution monitoring system that uses IoT technology to capture water images. Due to the low differences between classes and high differences within classes of captured water images, water image classification is a challenging task. The paper propose to use the ability of CNNs to extract highly distinguishable features to construct an attention neural network for classifying IoT captured water images. The proposed attention neural network encodes channel-wise and multi-layer properties to improve feature representation. It is also proposed a channel-wise attention gate structure and used it to construct a hierarchical attention neural network in local and global sense. The effectiveness of the proposed attention neural network was demonstrated through comparative experiments on a dataset of water surface images. The proposed neural network was applied as a key component of a water image-based pollution monitoring system, which helps users to monitor water pollution in real-time and take immediate action to address it. While the paper was innovative on using water image classification, the reliance on water image quality hinders especially for similar-looking water classes.

Adeleke et al (2023) aimed to develop and evaluate the efficiency of machine learning and the IoT at water storage stations. The researchers developed a system prototype and tested its performance using classifier and reliability matrices. The study considers physical and chemical parameters of water such as temperature, pH, turbidity, dissolved oxygen, total dissolved solids, oxidation-reduction potential, and electrical conductivity to evaluate the water pollutants present in drinking water. After analyzing the sensor data, the researchers used ANN and SVM machine learning algorithms to forecast the impurity level of the water measured. A water treatment method was also introduced to provide an automated corrective measure based on a specific amount of water contamination. The study collected water samples to create a training dataset for machine learning algorithms, and the performance of the ANN models was higher than that of the SVM models. The research concludes that AI and IoT are more efficient in remotely monitoring safe and harmful water conditions and an automated water treatment system offers considerable benefits in mitigating water pollution. The paper presented a strong case for multi-sensor combination for water quality analysis whilst limited to drinking water applications.

Gorgan-Mohammadi et al. (2023) explored the application of decision tree algorithms for predicting water quality parameters in Lake Erie. Their study compared four decision tree models: Classification and Regression Tree (CART), Chi-squared Automatic Interaction Detector (CHAID), Quick Unbiased Efficient Statistical Trees (QUEST), and C5, focusing on predicting dissolved oxygen and soluble phosphorus levels. Using a dataset of 327 samples, the CART model demonstrated superior performance in predicting actual values, with 214 cases for dissolved oxygen and 245 cases for phosphorus showing minimal deviation from observed data. The C5 model excelled in group classification, correctly predicting 256 phosphorus and 230 dissolved oxygen parameter groups. Their findings demonstrated that decision tree methods can effectively classify and predict water quality parameters with high accuracy and computational efficiency.

Study by Gai and Zhang (2023) introduced a differing approach to water quality prediction using a Momentum-optimized logistic regression algorithm (LRM). The study addressed limitations in traditional linear models by incorporating momentum to escape local optima during gradient descent. Testing on 4 datasets showed the LRM algorithm achieved a 1.11% improvement in prediction accuracy compared to existing methods. The model outperformed traditional algorithms like KNN in both convergence speed and steady-state error reduction.

Shadkani et al. (2021) evaluated three machine learning models for estimating suspended sediment load (SSL) in rivers: multi-layer perceptron (MLP), MLP with stochastic gradient descent (MLP-SGD), and gradient boosted tree (GBT). The research focused on two stations along the Mississippi River - St. Louis and Chester. The study used four evaluation criteria: Correlation Coefficient, Nash Sutcliffe Efficiency, Scatter Index, and Willmott's Index. The study concluded that MLP-SGD offers the most accurate model for SSL estimation, representing an advancement in sediment transport modeling techniques.

Illic et al. (2022) used Naïve Bayes algorithm for the classification of water quality within natural water sources in spanning 6 years in Vojvodina Province, Serbia.T he model incorporated nine water quality parameters: temperature, pH, electrical conductivity, oxygen saturation, biological oxygen demand, suspended solids, nitrogen oxides, orthophosphates, and ammonium. The model was trained on 48 samples and successfully predicted water quality classes in 64 out of 68 test cases, even with missing data.

Overall, machine learning has been widely used as a powerful tool in solving problems related to water environment such as predicting water quality, optimizing water resource allocation, and managing water shortages. However, several challenges still exist in fully applying machine learning techniques to evaluate water quality. One of the challenges is the dependence of machine learning on large amounts of high-quality data, which is often difficult to obtain due to cost or technology limitations in water treatment and management systems. Another challenge is the complexity of the conditions in real-world water treatment and management systems, which limits the applicability of current algorithms to specific systems.

2.5 RESEARCH GAP

While machine learning and IoT have emerged as vital tools for water quality monitoring, several key research gaps remain: There is a limited integration of comprehensive IoT sensor arrays with advanced ML algorithms for real-time river water quality monitoring, particularly in developing regions like Malaysia. Furthermore, while various projects have ventured and portrayed the viability of IoT, there is still a lack of studies comparing the performance and cost-effectiveness of low-cost IoT sensors to reference devices for water quality parameters. On the another aspect, most of ML application tends to be very situation and domain specific ,thus there is a need of research on scalable and adaptive ML models capable of handling the complex, non-linear relationships in water quality data across diverse environmental conditions.

This research aims to address these gaps by developing and implementing a cost-effective, integrated IoT-ML system for comprehensive river water quality monitoring in Malaysia. The study will focus on the Pusu River as a case study, with the goal of improving existing water and river management practices to mitigate environmental degradation caused by rapid development and urbanization.

CHAPTER THREE:

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter presents the methodology of our proposed water quality system by combination of IoT based sensor system and platform and classification algorithms and explains how it has been developed from the parameter sensor device set up to data collection to deployment and describes the overall setup that has been developed for testing of the system. Here, we introduce the data description, collection, cleaning and annotation process. We present how the data collection and dataset were done in two iterations in order to improve the system's classification capability and how the overall deployment and creation of the interface for the water monitoring system has been set in place. For this study, the problem involves river water quality monitoring; the amount of organic and inorganic pollutants released into the environment through waste and wastewater has significantly increased because of increased industrialization and human activity. These compounds, especially heavy metals, may be detrimental to the general public's health. Proper monitoring and management can be complicated due to the array of parameters to set up and consider for analysis of the water conditions in river. Thus, the research serves to be the solution for a feasible solution to the problem at hand.

3.2 RESEARCH METHODOLOGY

The initial phase of this research project involves a comprehensive literature review to establish a clear set of objectives. This phase will be accompanied by a meticulous selection process for both sensors and IoT modules, leveraging insights gained from the literature review. Concurrently, various machine learning models will undergo rigorous analysis utilizing pre-existing classified datasets.

Upon the completion of the precise sensor selection phase, the subsequent steps involve the establishment and configuration of an IoT framework. Leveraging technologies such as NodeRed Arduino components, the IoT infrastructure will be meticulously set up. A preliminary trial will commence to collect data, strategically conducted across diverse river profile points and timed intervals. Simultaneously, an exhaustive exploration for the most optimal ML model will ensue, employing the pre-existing classified dataset as a benchmark for testing and selection.

The third pivotal phase revolves around data refinement, ensuring the elimination of any extraneous or irrelevant information sourced during the collection process. This meticulous data cleaning procedure is crucial to eliminate potential inaccuracies and discrepancies in the final analysis. The overarching goal here is to facilitate precise feature extraction, laying the groundwork for a robust and reliable water quality analysis.

Additionally, during this stage, techniques for data cleaning and preprocessing will be rigorously applied to ensure the integrity and accuracy of the dataset. Techniques such as outlier detection, normalization, and imputation will be employed to enhance the dataset's quality and prepare it for subsequent analysis.

Throughout these meticulously orchestrated steps, the focus remains on not just data collection and sensor deployment but on the strategic utilization of cutting-edge technologies, ensuring a comprehensive and accurate assessment of water quality parameters for effective environmental management and conservation efforts.

The primary focus of this research can be summarized as follows:

1. Generate a water quality dataset consisting of 3 classes from potable water, river flowing water and puddle water with specific focus on River Pusu water bodies.

2. Develop and train a new algorithm to detect as well as classify the three different water classes.

3. Integrate the trained water quality detection model with the sensor collection interface on Node-Red.

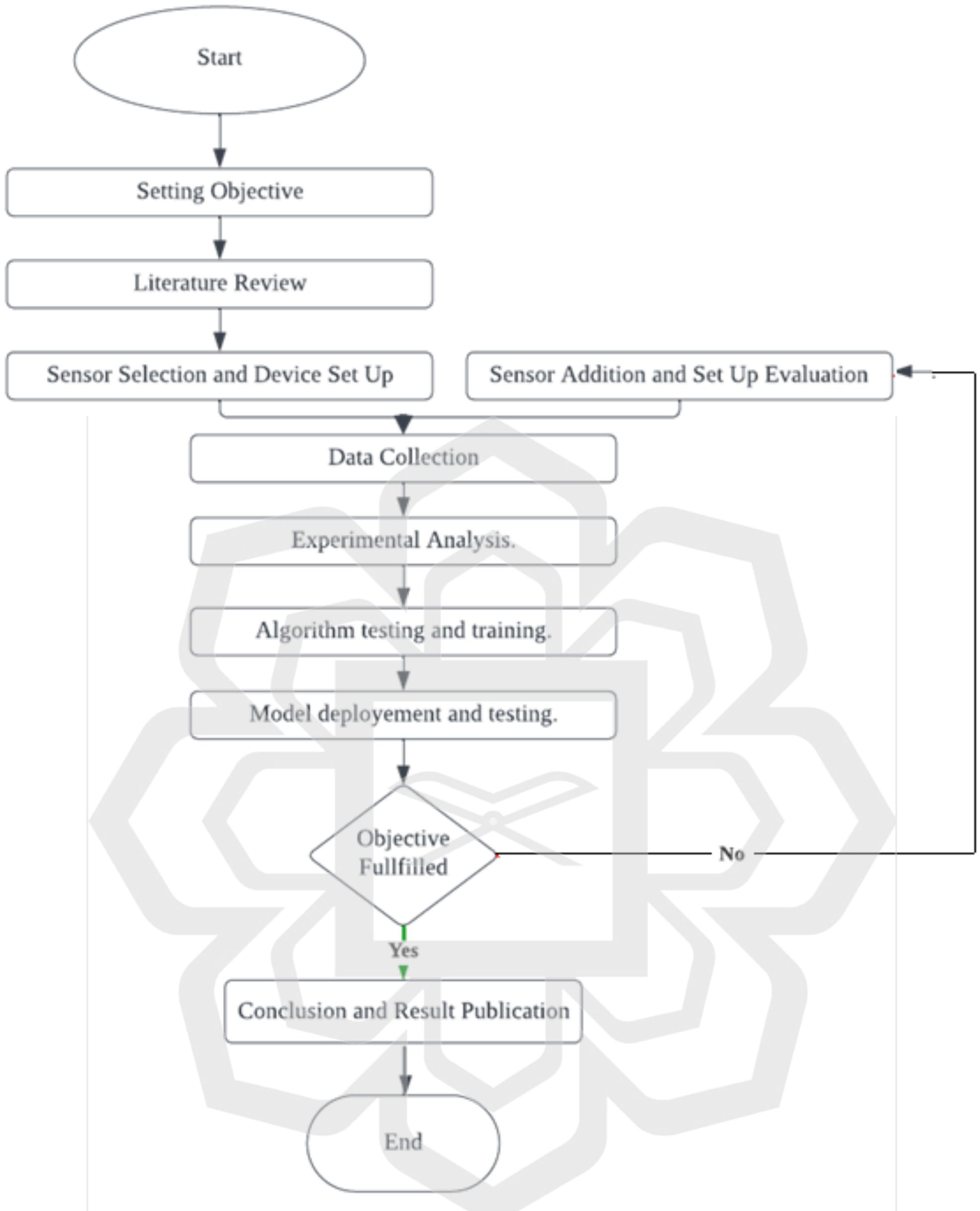


Figure 3.1: Research Flowchart.

3.3 SYSTEM ARCHITECTURE:

Before commencing any data collection, meticulous attention was directed towards devising a device structure primed to accommodate and scrutinize water quality with utmost precision. The foundation of this structure lay in the selection of a microcontroller, where the Arduino UNO ATmega328 emerged as the linchpin. This choice was driven by its cost efficiency and, notably, Arduino's inherent adaptability in seamlessly integrating diverse sensor systems, a pivotal factor in optimizing our data collection capabilities.

