Revolutionizing Water Quality Monitoring: TheIntersection of Machine Learning and IoT for Enhanced Detection

Dr.Katta Sugamya
Assistant Proffessor:Department of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad,India

Anusha Bandaru
Undergraduate:Department of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad,India

Charitha Gajarla
Undergraduate:Department of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad,India

Manasa Bedadha
Undergraduate:Department of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad,India

***Abstract*—Water quality prediction is an important task in ensuring the safety and sustainability of water resources. With the increasing population and industrialization, the quality of water is becoming a major concern worldwide. Water quality prediction models have been developed to predict the concentration of pollutants in water bodies based on various factors such as temperature, pH, dissolved oxygen, and other chemical and physical parameters. These models use statistical and machine learning techniques to analyze historical data and forecast future water quality. The accuracy of water quality prediction models can be improved by incorporating real-time monitoring data, remote sensing data, and meteorological data. In this project Machine learning algorithms such as random forest and decision tree algorithms are used to find the potability. The use of pH, turbidity and TDS sensors help in real time sensing of data.In this project cloud plays a major role to collect and provide the data for analysis and prediction.**

I. INTRODUCTION

When it comes to pollution, global warming, and other environmental disasters, there is no safe drinking water for the world's pollutants in the 21st century. There are many obstacles to real-time water quality monitoring today, including limited water resources, an ever-increasing population, and so on and so forth. As a result, new methods are needed to monitor the water quality in real time. Some of the parameters determining the water quality includes:

- pH: The parameters determining the purity of the water in order to determine the amount of hydrogen ions present in a solution, the pH scale must be used. It shows if the water's pH is basic or acidic. Alkaline water has a pH greater than or equal to 7, whereas acidic water has a pH lower or equal to 7.The pH scale value varies in between 0 and 14. The permissible level for pH ranges between 6.5 and 8.5 for drinking water.
- Turbidity: A measure of how many invisible particles are in a given volume of water is known as Turbidity. As turbidity increases, the risk of diarrhea increases. Water is clean when the turbidity is low.
- TDS: This refers to total dissolved solids in water. TDS is the total amount of inorganic and organic compounds dissolved in water. TDS levels in drinking water are tolerable up to 500 ppm.
- Sulphate: It is a naturally occurring chemical in water, although excessive concentrations can have a laxative effect. The maximum allowable amount for drinking water is 250 ppm.
- Chloramines: Chloramines are a disinfectant that is used in water treatment to destroy dangerous microorganisms. Up to 4 ppm is considered appropriate for drinking water.
- Hardness: The presence of minerals in water, such as calcium and magnesium, is referred to as hardness. Scaling and appliance damage can result from high amounts of hardness.
- Organic Carbon: The total amount of organic matter present in water that can alter both the taste and smell of water is determined by its organic carbon content. Up to 2 ppm is the acceptable range for drinking water.
- Conductivity: Water's conductivity, especially is determined by the amount of dissolved minerals, provides a factor in how well it conducts electricity. Poor water quality may be proven by high conductivity levels.
- Trihalomethanes (THMs): When chlorine is utilized to disinfect water, THMs are potential side effects of the process. The health of an individual can be harmed by high levels of THMs

Due to rapid Urbanisation and Industrialisation water quality

has been worsen in a rapid rate that has resulted in various diseases. Majority of the water resources provide adequate water for household, irrigation, industries etc., where water quality has become major concern. The purpose of water consumption maybe consumption of water for daily usage, food/goods preparation in industries, irrigation, agriculture etc., that is mainly depend on oxygen levels, solids dissolved, pH etc. Water quality detection using quality sensors can be applied in many areas like:

- Homes – The water must be potable and meet the parameter criteria for human beings to consume it on a daily basis.
- Industries – To produce different kinds of goods in industries requires the usage of water with differing values in its properties.
- Agriculture – It varies water quality parameter values such as TDS (less than 450 mg/L) and varying conductivity and salinity features.
- Wild Life and Fisheries – Water must contain free Ammonia of 1.2 mg/L or less etc., which requires water quality prediction.

