WATER QUALITY MONITORING SYSTEM

Project report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

By

Rohit Raj (191010)

UNDER THE GUIDANCE OF

Dr Rajiv Kumar

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

May 2023

TABLE OF CONTENTS

DECLARATION

We hereby declare that the work reported in the B.Tech Project Report entitled **"WATER QUALITY MONITORING SYSTEM"** submitted at Jaypee University of Information Technology, Waknaghat, India is an authentic record of our work carried out under the supervision of Dr Rajiv Kumar, Professor and Head. We have not submitted this work elsewhere for any other degree or diploma.

Rohit.R.

Rohit Raj

191010

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

08/05/2023

(Supervisor Signature)

Dr Rajiv Kumar

Professor and Head

Electronics and Communication Department

Jaypee University of Information Technology, Waknaghat

Date: 08/05/2023.

ACKNOWLEDGEMENT

I would want to convey my heartfelt gratitude to Dr Rajiv Kumar, my guide, for his invaluable advice and assistance in completing my project. He was there to assist me every step of the way, and his motivation is what enabled me to accomplish my task effectively. I would also like to thank all the other supporting personnel who assisted me by supplying the equipment that was essential and vital, without which I would not have been able to perform efficiently on this project.

I would also want to thank the Jaypee University of Information Technology for accepting my project in my desired field of expertise. I would also like to thank my friends and parents for their support and encouragement as I worked on this assignment.

LIST OF ACRONYMS AND ABBREVIATIONS

TDS	Total Dissolved Solids
ML	Machine Learning
IoT	Internet of Things
SVM	Support Vector Machine
RF	Random Forest
XGB	Extreme Gradient Boost
LR	Logistic Regression
ANN	Artificial Neural Network
kNN	k-Nearest Neighbour
DT	Decision Tree
NTU	Nephelometric Turbidity Unit

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

A water quality monitoring system can aid in preserving the environment, ensuring the security of nearby water sources, and fostering economic growth in rural areas. As a result, this will help to develop a system here that employs Internet of Things and Machine Learning to monitor the quality of water. This paper discusses the characteristics of water to let us know whether it is fit for human consumption or not. The sensors dipped in water samples acquired from wells, lakes, rivers, ponds, or other places are used to inform the development of an effective model made up of TDS, pH and turbidity sensors. The data will be delivered from the sensors as soon as they are received to the IDE, where it will then be sent to the cloud server. The model effectively accounts for test tables, where 1 indicates the water is fit for drinking and 0 indicates the water is not. The values were classified differently using Machine Learning models like SVM, RF and XG Boost method. Training data is pre-processed before being fetched from the cloud. Over that data, machine learning models like Support Vector Machine, Random Forest & Extreme Gradient Boost has been implemented. The maximum accuracy of 95.12% was observed using XG Boost. After testing, we will be able to determine whether the water is fit for drinking using the binary indicators of 1 and 0, where 1 indicates the water is fit for drinking and 0 indicates the water is not.

CHAPTER – 1

INTRODUCTION

1.1 Introduction

According to the World Health Organization (WHO), 368 million people use unprotected wells and springs and 122 million people gather untreated surface water from lakes, ponds, rivers, and streams. This means in 2020, approximately 2 billion people were without access to safe water. Cholera, diarrhoea, dysentery, Hepatitis, typhoid, and polio are just a few of the diseases that can spread as a result of contaminated water and inadequate sanitation. People are exposed to avoidable health risks when water and sanitation infrastructure is inadequate, poorly maintained, or managed improperly. Poor income, technological advancements, internal community management, water contamination from agricultural chemicals, industries, and waste disposal all contribute to rural villagers limited access to clean drinking water. Since 2000, more than 50% more city dwellers lack access to safely regulated drinking water. TDS and Turbidity are one the major qualities of water, as the drinking water should have a proper amount of TDS and turbidity in it[1].

In rural areas, the water quality of wells and ponds is assessed using two crucial parameters: pH and turbidity. The effectiveness of water treatment procedures, the taste and odour of drinking water, the corrosion of infrastructure, and the health and survival of aquatic species can all be impacted by the pH of water, which is a crucial parameter. Turbidity is a unit used to describe how cloudy or hazy a liquid, results suspended particles. Turbidity is a crucial metric that tells us about the transparency of the water. Turbidity can change the physical, chemical, and biological properties of water. The presence of suspended particles like silt, clay, and organic matter can cause a variety of issues, including high levels of turbidity in water. A solution's acidity or alkalinity can be determined by its pH value, which ranges from 0 to 14, pH readings below 7 signifies acidity while above this value signifies alkalinity whereas pH 7 is regarded as neutral. To guarantee that the water is safe and fit for use in rural regions where wells and ponds are frequently used as sources of drinking water, it is crucial to routinely test the pH and turbidity of the water. Indicators of the presence of pollutants or other contaminants in the water, which may have detrimental effects on both human health and the ecosystem, include high levels of turbidity or low or high pH. Frequent monitoring of these characteristics can aid in spotting possible issues before they become serious and enable the proper measures to be done to safeguard the water supplies in rural regions.

A high TDS level in water can be an indication of dangerous pollutants or heavy metals, which can have an adverse effect on both human health and the environment. The flavour, odour, and appearance of water can also be impacted by high TDS levels, making it less appealing for drinking and other uses whereas turbidity is a measurement of how murky or hazy the water is. The presence of suspended particles like dirt, algae, or organic materials is the source of it. When turbidity levels are high, it may be a sign that the water is contaminated with contaminants, bacteria, or other dangerous microorganisms that can lead to diseases that are spread by water. Moreover, turbidity can reduce the efficiency of water treatment procedures, making it more challenging to maintain clean water and eliminate contaminants [2].

TDS and turbidity can have negative health effects when present in high levels in drinking water. High TDS levels can lead to mineral imbalances in the body, while high turbidity levels can harbour harmful pathogens. As such, it is important to ensure that drinking water meets the acceptable limits for both TDS and turbidity as recommended by local authorities. Low TDS can result in bacterial infiltration, bad taste, and bad odour. It can cause a lack of vital minerals, which are necessary for good health. It may lead to high levels of heavy metals in drinking water, it may also reduce the efficacy of water treatment and increase bacterial contamination and can cause dehydration, especially in hot environments or after strenuous activity. High turbidity can reduce water clarity, making it murky or discoloured. It can reduce aesthetics, making water unpleasant to drink which also increases the risk of waterborne illnesses. It can cause clogging of water treatment equipment and can lead to corrosion. It can reduce efficiency and quality of industrial processes.

Total Dissolved Solids, or TDS, is a crucial indicator of water quality. It refers to the total amount of minerals, salts, metals, and other compounds that are dissolved in water, both organically and inorganically. Both natural resources like rocks and soil as well as human endeavours like farming, manufacturing, and urban growth can provide these materials. High TDS levels in water can have an impact on its flavour, aroma, and appearance as well as on how suitable it is for different uses, including drinking, irrigation, and industrial activities. For instance, water with a high TDS level may taste bitter or salty and may not be suitable for use in cooking or drinking. Similar to this, irrigation with water with a high TDS can cause salt in the soil and lower agricultural yields. TDS may also be a sign of other water pollutants, including nitrate, fluoride, arsenic, and chloride. Monitoring TDS levels on a regular basis can aid in spotting potential contamination problems and ensuring that water is suitable for its intended purpose. TDS concentrations can also have an impact on the efficiency of water purification techniques like distillation and reverse osmosis, which are intended to remove dissolved particles from water. The effectiveness of these processes can be hampered by high TDS levels, which also make successful water treatment more time- and energy-intensive. TDS is a significant factor in assessing the quality and usability of water for a variety of uses. TDS levels should be regularly monitored and managed to assist maintain the sustainability and safety of water supplies.

Turbidity is a term that describes the amount of cloudiness or haziness in water that is brought on by the presence of suspended particles like clay, silt, organic debris, or microscopic organisms. These particles scatter and absorb light, obstructing its passage through the water and giving it a foggy or murky appearance. In many applications, such as drinking water, wastewater treatment, industrial operations, and environmental monitoring, turbidity is a crucial metric used to assess the water quality.

Following are some explanations for why turbidity in water is important:

Water quality indicator: Turbidity is considered an essential water quality indicator. High turbidity levels may indicate the presence of harmful contaminants, such as bacteria, viruses, and parasites, that can cause waterborne illnesses. Thus, measuring turbidity is an effective way to determine whether water is safe for consumption or other uses.

Impact on aquatic life: Turbidity levels can significantly affect aquatic ecosystems. High turbidity can reduce the amount of light that penetrates the water, inhibiting the growth of aquatic plants and the photosynthesis process, which is essential for the survival of many aquatic organisms. Additionally, high levels of turbidity can clog the gills of fish and other aquatic animals, making it difficult for them to breathe.

Wastewater treatment: Turbidity is an essential parameter in wastewater treatment. High turbidity levels in wastewater can indicate the presence of solids, such as organic matter and pathogens, which can negatively impact the treatment process. Additionally, high turbidity can interfere with disinfection processes, making it difficult to effectively remove harmful contaminants from wastewater.

Industrial processes: Turbidity is also a crucial factor in many industrial processes, such as food and beverage production, pharmaceuticals, and chemical manufacturing. High turbidity levels can negatively impact product quality, increase production costs, and result in regulatory non-compliance.
Environmental monitoring: Turbidity is a critical parameter in environmental monitoring. It is used to determine the impact of human activities on water quality, such as erosion, construction, and land use changes. High turbidity levels in water bodies can negatively impact recreational activities, such as swimming, boating, and fishing, and can harm wildlife habitats.

A solution's acidity or basicity is determined by its pH. It is expressed on a scale from 0 to 14, where 0 represents the most acidic solution, 7 is neutral, and 14 is the most basic. In the context of water, pH is an important parameter that has significant implications for both human health and the environment. The solubility of various chemicals is one of the main reasons pH in water is significant. The solubility of many compounds and minerals varies with pH. For instance, excessive acidity can cause metals to dissolve in water and contaminate sources of drinking water. On the other side, high pH levels can result in the precipitation of minerals like calcium and magnesium, which can cause scale to form in pipes and other water-related infrastructure. The effect of pH on aquatic life is another crucial part of water chemistry. Numerous fish species and other aquatic organisms are sensitive to pH level changes. These organisms may face physiological stress in excessively acidic or basic water, which may result in stunted growth, poor reproductive success, and even death. The solubility of nutrients and other molecules in water can also be impacted by pH variations, and this can have a cascade effect on the entire food chain. In addition to having an impact on aquatic life, pH is crucial for the distribution and treatment of water. Maintaining a pH level that is suitable with the various treatment procedures and materials used in the distribution system is essential for the efficient treatment and distribution of water. As an illustration, high acidity levels can erode pipes and other infrastructure, causing leaks and other issues. The precipitation of minerals and other compounds can also be triggered by high pH levels, which can clog pipes and lessen the effectiveness of treatment procedures. The pH of water is a crucial variable that affects both the environment and human health in a significant way. It is a crucial parameter to take into account in any situation involving water because of its impacts on the solubility of different compounds, impact on aquatic life, function in the treatment and distribution of water, and usage in monitoring water quality. We can ensure that our water supplies are safe and wholesome for future generations by being aware of the significance of pH in water and taking the proper steps to maintain acceptable levels [4].