Our initial foray into data collection and experimentation hinged upon the incorporation of four core water parameter sensors. Each sensor was selected based commercial availability and cost-effectiveness, Arduino compatibility and integration capabilities and finally measurement accuracy within required ranges for water quality assessment. Figure 3.2 the experimental setup. Each sensor played a pivotal role in capturing specific facets of water quality:

1. Temperature Sensor: This sensor provided crucial insights into water temperature variations, a key determinant impacting aquatic life and biochemical processes.
2. pH Sensor: Vital for gauging the acidity or alkalinity of the water, the pH sensor offered invaluable data for assessing water suitability for diverse applications.
3. Turbidity Sensor: Addressing the clarity and cloudiness of water, the turbidity sensor served as a crucial indicator of particulate matter and sediment levels, pivotal for ecosystem health.
4. TDS (Total Dissolved Solids) Sensor: This sensor was instrumental in measuring the concentration of dissolved substances within the water, aiding in understanding its overall purity and suitability for various uses.

And for the second data collection and testing, we included:

5. Dissolved Oxygen Sensor: A critical component in evaluating water quality for sustaining aquatic life, the dissolved oxygen sensor provided insights into the amount of oxygen available in the water, vital for supporting aquatic organisms.

Table 3.1: Sensor Details.

Sensor	Brand/Product ID	Range of Measurement.	Accuracy
Temperature	DS18b20	-55 – 125 C	±0.5°C
pH	Gravity: Analog pH Sensor/Meter Kit V2 by DFRobots	0-14	±0.1@25℃
Turbidity	Gravity-Analog Turbidity Sensor For Arduino	0-4.5 V	N/A
TDS	SEN0244 by Dfrobot	0 to 1000 ppm	± 10% F.S. (25 ℃)
Dissolved Oxygen	Gravity: Analog Dissolved Oxygen Sensor by DFRobots	0-20 mg/L	N/A

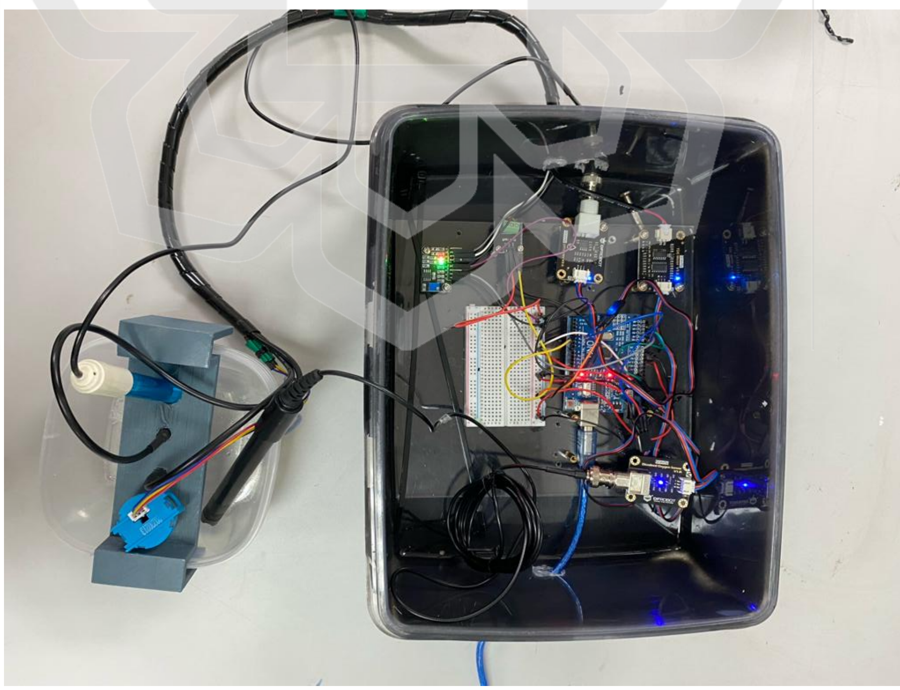


Figure 3.2: Device Complete Set-up with 5 parameters

3.3.1 NODE-RED IOT SETUP

To aid the data collection and storage, several IoT based tools have been procured. Node-RED nodes served as the building blocks for the graphical interface and the mediatory to allow for the data to be transferred over to Firebase servers.

Node-RED serves as an exceptional tool for constructing a server dedicated to water quality monitoring in IoT systems due to its versatile and intuitive visual programming interface. Its node-based architecture simplifies the process of integrating various sensors, IoT devices, and data streams, allowing seamless communication and data collection from diverse sources involved in water quality analysis. With its extensive library of pre-built nodes, Node-RED expedites the development of data collection workflows, enabling users to effortlessly configure data processing, aggregation, and analysis tasks. Its flexibility and support for numerous protocols make it adaptable to different sensor technologies used in water quality monitoring, ensuring compatibility and ease of integration. Moreover, Node-RED's web-based dashboard creation tools empower users to visualize collected data in real-time, facilitating efficient monitoring and decision-making for water quality management systems. Figure 3.3 displays the nodes and module connection required to

build the data collection mechanism using IoT.

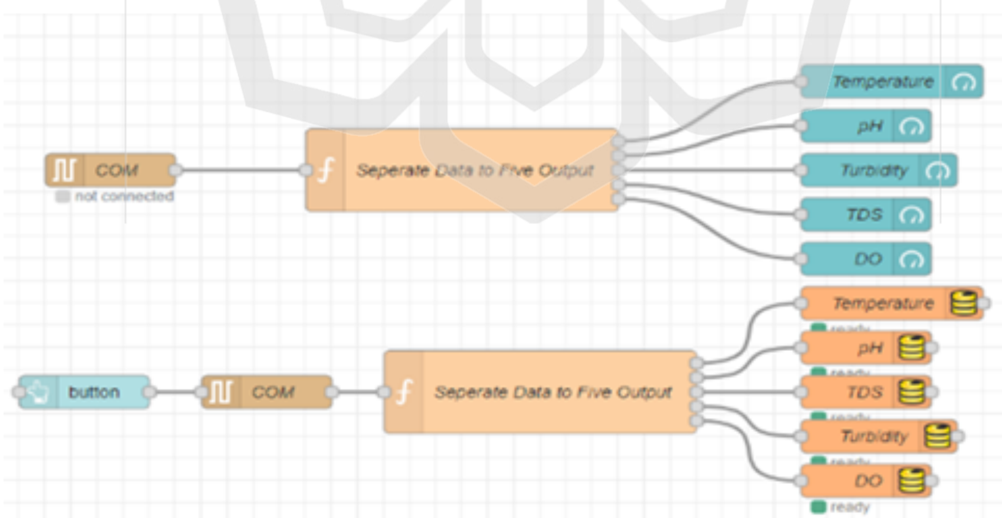


Figure 3.3: Basic Node-RED node structure for the 5 paramters data collection.

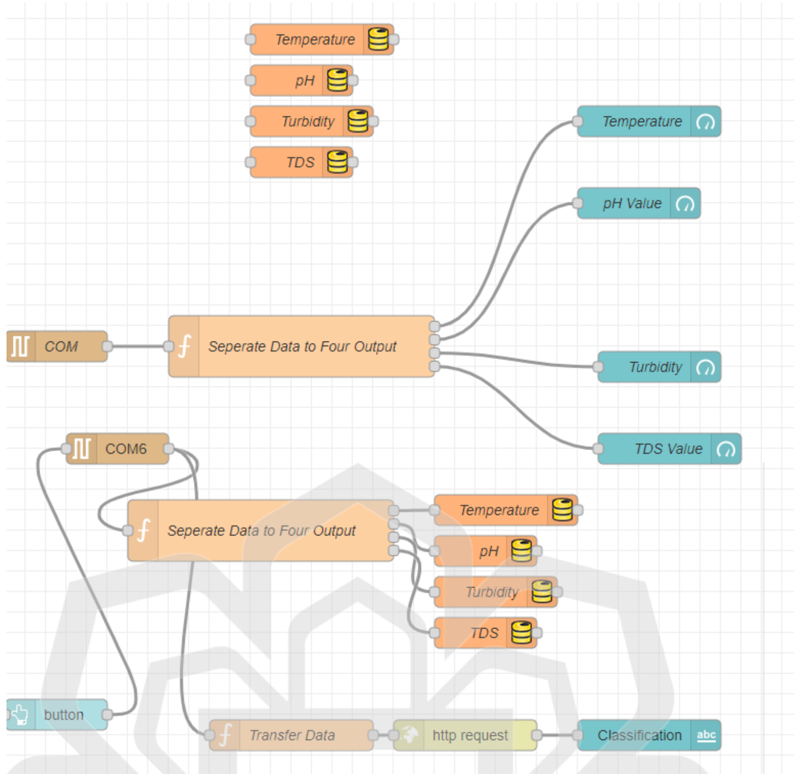


Figure 3.4: Set-up with the Machine Learning deployment and classification added for the 4-feature data collection.

Firebase was used as the data server host. Firebase is a robust tool for handling data as a server due to its array of powerful features and ease of use. Firstly, its real-time database functionality enables instantaneous updates and synchronization across multiple devices, which is highly beneficial for applications requiring live data feeds, such as IoT-based water quality monitoring systems. The real-time nature of Firebase ensures that changes made by sensors or devices are instantly reflected across the system, providing up-to-date information for analysis and decision-making. We attached an app script that allowed for data to be transferred over to a Google Sheet that would allow for easier extraction of data and data processing for machine learning aspects.

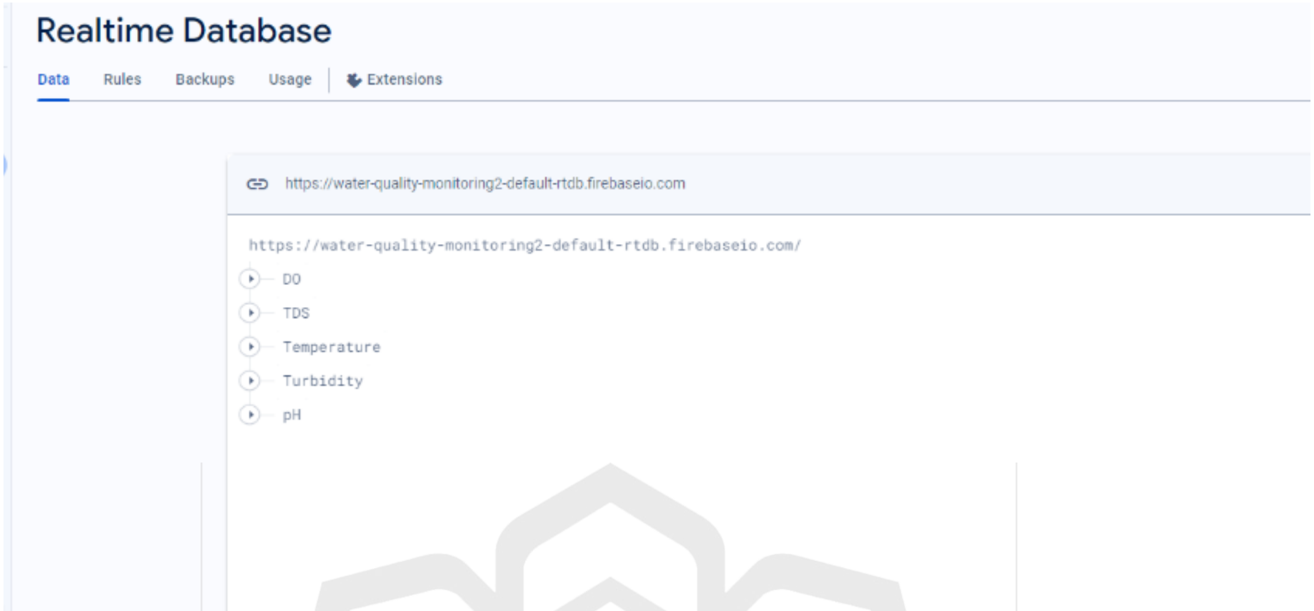


Figure 3.5: Firebase Server

3.4 DATA COLLECTION

Data collection involves the process of gathering crucial information about specific variables of interest crucial for your research purposes. It serves the purpose of training systems, examining hypotheses, and assessing results. Regardless of the academic domain, precise data collection holds utmost importance in preserving research integrity, acquiring precise analytics, and deriving meaningful conclusions.

