Enhanced Environmental Surveillance: IoT-Driven Classification of Water Pollution Levels using Machine Learning

Md. Saymon Ahammad[1], Sadia Akter Sinthia[2], Md. Nurul Afsar Ikram[3]
[1,2,3] Department of CSE, Daffodil International University, Dhaka, Bangladesh

Abstract

Water is the essence of life , and stands as the cornerstone of our existence. Accurate classification of water pollution levels is pivotal for effective environmental management and conservation efforts. This research delves into the critical task of accurately classifying water pollution levels, recognizing its pivotal role in environmental management and human well-being. Leveraging IoT technology and machine learning , we analyze a comprehensive dataset from various water sources. Encompassing 296672 entries with attributes like time , turbidity ,TDS and pollution category Clean, Polluted , Moderate Polluted. Our evaluation reveals exceptional accuracy rates across multiple algorithms , with Gradient Boosting , Random Forest and Decision Tree achieving 99.99% and SVM reaching 99.4%. Naive Bayes attained 97.76% accuracy. These findings underscore the urgency and efficacy of IoT-driven environmental monitoring , emphasizing the imperative of safeguarding our water resources for the sustenance of life and ecosystems.

Introduction

Water is essential for all kinds of life and may be found in many natural manifestations, such as the ocean, rivers, lakes , clouds , rain ,snow and fog [1].Technically, water that is chemically pure does not occur naturally for a significant duration. Water quality is an essential component of both public health and environmental health, as it supports biodiversity, maintains ecosystem integrity, and protects human life.The water we eat or use in municipal or industrial operations must adhere to particular specifications in order to ensure its quality. As an example the Environmental Protection Agency (EPA) has established legally binding thresholds for over 90 distinct substances that may be present in water [2]. These restrictions are essential to guarantee the preservation of drinking waters purity by preventing the presence of harmful substance that may lead to health problems or the emergence of waterborne illness.Physical , chemical and biological attributes of water comprise its quality , which dictate its appropriateness for a variety of applications-including agricultural,imbibing , industry and recreation.

The physical water quality matrix consists of many indicators. Indicators like turbidity and TDS are on of them. Turbidity is the cloudiness or haziness of a fluid generated by a huge number of individual particles that are usually undetectable to the human eye, similar to smoke in the air [3].TDS represents the concentration of dissolved particles or solids in water.TDS is made up of inorganic salts including calcium,magnesium.chlorides,sulfates and bicarbonates as well as many other inorganic compounds that readily dissolve in water. These indicators can be measured using sensors and other tools. This indicators are need to be measured as human are more likely to get gastrointestinal illness when drinking water is turbid.This is because pollution like viruses and bacteria can cling to suspended solids.These solids can interfere with disinfection procedures.Turbidity sensors detects fluid cloudiness by evaluating the light scattering or absorption caused by suspended particles. After calibration , the sensor is submerged in the sample, producing light and measuring its interaction with particles to calculate turbidity.This data is then

utilized for monitoring or control , with regular maintenance and quality checks to ensure correction[4]. TDS sensors work by measuring the electrical conductivity of water, which varies depending on the presence of dissolved ions and particles. When submerged in water, the sensor monitors conductivity ,which increases in tandem with dissolved solids. This rise happens because dissolved solids have electric charges. The sensor estimates the TDS content in the water by examining its conductivity, which is commonly quantified in parts per million (ppm) or milligrams per liter (mg/L). These sensors are critical in a variety of industries, including agriculture , aqua culture , industrial operations and environmental monitoring , where accurate water quality evaluation is required for peak performance and regulatory .In this case machine learning algorithm assess water quality by extracting relevant features from turbidity and TDS sensing data, preprocessing it selecting appropriate models such as Gradient Boosting,Random Forest,Decision Tree,SVM ,Naive Bayes , training model on historical data , evaluating its performance and deploying it for real time monitoring [5]. These algorithms enable proactive control of water quality by detecting patterns and correlations in sensor data, allowing for timely actions to maintain safe and clean water supplies.

Understanding the necessity of water quality monitoring in this research, we want to make a device using some IoT components for collecting real time data. Our main goal is to classify the water pollution levels. For that, we will collect water from many sources and check or measure the contaminants mixed in the water.