A dataset with pH and turbidity water quality metrics, along with their accompanying water quality labels, can be used to train using Extreme Gradient Boost model, comprising of Support Vector Machine and Random Forest. Based on fresh pH and turbidity measurements, the model will then be able to monitor water quality. Since they can manage non-linear correlations between the input characteristics & the output labels that are less prone to overfitting than other tree-based models.[4] A comparable dataset can also be used to train an SVM model, which outputs water quality labels and uses pH and turbidity as input features. SVMs are especially helpful when there is a distinct boundary dividing different water quality classes since they seek out the best separation hyperplane in a high-dimensional space between two classes. SVMs are typically more resistant to data noise than some other algorithms. To increase overall accuracy, ensemble combines the overall at method loss of various machine learning models. With the same dataset, for instance, an ensemble of SVM and random forest models might be trained, and their individual forecasts may be integrated to yield a more precise overall prediction. Random Forest and SVM are powerful ensemble learning techniques for TDS and turbidity prediction, as they can handle both classification and regression difficulties and manage missing data and outliers [5].

1.2 Problem Statement

Poor water quality for drinking can cause a wide range of health issues and can have significant negative impacts on both individuals and communities. Therefore, a webpage is being created that will inform the user of whether the water is fit for drinking or not by using a model created using various machine learning techniques. The following are some of the primary problems brought on by drinking water of poor quality:

Waterborne diseases: Poor water quality can cause a variety of waterborne diseases such as cholera, typhoid, hepatitis A, dysentery, and diarrhoea. These diseases can cause severe dehydration and can be fatal, especially in children and people with weakened immune systems.

Chemical contamination: Poor water quality can also lead to chemical contamination of drinking water. Chemicals such as lead, arsenic, and mercury can leach into the water supply from natural sources or from industrial and agricultural activities. Long-term exposure to these chemicals can lead to serious health issues such as cancer, neurological disorders, and developmental delays.

Microbial contamination: Microbial contamination is another significant issue caused by poor water quality. Bacteria, viruses, and other microorganisms can contaminate the water supply, leading to illnesses such as gastroenteritis, hepatitis, and other infections.

Nutrient pollution: Nutrient pollution can also be a problem in areas where fertilizers and animal waste runoff enter the water supply. Nutrient pollution can lead to the growth of harmful algae blooms that can produce toxins that are harmful to humans and animals.

Environmental impact: Poor water quality can also have negative impacts on the environment. Polluted water can harm aquatic life and disrupt ecosystems. It can also impact the quality of soil, crops, and natural resources that rely on clean water for growth and survival.

Economic impact: Poor water quality can have significant economic impacts on communities. It can lead to increased healthcare costs due to water-related illnesses, lost income due to missed work or school, and decreased property values due to contaminated water supplies.

Finally, drinking water with poor quality can have negative effects on your health, the environment, and your finances. To guarantee that their water supply is secure and uncontaminated, it is crucial for individuals and communities to take action. This can be done by conducting routine water tests, treating and filtering the water appropriately, and managing natural resources sustainably [6].

1.3 Objective

- The project involves a thorough literature review.
- The aim is to develop a more efficient and accurate model for water quality.
- The project will deal with the aspects of water that must be considered in order to calculate water quality.
- Various machine learning techniques will be used to thoroughly suspect all the parameters.
- The project will contribute to the broader research efforts in IoT, specifically in the area of IoT and ML.

1.4 Methodology

The complete system of Water Quality Monitoring System for rural areas using IoT and Machine Learning has been illustrated in this paper. The proposed methodology is a combination of hardware and software. For real time implementation, various samples from the wells and ponds of Richhana village, Solan was collected with the help of microcontroller (ESP 32) & sensors, pH and turbidity values were evaluated, which were validated using a model designed by Kaggle dataset. Machine learning techniques were be applied to the data. SVM, RF and Ensemble method. utilising Extreme Gradient Boost were both used in this work. The classification is done on the basis whether the water is fit for human consumption namely pH and turbidity as attributes. TDS and turbidity values were assessed using sensors and a microcontroller (ESP 32), and a model created using a Kaggle dataset was used to validate the results.

The proposed models were created using ML models. TDS and Turbidity parameters are used as attributes to classify whether the water is fit for human consumption. The Turbidity and TDS sensors were submerged in water, allowing the microcontroller (ESP32) to obtain real-time information. Before transferring the data to the ThingSpeak Cloud server, the Arduino IDE will first show the data. To assess whether the data is linear or variable, the data will be graphed. The data will then be analysed using machine learning methods. Support vector machines and random forests were both used in this study. Using the two water characteristics, Extreme Gradient Boost has been utilized in this instance to combine both approaches, giving exact information on whether the water is fit for human consumption. The Extreme Gradient Boost model, which consists of SVM and RF, has been trained using a dataset that contains water quality metrics for TDS and turbidity along with the labels that correspond to those metrics' water quality. They are less prone to overfitting than other tree-based models because they can handle non-linear correlations between the input features and the output labels.

1.5 Organisation

The project is organized into several chapters, starting with the Introduction in Chapter 1, which provides a brief overview of the topic and its relevance. This chapter also covers the problem statement, methodology, and objectives of the study.

Chapter 2, titled Literature Survey, reviews the previous studies conducted in this field, with a focus on Internet of Things, Machine Learning techniques, Random Forest, and Support Vector Machine

Chapter 3, the Methodology, explains the project's methodology in detail, including the data receiving, the model training, and the testing phase. It also includes a thorough discussion of the experimentation, and the results of the project.

Chapter 4, Results and Discussion, presents the findings of the experiment and provides a detailed discussion of the results. This chapter also includes a comparison of the different model configurations and discusses the impact of various parameters on the accuracy of the models.

Chapter 5, Conclusion explain that more parameters can be added to the project to know about the clarity of the water, as the more parameters, the more precise the water will be.

Finally, all the sources, including research studies, real-time values, algorithms, and others, are referenced in the last section of the project.

CHAPTER - 2

LITERATURE REVIEW

2.1 Literature Review

The algorithm proposed by Amir *et al.* [7] studied how well artificial intelligence methods like ANN, GMDH, and SVM predicted Tireh River in Iran's water quality components. The best performance was demonstrated by the tansig, RBF transfer, and kernel functions. The GMDH model performed as expected, however its accuracy lagged behind that of ANN and SVM. In this study, the effectiveness of artificial intelligence methods such as GMDH, SVM, and ANN was assessed in order to forecast the water quality elements of the Tireh River (Iran). In order to achieve this, the majority of dataset-related well-known components were gathered, including pH, So4, Na, Ca, Cl, Mg, Hco3, etc. Results showed that the applied models performed satisfactorily in predicting the various elements of water quality, although the SVM had the best performance. Reviewing the SVM's structure revealed that the RBD's role as the kernel function was responsible for the best accuracy. The ANN's results showed that its accuracy is sufficient for practical uses. Tansig was associated with the tested transfer function's best performance. GMDH-related models have the lowest accuracy.

The algorithm proposed by Yingyi *et al.* employed feedforward, recurrent, and hybrid architectures in artificial neural network (ANN) for the prediction of water quality. Five output strategies they summarised& consequently discovered that the ANN models could handle various modelling issues in rivers, lakes, reservoirs, and Waste Water Treatment Plants (WWTPs).Water chemistry and water quality are examined. Water quality parameters discussed include oxygen content, pH, nitrates, heavy metals, trace elements, hardness, odour and colour. Water quality standards from Europe and the USA, and their relation to human and aquatic health are examined.

The algorithm suggested by Umair *et al.*[8] looked at how well ANN, GMDH, and SVM performed as artificial intelligence techniques for predicting Tireh River in Iran water quality components. Results showed that tansig and RBF transfer and kernel functions had the best performance. GMDH model had acceptable performance, but its accuracy was slightly less than ANN and SVM. In order to calculate the water quality index (WQI) and the water quality class (WQC), the research investigates supervised machine learning algorithms. Due to their adaptive respiratory physiologies and high reproductive rates, Limnodrilus hoffmeisteri and Tubifex tubifex predominate in anaerobic environments. The three most effective input parameters used in the suggested methodology are gradient boosting, polynomial regression, and multi-layer perceptron. The suggested methodology validates its usage in real-time water quality detection systems by achieving reasonable accuracy with a small number of parameters. This study sought to predict water quality using a small set of characteristics with inexpensive sensors because water quality parameter sensors are expensive. Our initial analysis took into account four variables: temperature, turbidity, pH, and total dissolved solids. Gradient boosting was the most effective technique when using the regression algorithms, with an MAE of 1.9642, MSE of 7.2011, RMSE of 2.6835, and RSE of 0.7485. We predicted the water quality class (WQC), which was given to samples based on their previously determined WQI, using classification algorithms. The classification process used the same criteria as in the preceding section. The same four criteria were first taken into account. With an accuracy of 0.8507, precision of 0.5659, recall of 0.5640, and F1 score of 0.5649, we discovered that MLP outperformed the other algorithms in this situation.

For the comparison of water quality, Najah *et al.* [9] presented linear regression models (LRM), multilayer perceptron neural networks (MLP), and radial basis function neural networks (RBF-NN). This result shows that RBF-NN models are more accurate than LRM and MLP. High water quality is necessary for the management of the Johor River Basin in Malaysia's Johor state. A recent method called artificial neural networks is capable of making precise predictions about the future. The most precise and trustworthy method for handling substantial amounts of non-linear, nonparametric data is RBF-NN modelling. Environmental protection greatly benefits from the modelling and forecasting of water quality. The development of a model using cutting-edge AI algorithms can be utilised to assess the water quality in the future. The WQI was predicted using this proposed methodology using the advanced artificial intelligence algorithms NARNET and LSTM models. Furthermore, to categorise the WQI data, machine learning methods like SVM, KNN, and Naive Bayes were applied. The proposed models were assessed and looked at using some statistical factors. According to the results, the NARNET model performs marginally better than the LSTM model for the WQI prediction based on the acquired R value. When compared to KNN and Naive Bayes algorithms, the SVM method has, however, produced the WQC prediction with the highest degree of accuracy.

Hu and Ma[10] employed two distinct hybrid decision-tree-based machine learning models, including Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), that took advantage of the data denoising technique. The results show that CEEMDAN-RF offers the most accurate forecasts for temperature, dissolved oxygen, and specific conductance. CEEMDAN-XG Boost provides the best predictions for pH, turbidity, and fluorescent dissolved organic matter.