3.4.1 DATASET 1 (4 PARAMETERS)

In our first data collection, we gathered water quality data for four key parameters: temperature, pH, TDS, and turbidity. The data was classified into three distinct categories based on water quality characteristics and common usage pattern:

1) Potable Water:

    a. Defined as water safe for human consumption:

    b. Sources included:

        i. Commercial bottled water (mineral, spring, sparkling, distilled)

        ii. Campus tap water from Wudhu area

        iii. Filtered water from drinking fountains

    c. This category represents the baseline for safe water quality parameters and serves as the control group for comparison.

2) Flowing River Water:

    a. Focused exclusively on Pusu River

    b. Sampling points strategically selected to capture:

        i. Areas of high human activity within campus

        ii. Upper stream locations with fast-flowing water

        iii. Lower stream points with reduced flow rates

        iv. Multiple collection points ensured comprehensive representation of river conditions. Figure 3.7 pinpoints the different water collection locations through the river.

3) Stagnant Water:

    a. Collected from disconnected water bodies.

    b. Sources included:

        i. Rain-formed puddles

        ii. Construction site water accumulation

        iii. Roadside depressions and potholes

        iv. Drainage pools

c. Figure 3.6 (b) shows an example of a puddle water collected.

This three-tier classification system was chosen because it:

- Represents distinct water quality conditions (potable, flowing, stagnant)
- Aligns with common water management categories (UNEP, 2024).
- Provides clear separation of water quality characteristics.
- Reflects real-world water quality monitoring needs.

The classification approach ensures comprehensive coverage of water quality variations while maintaining clear boundaries between categories for effective model training and validation.

For each sample, exactly 500 ml of water was collected in a 2-liter container. The sensors were then placed into this container to ensure consistent and accurate measurements across all samples. This standardized collection method helped minimize variability in sensor readings.

(a) (b) (c)

Figure 3.6: Data collection from: (a) a point from River Pusu lower stream (b) a puddle (c) a point from River Pusu upper stream.

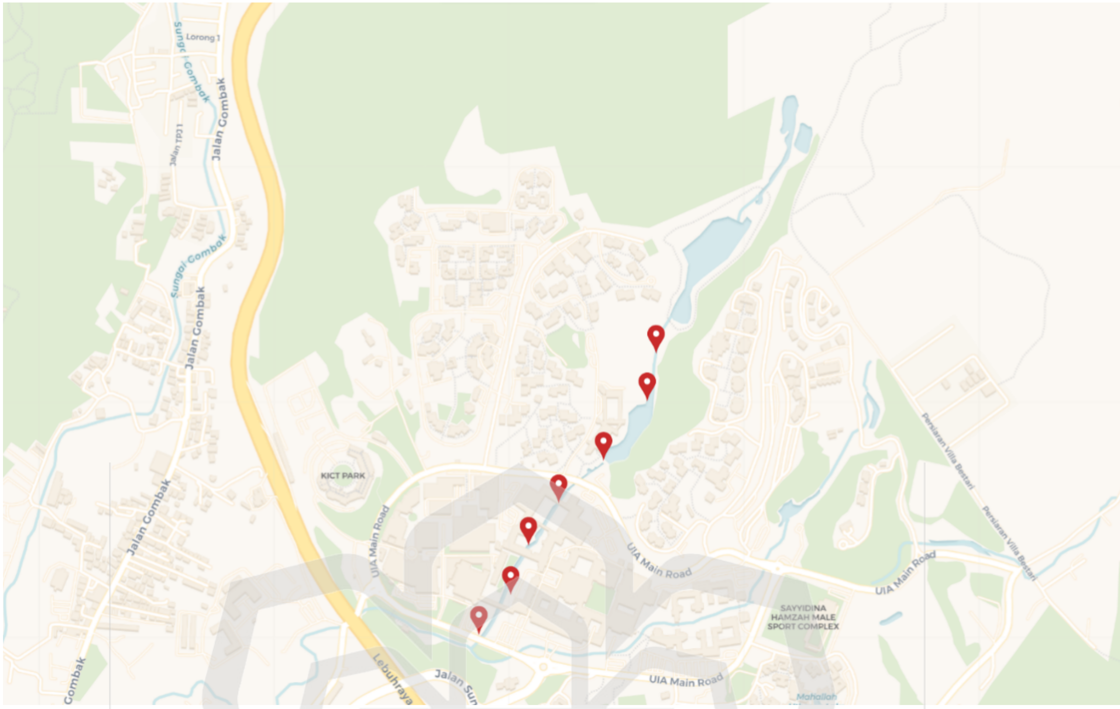


Figure 3.7: Focus points for data collection along the River Pusu that flowed by the IIUM campus.

3.4.1 DATASET 2 (5 PARAMETER)

Second dataset collection expanded upon the initial dataset by incorporating a dissolved oxygen sensor alongside the four existing parameters (pH, temperature, turbidity, and TDS). This addition was implemented to enhance the system's ability to differentiate between the three water quality classes, particularly given the overlapping characteristics observed in Dataset 1's analysis.

The dataset this time for river flowing water and puddle water only relied on the use of water within the River Pusu source. Potable water only relied on the use of water within bottled drinking water and filtered water.

This refined sampling strategy aimed to:

- Reduce data variability from multiple sources
- Improve classification accuracy through more standardized sample collection
- Create a more balanced dataset for training the machine learning models
- Establish clearer boundaries between water quality classes

The addition of dissolved oxygen as a parameter proved significant, as it showed a strong correlation (approximately 62%) with water potability, complementing the existing parameters and improving the overall classification accuracy.

3.5 MACHINE LEARNING

Given the classification-centric nature of this project, where the task revolves around categorizing observations into three distinct water quality classes, a thoughtful selection of multinomial classification algorithms was undertaken. The choice of these specific algorithms was guided by recommendations from prior studies and their prevalent use within the literature of water quality monitoring applications.

The chosen algorithms underwent rigorous testing and evaluation to ascertain their effectiveness in handling this specific classification task. The array of algorithms included:

1.Random Forest Classifier: Renowned for its ensemble learning technique, this algorithm harnesses the strength of multiple decision trees, making it robust and well-suited for complex classification tasks. Selected based on Shen et al. (2020)'s successful application in predicting nitrogen and phosphorous concentrations across 62,495 stations.

2. KNN (K-Nearest Neighbors): Employing a proximity-based approach, where classification is determined by the similarity to neighboring data points, KNN offers

simplicity and effectiveness, often used as a baseline in classification tasks. Demonstrated effectiveness in Gai and Zhang (2023)'s comparative study for agricultural water quality prediction.

3. Decision Tree: Characterized by a tree-like structure, Decision Trees excel in interpreting and representing data relationships, providing clear decision paths based on feature. Chosen based on Gorgan-Mohammadi et al. (2023)'s successful implementation with CART showing superior performance in Lake Erie studies.

4. Gradient Boosting Classifier: Known for its sequential training of weak learners, gradually improving predictive performance, Gradient Boosting is adept at handling complex relationships within data. Proven effective in Shadkani et al. (2021)'s study of suspended sediment load estimation.

5. Naïve Bayes Classifier: Based on Bayes' theorem, this algorithm assumes independence among features, showcasing efficiency in processing large datasets while maintaining reasonable accuracy levels. Chosen based on Illic et al. (2022)'s successful implementation with 94% accuracy in natural water source classification.

6. Logistic Regression: Despite its name, Logistic Regression serves as a reliable classification algorithm, particularly suitable for binary classification tasks, but adaptable for multinomial classifications as well. Demonstrated effectiveness in Gai and Zhang (2023)'s comparative study for agricultural water quality prediction.

7. Support Vector Machine (Polynomial): Leveraging a kernel-based approach, the Support Vector Machine with a polynomial kernel aims to identify optimal decision boundaries, effective in capturing complex relationships among features. Selected based on Guo et al.

(2015)'s comparative study showing superior performance in wastewater treatment prediction.

8. Multi-Layer Perceptron Classifier: A form of artificial neural network, the Multi-Layer Perceptron is capable of learning complex patterns in data, offering adaptability and versatility in handling intricate classification tasks. Demonstrated high accuracy in Shadkani et al. (2021)'s river monitoring application

The goal was to identify the algorithm or combination of algorithms that best suit the intricacies and nuances of this specific water quality monitoring application. These models were selected based on key criteria from the literature: Proven accuracy in water quality classification and the ability to handle multiple water quality parameters. Thus, the selection ensures proper evaluation across different algorithmic approaches, addressing both simple and complex classification scenarios in water quality monitoring.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter presents the key and iterative results obtained to evaluate the performance of a water quality monitoring system using Internet of Things for data collection and Machine learning models for data analysis and classification. The analysis covers experiments that varied in the number of features collected, evaluated under both individual and comparative metrics. Through examination of the varying parameters, the study provides insights into the general efficacy and robustness of the IoT and ML framework. This comprehensive approach helps determine whether the system can perform effectively as an optimal mechanism for water quality monitoring.

4.2 TRAINING DATASET 1

The first dataset was collected using four sensor parameters: pH, temperature, turbidity, and total dissolved solids (TDS). Initially, 877 data entries were gathered across three water quality classes. A rigorous data cleaning process was implemented to ensure data quality and reliability, which included:

- Removal of duplicate entries
- Elimination of null values
- Filtering of out-of-range measurements based on sensor specifications
- Removal of faulty readings due to sensor malfunction or calibration issues.

After the data preprocessing, the final dataset comprised 526 valid entries distributed across three distinct water quality classes: potable water, flowing river water, and stagnant puddle water.

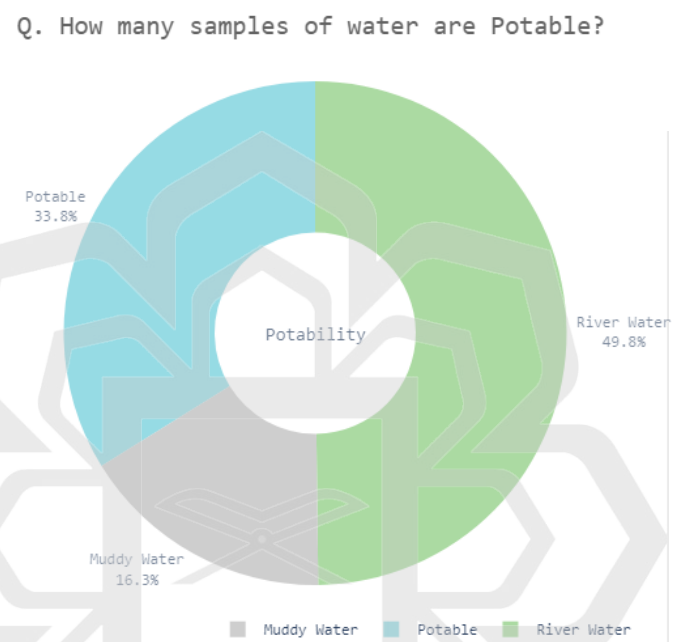


Figure 4.1 The distribution of the dataset

Figure 4.1 displays the distribution of the dataset by the potability class where there is certain level of class imbalance as river water represented over 45%, whereas potable water represented 33.8% (mild imbalance) and muddy water represented 16.3% (moderate imbalance This disproportionate representation stems from the study's primary focus on river water sampling, particularly from River Pusu. The class imbalance poses potential challenges for model training and evaluation, as the overrepresentation of river water samples could lead to biased predictions and affect the model's ability to accurately classify minority classes.