II. Literature Survey

Unlike the machine learning technique traditional way of monitoring water quality provides a high rate of accuracy.But,The traditional way takes longer time to produce output and also require more human power to gain results.The machine learning technique helps to analyze the data and produce quicker potability output.The IOT devices provide real time monitoring of data which helps to monitor day-to-day water quality.Articles on different techniques used for water quality monitoring and prediction can be found on vasrious platforms.Some of them being:

”Water-Quality-Analysis using Machine Learning,” a study on machine learning-based water potability prediction, was published by R. Akshay, G. Tarun, P. U. Kiran, K. D. Devi, and M. Vidhyalakshmi. To determine if the water is potable or not, the machine learning model is applied. Split the data into training and testing datasets, with the ratios being 8:2 respectively, and then run the model selection. These several classifiers that are considered include the Decision Tree,Support Vector Classifier (SVC), Random Forest, GaussianNB, and XGBoost. Data preparation is the process of transforming acquired data into a format that a machine learning algorithm can utilise. As a result, these structures are present in the prototype. For sentiment analysis, words or word clusters are extracted from tweets.

A study by M. Munara, N. Kumar and K. Shanmugam in paper ”Recommending IoT based Real-time Water Quality Monitoring System in Malaysia”, through specific sensor nodes, the IoT-based water quality monitoring system would enable real-time collecting of water parameters including temperature, pH level, and turbidity. People who are motivated to act will be informed in the appropriate manner. If the water parameters go over certain criteria, it will send an SMS (or email) to the appropriate authorities, and it will only allow users on the authorized list access to the system. They are all quite well-liked in the IoT development community. Therefore, the user will access the cloud server to obtain data. This method makes data transmission secure and puts less strain on the ESP8266 gadget.

”IoT Based Water Quality Monitoring System,” by A. Roy, S. Mukhopadhyay, and S. Roy, is another study on using IOT to monitor water quality.The major goal of this work is to monitor several water parameters, including pH levels, water turbidity, temperature, the amount of dissolved oxygen in a given volume of water, conductivity, etc. All sensor modules, including those for temperature, turbidity, pH, and water flow, are included in the sensing step. However, due to the analogue nature of all sent signals, these are transformed into digital values. The collected data is then delivered to the IoT cloud server. The Raspberry Pi 4 has an integrated wireless data transfer module that enables the microcontroller to wirelessly communicate data to the cloud during the cloud stage.

III. Data Analysing and Preprocessing

The dataset regarding water quality is taken from KAGGLE and then the data is pre-processed. In pre-processing we take only the required factors to analyze water quality and fill the null values with the median values. Then we split the data set to train and test with the Random Forest and decision Tree Algorithms. The real time sensed data is used to detect the water portability based on the trained criteria.

TABLE I
Description and Permissible range for Dataset variables

Parameter	*Description(in water)*	*Acceptable Range*
pH	basic or acidic nature	6.5 to 8.5
Hardness	Soap precipitation capacity	120 to 170 mg/L
TDS	Total dissolved solids	500 to 1000 mg/L
Chloramines	Chlorine level	upto 4 mg/L
Sulfate	Sulfate concentration	3 to 30 mg/L
Conductivity	Electrical conductivity	upto 400 µS/cm
Organic carbon	total organic carbon	2 to 4 mg/L
Trihalomethanes	Chemicals in treated water	upto 80 ppm
Turbidity	Light emitting properties	upto 5 NTU

a) Data Collection: : The water potability dataset was obtained from the Kaggle website’s machine learning repository. There are 9 variables in all, including the outcome, Potability, which has 3276 occurrences. Except for the response, all variables are numerical. The dataset has been split into two categories: 0 for non-potable and 1 for potable. Only 40% of the data is drinkable, whereas 60% is not. The dataset contains missing values that require specific treatment. Each variable has a range that can be deemed safe drinking water.

b) Data pre-processing:: In data pre-processing we first find out which parameter have null values and how many

count of number of rows having null values. We then fill the null valued parameter with mean or median or respective ideal values. Then in pre-processing plotting graphs and visual representation of the data will help us to understand and analyze the parameters. Figure provides the overview about

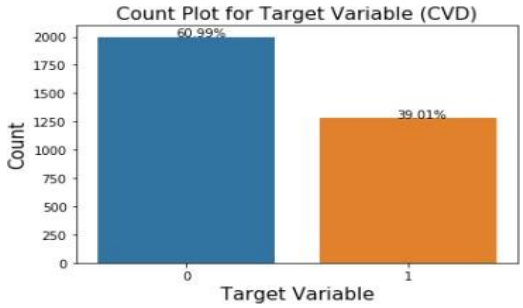

Fig. 1. Count plot for potability parmeter

the dependent parameter count i.e,. Potability parameter.The count of water being potable is 39.01% whereas water being non potable is 60.99%.This graphical representation helps us to thoroughly understand and analyze potability parameter.

c) Data Splitting:: In every machine learning algorithm, the dataset taken will be split into training and testing dataset. In this project, the dataset spilt ratio for training and testing is 8:2, where the total number of rows are 3276. In each of this split dataset we have two types of attributes one is independent attribute and the other is dependent attribute (final Output column).The training dataset will be used to train the machine learning model and the testing dataset will be used to find the accuracy in the prediction.