- We have selected two features of water that are able to indicate the water pollution level; TDS and Turbidity . To collect data on those two features, we have selected two sensors TDS Sensor ,Turbidity Sensor.
- We are using an Arduino Uno R3 for integrating all sensors and devices. After designing the whole device, we moved forward with collecting data, and when we are done with data collection, we have moved forward to data preprocessing, followed by the analysis of the dataset using machine learning algorithms.
- The collected data set has 296672 data points which are collected from many water sources.
- We have applied four machine learning algorithms and achieved good results.

Through the utilization of machine learning algorithms, turbidity and TDS sensing data can be efficiently employed to evaluate the quality of water across a range of applications. This empowers proactive management and intervention strategies to guarantee the provision of safe and hygienic water supplies.

Literature Review

Advancement in sensing technologies, including embedded and wearable devices have facilitated the development in water monitoring systems. Recent research literature offers critical assessments that align with our investigation.

Elvin et al. (2024) applied machine learning methods and hyperparameter optimization to enhance the water Quality Prediction process. Researchers utilize extreme Gradient boost (XGBoost) algorithms and also evaluate other algorithms like random forest classifier, decision tree classifier, adaptive boosting classifier, support vector machine, Naive Bayes and extra tree classifier for comparison. Dataset retrieved

from the Kaggle site 21 input attributes—chemical components found in the water area—were included in the total 8,000 data points. The performance of the XGBoost model was superior , achieving 97.06% accuracy. Also Recall is at 81.5%, F1-score is at 87.4%, and precision is at 94.22% [6]

Siahaan et al. (2019) presents the development of a turbidity measuring device and its testing using ink water samples. The device accurately determines turbidity levels by correlating the voltage values with NTU values. Results show the device effectively measures turbidity , with clean water registering close to NTU and dark water near 500 NTU. The equation (y= -11.25x+506.67) is used to convert voltage to NTU values.Overall , the device proves capable of assessing water turbidity levels accurately and can be utilized for practical applications. Here environmental factors impacting turbidity measurements in natural water bodies are not considered , raising questions about the devices applicable in practical settings [7] .

Mohiddin et al. (2020) presents a monitoring system employing pH , turbidity and temperature sensors with an Arduino UNO as the main controller. It continuously evaluates water quality , transmitting data to an Excel Sheet and displays portability status on an LCD screen , ensuring the safety of purified water for consumption. The study presents one hour monitoring data of Acid , Distilled water , Base solutions. The result indicates acid nature for lemon water (ph 4.8) , base for sodium hydroxide (ph 8.4) and acceptable levels of ph 7.8 , turbidity (87) and TDS (0.7) for distilled water , facilitating low - cost , portable water quality assessment in rural areas. In this process short duration of monitoring is applied which limits the assessments of long term water quality trends [8].

Feng et al. (2020) present a water quality monitoring system utilizing MCU and Bluetooth technology centered on an Arduino development board with sensors for ph , turbidity conductivity, and temperature. Data is transmitted to smartphones via Bluetooth, alerting users to abnormal parameters. test results confirm the system timely and accurate measurements of water quality parameters, demonstrating stable performance suitable for various monitoring scenarios. The measured values for temperature, ph , TDS and turbidity were compared to actual values. Temperature had a 0.4% error , ph had error of 3.3 % for 3.37 ph and 0.9% for 9.08 ph .TDS had errors of 7.3% for 278 ppm and turbidity errors of 3.3% for 386.66 NTU. By analyzing this study, limitations found like variables such as ambient temperature fluctuations , variations in light intensity impacting reading or interference from nearby sources could potentially affect the accuracy and reliability of the measurements obtained by the monitoring system [9].

Pantjawati et al. (2020) , developed an IOT system for real time monitoring of Citarum River water quality , aiming to analyze it against WHO standards , monitoring pH, turbidity and TDS two points along the river, result indicates significantly altered water quality post-factor sewer , with pH dropping from 5.281 to 2.435, turbidity from 1118.768 NTU to 900.65 and TDS from 134.44 ppm to 247.625.Despite TDS levels below standards , the overall water quality remains inadequate , emphasizing the system’s utility for continues monitoring and assessment. By studying this paper gpas or limitations found like monitoring only two points along the Citarum River may not provide a comprehensive understanding of water quality variations throughout the entire river system [10].