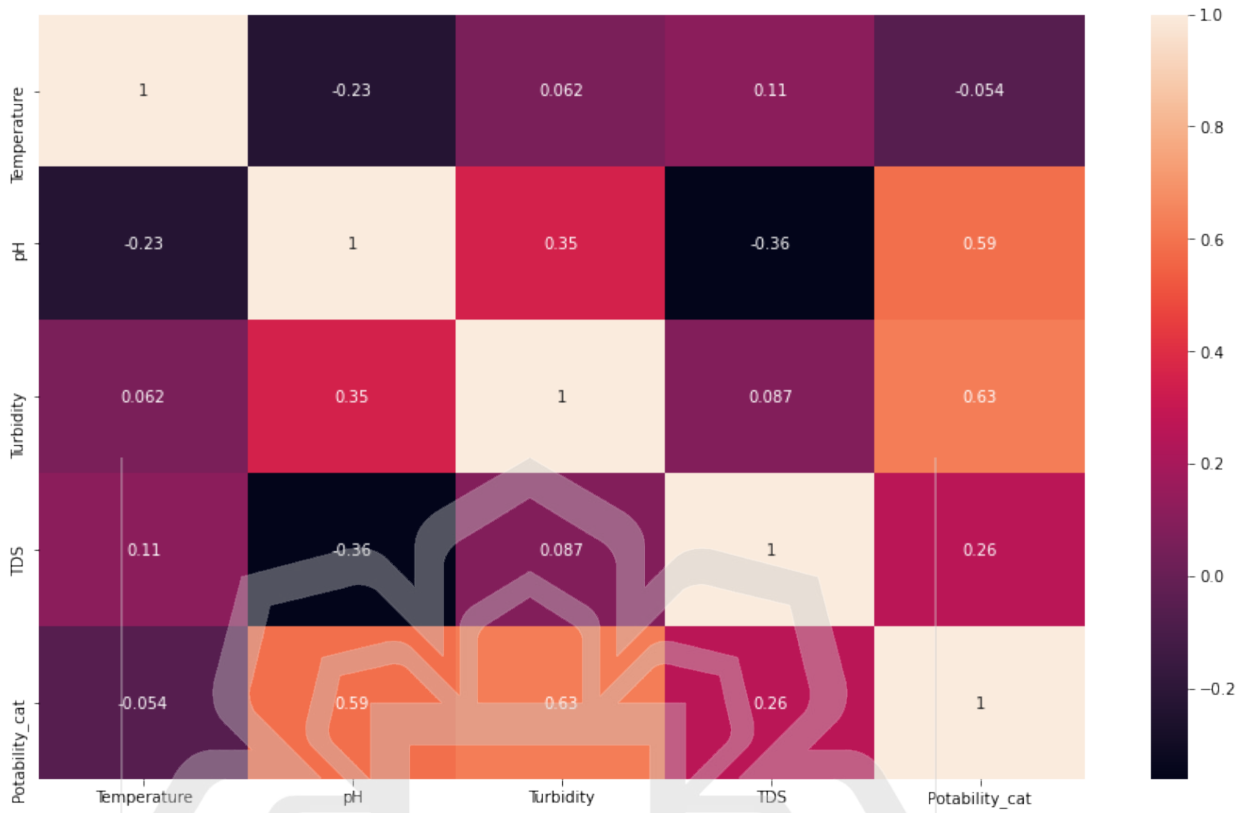


Figure 4.2 Correlation Matrix for dataset 3.

Feature selection is a crucial process in machine learning that involves choosing the most relevant and informative parameters (features) while excluding redundant or less useful ones. In the context of water quality monitoring, feature selection helps optimize model performance and computational efficiency. Figure 4.2 display the correlation matrix of each parameter to one another as well as to the classification of the water quality classes themselves. The correlation matrix offers insights into the influence of each parameter on water quality. An important observation is the lack of correlation between the temperature sensor readings and other features, as well as potability itself. This can be attributed to the multifactorial nature of temperature, which varies significantly based on factors such as collection point and time of day, irrespective of water potability. Consequently, the temperature feature was excluded from model training and testing due to its limited predictive value.

Conversely, both turbidity and pH exhibit a strong overall correlation of approximately 50% with potability, indicating their significance in classification tasks. However, total dissolved solids (TDS) show a comparatively lower correlation of 23% with potability. Furthermore, there is minimal correlation observed between each feature and other features in the dataset. These findings underscore the unique contribution of each parameter in assessing water quality and highlight the importance of feature selection in model development.

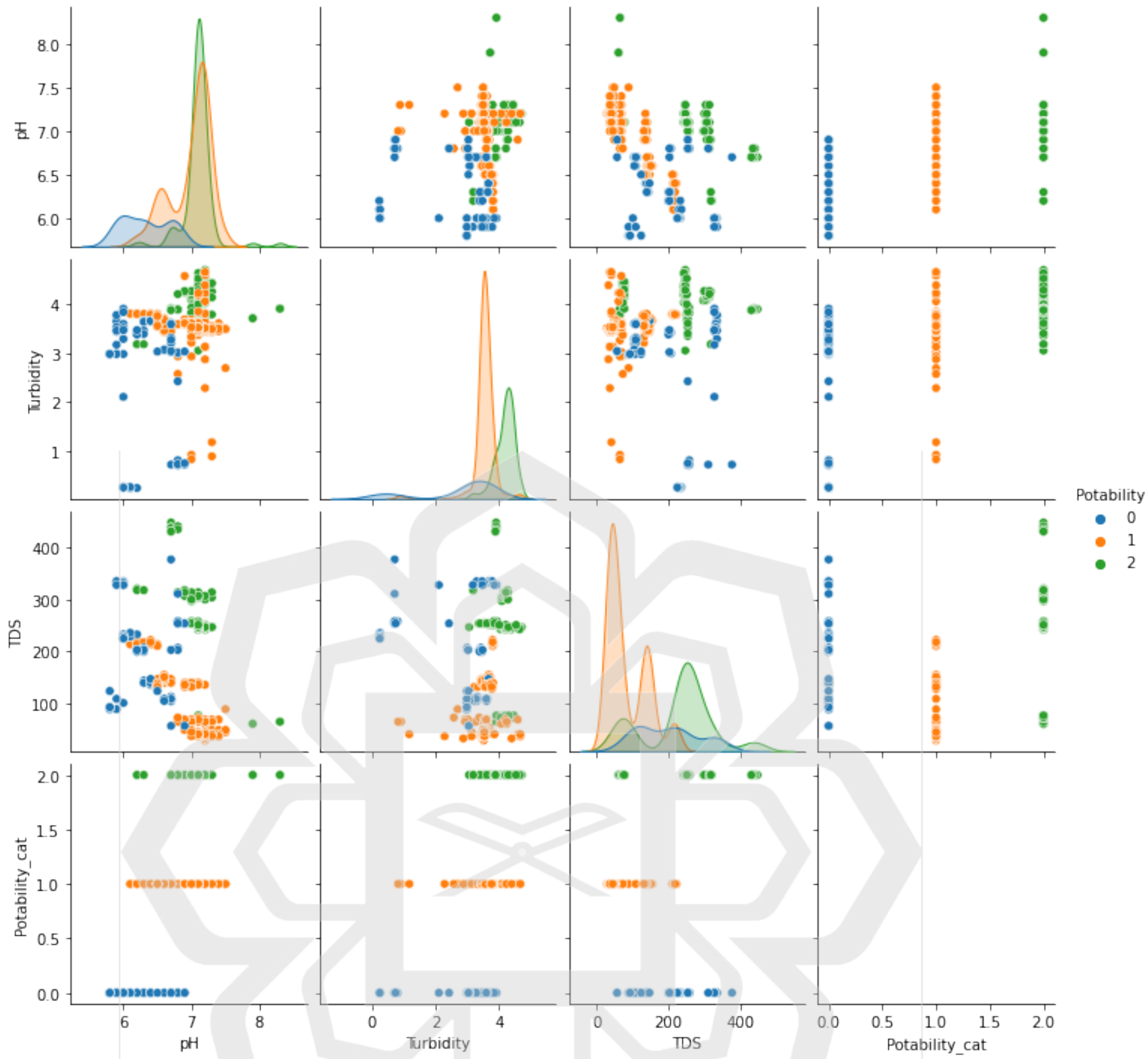


Figure 4.3: Scatter Plot of each parameter against another.

Figure 4.2 displays a scatter plot matrix reveals important patterns in the distribution of water quality parameters across different classes. The diagonal displays density distributions for each parameter, while off-diagonal plots reveal pairwise relationships between parameters, with colors distinguishing between muddy water (0), river water (1), and potable water (2).The figure highlights overall how the scatter plot of the distribution of the classes are scattered in each feature in order to understand the class overlap and allow for the models to better distinguish each class.

The pH versus Turbidity relationship demonstrates the clearest class separation, with potable water forming tight clusters between pH 6.5-7.5 and low turbidity values. River water creates a compact cluster with moderate pH and turbidity levels, while muddy water exhibits high variability in both parameters. TDS relationships show distinct banding patterns across classes, with potable water maintaining consistent TDS values, river water showing moderate TDS with some overlap, and muddy water displaying the highest variability.

Each class exhibits distinct characteristics: potable water forms tight clusters across parameters with consistent pH ranges (6.5-7.5) and low turbidity values; river water creates compact clusters with intermediate values for most parameters, showing some overlap with both potable and muddy water; and muddy water demonstrates the highest variability across all parameters, forming dispersed clusters with significant overlap with river water in some parameter combinations.

4.2.1 MODEL PERFORMANCE FOR DATASET 1

In order to evaluate the performance of the selected algorithms, we need to look at how the metrics of evaluation including:

1. Classification Accuracy: The concept of classification accuracy involves determining the proportion of correctly predicted instances among the total input samples, providing a straightforward measure of model effectiveness. The model accuracy score is the most comprehensive evaluation of its performance.
2. Confusion Matrix: The confusion matrix offers a detailed breakdown of the model's performance. In binary classification scenarios, it categorizes predictions into true

positives, true negatives, false positives, and false negatives, providing a granular view of the model's predictive capabilities. From this matrix, accuracy can be derived, offering deeper insights into the model's predictive accuracy.

3. Precision, Recall and F1 score: Precision quantifies the ratio of true positives to all instances predicted as positive, serving as a gauge of the accuracy of positive identifications. It's computed by dividing true positives by the sum of true positives and false positives. Recall, on the other hand, measures the proportion of correctly identified positive instances out of all actual positive instances, indicating the model's ability to capture true positives. It's calculated by dividing true positives by the sum of true positives and false negatives. The F1 score is a metric used to evaluate the performance of a classification model. It combines both precision and recall into a single value, providing a balance between these two metrics. The F1 score is calculated as the harmonic mean of precision and recall, and it is particularly useful when there is an imbalance between the classes in the dataset.
4. Decision Boundary: A decision boundary serves as a delineating line or surface that segregates distinct classes or categories within a classification problem. It not only defines the boundaries between classes but also offers insights into the model's classification accuracy. A fuzzy or indistinct boundary might indicate suboptimal modeling, while an overly intricate boundary could signal overfitting, where the model excessively tailors to the training data.

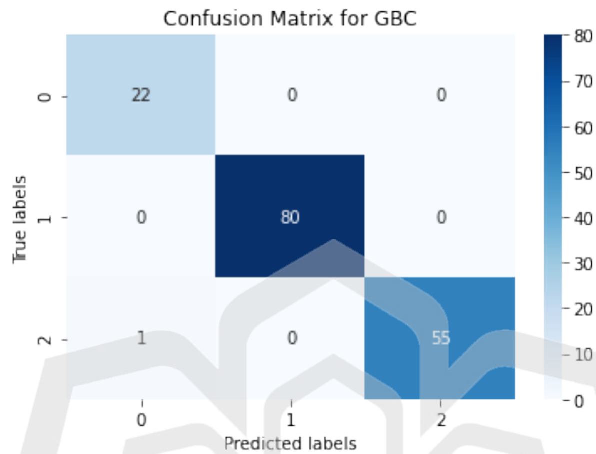
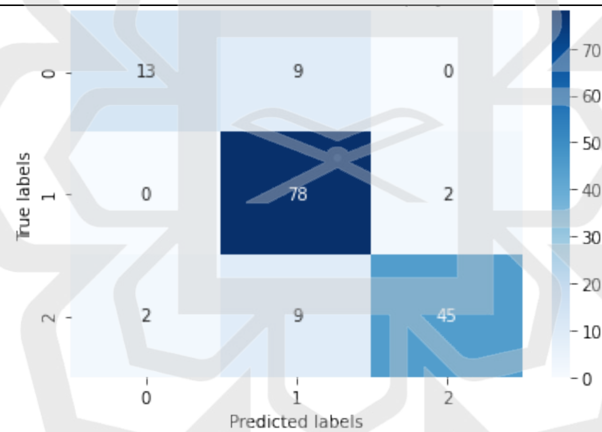
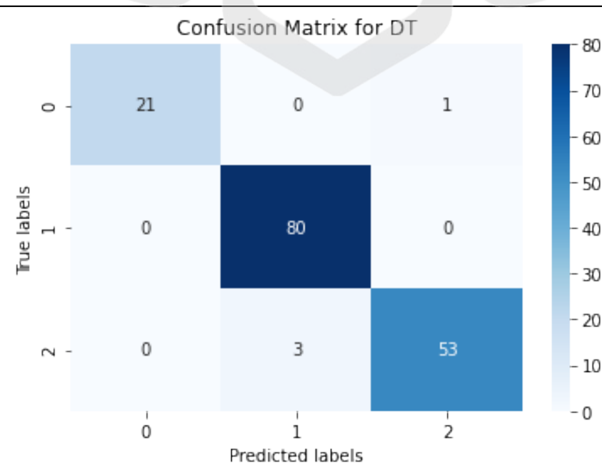
Table 4.1: Accuracy Scores Across Models for dataset 1.