Chen *et al*.[11]made important contributions to water pollution and regulation, including drinking water regulation, water quality standards, and policy instrument choice. It employed 10 learning models for huge data from the Chinese river (7 classical and 3 ensemble models). The results demonstrate that learning models may outperform other models with larger data sets in the prediction of water quality. Water quality indices (WQI) are used to portray data in a straightforward manner and are an important criterion for integrated environmental management and sustainable development. Furthermore, two significant water parameter sets with high specificity for the prediction of water quality were identified and confirmed by Decision Tree, Random Forest, and Deep Cascade Forest.

The algorithm proposed by Huber *et. al*.[12] focuses on both urban and rural methodologies, including straightforward techniques like constant concentration, regression, statistical, and loading functions as well as sophisticated models like SWMM, HSPF, STORM, CREAMS, and SWRRB.

Duda *et. al*.[13] waste discharges were found to be the most likely reason for the poor water quality in two urban streams. Additionally, comparisons with an upstream, forested control reach were made.

McGrane[14] describes leaky infrastructure contributes to recharge and runoff rates, while contaminants pose new challenges to monitoring and treatment. Water quality classification (WQC) and water quality index (WQI) prediction systems have been created. The models have been assessed using Naive Bayes, NARNET, LSTM, SVM, and K-NN. Results demonstrated that the suggested models are capable of correctly predicting WQI and categorising the water quality.

Samples from four water sources and eight house connections were compared to WHO guidelines which showed 50% to 62.5% of samples having bacteriological contamination before and 75% after the monsoon in the tests performed by Aziz *et. al*[15].

Koc *et. Al*[16]. stated that River water pollution in the yna River of Northern Poland divides the urban areas into urban river continuum.

Ruan *et. al*.[17] proposed an algorithm in which landscape indicators can explain water chemistry and bacterial concentrations in streams, with 5-m resolution imagery explaining 50% of variability as depicted in the experiments. Water is essential for human existence, but water pollution has become a serious problem. To control water pollution, a model that predicts water quality is needed. This study proposes a model based on four water parameters, using multiple regression and artificial neural networks.

Rieger *et. al.*[18] allium cepa was used to examine the genotoxic, mutagenic, and cytotoxic potential of surface waters in urban streams. All samples had varying degrees of environmental damage, according to physicochemical analysis, but downstream samples had the biggest influence.

Due to the fact that MLP-NN models needed a lot of hidden neurons and hidden layers, they also had sluggish convergence during training. RBF-NN was consequently used to get around this problem. RBF-NN was used to further improve the prediction of water quality metrics. We found that the RBF-NN trains more quickly than the MLP-NN and converges to a solution with less computing complexity. Additionally, the RBF-NN increased water quality parameter prediction accuracy while requiring little computational effort. When compared to other traditional modelling techniques. An innovative technique that can discern complex non-linear correlations between input and output data is the artificial neural network (ANN). Using three alternative model techniques—LRM, multilayer perceptron neural networks (MLP-NN), and RBF-NN—this study examined which of the water quality measurements along the Johor River basin on two different levels of water bodies made the best predictions. This book focuses on significant elements that have an impact on water quality as a result of urbanisation along rivers. The characteristics that were looked at were EC, total dissolved solids (T.D. solids), and turbidity. Compared to the LRM standard model, the MLP-NN performed better. In 97 watersheds in the Southeast Atlantic region of the USA, water quality metrics were predicted using machine learning techniques, and experiments showed that the RF approach was simpler to train and more resistant to model overfitting by Mishra *et. Al*[19]..

The purpose of the study is to provide an overview of the state-of-the-art for IoT-based smart water quality monitoring systems (IoT-WQMS) for household applications. It investigates typical WQM metrics, their safe drinking water limits, associated smart sensors, a critical evaluation, and the ratification of modern IoT-WQMS. Faecal coliform count, biochemical oxygen demand, pH, NO3, PO4, temperature variation from equilibrium, turbidity, total solids, and percent saturation of dissolved oxygen are five water quality indices that are compared in this research. All five indices have strong correlations with expert assessments, although Harkins' index is the least reliable and needs to be updated whenever new information becomes available. Although the high density aggregated construction had a potential to increase E. coli pollution, it was more successful in lowering TP and NO3-N concentrations by Kim *et. al.*[20].

2.2 Issues to be addressed due to poor quality of water

Water that is of poor quality can have a variety of negative consequences on the environment and human health. Here are some of the main consequences of water that is of poor quality:

Health Risks: Numerous health risks can result from water of poor quality. Waterborne diseases including cholera, typhoid, dysentery, and hepatitis A can all be brought on by bacteria, viruses, and other pathogens that are present in contaminated water. Lead, arsenic, and mercury exposure can also result in major health issues like cancer, neurological abnormalities, and developmental delays.

Environmental Impact: Water of poor quality can have a serious negative effect on the ecosystem. Water pollution can affect aquatic life, destabilise ecosystems, and impair plant and animal habitats. Degradation of soil and other natural resources that depend on clean water for growth and survival can also be brought on by pollution.

Economic Impact: Water of poor quality can have serious negative economic effects. Due to illnesses brought on by contaminated water, higher healthcare bills, lost wages from missed work or education, and decreased property values are all possible consequences. Tourism and other industries that depend on clean water resources may potentially suffer as a result.

Food Safety: Water of poor quality can potentially affect the safety of food. When food is produced and processed, contaminated water can encourage the growth of dangerous bacteria, which can result in foodborne illnesses.

Climate Change: Water quality can be impacted by climate change. Water availability, water quality, and the distribution of water resources can all be impacted by variations in temperature and rainfall patterns.

Political Instability: Political instability can also result from poor water quality. Pollution and a lack of water can cause population displacement, migration, and conflicts over resources[21].

The environment, the economy, and human health can all be significantly harmed by low-quality water. Water resources must be safeguarded and conserved for future generations, so it is crucial that individuals, governments, and companies take action. This can be accomplished through using natural resources responsibly, investing in water technology and infrastructure, and raising public knowledge of water conservation and pollution prevention.

CHAPTER - 3

METHODOLOGY

3.1 Descriptions of the Methodologies

Systems for monitoring water quality are crucial for ensuring the security and well-being of societies, ecosystems, and economies. The development of automated and intelligent water quality monitoring systems that can help identify pollution sources, spot abnormalities, and forecast future water quality levels is made possible by machine learning techniques.
Using machine learning techniques, the following methodology has been applied to create a model for monitoring water quality:

1. Identify Key Parameters: Finding the important metrics to monitor is the first stage in creating a machine learning-based water quality monitoring system. These could be chemical parameters like nutrient levels, heavy metal concentrations, and organic pollutants, as well as physical parameters like pH, TDS, and turbidity.
2. Collect Data: Data on these characteristics must be gathered from a variety of sources, including water quality sensors, remote sensing data, and laboratory examination, once the critical parameters have been identified. Databases can also be used to extract historical data to establish a baseline and spot trends and patterns.
3. Data Pre-processing: Pre-processing is necessary to get rid of outliers, mistakes, and missing values from the acquired data. Data normalisation, data cleaning, and data transformation may all be part of this process.
4. Feature Selection and Extraction: While feature extraction involves turning raw data into useful features that can be used in machine learning models, feature selection entails choosing the parameters that have the greatest influence on water quality.
5. Machine Learning Model Selection and Development: For the purpose of monitoring water quality, a variety of machine learning models, such as decision trees, random forests, support vector machines, and neural networks, can be utilised. The chosen model should take into account the features of the data, the complexity of the issue, and the desired results. The machine learning algorithm is calibrated to forecast levels of water quality using past data.
6. Model Evaluation: The efficiency of the machine learning model is evaluated using metrics including accuracy, precision, recall, and F1 score. To make sure the model can generalise to fresh data, cross-validation techniques can also be used to test it[22].
7. Deployment: The created machine learning model can be used as a real-time water quality monitoring system to continually track water quality levels, spot abnormalities, and forecast future water quality levels. To give stakeholders useful information, the system can be connected with dashboards and data visualisation tools.

In conclusion, essential parameters must be identified, data must be gathered, pre-processed, features must be extracted, models must be chosen and developed, models must be evaluated, and finally the system must be deployed. The system can aid in the sustainable management of water resources by assisting in the identification of pollution sources, the detection of anomalies, and the prediction of future water quality levels.

3.2 Machine Learning Techniques Used:

1. **Support Vector Machine:**
   The SVM[23] maps the data in a high-dimensional space where the model draws a straight line, called a hyperplane, to divide the data into several classes, resulting in support vectors that aid in the prediction of the target labels. The SVM classifier is expressed by Eq. and SVM classification for dual formation is expressed.

$$min_{f\xi_i}\|f\|_k^2 + C\sum_i^l \xi_i y_i \;\; f(x_i) \geq 1 - \xi_i, for\ all\ i \;\; \xi_i \geq 0 \;\;)$$

$$min_\alpha \sum_a^1 \alpha_i - \frac{1}{2}\sum_n^1 \sum_n^1 \alpha_i\alpha_j y_i y_j K(x_i, x_j) \quad 0 \leq \alpha_i \leq C \;, for\ all\ i; \;\; \sum_{i=1}^l \alpha_i y_i = 0$$

In Eq, ξ_i are slack variables and they measure the error produced at point (x_i,y_i) &α_i is the Langlier's multiplier.

Here are some of the primary traits and qualities of SVM:

- Binary classification: The SVM method divides data points into two classes according to their features.
- Maximum Margin Classifier: The SVM identifies the hyperplane that maximises the margin between the two classes, hence lowering the generalisation error.
- Kernel trick: SVM can deal with nonlinear decision boundaries by using a kernel function to project the input data into a higher-dimensional feature space.
- Regularisation parameter: To balance the trade-off between maximising the margin and minimising the classification error, SVM employs a regularisation parameter (C).
- Support vectors: against obtain the ideal hyperplane, SVM only considers the data points that are on the margin or violate it (support vectors), which increases the algorithm's efficiency and robustness against outliers.

2. **Random Forest:**

   In order to forecast the target labels, Random Forest[24] builds a decision tree for each piece of training data, averages those decision trees, and then allows users to vote on their favourite prediction outcome. Random Forest utilised is expressed.

$$\underline{r}(X) = E_{\theta}[r_n(X,\theta)] = E_{\theta}\left[\frac{\sum_{i=1}^{n} Y_i \, 1_{[X_i \in A_n(X,\theta)]}}{\sum_{i=1}^{n} 1 * 1_{[X_i \in A_n(X,\theta)]}} 1_{E_n(X,\theta)}\right]$$

   In Eq,$r_n(X,\theta)$ is the randomized tree of rectangular cell of the random partition containing $E_n(X,\theta)$ trees.