Bineet Kumar et al. (2020), distributed his work, which was mainly divided into two phases for the water quality monitoring system. Phase 1 is Sensing groundwater from different areas through overhead tanks based on parameters like PH , Water level, turbidity, conductivity, and TDS . Phase 2 is sensing the data store on Cloud and using Machine learning algorithms to Monitor Water Quality and create a framework for showing predicted Water Quality. In this research , CART (Classification and Regression Trees) performs the best among other algorithms, achieving the highest accuracy level of 98.67%. In the near future, decision tree algorithm research's scalability, efficiency, and real-time capabilities may be

expanded through data stream processing within the Spark framework. Limitations of this work is the focus on sensing groundwater from overhead tanks, potentially neglecting other water sources and quality parameters , thus limiting the comprehensiveness of water quality monitoring [11].

Mohammad Salah Uddin et al. (2019) Implement a system that monitors real time river water Quality by analyzing threshold value and based on it displayed comments “Good” or “BAD”. Monitoring Real time river Water Quality displaying the result through sensed pH, temp, turbidity, and ORP values. It continuously senses the values of pH, temp, turbidity, and ORP and the resulting values are displayed. Analysis through Spark MLlib,Deep learning neural network models, Belief Rule Based (BRB) system and was also compared with standard values. Due to budget issues it can't be used in Large scale or local area water Monitoring Systems [12].

Varsha et al. in (2021) proposed a cost effective and efficient IoT based smart water Quality monitoring system. Using IoT devices and sensors , tested many water samples based on 6 parameters : PH , Turbidity , Conductivity , Carbon dioxide , Humidity and Temperature to collect the data .Data stored on cloud. Lastly , this developed model is tested with 3 water samples based on those 6 parameters to classify if the water is drinkable or not . Due to some limitations, it can't use many advanced sensors and also can't add immediate response through SMS notification [13].

Umair et al. (2020) discussed many traditional methods and found best solutions to monitor water quality through Internet of Things (IoT) and machine learning to detect anomalies . Implementing IOT based technology based on these parameters PH, Turbo Temperature, Cl, EC, DO, TH, TSS, TDS, BOD, COD, FC, and TC analyze the Water Quality. By sensing parameters measurements data stored on cloud and later it is used to build the model.Also there is a dashboard on the system that visualizes the data of water Quality. In anomaly detection , KNN algorithm has been used to classify water Quality into three classes respectively class 0 , class 1 and class 2 . Research showed a weakness in the approaches for using machine learning.May determine water quality using less parameters, which is easily implementable in an inexpensive IoT system with the fewest possible parameter sensor numbers [14] .

A. J. Ramadhan et al. (2020) proposed system is based on Wireless Sensor Network (WSN) and Internet of Things (IoT) to perform Supply Water Monitoring in Najaf,Iraq and issue timely warning through SMS and Emails. Data Collected from five different Water Stations in Najar,Iraq. This System Monitor water Quality based on parameters are water pH level, Temperature,nitrate, chloride, and dissolved oxygen concentration; turbidity; oxidation-reduction potential (ORP); conductivity or total dissolved solids (TDS) and sodium content. Test results obtained at the five stations based in Najaf revealed poor quality of drinking water coupled with use of inefficient water-purification systems. Due to political issues and tight regulation to buying electrical components, this proposed system affordability [15].

Sathish et al. (2020) proposed system consists of several sensors to measure various parameters such as pH value, the turbidity in the water, level of water in the tank, temperature and humidity of the surrounding atmosphere to monitor the Quality of water. Sensing Value store on cloud by using IoT based Thingspeak application to Monitor the Water Quality based on those parameters and The experimental setup consists of an MCU with a sensor network that takes samples for every 10s from the water storage tank and the parameters are displayed on the Arduino IDE serial display. Later Store on Thingspeak server for monitoring Real Time Water Quality. Work Limitations is that, more parameters like electrical conductivity, free residual chlorine, nitrates, and dissolved oxygen in the water can be used to carry out best results [16].

Nasir et al. (2022) evaluated an AI algorithm to develop an approach for producing consistently accurate water quality classifications and predictions. Water quality data was collected from several Indian states

in between 2005 to 2014. A total of 1679 samples were tested based on dissolved oxygen (DO), pH, conductivity, biochemical oxygen demand (BOD), nitrate, fecal coliform, and total coliform parameters.The CATBoost model achieved the highest accuracy (94.51%) for individual models, and stacking ensemble models achieved 100% accuracy.Furthermore, for more effective water monitoring, additional classifiers, neural networks, and other AI approaches can be applied to the datasets [17].