IV. METHODOLOGY

A. System architecture

The overall working of this system is based on Ph sensor,Arduino, Thingspeak,TDS Sensor, Turbidity Sensor, WI-FI Module, Thingspeak, Random Forest and Decision tree Algorithm.

A straightforward system design that will perform as indicated in Figure 2 serves as a demonstration of the whole project architecture. The method

will be done via a wireless connection for remote monitoring. The ESP8266 Wi-Fi module will first send the sensor data collection to the cloud server database. All calculations and monitoring data will thereafter be kept in the cloud. Therefore, the user will access the cloud server to obtain data. This method makes it possible to transmit data securely while putting less strain on the ESP8266 module.

The three sensors are connected to the Arduino Analog pins to fetch the sensed data. The Arduino is connected to WI-FI Module to provide internet. The Wi-Fi-Module uses TCP connection to connect with the Arduino. Code for collecting and transmitting the sensed data is written in the Arduino IDE. Once the Arduino collects the data it must send the data to the Thingspeak platform. Before sending the data, we use the write key of the Thingspeak platform to connect the Arduino

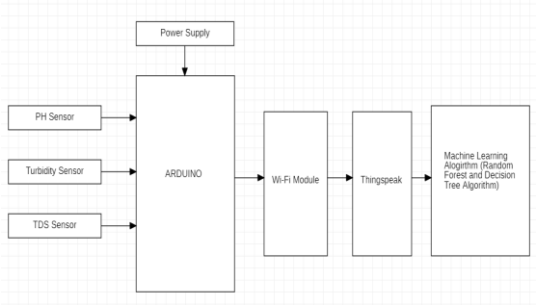

Fig. 2. Work flow diagram for monitoring water quality.

with Thingspeak. Once the connection is established the real time sensed data is transferred to Thingspeak. The Thingspeak platform is mainly useful to analyse the stored data.

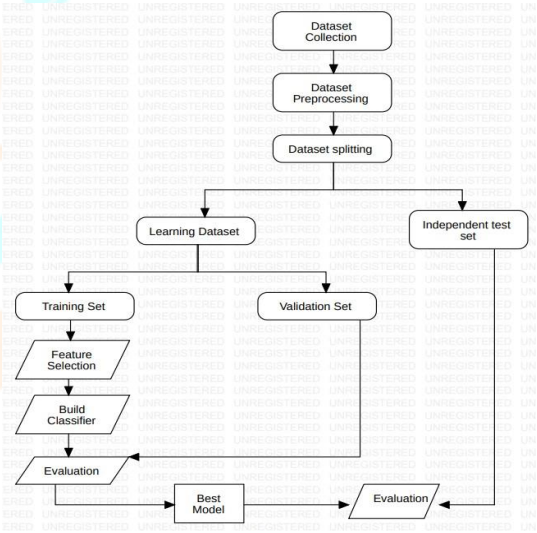

Fig. 3. Machine learning flow chart

Figure 3 gives a brief on how we build the machine learning model to predict the water potability based on the parameters i.e,. the data sensed.The algorithms are mainly based on the importance of features.Based on the feature importance the classification algorithms are built and evaluated with the already available dependent parameter.In this case the tenth parameter named potability is the validation set.The best model is taken based on the accuracy and used to evaluate real time sensed data for more precise and accurate results.