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

Overall the datasets performed above 85% accuracy over all the models, which can be attributed to the size of the dataset and existing clusters amongst the classes. Gradient Boosting performed the best in terms with 99.3% accuracy and similarly other ensemble models like random forests and along with non-linear models like decision trees and KNN, tend to perform exceptionally well in classifying water potability across multiple classes. Linear models such as logistic regression and support vector machines struggled more with the non-linear relationships in the data, potentially impacting its classification accuracy in this multiclass scenario.

For next comparison, we only considered models with accuracy score of over 95%.

Table 4.2: Confusion Matrix for Top 5 models in Dataset 1

Model	Confusion Matrix	Performance
Gradient Boosting		The model performs well overall with minimal misclassifications.
Random Forest		The model shows high accuracy with no misclassifications in classes 0 and 1, but there are some misclassifications in class 2.
Decision Tree		The model performs well overall with a few misclassifications in classes 0 and 2.

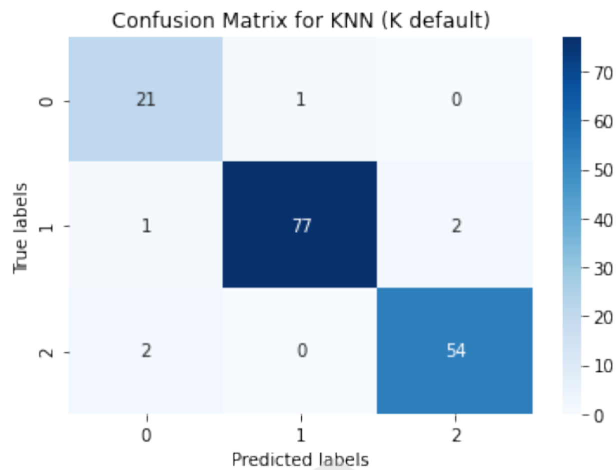
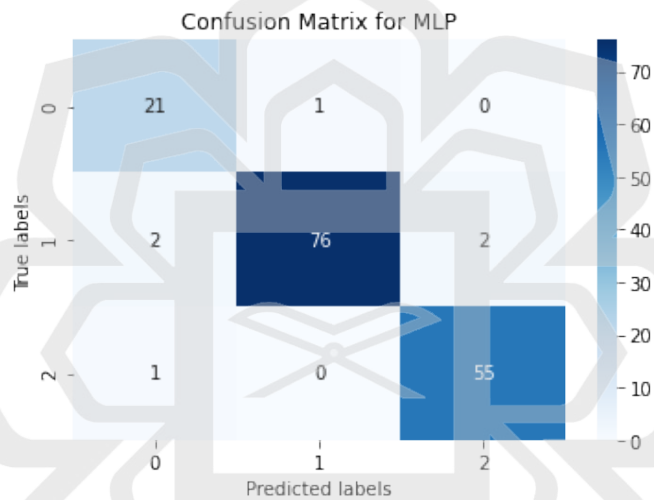
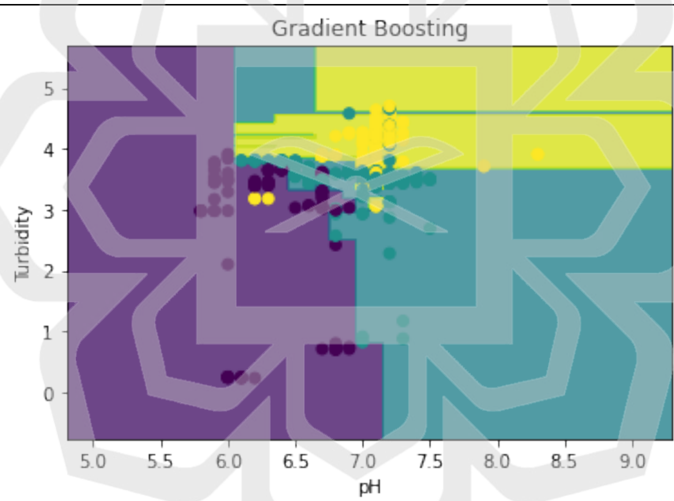
<table>
<tr><td>KNN</td><td>
</td><td>Like RF and GB, the model performs well overall, with a few misclassifications in classes 1 and 2.</td></tr>
<tr><td>Multilayer Perceptron</td><td>
</td><td>The model demonstrates good performance with some misclassifications in classes 1 and 2.</td></tr>
</table>

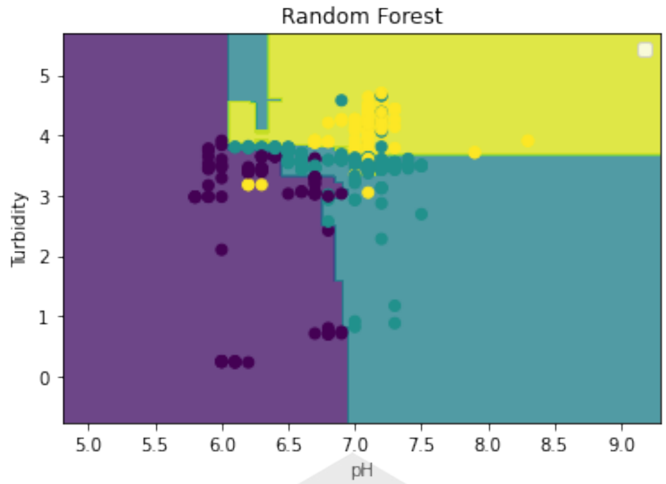
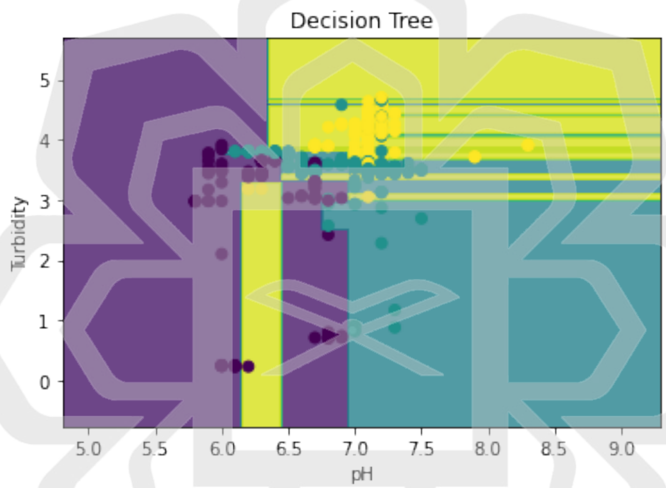
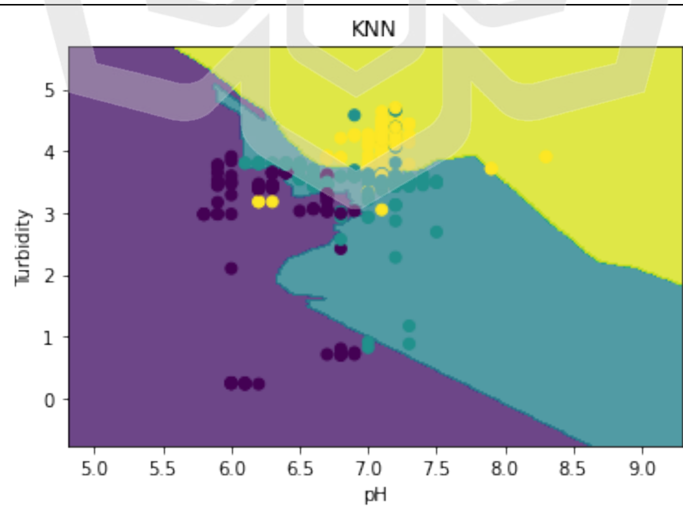
Table 4.3: Average Precision, Recall and F1 Score for Top 5 models in Dataset 1.

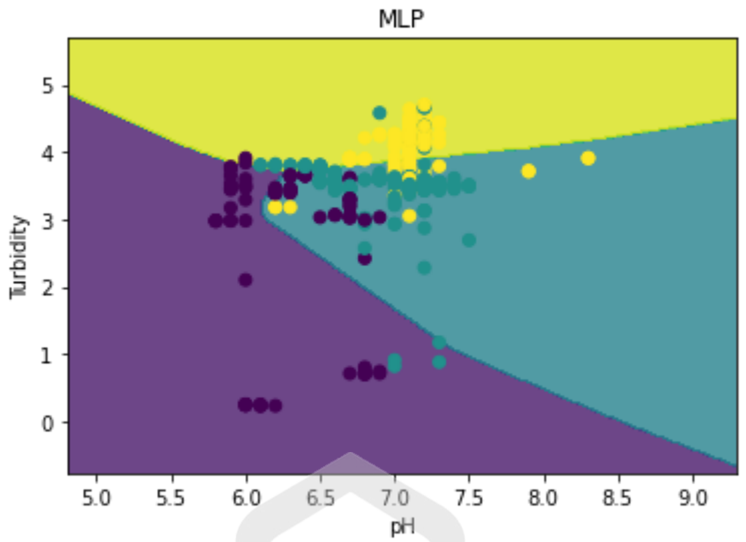
	Precision	Recall	F1 Score
Gradient Boosting	0.9866666667	0.9933333333	0.99
Random Forest	0.9666666667	0.9766666667	0.97
Decision Tree	0.9766666667	0.985	0.98
KNN	0.9766666667	0.985	0.98
MLP	0.9733333333	0.9822222222	0.9766666667

Overall, all models exhibit high precision, recall, and F1 score, indicating their effectiveness in correctly classifying instances across multiple classes. Gradient Boosting achieves the highest scores, followed closely by Decision Tree, KNN, and MLP, while Random Forest performs slightly lower but still demonstrates strong performance. These metrics provide valuable insights into the models' classification capabilities and can guide the selection of the most suitable model for the specific classification task.

Table 4.4: Decision boundaries for Top 5 models for dataset 1.

Model	Decision Boundary	Performance
Gradient Boosting		The boundary has several instances of being overfitted.

Random Forest		Overall, a general and defined decision boundary without much overfitting.
Decision Tree		Very overfitted.
KNN		Well defined with some areas of overfitting.

Multilayer Perceptron		Very generalized.

The decision boundaries enable us to visualize how models delineate and separate different classes concurrently. Both Random Forest and KNN demonstrate the most accurate boundary definitions, effectively avoiding overfitting to individual data points. For instance, the scatter plot of turbidity against pH was chosen due to the strong correlation between these features and water potability classifications.

Overall, taking into account all the matters and evaluation metric, random forest algorithms and it's trained model was selected for deployment.

4.3 TRAINING DATASET 2

The second training dataset was developed as an address to the limitations identified in the initial dataset. This dataset incorporated a dissolved oxygen sensor alongside the original four parameters (pH, turbidity, TDS, and temperature), expanding the system's capability to detect water quality variations. The data collection methodology was refined to ensure statistical robustness, with approximately 350 data points gathered using a balanced sampling approach across all three water quality classes. To minimize sensor-based fluctuations and enhance measurement reliability, each data point represents the average of three consecutive readings taken from the same sample within a consistent time frame. This averaging technique effectively reduced random measurement errors and improved the overall data quality, providing a more stable foundation for machine learning model training and validation.