Table 1. Summary Table of Previous Works

Author and Year	Dataset	Best Model	Best Accuracy	Limitations
Elvin et al. (2024)	Kaggle Dataset	XGBoost	97.06%	The authors used a very small dataset.
Siahaan et al (2019)	Ink Water Samples	-	-	Lack of consideration for environmental factors impacting turbidity measurements in natural water bodies
Mohiddin et al. (2020)	one hour monitoring data of Acid, Distilled water , Base solutions	-	-	Short duration of monitoring limits assessment of long-term water quality trends
Feng et al. (2020)	-	-	-	Ambient temperature fluctuations, variations in light intensity, and interference from nearby sources may affect measurement accuracy and reliability
Pantjawati et al. (2020)	-	-	-	Monitoring only two points along the river may not provide a comprehensive understanding of water quality variations throughout the entire river system
Bineet Kumar et al. (2020)	Ground Water	CART	98.67%	Neglect of other water sources and quality parameters, potentially limiting the

				comprehensiveness of water quality monitoring
Mohammad Salah Uddin et al. (2019)	River Water	-	-	Budget constraints limit usage in large-scale or local area water monitoring systems
Varsha et al. in (2021)	-	-	-	Inability to use advanced sensors and lack of immediate response through SMS notification
Umair et al. (2020)	-	KNN	-	Potential limitation in determining water quality using fewer parameters, which may limit implementation in inexpensive IoT systems with fewer sensor numbers
A. J. Ramadhan et al. (2020)	-	-	-	Affordability issues due to political issues and tight regulations on purchasing electrical components.
Sathish et al. (2020)	-	-	-	Lack of additional parameters such as electrical conductivity, free residual chlorine, nitrates, and dissolved oxygen, which may affect the accuracy of water quality monitoring.
Nasir et al. (2022)	Indian States Data from 2005-2014	CATBoost	94.51%	None specified, but additional classifiers, neural networks, and other AI approaches are suggested for more effective water monitoring.

Gap Analysis : The gap analysis of existing research on water quality monitoring systems identifies various areas for improvement in terms of efficacy and applications.Existing studies sometimes disregard environmental elements such as temperature swings and light intensity variations,which are critical for accurate assessment . Furthermore , there is a tendency to focus on short term monitoring rather than long term evaluation.Scalability and real time capabilities remain areas for improvement , as does the need for more broad dataset coverage and analysis Most of the previous work is working with less amount of data.

Methodology

This part will provide an overview of the entire study approach, from data collecting to data preprocessing, data balancing, train-test split, machine learning models, evaluation, and result analysis.