Here are some of the primary traits and qualities of Random Forest:

- Ensemble method: Random Forest is an ensemble method that combines various decision trees to increase the prediction's robustness and accuracy.
- Random sampling: Random Forest builds several decision trees from various subsets of data using random sampling with replacement, which lowers overfitting and raises generalisation error.
- Random feature selection: Random Forest divides each node of the decision tree using random feature selection, which further boosts diversity and lessens correlation between the trees.
- Voting system: To aggregate all of the decision trees' predictions into one, Random Forest employs a voting system (majority vote for classification, mean for regression).
- Hyperparameters: A few of the hyperparameters that can be changed in Random Forest to enhance model performance are the number of trees, the depth of each tree, and the bare minimum number of samples required to split a node.

3. **kNN:**

   For validation, kNN[25], model are employed. In kNN algorithm, k represents the number of clusters that model builds. The Elbow approach is used to estimate k's value. For example, if k is set to 7, the model creates 7 clusters and, using those clusters, predicts the value of the target labels.

   kNN is expressed

$$min_w||X^T W - Y||\frac{2}{F} + \ \rho_1 R_1(W) + \ \rho_2 R_2(W) + \ \rho_3 R_3(W)$$

In Eq., W∗ is the reconstruction weights, ρ is the normalization vector, X and Y denotes the set of training and testing data points and R represents each attributes of dimensionality d.

Here are some of the primary traits and qualities of kNN:

- Distance-based: The kNN algorithm exploits the separations between data points to generate predictions.
- Non-parametric: Because kNN is a non-parametric algorithm, it can handle both linear and nonlinear interactions and does not presume a certain data distribution.
- Lazy learning: Learning implicitly from training data is not possible using the kNN method, which is a lazy learning technique. Instead, it keeps track of all the training data and determines how far the test data and training data are apart at prediction time.
- Hyperparameter: The number of nearest neighbours to take into account while making a prediction is indicated by the hyperparameter k of the kNN algorithm.
- Regression: In the case of regression, a test data point's k closest neighbours are determined based on their distance, and the test data point's target value is determined by averaging the target values of the k neighbours.

4. **Logistic Regression:**

The logistic regression[26] model takes data factors as its foundation; it uses one attribute to determine its dependence on another attribute and then computes their values in a finite number of iterations to forecast the values of the target labels. Logistic Regression is expressed

$$\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

In Eq, x is the independent variable. The probability of recurrence and is the binary outcome of an attribute x.

Here are some of the primary traits and qualities of Logistic Regression:

- Model in linear form: Using a linear combination of input features, logistic regression may predict whether an input will fall into a particular class.
- Sigmoid function: The linear combination of input features is transformed by the sigmoid function used in logistic regression into a probability value between 0 and 1, which signifies the likelihood that the input belongs to the positive class.
- Decision boundary: The decision boundary of logistic regression is the threshold value above and below which an input is categorised as belonging to the positive class and the negative class, respectively.
- Maximum likelihood: Logistic regression looks for the parameters that get the predicted probabilities the closest to the actual class labels. This means it maximises the likelihood of the observed data under the model.
- Regularization: L1 or L2 regularisation can be used in logistic regression to increase generalisation performance and reduce overfitting.

5. Decision Tree:

In Decision Tree[27], Gini index is determined. After the evaluation of Gini index's value, decision trees use it to choose the root node, as the entire tree is built. The created tree provides us with the projected label values based on the output. Decision Treeis expressed.

$$Gini\ (D) = 1 - \sum_{j=1}^{n} p_j^2$$

In Eq, p_j^n is relative frequency of class j in D.

Here are some of the primary traits and qualities of Decision Tree:

- Tree-like structure: A decision tree is made up of nodes that each reflect decisions based on input features and branches that each represent potential values for those features.
- Root node: The topmost node of the tree, or the root node, serves to represent the whole input space.

- Leaf node: The output values or class labels are represented by the leaf nodes, which are the bottommost nodes of the tree.
- Splitting criterion: In order to maximise the purity or homogeneity of the output subsets, the decision tree method uses a splitting criterion to decide which feature and value to use to split the input space at each node.
- Multi-class classification: By utilising one-vs-all or one-vs-one techniques, decision trees can be expanded to address situations involving multi-class categorization.

6. **Extreme Gradient Boost:**

The XG Boost[28] ensemble model combines the loss functions of one or more models and then utilises the pooled training data to forecast the values of the target labels. This technique, known as gradient boosting, dramatically increases experiment accuracy. XG Boost classifier is given.

$$L^t = \sum_{i=1}^{n} 1 * l(y_i , \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

In Eq., $\hat{y}_i^{(t-1)}$i be the prediction of the i^{th} instance at the t^{th} iteration, we must add f_t in order to reduce the goal.

Here are some of the primary traits and qualities of XGBoost:

- A machine learning method called XGBoost is used for supervised learning tasks like classification, regression, and ranking.
- Extreme Gradient Boosting is an ensemble learning technique based on decision trees that combines a number of weak learners to produce a strong one.
- Machine learning contests frequently employ XGBoost because it is quick, scalable, and accurate.
- The technique is implemented in C++ and offers APIs for Python, R, Java, and Scala, among other computer languages.
- Decision trees are incrementally added to the model using XGBoost's gradient boosting architecture to reduce the loss function and boost prediction precision.

3.3 Hardware Description

Here are some of the hardware elements that have been used to test the water quality using different parameters:

1. **ESP 32 Microcontroller**:

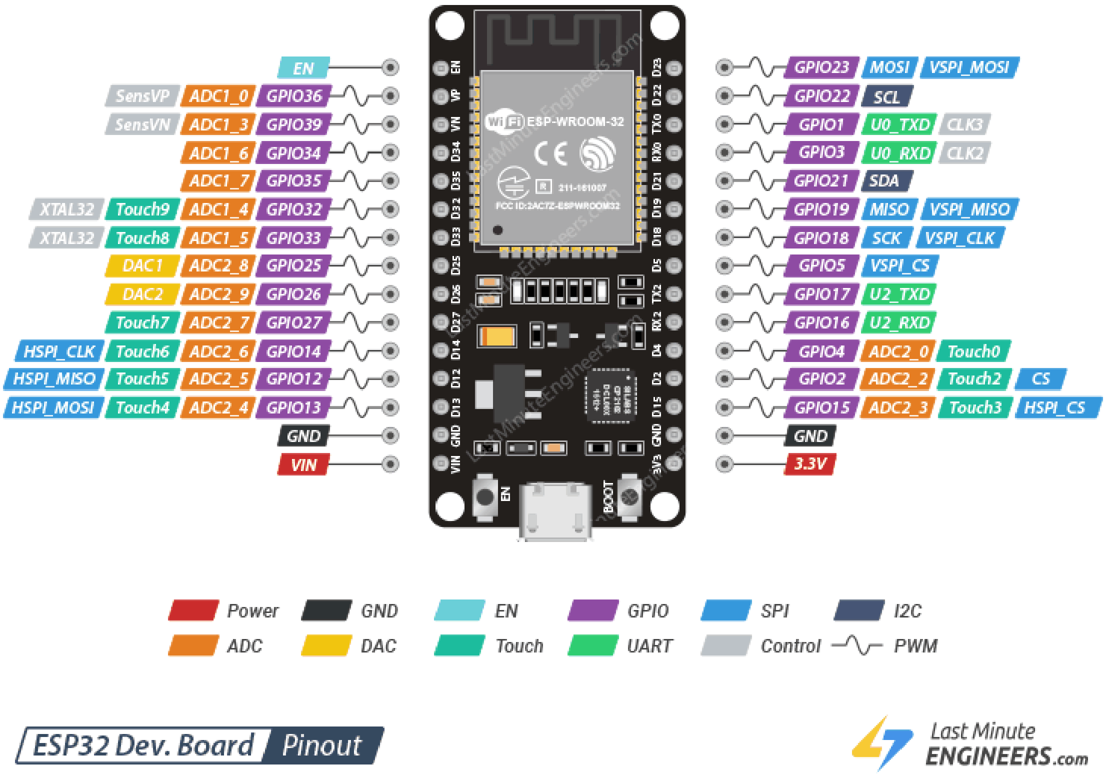


Fig 1. ESP32[29]

The ESP32 is a well-liked system-on-chip (SoC) microcontroller that is frequently used in Internet of Things (IoT) and other embedded systems applications. It is based on a chip called the ESP8266.

The ESP32 has the following salient qualities and features:

- Dual-core processor: The ESP32 has a dual-core processor with a maximum 240MHz clock speed. As a result, it can handle more complicated tasks and applications than the ESP8266.
- Connectivity via Wi-Fi and Bluetooth: The ESP32 has built-in Wi-Fi and Bluetooth features that make it simple to connect to wireless networks and other devices.
- Interfaces for peripheral devices: The ESP32 has a variety of interfaces for external devices, including GPIO, I2C, SPI, UART, and others. This makes it appropriate for a variety of uses and applications.
- Low power requirements: Because of its power-saving design, the ESP32 is perfect for battery-powered applications and other low-power use cases.

- Integral security features: The ESP32 has hardware-accelerated encryption and decryption, support for SSL/TLS encryption, and other integrated security measures.
- IoT devices: The ESP32 is frequently used in IoT devices, including wearables, sensors, and smart home gadgets.
- Robotics: The ESP32 is a viable choice for robotics projects and applications due to its robust CPU and peripheral ports.
- Industrial automation: The ESP32 is a great choice for use in industrial automation and control systems because to its low power consumption and integrated connectivity.
- Audio and multimedia applications: The ESP32 is appropriate for use in audio and multimedia applications since it supports digital signal processing (DSP) and other multimedia features.

2. TDS Sensor:

Fig 2. TDS Sensor [30]

The concentration of dissolved solids in a liquid is measured using an electrical device called a TDS (Total Dissolved Solids) sensor. The following are some of the main traits and qualities of TDS sensors:

- Principle of measurement: To determine the number of dissolved particles in a liquid, TDS sensors commonly employ electrical conductivity. A modest electrical current is used to do this, and the conductivity that results is measured.
- Range: Depending on the sensor, TDS sensors can measure concentrations at levels as low as a few parts per million (ppm) and as high as several thousand ppm.
- Calibration: Calibration is usually necessary before using TDS sensors. In order for the sensor to deliver an accurate measurement, the conductivity of a known solution must be measured.
- Temperature adjustment: To take into account temperature variations that may alter conductivity measurements, TDS sensors may have temperature adjustment.
- Type of probe: There are various probe types that can be used with TDS sensors, including immersion probes, flow-through probes, and in-line probes. Depending on the application and the liquid being tested, a certain type of probe will be utilised.
- Precision and accuracy: Depending on the sensor type and measurement range, TDS sensors' precision and accuracy can differ. Some sensors may be accurate to within +/- 2%, while others may be accurate to within +/- 5%.