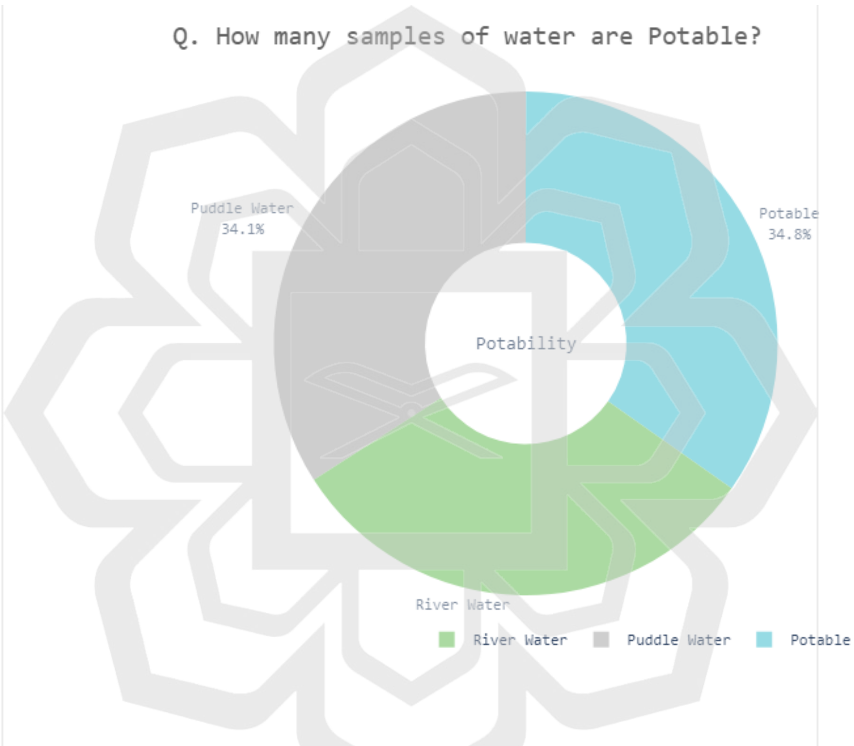


Figure. 4.4: Class Distribution of Dataset 2

In Figure 4.4, the distribution of the dataset by potability class is illustrated. Notably, the dataset exhibits a balanced class distribution ratio of 1:1:1, ensuring each class is represented equally. This balanced representation facilitates more accurate modeling, as challenges associated with class imbalance are mitigated in this scenario.

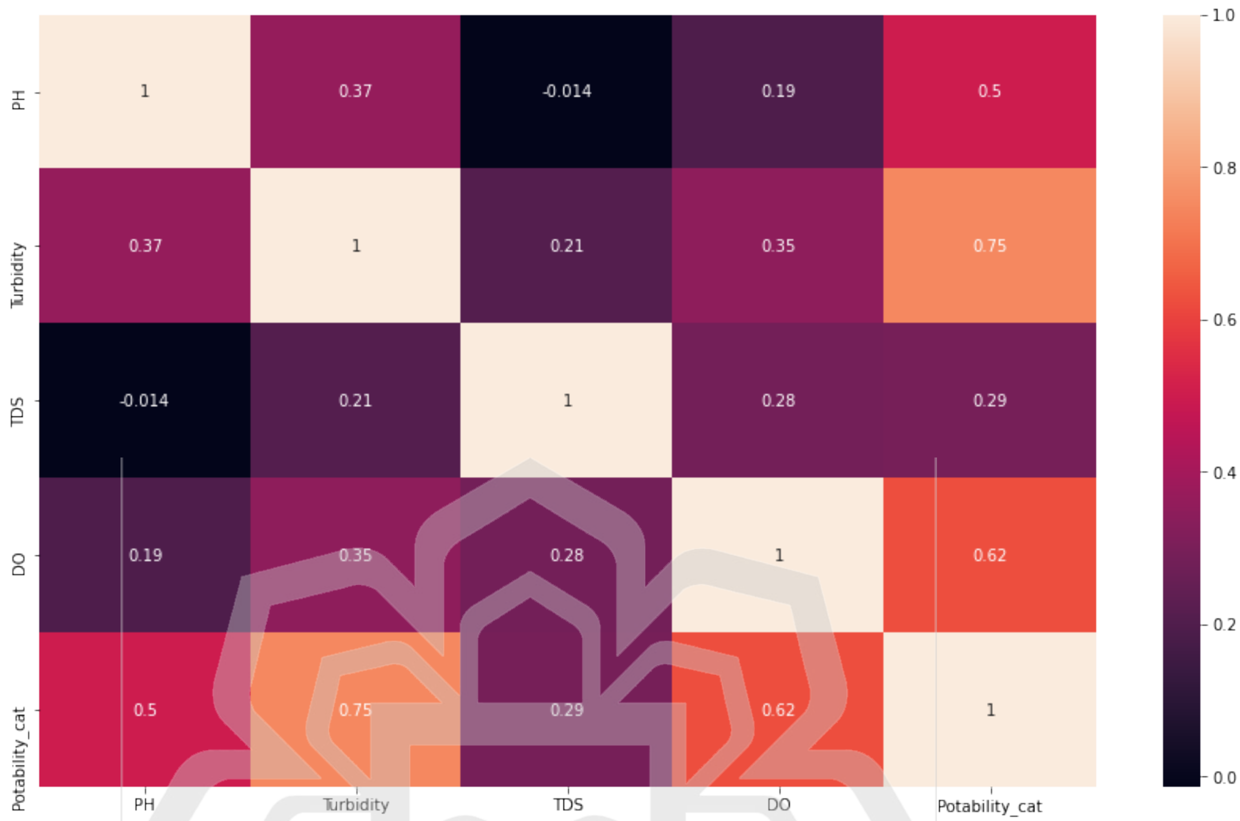


Figure 4.5 Correlation Matrix for dataset 2.

Once again, the correlation matrix underscores the potential significance of various parameters in determining water quality. Notably, pH and turbidity exhibit robust correlations with potability, standing at 50% and 75%, respectively. Additionally, the Dissolved Oxygen (DO) sensor demonstrates a notable correlation of approximately 62% with potability, further enriching the dataset's predictive capabilities. However, Total Dissolved Solids (TDS) maintain a relatively lower correlation at 29%. The inclusion of the DO sensor enhances the dataset by providing three features with significant correlations, thereby contributing to a more comprehensive understanding and prediction of water quality. Consistent with previous experiments, temperature shows negligible correlation and is consequently excluded from further analysis.

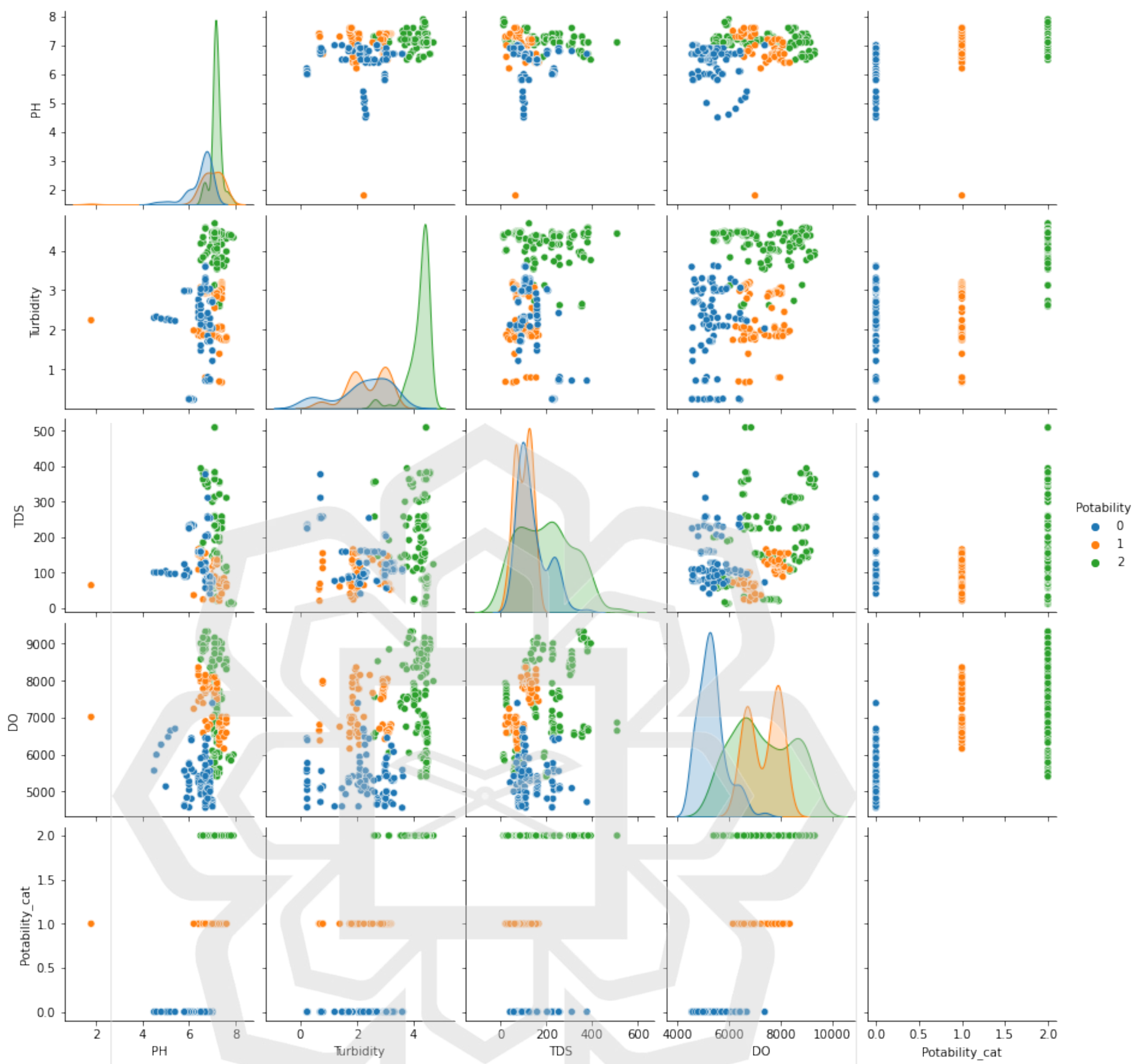


Figure 4.6: Scatter Plot of each feature against another.

Figure 4.6 highlights the key clusters each class can exhibit within the scatter plots. Most notable being the turbidity against dissolved oxygen sensor reading. With both sensor readings giving high correlation with potability, the sensors readings are as result providing the clearest region of clusters for each class with the minimal overlap between clusters. Furthermore, there can be seen certain overlap of clusters in other plots such as those for pH against DO and so on however this emphasizes the importance of inclusion of all these features in order to best extract a well learnt model.

4.3.1 MODEL PERFORMANCE FOR DATASET 2

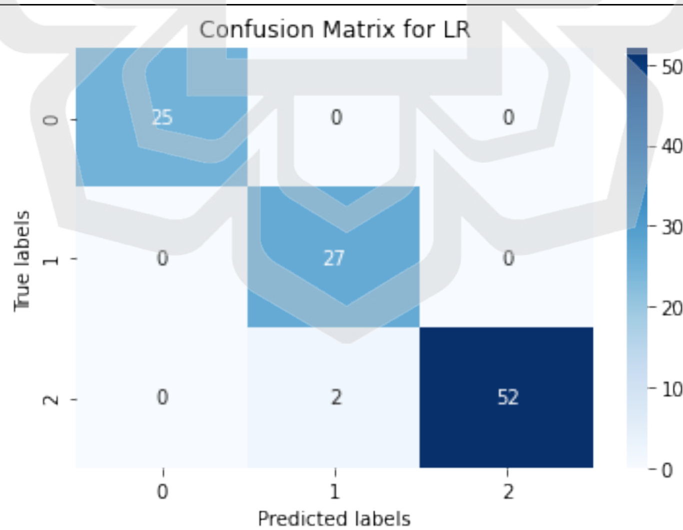
Table 4.4: Accuracy Scores Across Models for dataset 2.