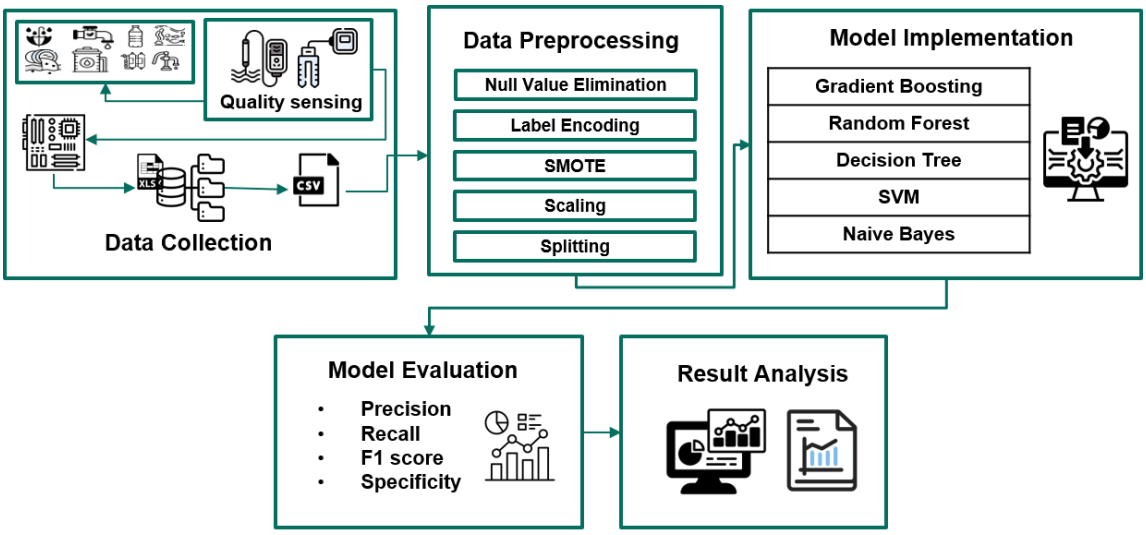


Figure 1 : Proposed Methodology

Data Collection Process

As we are working on classification of water pollution levels, at the beginning of the project, we want to make an IoT device using some IoT components for collecting real time data. Our main goal is to classify the water pollution levels.

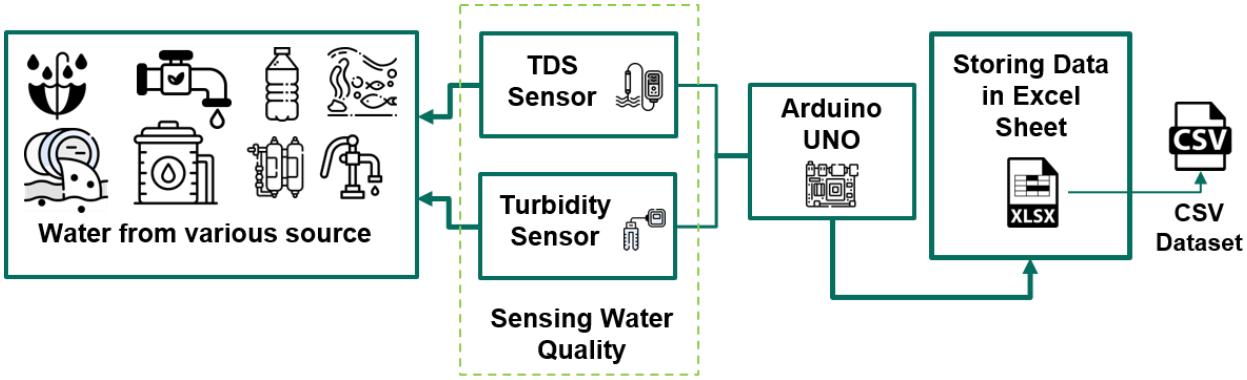


Figure 2 : Data Collection Process

- **IoT Tools Required For Data Collection**

  Arduino Board which is a microcontroller Platform used for building and prototyping. Turbidity Sensor to measure the clones or haziness of liquid , indicating the level of suspended particles of solids in the water. TDS Sensor to measure the concentration of dissolved solids in water. Electronic Breadboards are reusable solderless devices to build or to test electronic circuits. Cables like Female-to-male , Male-to-male cables are used in the projects among Arduino , sensors and breadboard. To store the sensing data we use Excel Sheets for data storing from Arduino UNO inputs.

  A. Internet of things (IoT) devices

  1) Arduino Uno R3 : The Arduino Uno R3 is a popular microcontroller board known for its versatility, simplicity, and robustness in a wide range of applications. Built around the Atmega328P microcontroller, it has 14 digital input/output pins, 6 of which support PWM capabilities, as well as 6 analogue input pins, allowing for smooth interfacing with a variety of sensors and peripherals. With USB programming and power connectivity, as well as compatibility with a wide range of shields and sensors, the Uno R3 provides an accessible platform for both beginners and specialists to construct new projects ranging from basic LED control to sophisticated IoT solutions. Its modest form factor, combined with the rich Arduino environment, which includes libraries, tutorials, and an active community, renders it an indispensable tool in the world of embedded systems. The Arduino Uno R3 can be powered by an external power source or a USB connection. It operates best with input voltages between 7 and 20 volts, however 7 to 12 volts is the suggested range. 32 KB of flash memory, 2 KB of SRAM, and 1 KB of EEPROM are available on the ATmega328P microcontroller for storing code and non-volatile data, respectively. The board has an integrated USB-to-serial converter for connectivity, making it simple to communicate with a computer for data exchange and programming. For firmware reprogramming, it also has a 6-pin ICSP (In-Circuit Serial Programming) header. [18]