- Application areas: TDS sensors are utilised in a variety of processes, such as industrial operations, food and beverage manufacturing, water treatment, aquaculture, hydroponics, and many other processes.
- Maintenance: To provide reliable readings, TDS sensors need to be regularly maintained. This can entail routinely calibrating the sensor and frequently cleaning the probe and sensor cartridge.
- Output: Depending on the type of sensor, TDS sensors can produce analogue or digital output signals. RS232, RS485, and USB are examples of digital output signals.
- Adaptability: Depending on the sensor type, TDS sensors may be adaptable to many controller types, including microcontrollers and PLCs.

3. Turbidity Sensor:

Fig 3. Turbidity Sensor [31]

A turbidity sensor is an electronic tool that assesses the concentration of suspended particles, such as solids or bacteria, in a liquid. The following list of turbidity sensors' main attributes and traits:

- Measuring principle: The number of suspended particles in a liquid is commonly measured using turbidity sensors using either light absorption or light scattering.
- Range: Nephelometric turbidity sensors can measure a wide range of turbidity levels, from a few NTU to several thousand NTU.
- Calibration: Turbidity sensors need to be calibrated before use. In order to get an accurate reading, the sensor must be adjusted after measuring the turbidity of a known solution.

- Sensitivity: Depending on the sensor type and the measuring range, turbidity sensors' sensitivity can change. Depending on the sensor, its sensitivity could range from 0.01 NTU to 1 NTU.
- Application areas: Turbidity sensors are employed in a variety of applications, such as water and wastewater treatment, environmental monitoring, and the manufacture of food and beverages.
- Probe type: Turbidity sensors may use immersion probes, in-line probes, or flow-through probes, among other types of probes. Depending on the application and the liquid being tested, a certain type of probe will be utilised.
- Maintenance: To obtain reliable readings, turbidity sensors need to be regularly maintained. This can entail routinely calibrating the sensor and frequently cleaning the probe and sensor cartridge.
- Output: Depending on the sensor type, turbidity sensors can provide analogue or digital readings. A few examples of digital output signals are RS232, RS485, and USB.
- Compatibility: Depending on the sensor type, turbidity sensors may be compatible with various controller types, such as microcontrollers or PLCs.
- Accuracy and precision: Depending on the sensor type and the measurement range, turbidity sensors' accuracy and precision can change. Some sensors may be accurate to within +/- 2%, while others may be accurate to within +/- 5%.

4. pH Sensor:

Fig 4. pH Sensor[32]

A pH sensor is an electronic device that gauges a liquid's acidity or alkalinity. The following are some essential pH sensor qualities and features:

- Measuring principle: The concentration of hydrogen ions (H+) in a liquid is commonly measured by pH sensors using a glass electrode or a combination electrode. The voltage differential between the electrode and a reference electrode is measured to accomplish this.
- Range: A liquid's pH can be measured with a pH sensor on a scale from 0 to 14. While values lower than 7 are considered acidic and higher than 7 are considered alkaline, a pH of 7 is considered neutral.
- Calibration: pH sensors must be calibrated before being used. The pH of a known solution must be tested, and the sensor must be modified, in order to obtain an accurate value.
- Temperature adjustment: pH sensors may have temperature adjustment to take into consideration temperature variations that may have an impact on pH readings.
- Type of probe: There are various probe types that can be used with pH sensors, including immersion probes, in-line probes, and flow-through probes. Depending on the application and the liquid being tested, a certain type of probe will be utilised.
- Maintenance: To get reliable readings, pH sensors need to be regularly maintained. This can entail routinely calibrating the sensor and frequently cleaning the probe and sensor cartridge.

- Application areas: pH sensors are employed in a variety of applications, such as water treatment, the manufacture of food and beverages, medical equipment, and environmental monitoring.
- Output: Depending on the sensor type, pH sensors can produce analogue or digital output signals. A few examples of digital output signals are RS232, RS485, and USB.
- Compatibility: Depending on the sensor type, pH sensors may be compatible with many controller types, including microcontrollers and PLCs.
- Accuracy and precision: Depending on the sensor type and measurement range, pH sensors' accuracy and precision can change. Some sensors may have a +/- 0.1 pH unit accuracy, while others may have a +/- 0.01 pH unit accuracy.
- Response time: Depending on the type of sensor and the measurement range, the pH sensor's response time can change. Some sensors might respond in just a few seconds, while others might take many minutes to respond.

3.4 Software Description

These software components were used to test the water quality for several metrics:

1. **Arduino IDE:**

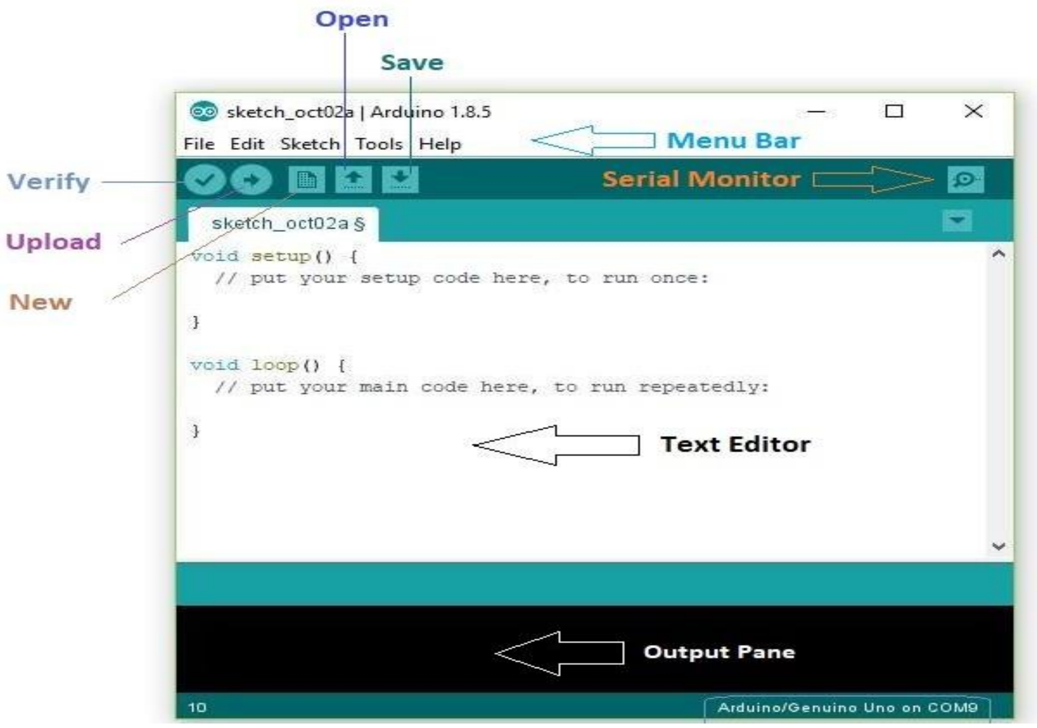


Fig 5. Arduino IDE [33]

Code for Arduino boards is created using the software tool known as the Arduino Integrated Development Environment (IDE). The Arduino IDE has the following salient qualities and features:

- User interface: Users can write, build, and upload code to an Arduino board using the Arduino IDE's user-friendly interface.
- Code editor: The integrated development environment (IDE) has a code editor with tools including syntax highlighting, auto-completion, and code folding.
- Library manager: This feature of the IDE makes it simple for users to add, remove, and manage libraries for their projects.
- Serial monitor: The IDE comes with a serial monitor that enables users to transmit and receive data from an Arduino board over a serial connection.
- Board manager: The IDE comes with a board manager that lets users manage and install board definitions for various Arduino boards.
- Sketchbook: The IDE comes with a sketchbook where users can store and arrange their code files.
- Support for several operating systems: The Arduino IDE is available for Windows, Mac OS X, and Linux.
- Open-source: The Arduino IDE is software that, with certain restrictions, allows users to edit and redistribute the code.

- Plugins from other developers: The Arduino IDE can be enhanced by third-party plugins from other developers to bring new features like code analysis, debugging, and visual programming.
- Support for many programming languages: The Arduino IDE has third-party plugins that enable programming in Python and JavaScript in addition to C and C++.
- Integration with other tools: To offer a full development environment, the Arduino IDE may be integrated with other tools like version control systems, debuggers, and simulators.
- Support from the community: To assist users in learning how to use the IDE, the Arduino community offers support and resources like tutorials, forums, and examples.

2. ThingSpeak (Cloud):

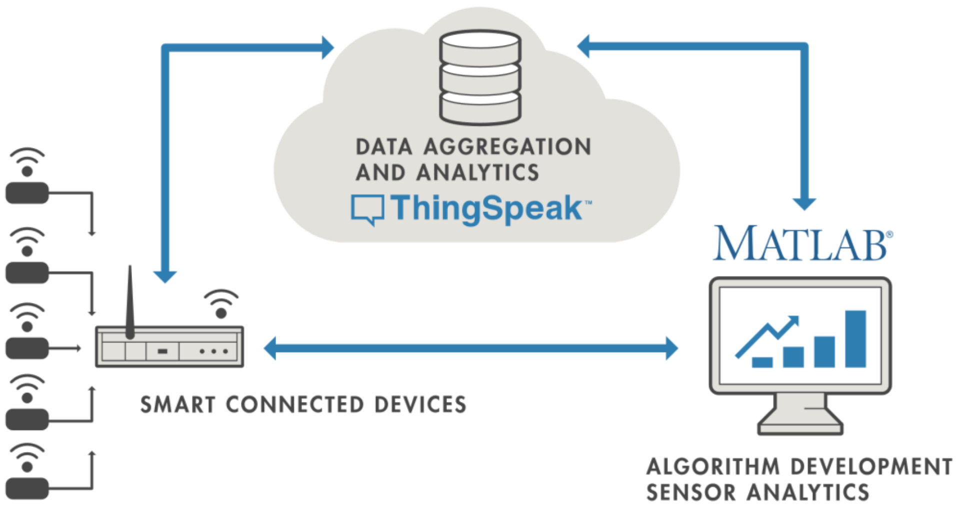


Fig 6. ThingSpeak[34]

A platform for Internet of Things (IoT) applications called ThingSpeak is hosted in the cloud. Following are a few of ThingSpeak's salient qualities and traits:

- Data gathering: With the help of ThingSpeak, users may gather and store data from IoT gadgets like sensors and actuators in the cloud.
- Data sharing: Using APIs (Application Programming Interfaces), ThingSpeak users can openly or privately share their data with other users.