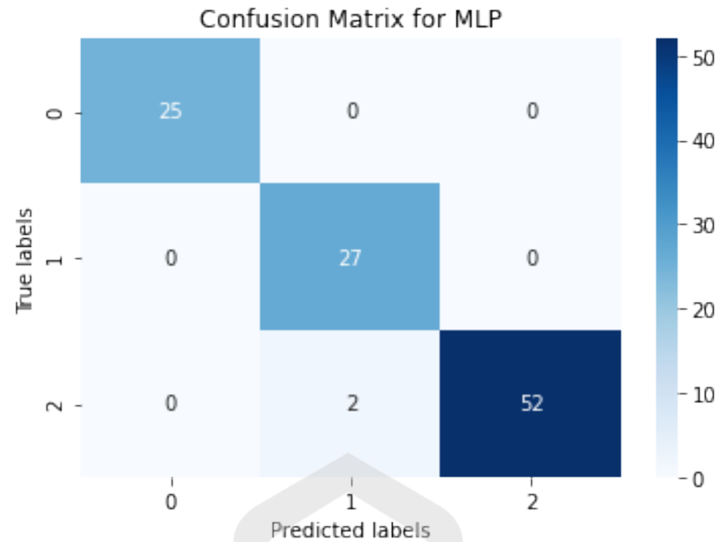
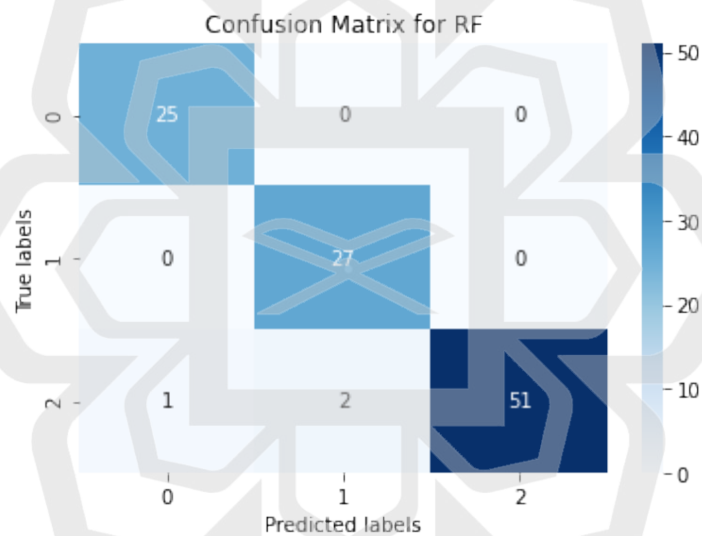
Model	Accuracy_score	
1	Logistic Regression	0.981132
2	Multilayer Perceptron	0.981132
3	Random Forest	0.971698
4	KNN	0.971698
5	Gradient Boosting	0.971698
6	Naive Bayes	0.971698
7	Decision Tree	0.943396
8	Support Vector Machines	0.905660

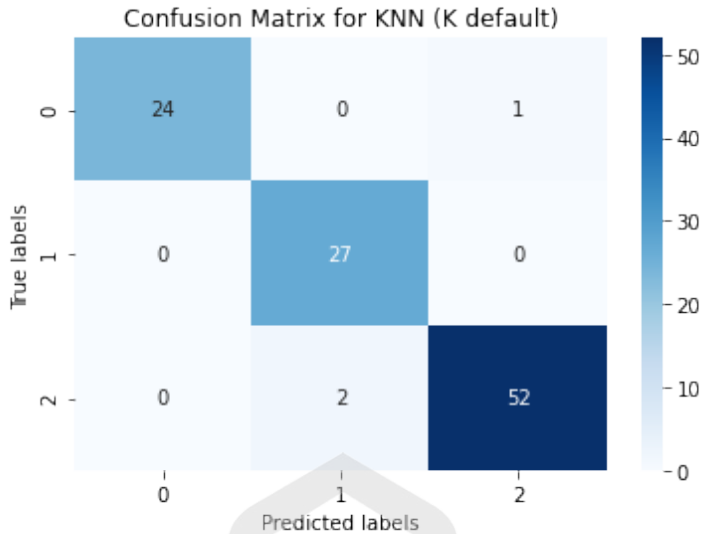
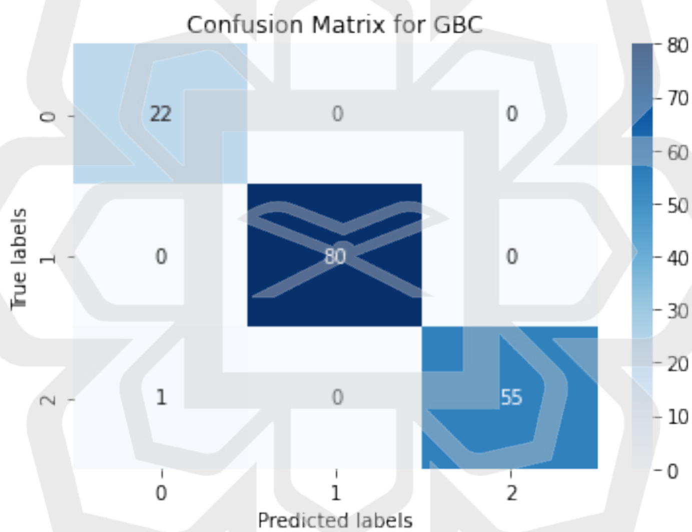
The dataset provided a general improvement overall within the models, with every model performing over 90% accuracy score, however, there has been a shuffle in terms of how the models ranked against one and another. Logistic regression and MLP achieved the highest accuracy scores of 98.11%. These models demonstrate strong performance could be attributed to to the ability in capturing non-linear relationships between features and the target variable. This flexibility allows them to effectively model complex patterns present in the dataset, including the relationships between water quality parameters and potability. Random Forest, KNN (K-Nearest Neighbors), Gradient Boosting, and Naive Bayes performed very same at 97.16%. Comparatively these models also performed strongly with the first dataset provided. Random Forest and Gradient Boosting are ensemble methods that combine multiple weak learners (decision trees in this case) to form a strong classifier. This ensemble approach often leads to robust and accurate predictions by mitigating the weaknesses of individual models whereas KNN is a non-parametric algorithm that makes predictions based on the majority class of its k-nearest neighbors. Its localized approach can be effective in capturing complex decision boundaries and handling nonlinear relationships in the data. Decision Tree in comparison underperformed from the first dataset while SVM remains the model that consistently is unable to perform in both datasets

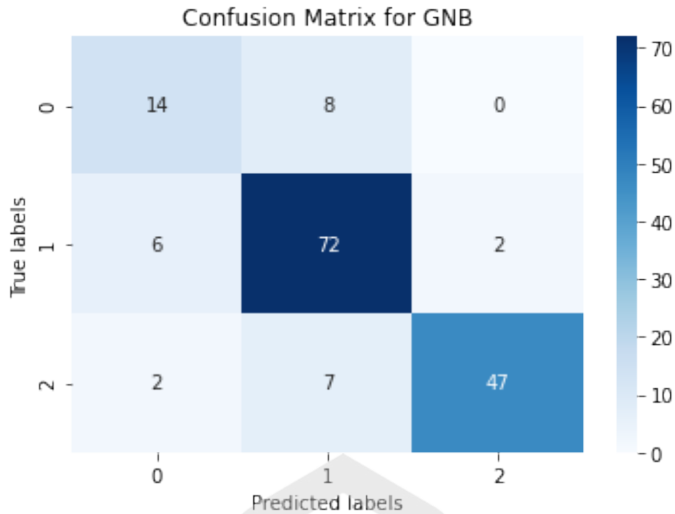
For next comparison, we only considered models with accuracy score of over 95%.

Table 4.5: Confusion Matrix for Top 6 models in Dataset 2

Model	Confusion Matrix	Performance
Logistic Regression		Logistic Regression achieves high accuracy with no misclassifications in Classes 0 and 1. However, there are two false positives in Class 2

Multilayer Perceptron		MLP performs similarly to Logistic Regression, with no misclassifications in Classes 0 and 1 but two false positives in Class 2.
Random Forest		Random Forest demonstrates strong performance with no misclassifications in Classes 0 and 1. There is one false positive in Class 2.

KNN		The KNN model performs well overall, with high numbers of true positives in each class. There is one misclassification in Class 0 and two false positives in Class 2.
Gradient Boosting		Gradient Boosting achieves high accuracy with no misclassifications in Classes 0 and 1. However, there are three false positives in Class 2.

Naïve Bayes		The Naive Bayes model demonstrates generally good performance, with high numbers of true positives in each class.

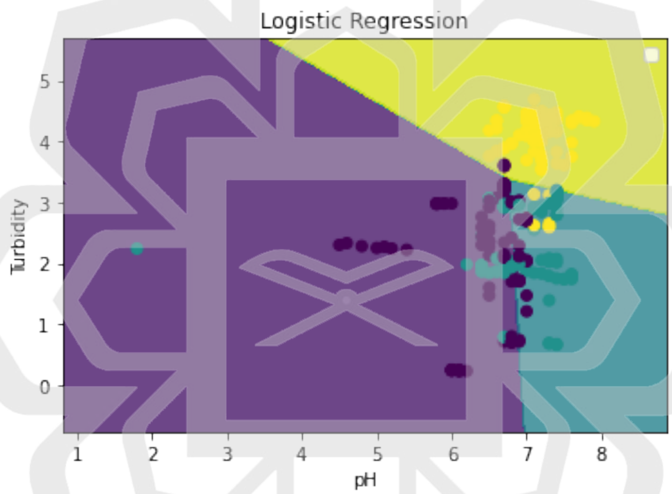
Overall, all models demonstrate strong performance with high numbers of true positives in each class. However, false positives in Class 2 are observed in Logistic Regression, MLP, and Gradient Boosting, indicating some difficulty in accurately classifying instances in this class. Further analysis may involve examining other performance metrics such as precision, recall, and F1 score to gain a more comprehensive understanding of each model's performance.

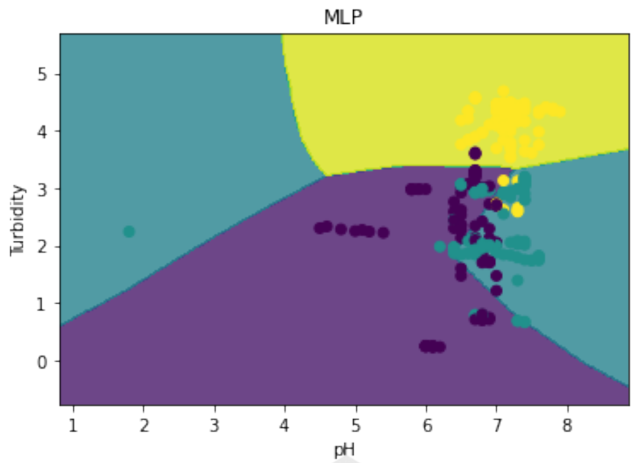
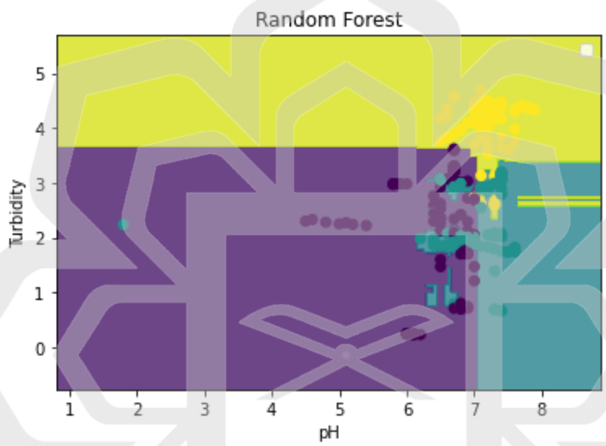
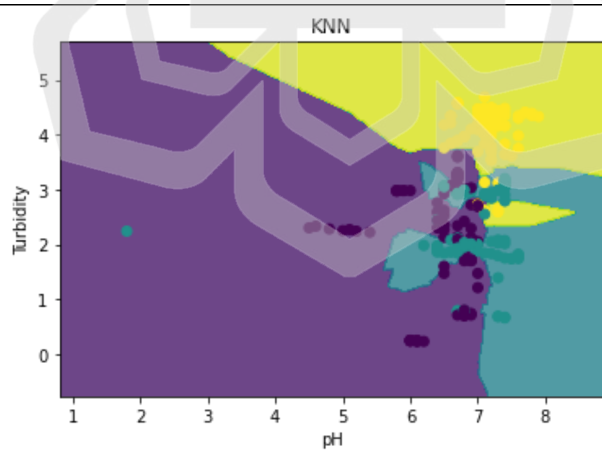
Table 4.6: Average Precision, Recall and F1 Score for Top 6 models in Dataset 2.