B. Algorithms used for Analysis

a) Decision Tree: :It is a supervised learning approach for decision assistance that employs a decision tree-like model. Applied to classification and regression issues. Typically, specialised data analysis approaches handle just one type of variable, while DT can handle both numerical and categorical data. It can also handle missing values automatically. Each tree has one parent node and one or more child nodes. When compared to traditional categorising approaches based on maximum livelihood theory concepts, decision trees have greater

advantages. The decision tree classification is non-parametric, and no assumptions about the distribution of input data are required. Finally, the structure and framework of the decision tree are basic and easy to understand. DecisionTreeClassifier (): To create a decision tree model using Python's Machine Learning framework, use the DecisionTreeClassifier () method. The DecisionTreeClassifier () function looks like this: DecisionTreeClassifier (criteria = 'gini', random_state = None, maximum depth = None, minimum samples leaf = 1)

Here are a few important parameters:

- criterion: In a decision tree classification, the split's quality is gauged using this criterion. It is 'gini' by default and also supports 'entropy'.
- max_depth: After the decision tree has been enlarged, this is done to give it the most depth possible.
- min_ samples_ leaf: Using the min samples _leaf argument, you can specify the bare minimum number of samples that must be present at a leaf node.

b) Random Forest: : A group of decision trees that come together to form one tree is Random forest.It is widely recognised as one of the finest classification algorithms for dealing with overfitting and producing more exact results. The supervised learning approach includes the well-known learning algorithm Random Forest. It can be used to address issues with regression and classification in machine learning. It is founded on the idea of ensemble learning, a method for merging many classifiers to address challenging issues and enhance model performance. As its name suggests, Random Forest is a classifier that employs many decision trees on various subsets of the supplied dataset and averages them to improve the accuracy of the dataset's predictions.In place of relying just on one decision tree, the random forest makes use of each decision tree's predictions and predicts the result based on the votes of the majority of projections.
Below are a few of the often-adjusted hyperparameters.

- N_estimators: The parameter N_estimators aids in determining the number of trees in the forest. The more trees there are, the more robust the aggregate model will be, but this will require more computing power.
- max_depth: The level count for each tree is limited by this parameter. The likelihood of taking into account additional features in each tree is increased by adding more tiers.
- max_features: This option allows us to limit the number of features that can be taken into account at each tree. bootstrap = This would allow us to choose whether to sample data points with replacement or without it.
- max_samples – This determines the proportion of data from the training set that should be used for training. Since the samples that are not utilised for training (out of bag data) can be used to evaluate the forest and it is desirable to use the whole training data set for training the forest, this parameter is typically left alone.

V. Result and Discussion

In this project we have obtained the result with IOT Sensors,Thingspeak and machine learning algorithms.Figure 4 displays the IOT setup where 3 sensors are used as a prototype.The 3 sensors are pH, tds and turbidity.All the sensors, lcd display, power supply and Wi-Fi module are connect to Arduino uno board.Through Wi-Fi module data is sent to cloud storage i.e,. Thingspeak.

Figure 5 provides a graphical representation about the real

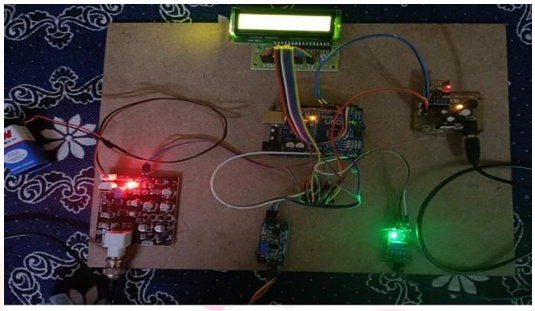

Fig. 4. IOT Environment for monitoring water quality

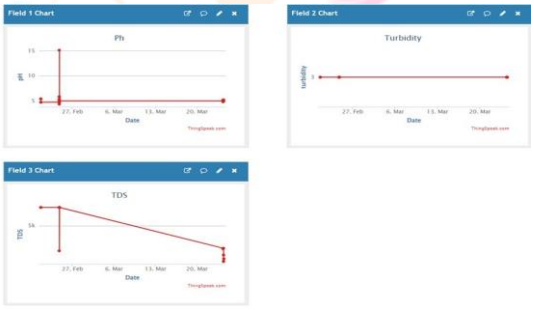

Fig. 5. Data Collected in ThingSpeak

time sensed data stored in cloud.In this project the cloud is Thing speak.The first graph is for pH i.e,. field 1.The second field graph is Turbidity and the third field graph being tds(total dissolved solids).

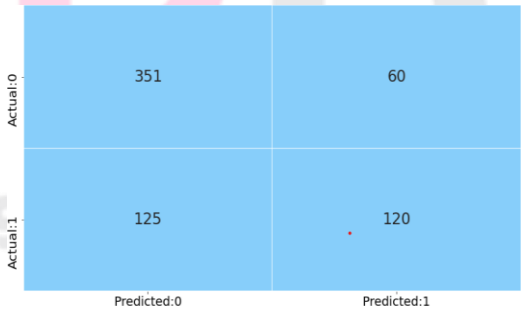

Fig. 6. Confusion matrix for Decision Tree Classifier.