- Integration with other platforms: To provide further functionality, ThingSpeak may be integrated with other platforms like MATLAB, IFTTT (If This Then That), and Twitter.
- Channels: Channels are containers for the data from a particular device or application, and ThingSpeak groups data into them.
- API: ThingSpeak offers an API that enables users to transmit and receive information from their channels as well as carry out additional actions like adding and removing channels.
- Widgets: To display real-time data from channels on websites or mobile apps, ThingSpeak offers widgets.
- Alerts: ThingSpeak users can configure alerts to be sent out when specific criteria are satisfied, such as when the temperature rises above a predetermined level.
- Security: ThingSpeak ensures secure connection between devices and the cloud by using HTTPS (Hypertext Transfer Protocol Secure). Additionally, it offers resources for administering authentication and access control.
- Community support: ThingSpeak offers a user base that helps newcomers to the platform by offering lessons, forums, and examples as well as assistance and resources.
- Scalability: ThingSpeak is appropriate for use in smart cities, agriculture, and industrial automation because to its capacity to scale up to handle enormous amounts of data from numerous devices.

3. Google Colab:

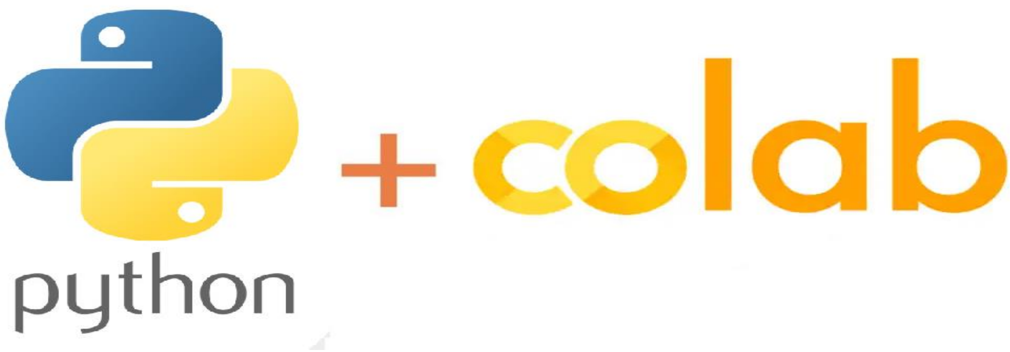


Fig 7. Google Colab[35]

A cloud-based platform for coding and data analysis, Google Colab—also called Google Colaboratory—is called Google Colab. The following list of Google Colab's essential qualities and features:

- Free to use: The user does not need to install or set up Google Colab on their computer in order to utilise it. Users can use their Google account to access Colab using their web browser.
- Editing together: Google Colab enables simultaneous editing by a number of individuals on the same notebook. Users can collaborate on projects and share notebooks with one another.
- Support for GPUs and TPUs: Google Colab gives users access to GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units), which can be utilised to speed up computations, especially for applications involving machine learning and deep learning.
- Integration with Google Drive: Users can save and access their notes and data files from Drive thanks to Google Colab's integration with Drive.
- Pre-installed libraries: NumPy, Pandas, TensorFlow, and PyTorch are just a few of the pre-installed libraries for data analysis, machine learning, and deep learning that come with Google Colab.
- Simple installation of additional libraries: Users do not need to install additional libraries on their local workstation in order to install them using conda or pip.
- Templates and code snippets: Google Colab offer templates and code snippets for typical activities, making it simple for users to get started with coding and data analysis.

- Computing in the cloud: Google Colab utilises the company's cloud infrastructure, which offers high-performance computing capabilities and enables users to run code that might be too computationally complex for their local workstation.
- Version control: Git version control is supported by Google Colab, allowing users to follow changes to their code and work together.
- Automatic save and backup: Google Colab automatically save and backups notebooks to make sure users don't lose their work in the event of an unanticipated malfunction or outage.
- Export options: Users of Google Colab have the ability to export notebooks in a number of different file types, including HTML, PDF, and Python scripts.

3.5 Flow Chart

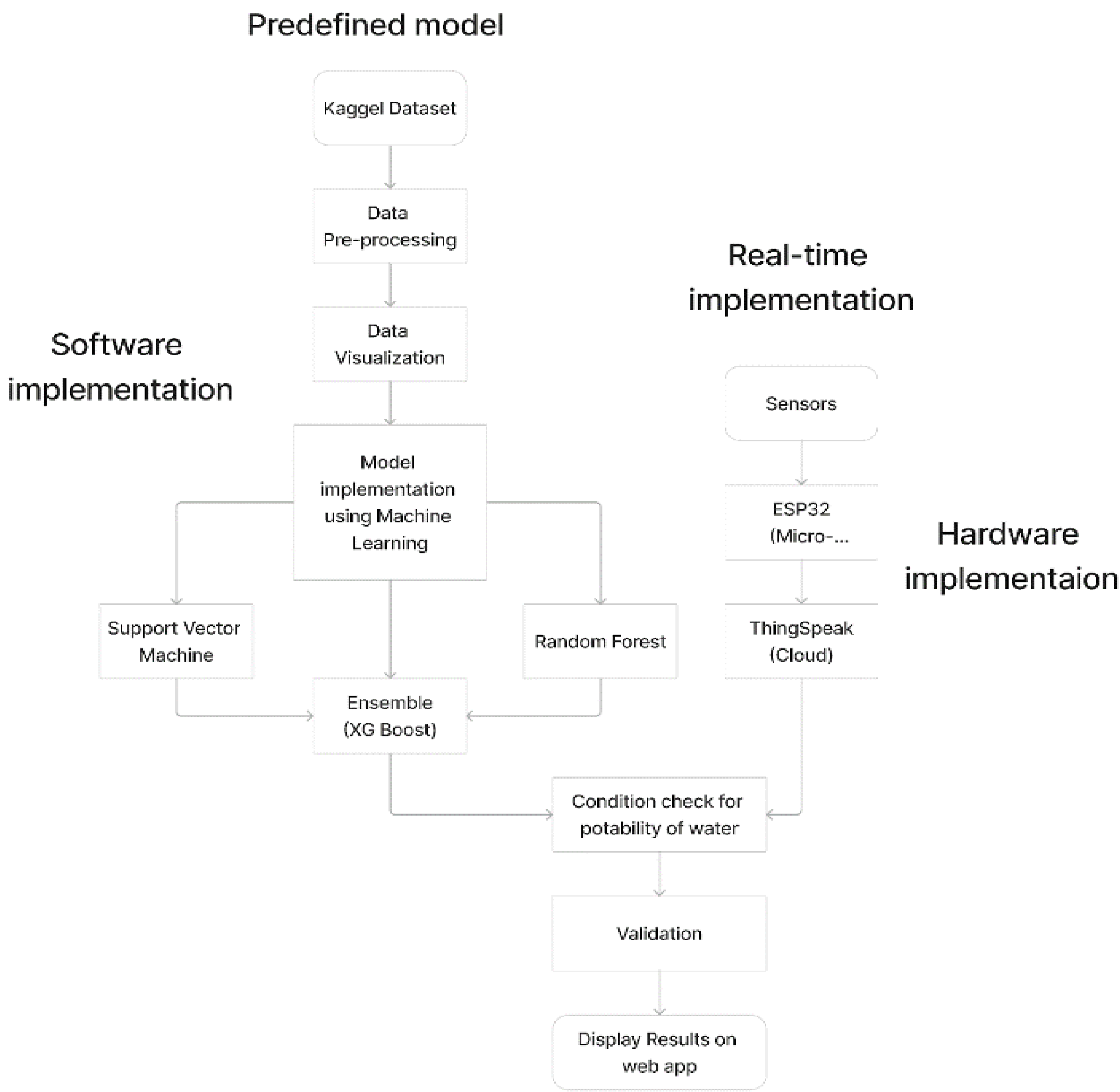


Fig 8. Flow Chart

The complete system of Water Quality Monitoring System for rural areas using IoT and Machine Learning has been illustrated in this flow chart. The proposed methodology is a combination of hardware and software. The flowchart of the entire proposed model is illustrated in Fig.8. For real time implementation, various samples from the wells and ponds of Richhana village, Solan was collected with the help of microcontroller (ESP 32) & sensors, pH and turbidity values were evaluated, which were validated using a model designed by Kaggle dataset [36]. Machine learning techniques were be applied to the data. SVM, RF and Ensemble method. utilising Extreme Gradient Boost were both used here.

CHAPTER – 4

RESULTS AND DISCUSSION

Water covers over 70% of the surface of the Earth. 97 percent of the water on Earth is found in the oceans. Freshwater makes up just 3% of the total. At the North and South poles, glaciers and ice have locked up nearly 2% of that meagre amount of freshwater. The world's lakes and rivers are filled largely with the 1% of freshwater that is left over, which is mostly groundwater. [37]. Therefore, for the quality check of water in rural areas, a model is proposed in this paper employing IoT and Machine Learning. The sensors are used by the microcontroller to collect data from the various water samples. After that, various machine learning techniques were applied to the sensor data.

4.1 Implementation using IoT

This project addresses the issues with water quality in rural locations where TDS, pH and turbidity have a substantial impact on the quality of the water. The model used for the implementation of Water Quality Monitoring System is depicted in Fig. 9 and is being used in real time with sensors for receiving and training purposes.

Fig 9. IoT implementation setup for detection of Water Quality

Fig.9 includes TDS, turbidity sensor, pH sensor, a microcontroller ESP32, and a glass. The different water samples were tested. The TDS, turbidity and pH values are shown in Fig 10, Fig 11 and Fig 12 respectively.

Fig.10 Turbidity Graph

Fig.11 pH Graph

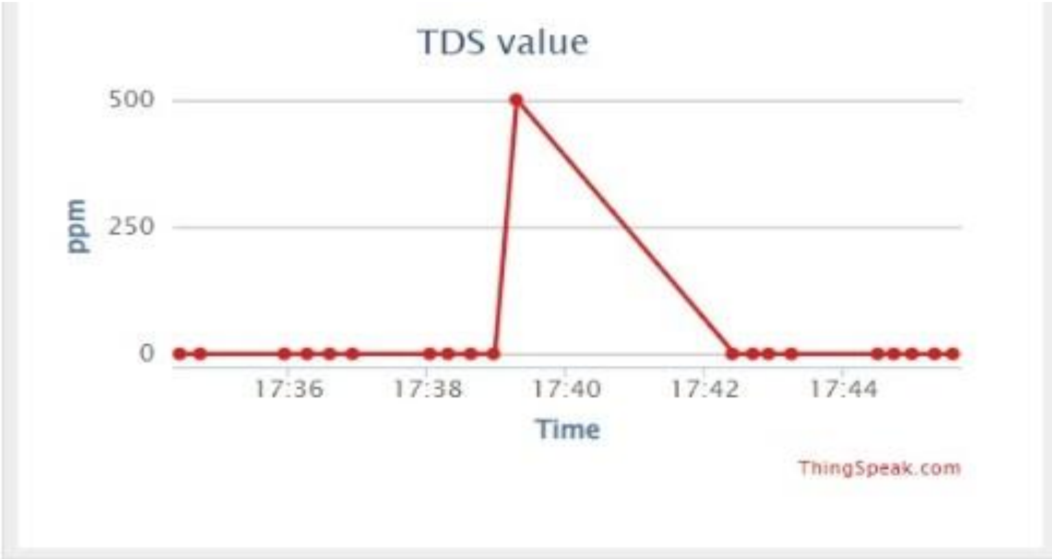


Fig.12. TDS Graph

Fig.10 displays various values obtained from turbidity sensor which shows that the values are in range which is safe for the consumption. Nephelometric Turbidity Unit, or NTU (ranging between 0 NTU – 1000 NTU), is the term used to describe the unit used to quantify a fluid's turbidity or the presence of suspended particles in water.