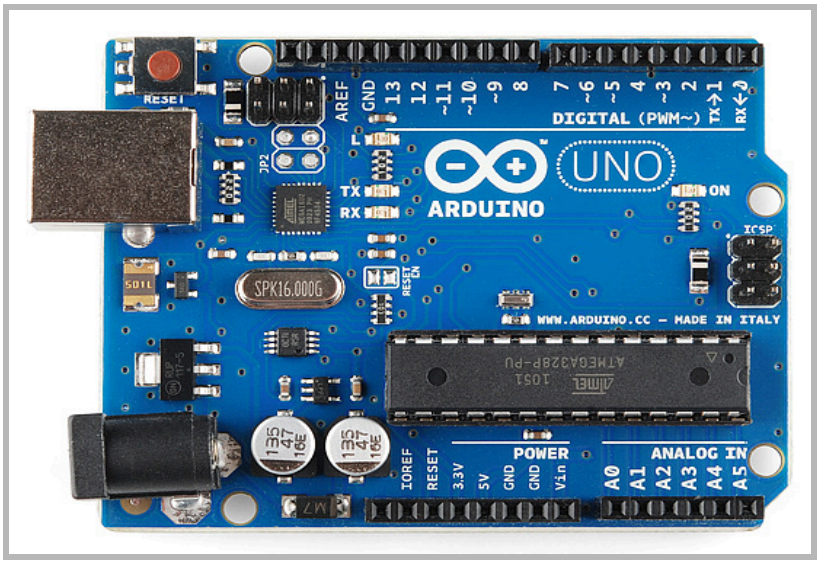


Figure 3 : Arduino Uno R3

2) Turbidity Sensor : Turbidity refers to the cloudiness or haziness of a fluid caused by tiny particles that are not visible to the naked eye, similar to smoke in the air. Turbidity testing is vital to determine water cleanliness. Turbidity in water is caused by suspended or dissolved particles that scatter light, creating a foggy or murky appearance. Particulate content includes sediment, organic and inorganic matter, soluble organic compounds, algae, and microorganisms. Turbidity is detected using specialized optical equipment. A water sample is illuminated, and the quantity of light scattered is measured.Nephelometric Turbidity Units (NTUs), which appear in a number of variations, are the unit of measurement. The amount of turbidity increases with light dispersion. High values of turbidity suggest low water clarity, while low values indicate high water clarity.Turbidity sensors typically have set of pins like VCC pin to supply 3.3v or 5v power depending on sensors specification, GND ground pin, SIG(Signal pin) , TX (transmit pin) , RX (receive pin). [19]

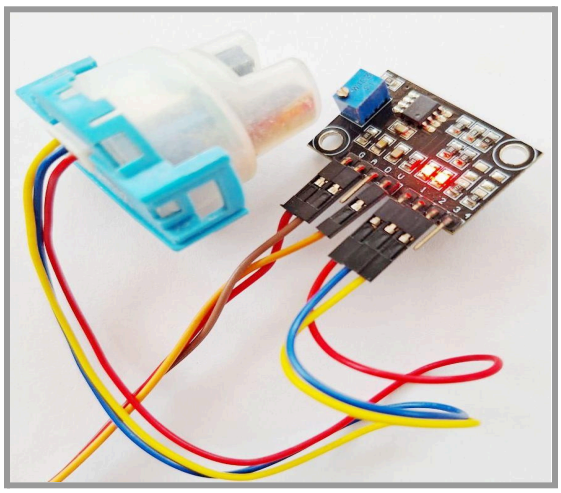


Figure 4 : Turbidity Sensor

3) TDS Sensor : The term TDS describes the volume of materials dissolved in a solution. These dissolved solids' concentration is determined using a TDS meter. These materials include calcium, salts, minerals, metals, and other organic and inorganic chemicals that raise a solution's electrical conductivity (EC). Consequently, TDS can be estimated using EC. TDS is often measured by boiling a sample of water until only the chemicals that remain are weighed. The unit of TDS is ppm, or milligrams per liter. Water with low TDS is considered more pure compared to water with greater TDS levels. Water purity is inversely related to TDS levels. Reverse osmosis water purification results in a TDS value of 0-10, while unpurified water might range from 20-100, depending on location. It has set pins for specifications like power GND of 0V for power supply, Power VCC of 3.3-5V for voltage supply , analog single output(0-2.3V), TDS prove connector ,Power indicator (LED). [20]