Figure 6 denotes the confusion matrix for decision tree classifier.The four main elements in confusion matrix being

True negative with value 351,True Positives of value 120,False positives being 60 and False negatives being 125.

The Random forest algorithm is built using the importance

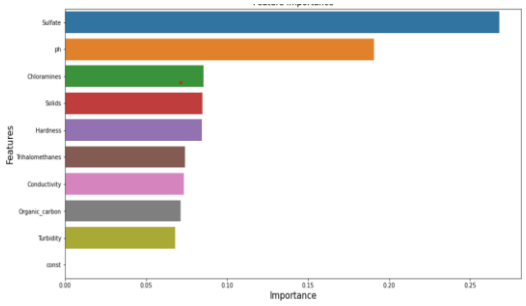

Fig. 7. Feature importance of parameters.

of features i.e,. importance of parameters taken in the dataset. The most important feature being Sulfate and least important being Turbidity.This importance of feature helps to gain output where the most important feature effects the most on result.And turbidity effects the least on result.Figure 7 depicts the bar graph for feature importance.

The above Confusion matrix(fig 8) is for random forest.Here

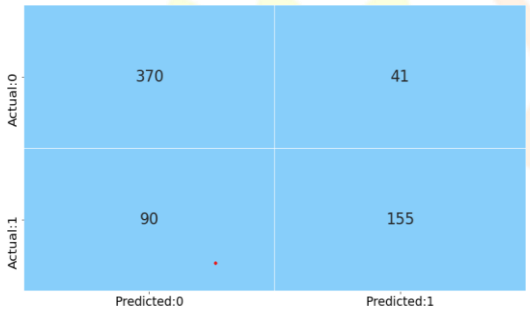

Fig. 8. Confusion matrix for Random Forest.

the true negative value is 370 while, true positive value being 155.The false positive value is 41 and false negative value is 90.

Conclusion

A water quality prediction system is used to monitor water turbidity, pH, and TDS etc.The technology is low-cost and does not necessitate the presence of a greater number of employees to keep track of water quality.Using this system to monitor other aspects of water quality such as dissolved solids, chloramines, organic carbons necessitate purchasing new sensors and rewriting the software.As a result, it can be used in numerous situations. Depending on the number of sensors the central controller is decided (for instance, for a prototype of three sensors Arduino Uno is suitable). It is possible to bring the environment to life through networking by putting sensors in the environment and install the software required in the system.

There are multiple ways to develop this project in future. Some of them being, sending an alert message to the user in different techniques such as SMS, mail, or thing tweet etc. The second one being Use of all parameters for predicting water potability. Since in this project there is only three factors taken into consideration but considering all factors helps to give better and accurate predictive solution.

References

[1] H. Yusuf, S. Alhaddad, S. Yusuf and N. Hewahi, "Classification of Water Potability Using Machine Learning Algorithms," 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 2022, pp. 454-458.

[2] A. Arora, V. Singh, M. K. Gourisaria and A. K. Jena, "Analyzing the Potability of Water using Machine Learning Algorithm," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2022, pp. 250-256.

[3] A. Roy, S. Mukhopadhyay and S. Roy, "IoT Based Water Quality Monitoring System," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1-4.

[4] F. B. Alam and F. T. Zohura, "Automated IoT-based water quality measurement and analysis tool," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1523-1526.

[5] R. Akshay, G. Tarun, P. U. Kiran, K. D. Devi and M. Vidhyalakshmi, "Water-Quality-Analysis using Machine Learning," 2022 11th International Conference on System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 13-18.

[6] A. Kadiwal, "kaggle," 25 April 2021. [Online]. Available: https://www.kaggle.com/adityakadiwal/water-potability.

[7] M. Munara, N. Kumar and K. Shanmugam, "Recommending IoT based Real-time Water Quality Monitoring System in Malaysia," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-5.

[8] A. Roy, S. Mukhopadhyay and S. Roy, "IoT Based Water Quality Monitoring System," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1-4.

[9] Thingspeak.com. (2017). Learn More - ThingSpeak. [online] Available at: https://thingspeak.com/pages/learn_more .

[10] E. Kuruvilla and S. Kundapura, "Performance Comparison of Machine Learning Algorithms in Groundwater Potability Prediction," 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), MANGALORE, India, 2022, pp. 53-58.