The readings were collected from the pH sensor. The values fluctuate between the permitted pH range for potable drinking are shown in Fig. 11.

Several values acquired from a TDS sensor are displayed in Fig. 12, which demonstrates that the readings are within a range that is safe for consumption. The TDS readings from the TDS sensor are shown in Fig. 12. The results vary within the acceptable TDS range for potable drinking water.

4.2 Implementation using ML

Depending on the test results, we may determine whether the water is drinkable or not depending on the values predicted by our proposed model. By importing the pre-trained model and then deploying it, we created a web application for our proposed model that presents real-time data about the water being tested together with potability values and indicates whether the water is potable or not [38].

The methods we employed in calculating the potability values, which are crucial for classifying and evaluating whether a body of water is fit for consumption. By applying the test values to a trained model, this is accomplished. The pre-training of the model was preceded by the data pre-processing, which included the elimination of extraneous data attributes. As a result, the model was trained exclusively on TDS, turbidity and pH. All null values that occurred in the data attributes were then eliminated.

The accuracy, F1 score and recall values of our proposed model is tabulated in Table 1.

Table 1 – Evaluation of water quality parameters of based on proposed methodology

Model	**Accuracy**	**F1 Score**	**Recall**
Random Forest	69.35	57.24	57.77
SVM	65.85	58.23	59.61
XG boost	95.12	94.74	94.06

The maximum accuracy of 95.12% is achieved using XG boost. Later for validation of our model, we have used logistic regression, kNN and decision tree. The evaluation parameters are tabulated in Table 2.

Table 2 - Comparison of the accuracy using different models

Model	**Accuracy**	**F1 Score**	**Recall**
Logistic Regression	64.17	40.51	81.98
Decision tree	65.58	52.26	55.55
Random forest	69.35	57.24	57.77
kNN (k = 7)	63.56	56.84	57.05
SVM	65.85	58.23	59.61

XG Boost	95.12	94.74	94.06

Table 2 compares the accuracy of SVM, kNN, Random Forest, Decision Tree, Logistic Regression & XG Boost. The models we utilised had high recall and precision along with F1 score. It demonstrates that of all the approaches, the XG Boost has the highest accuracy. Fig. 13 describes the XG Boost model as a Confusion matrix.

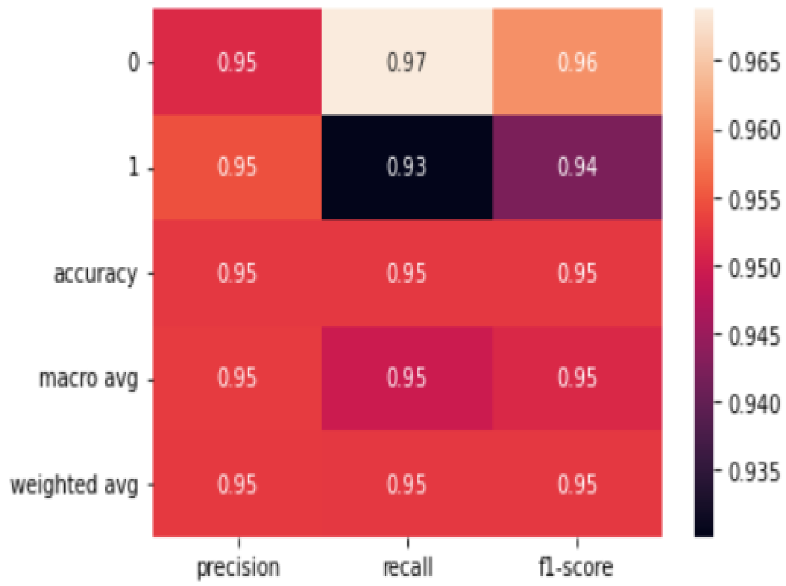
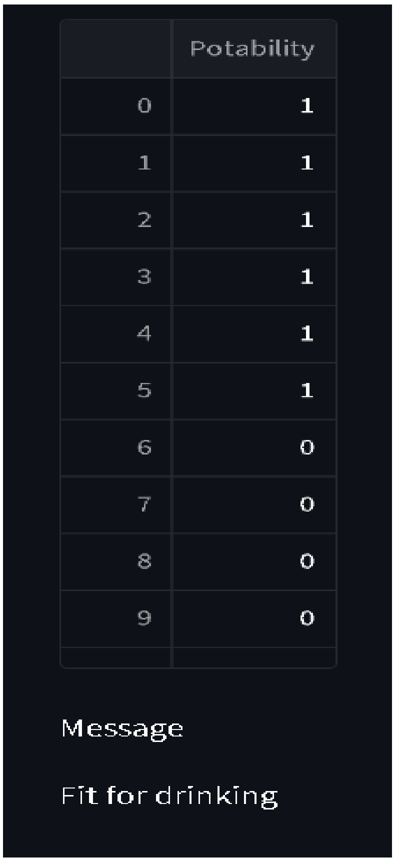


Fig 13: Confusion matrix of XG Boost model

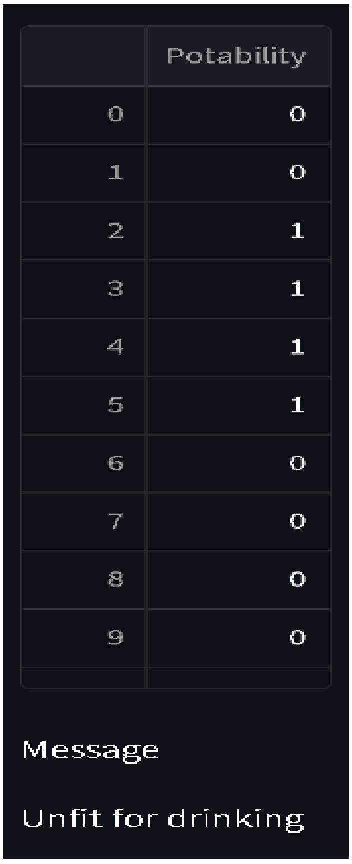
Fig.13 displays the plot of the confusion matrix, which includes accuracy, f1 score, recall, precision, and for each test label, which are 0 and 1. It is evident that the model effectively accounts for both test labels. We predicted potability values for 10 samples at a time interval of 0.5 seconds and whether the water is fit for drinking or not. Fig. 14 shows the real time implementation of our proposed model.



	Potability
0	1
1	1
2	1
3	1
4	1
5	1
6	0
7	0
8	0
9	0

Message

Fit for drinking



	Potability
0	0
1	0
2	1
3	1
4	1
5	1
6	0
7	0
8	0
9	0

Message

Unfit for drinking

(a) (b)

Fig. 14 – Potability: (a) Fit for Drink, (b) Unfit for Drink

Fig. 14 shows the predicted potability values for being tested is fit for drinking or not was calculated. The output displays two numbers, 0 and 1, respectively. Thus, 0 signifies that the water was unfit for human consumption, while 1 signifies that it was of good quality and fit for human consumption.

CHAPTER – 5

CONCLUSION

5.1 Conclusion

The potability of the water, which is determined by factors like turbidity, pH, and TDS, will be properly predicted by our suggested method, indicating whether the water is suitable for human consumption or not. In this project, we offer a model that fuses artificial intelligence with the internet of things. Multiple sensors and microcontrollers are utilised to assess the TDS, pH, and turbidity data that were validated since the model used machine learning techniques. More sensors can be added in the future to expand the range of properties (such temperature, pH, and dissolved oxygen) that our model can examine.

REFERENCES

[1] Drinking water Fact Sheet: https://www.who.int/news-room/fact-sheets/detail#:~:text=Contaminated%20water%20and%20poor%20sanitation,http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water%20Quality%20Monitoring%20Report%202005-06.pdf

[2] Ahref="http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/FILTRTAION%20PLANTS% 20REPOT-CDA.pdf

[3] Water Quality Monitoring Report for 2005-2006 at http://www.pcrwr.gov.pk

[4] www.undp.org/content/undp/en/home/presscenter/articles/2010/03/22/clean-water-for-a-healthyworld.html.

[5] Environmental-health/Aboriginal-environmental-health/Environmental-Health-Practitioner-Manual.pdf, available at https://ww2.health.wa.gov.au/media/Files/Corporate/general-documents.

[6] https://www.who.int/news-room/fact-sheets/detail/drinking-water#:~:text=Contaminated%20water%20and%20poor%20sanitation,individ.

[7] H. Amir, H. Nasrolahi, and A. Parsaie, "Water Quality Research Journal: Water Quality Prediction Using Machine Learning Methods," in WA Publishing, vol. 53, pp. 3–13, 2018.

[8] [8] U. Ahmed, R. Mumtaz, H. Anwar, A. Shah, R. Irfan, and Garc, "Effective water quality prediction using supervised machine learning," in MDPI, vol.11, pp. 10-22, 2019.

[9] [9] A.Najah, A.El-Shafie, O.Karim, H.Amr, and E.El-Shafie, "Application to artificial neural networks for water quality prediction," Neural Computing and Applications, in Springer, vol. 22, pp. 187-201, 2013.

[10] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," Chemosphere, Elsevier, vol. 171, 2020, pp. 126–169.

[11] K.Chen, H.Chen, C.Zhou, Y.Huang, X.Qi, R.Shen, F.Liu, M.Zuo, X.Zou and J.Wing, , Using several machine learning models based on big data to compare the effectiveness of predicting surface water quality and identify important water characteristics., *Water research, in Elsevier,* vol. 171, pp 114-454, 2020.

[12] Mehmood, S.; Ahmad, A.; Ahmed, A.; Khalid, N.; Javed, T. Drinking Water Quality in Pakistan's Capital City. Open Access Sci. Rep. 2013, 2.

[13] Huber and B. E. Reddy, "Using a hybrid ARIMA-ANN model with a moving-average filter, one can predict time series data.," Applied Soft Computing, vol. 23, pp. 27–38, 2014.