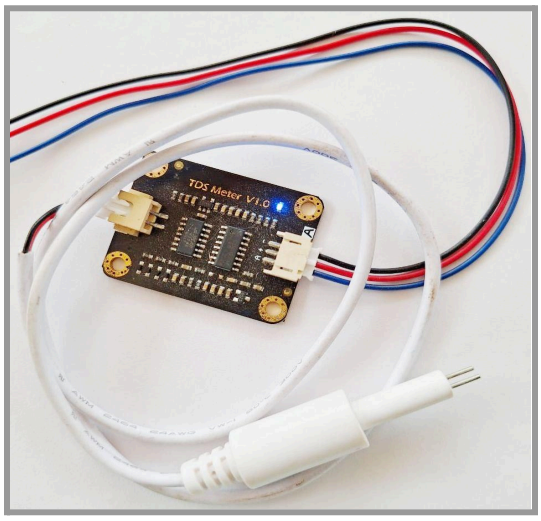
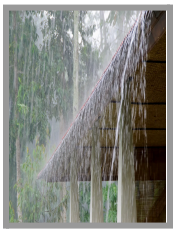


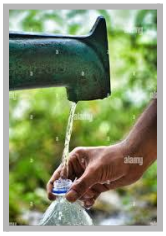
Figure 5 : TDS Sensor

B. Water Source for Data Collection

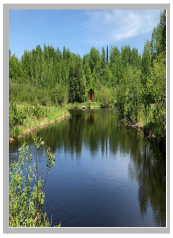
We have collected the data from Rain water ,Tap water ,Tube Well water, Lake water, Pond water and Filter water .


Rain Water


Tap Water


Tubewell Water


Lake Water


Pond Water


Filter Water

Figure 6 : Data Sources

After connecting sensors we have inserted code in Arduino UNO and inserted the sensors in our selected water sources for a maximum to minimum range of time. The data is being stored in numerical form as in the voltage unit for the turbidity sensor and in the PPM unit for TDS sensor.