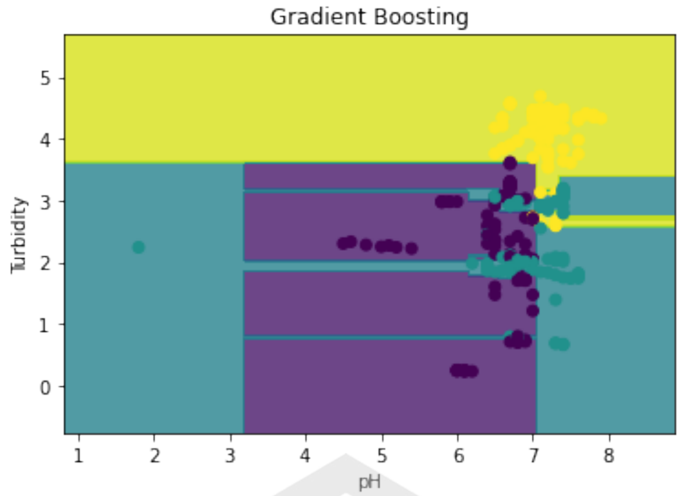
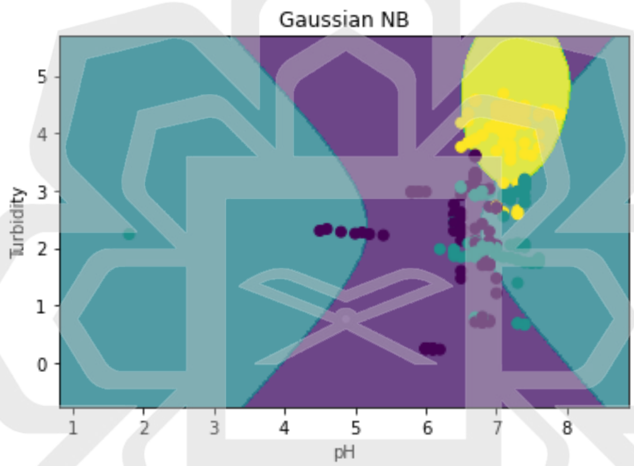
	Precision	Recall	F1 Score
Logistic Regression	0.94	0.94	0.94
MLP	0.97	0.98	0.97
Random Forest	0.96	0.98	0.97
KNN	0.97	0.97	0.97
Gradient Boosting	0.97	0.98	0.97
Naïve Bayes	0.97	0.98	0.97

Furthermore, all models exhibit high precision, recall, and F1 score, suggesting their effectiveness in accurately classifying instances. MLP shows slightly superior performance, but the differences between models are marginal.

Table 4.7: Decision boundaries for Top 6 models for dataset 2.

Model	Decision Boundary	Performance
Logistic Regression		The classes embody general class division without overfitting however some misclassification can occur in the overlap between puddle (0) and river flowing water (1)

<table>
<tr>
<td>MLP</td>
<td>
</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>
</td>
<td>Model is showing some instances of overfitting.</td>
</tr>
<tr>
<td>KNN</td>
<td>
</td>
<td>Due to the nature of the KNN mechanism, there are several clusters that is able to capture the various specific natures of each class.</td>
</tr>
</table>

Gradient Boosting		Overall many instances of overfitting
Naïve Bayes		A generalized shape and at instances is underfitting.

Overall, KNN is able to best capture the dynamic and nuance nature of the different classes without overfitting the data and since it performs well with the other evaluation metrics , does show to be reliable for deployment use.

4.4 DISCUSSION

With dataset 1, while there was certain level of class imbalance, served as the initial benchmark for model evaluation. The performance within the different metrics highlighted key steps in improving the experimentation. The correlation analysis revealed insights into the influence of different parameters on water quality, with turbidity and pH emerging as significant factors in potability classification. However, the temperature sensor exhibited negligible correlation and was thus excluded from further analysis. Within the scatter plot, it was possible to notice the overlapping nature of the 3 classes particularly within puddle and river flowing water and at instances river flowing water with potable water.

Upon training and testing, ensemble methods like Gradient Boosting and Random Forest, along with non-linear models such as KNN and Multilayer Perceptron, demonstrated superior performance, surpassing 95% accuracy. However. overall, random forest emerged as the optimal choice for deployment due to its robust performance across multiple evaluation metrics and its ability to handle non-linear relationships in the data effectively.

Dataset 2 was an extension from dataset 2 in addressing the limitations by incorporating the an additional feature with dissolved oxygen sensor and curating a more class balanced distribution. With the model performance, all models achieved over 90% accuracy, with Logistic Regression and MLP leading with 98.11%, showcasing the dataset's enhancement. While models exhibited comparable performance, Logistic Regression and MLP stood out for their ability to capture complex relationships in the data, followed closely by ensemble methods like Random Forest and Gradient Boosting, and

KNN. However, in a complete evaluation, KNN demonstrated the most nuanced and curated decision boundary, reflecting its ability to capture the dynamic nature of different classes without overfitting or underfitting.

4.5 CHAPTER SUMMARY

The evaluation of both datasets and ML models underscores the efficacy of IoT-enabled water quality monitoring systems coupled with ML for analysis and classification tasks. The findings not only validate the effectiveness of the proposed framework but also highlight the importance of feature selection, model evaluation, and iterative improvements for achieving accurate and reliable water quality assessment. Moreover, the selection of Random Forest for deployment in Dataset 1 and KNN in Dataset 2 underscores the significance of model selection tailored to specific datasets and classification requirements.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 CONCLUSION

In conclusion, leveraging the IoT mechanism, we efficiently generated two distinct real datasets gathering crucial water quality parameter information. After thorough cleaning, these datasets underwent rigorous training and testing using a variety of conventional classification algorithms. The first dataset encompassed four parameters, including pH, temperature, turbidity, and total dissolved solids, while the second dataset incorporated the dissolved oxygen sensor for further extension.

The datasets were tested against 9 different conventional classification algorithms ranging from ensemble models to supervised learning algorithms to artificial neural network algorithms. The algorithms were tested upon various evaluation metrics including confusion matrix, precision, recall, f1 score, decision boundary, and finally accuracy score. With dataset 1, we found that random forest proven to be the most optimal model due to its accuracy of 96%, negligible misclassification and balanced decision boundary. For dataset 2, KNN proved to be the most optimal model with an accuracy of 97 % and providing the most balanced projection of decision boundary.

The trained models were successfully deployed onto the Node-Red-based water quality monitoring interface. The interface connected with the water quality sensors displayed the sensor readings and upon pressing the input switch, was able to capture the real time reading and run by the deployed model and displayed the classified condition.

In summary, our developed device and architecture represent a practical and cost-effective solution for real-time water quality monitoring across various natural resources. Its potential to streamline water management processes in rivers, waters, and streams underscores its importance in promoting effective environmental management.

5.2 RECOMMENDATIONS

While our proposed system integrates proven state-of-the-art classification algorithms with cost-effective sensors and an efficient IoT data transmission network, there are opportunities for enhancement in future studies. The project primarily focused on studying the conditions of River Pusu, limiting the scope of data collection to this specific location. To broaden our understanding and ensure a more comprehensive model, we recommend expanding the study area and dataset to encompass a wider range of water bodies and environmental conditions. This can include the increase in the portability classes to be considered. Additionally, increasing the number of classes considered for classification can enhance the model's adaptability and accuracy in real-world scenarios. Furthermore, as the system's goal to be easily scalable, the scalability will be best displayed by increasing the feature sensors included such as chlorine and ammonia sensors. In conclusion, future studies should focus on broadening the geographical scope of data collection, increasing the diversity of classes considered, and expanding the sensor suite to ensure the robustness and scalability of the water quality monitoring system.

REFERENCE

Adeleke, I. A., Nwulu, N. I., & Ogbolumani, O. A. (2023). A hybrid machine learning and embedded IoT-based water quality monitoring system. Internet of Things, 22, 100774.

Agunbiade, M. O., Pohl, C., Heerden, E. Van, Oyekola, O., & Ashafa, A. (2019). Evaluation of fresh water actinomycete bioflocculant and its biotechnological applications in wastewaters treatment and removal of heavy metals. International Journal of Environmental Research and Public Health, 16(18), 1–15.

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

Baanu, Bharani & Babu, K.s. Jinesh. (2021). Smart water grid: a review and a suggestion for water quality monitoring. Water Supply.

Che Mahmud, N. A. (2021, March 11). *River water quality issues in Malaysia*. UMP News. Retrieved January 8, 2023, from https://news.ump.edu.my/experts/river-water-quality-issues-malaysia

Das B and Jain, P. C, "Real-time water quality monitoring system using Internet of Things," 2017 International Conference on Computer, Communications and Electronics (Comptelix), Jaipur, India, 2017, pp. 78-82.

Dong J, Wang G, Yan H, Xu J, Zhang X. (2015) A survey of smart water quality monitoring system. Environ Sci Pollut Res Int. Apr;22(7):4893-906.

Gai, R., Zhang, H. Prediction model of agricultural water quality based on optimized logistic regression algorithm. EURASIP J. Adv. Signal Process. 2023, 21 (2023).

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

Gorgan-Mohammadi, F., Rajaee, T., & Zounemat-Kermani, M. (2023). Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water. Sustainable Water Resources Management, 9(1).

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.

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J. P., ... & Cho, K. H. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences*, *32*, 90-101.

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.

Ilić, M., Srdjević, Z., & Srdjević, B. (2022). Water quality prediction based on Naïve Bayes algorithm. Water Science and Technology, 85(4), 1027-1039.

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.*

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.

N. Kumar Koditala and P. Shekar Pandey, "Water Quality Monitoring System Using IoT and Machine Learning (2018) ," *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, pp. 1-5.

Omer, N.H. (2019). Water Quality Parameters. *Water Quality - Science, Assessments and Policy*.

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.

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

Roy, Ritabrata. (2019). An Introduction to Water Quality Analysis. 6. 201-205.

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,*

Sensorex. (2021). The concern for total dissolved solids in water and how to measure it. https://sensorex.com/total-dissolved-solids-in-water/

Sensorex. (2022). The importance of dissolved oxygen in drinking water. https://sensorex.com/dissolved-oxygen-drinking-water/

Shadkani, S., Abbaspour, A., Samadianfard, S., Hashemi, S., Mosavi, A., & Band, S. S. (2021). Comparative study of multilayer perceptron-stochastic gradient descent and gradient boosted trees for predicting daily suspended sediment load: The case study of the Mississippi River, U.S. International Journal of Sediment Research, 36(4), 512–523

Stikker, A. (1998). Water today and tomorrow: prospects for overcoming scarcity. Futures, 30(1), 43–62

U.S. Environmental Protection Agency. (2025). Indicators: Dissolved oxygen. https://www.epa.gov/national-aquatic-resource-surveys/indicators-dissolved-oxygen

United Nations Environment Programme. (2024). FAQs on water quality. https://www.unep.org/explore-topics/water/what-we-do/world-water-quality-alliance-wwqa-partnership-effort/faqs-water

Walker, D. B., Baumgartner, D. J., Gerba, C. P., & Fitzsimmons, K. (2019). Surface Water Pollution. In Environmental and Pollution Science (3rd ed.).

Wang, J., Liu, X. D., & Lu, J. (2012). Urban River Pollution Control and Remediation. Procedia Environmental Sciences, 13(2011), 1856–1862.

Wisewell. (n.d.). Which water temperature is best for drinking? Wisewell., From https://www.wisewell.com/blogs/news/which-water-temperature-is-best-for-drinking

Wu, Y., Zhang, X., Xiao, Y., & Feng, J. (2020). *Attention Neural Network for Water Image Classification under IoT Environment. Applied Sciences, 10(3), 909.*

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.

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*.