[14] X. Duda, N. Hu, Z. Cheng, and H. Zhong, "Recovery of vibration data using compressed sensing,," Acta Physica Sinica, vol. 63, no. 20, pp. 119–128, 2014.

[15] McGrane, "Using a multimedia two-model approach, assess non-point source pollution and river water quality.," Journal of Hydrology, vol. 409, no. 3-4, pp. 583–595, 2011.

[16]] J. Aziz, N. Liu, M. Wang, and K. Yan, "Application of the WASP model to the validation of the delineation of reservoir drinking water source protection areas," in 2010 3rd International Conference on Biomedical Engineering and Informatics, pp. 3031–3035, Yantai, China, October 2010.

[17] I. R. Koc and H. K. Bach, "MIKE 21: A system for modelling estuaries, coastal waters, and the oceans," Environmental Software, vol. 7, no. 4, pp. 229–240, 1992.

[18] D. F. Ruan, J. W. Labadie, T. G. Sanders, and J. K. Brown, "The improvement of water quality in hydropower system operations," Water Resources Research, vol. 34, no. 3, pp. 471–483, 1998.

[19] L. Rieger, C. Zhang, C. Hu, and G. Jiang, "Using the grey system to evaluate the quality of drinking water: a case study of Jiaozuo, China," in 2009 IEEE International Conference on Grey Systems and Intelligent Services.
[20] S. Mishra, A. Rajput, and G. Sanjeev, "Building a model for water quality using a decision tree technique," Journal of Data Mining and Networking in Binary, vol. 4, pp. 25–28, 2014.
[21] Kim, C. Yu, Z. Hu et al., "An accurate water quality prediction method for intelligent mariculture using a deep Bi-S-SRU learning network," IEEE Access, vol. 8, pp. 24784–24798, 2020.
[22] H. Liao and W. Sun, "Chao Lake's water quality is forecasted and evaluated using an upgraded decision tree method," Procedia Environmental Sciences, vol. 2, pp. 970– 979, 2010.
[23] V. Akkula, Tutorial on support vector machine (SVM),*School of EECS, Washington State University*, vol. 37, pp-3, 2006
[24] G. Biau, The Journal of Machine Learning Research, "Analysis of a Random Forests Model,", *in JMLR. org*, vol.13, pp 1063-1095, 2012
[25] S. Zhang, X. Li,M. Zong, X. Zhu and D. ChengACM Transactions on Intelligent Systems and Technology, Learning k for KNN classification *(TIST),ACM New York, NY, USA*, vol. 8, pp-1-19, 2017
[26] S. Lemeshow and D. Hosmer, Applications of logistic regression analysis to ophthalmic research in the American Journal of Ophthalmology, published by Elsevier, vol. 147, pp. 766-767, 2009.
[27] S. Kotsiantis, Recent information about decision trees from Artificial Intelligence Review, *in Springer*, vol. 39, pp. 261-283, 2013.
[28] T. Chen, T. He , M. Benesty and V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou and others, Extreme gradient boosting is known as Xgboost., *in Microsoft*, vol. 1, pp. 1-4, 2015.
[29] https://www.google.com/search?q=Esp32&sxsrf=ACEIAEMgUIABCABDIFcyeBwUaA:1683282028515&source=lnms&tbh=729&dpr=1.26
[30] https://www.google.com/search?q=tds&tbm=isch&ved=2ahUKEwiaqLid-t3-AhUsDbcAHXUECoAQ]s_lcp= PwXEdcA_yK2ScKhSK0JzDQYk-FCAA =1520
[31] AQrAAQE&sclient=img&ei=jdhUZNKaK4Wbz7sP6ICJ8A8&bih=729&biw=1520
[32] https://www.google.com/search?q=ph+sensor&tbm=isch&ved=2ahUKEwip2amx-t3-i1pbWc&sclient=img&ei=m9hUZOmfHem4z7sPq5Wn-Aw&bih=729&biw=1520
[33] https://www.google.com/search?q=arduino+ide&tbm=isch&ved=2ahUKEwjNjsbf-t3-AckEkgEHMC4xLjAuMZgBAKABAaoBC2d3cy13aXotaW1nwAEB&sclient=img&ei=_NhUZM2gGdeqz7sPrtqTiAQ&bih=729&biw=1520
[34] https://www.google.com/search?q=thingspeak&sxsrf=APwXEdfvF0p5M9o1sZ3J9CnAc9dEoPCAAQigUQ6gIQtIAccmkgEJMC4xLjQuOC0ymAEAoAEBsAEUyAEKwAEB2gEGCAEQARgB&sclient=gws-wiz-serp
[35] https://www.google.com/search?q=colab&tbm=isch&ved=2ahUKEwjE6r2D-93-AhUeGbcAHevOAGgQ2-cCegQIABAA&oq=colab&gs_lcp=CgNpb1nsAEKwAEB&sclient=img&ei=R9lUZISrL56y3LUP652DwAY&bih=729&biw=1520
[36] https://www.kaggle.com/datasets/adityakadiwal/water-potability.
[37] T. Aldhyani, M. Al-Yaari, H. Alkahtani, M. Mashael and others, Water quality prediction using artificial intelligence algorithms, *in Hindawi*, vol. 2020, 2020.
[38] M. Azrour, J. Mabrouki, G. Fattah, and A. Guezzaz and F. Aziz, Machine learning algorithms for efficient water quality prediction, *in Springer*, vol. 8, pp. 2793-2801, 2022

APPENDIX

1. **Import Function for the data received**

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
```

```
[ ] train_df = pd.read_csv("water_quality.csv")
    train_df
```

2. **Random Forest**

```
[ ] from sklearn.ensemble import RandomForestClassifier
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(x_train, y_train)
```

3. **Support Vector Machine**

```
[ ] rf_preds = rf_model.predict_proba(x_test)
    svm_preds = svm_model.decision_function(x_test)
```

```
[ ] svm_model = SVC(kernel='rbf', C=1.0)
    svm_model.fit(x_train, y_train)
```

4. **Extreme Gradient Boost**

```
[ ] import xgboost as xgb
    import numpy as np
    xgb_model = xgb.XGBClassifier()
    xgb_model.fit(np.column_stack((rf_preds[:, 1], svm_preds)), y_test)
```

PLAGRISM REPORT

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 08/05/2023

Type of Document (Tick): PhD Thesis M.Tech Dissertation/ Report B.Tech Project Report Paper

Name: Rohit Raj Department: Electronics and Communication Enrolment No: 191010

Contact No.: 9919912061 E-mail.: 191010@juitsolan.in

Name of the Supervisor: Dr. Rajiv Kumar

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): WATER QUALITY MONITORING SYSTEM

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 57
- Total No. of Preliminary pages = 51
- Total No. of pages accommodate bibliography/references = 6

Rohit. R

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found Similarity Index at 10.(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

ROHIT RAJ 191010

ORIGINALITY REPORT

10%	7%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	www.ncbi.nlm.nih.gov Internet Source	1%
3	link.springer.com Internet Source	1%
4	Submitted to Liverpool John Moores University Student Paper	1%
5	www.researchgate.net Internet Source	<1%
6	Submitted to Lovely Professional University Student Paper	<1%
7	Submitted to Texas A&M International University Student Paper	<1%
8	Submitted to Prospect High School Student Paper	<1%

9	Submitted to Rutgers University, New Brunswick Student Paper	<1%
10	Submitted to Florida International University Student Paper	<1%
11	Raavi Akshay, Gadiraju Tarun, Pinapothini Uday Kiran, K Durga Devi, M. Vidhyalakshmi. "Water-Quality-Analysis using Machine Learning", 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), 2022 Publication	<1%
12	Submitted to University of Teesside Student Paper	<1%
13	Juncheng Guo, Yafei Sang, Peng Chang, Xiaolin Xu, Yongzheng Zhang. "Chapter 16 MGEL: A Robust Malware Encrypted Traffic Detection Method Based onEnsemble Learning with Multi-grained Features", Springer Science and Business Media LLC, 2021 Publication	<1%
14	Submitted to University of Technology, Sydney Student Paper	<1%
15	www.ijert.org Internet Source	<1%

16	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
17	Submitted to Coventry University Student Paper	<1 %
18	Submitted to Karunya University Student Paper	<1 %
19	idoc.pub Internet Source	<1 %
20	Submitted to Gitam University Student Paper	<1 %
21	Submitted to Hillel Academy Student Paper	<1 %
22	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1 %
23	Submitted to The Robert Gordon University Student Paper	<1 %
24	dspace.univ-bouira.dz:8080 Internet Source	<1 %
25	nrl.northumbria.ac.uk Internet Source	<1 %
26	www.ameerpet.org Internet Source	<1 %

27	Submitted to Central Queensland University Student Paper	<1 %
28	Shazia Perveen, Amar-Ul-Haque. "Drinking water quality monitoring, assessment and management in Pakistan: A review", Heliyon, 2023 Publication	<1 %
29	Submitted to University of Queensland Student Paper	<1 %
30	ceur-ws.org Internet Source	<1 %
31	Submitted to CINEC Campus Student Paper	<1 %
32	S. Nashat, A. Abdullah, S. Aramvith, M.Z. Abdullah. "Support vector machine approach to real-time inspection of biscuits on moving conveyor belt", Computers and Electronics in Agriculture, 2011 Publication	<1 %
33	Submitted to University of Bolton Student Paper	<1 %
34	Submitted to South Bank University Student Paper	<1 %
35	d-nb.info Internet Source	<1 %

36	ebin.pub Internet Source	<1%
37	res.mdpi.com Internet Source	<1%
38	Submitted to City University of Hong Kong Student Paper	<1%
39	ijates.com Internet Source	<1%
40	seminole.wateratlas.usf.edu Internet Source	<1%
41	www.hindawi.com Internet Source	<1%
42	"Mobile Radio Communications and 5G Networks", Springer Science and Business Media LLC, 2023 Publication	<1%
43	M Campolo. "Water quality control in the river Arno", Water Research, 2002 Publication	<1%
44	npl971975.wordpress.com Internet Source	<1%
45	www.aimspress.com Internet Source	<1%
46	www.research-collection.ethz.ch Internet Source	<1%

47	Yingying Hu, Shuhua Yin, Weihua Xiao, Baodeng Hou, Xuelei Zhang. "Spatial scale effects of landscape patterns on water quality changes in drinking water source protection areas: a case study in the Guishui River Basin, China", Arabian Journal of Geosciences, 2022 Publication	<1 %
48	docplayer.net Internet Source	<1 %
49	dokumen.pub Internet Source	<1 %
50	dspace.iiuc.ac.bd:8080 Internet Source	<1 %
51	dyuthi.cusat.ac.in Internet Source	<1 %
52	mansci-web.uai.cl Internet Source	<1 %
53	repository.futminna.edu.ng:8080 Internet Source	<1 %
54	www.frontiersin.org Internet Source	<1 %