Circuit Diagram

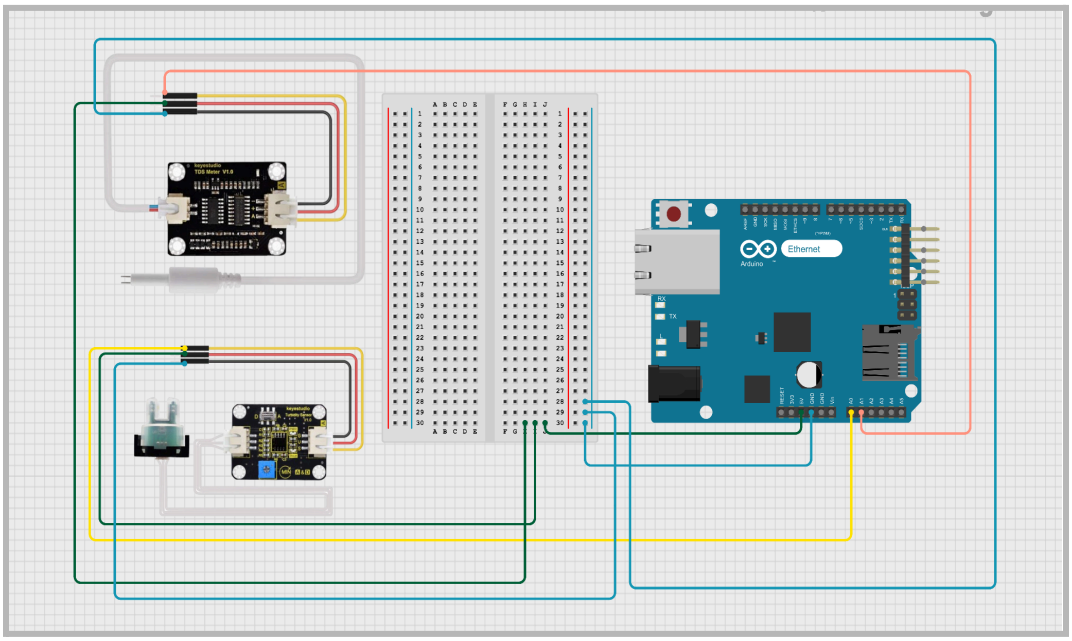


Figure 7 : Circuit Diagram

Figure 7 shows the Arduino Uno R3 on the right side. This microcontroller board has numerous pins for power, ground, and digital/analog input-output (I/O) functions. The configuration contains a breadboard in the center, which allows for solderless connections between components. The TDS sensor is linked to the top left of the breadboard, while the turbidity sensor is connected to the bottom left. The Arduino's power and ground lines are distributed throughout the breadboard rails, allowing for simple access to many components. The Arduino's 5V pin connects to the red power rail of the breadboard, while the GND pin connects to the blue ground rail. The TDS sensor's signal output, which is normally a single wire, is linked to one of the Arduino's analog input pins (A0). This enables the Arduino to read the voltage that corresponds to the TDS values. Similarly, the turbidity sensor's signal output is linked to an additional analog input pin (A1) on the Arduino to record turbidity data. Depending on the project's specific requirements, the schematic may incorporate additional pins for communication or control signals. The overall circuit diagram depicts a realistic implementation of IoT technology in environmental monitoring, with the sensors and Arduino microcontroller integrating to provide valuable data on water quality.

IoT Device Prototype

Our prototype is an overall setup that includes an Arduino Uno R3 microcontroller, a Total Dissolved Solids (TDS) sensor, and a turbidity sensor. The Arduino Uno R3 serves as the system's central processing unit, communicating with the sensors to monitor water quality data.

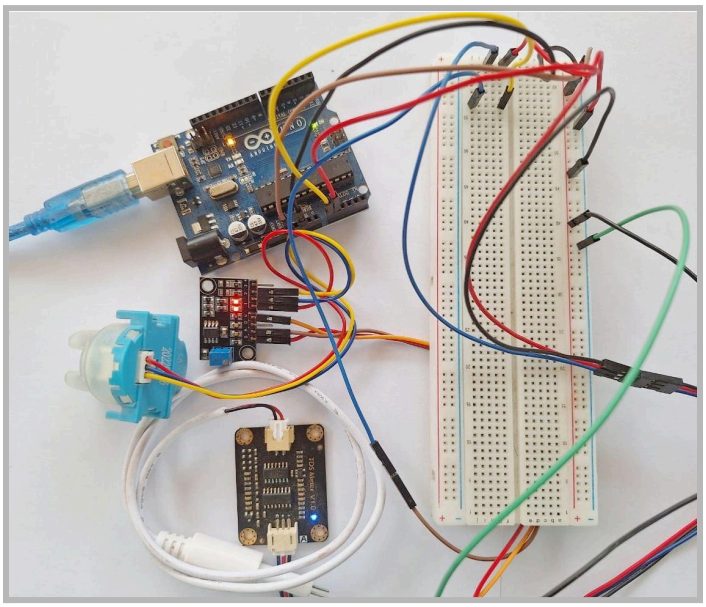


Figure 8 : Water Pollution Level Monitoring IoT Device Prototype

Pseudocode

To establish connectivity between IoT devices an Arduino sketch is necessary.The proposed system utilizes the following pseudocode :

1. Define turbidityPin as analog pin A0
2. Define tdsPin as analog pin A1
3. Define kValue as 1.0 // Calibration factor for TDS sensor
4. Function setup():
5. Initialize serial communication at baud rate 9600
6. Function loop():
7. Read analog value from turbidityPin and store it as turbidityValue
8. Read analog value from tdsPin and store it as tdsValue
9. Convert turbidityValue to turbidity in NTU (Nephelometric Turbidity Units) using convertToTurbidity function
10. Convert tdsValue to TDS (Total Dissolved Solids) in ppm (parts per million) using convertToTDS function
11. Print turbidityValue and tdsPPM separated by a comma, followed by a newline character
12. Wait for 50 milliseconds
13. Function convertToTurbidity(analogValue):

14. Convert analogValue to voltage using the formula: voltage = analogValue * (5.0 / 1024.0)
15. Convert voltage to turbidity in NTU using the formula: ntu = -111.25 * voltage + 506.67
16. Return ntu
17. Function convertToTDS(analogValue):
18. Convert analogValue to voltage using the formula: voltage = analogValue * (5.0 / 1024.0)
19. Convert voltage to TDS in ppm using the formula: tdsPPM = (voltage / kValue) * 1000.0
20. Return tdsPPM

Dataset Description

The dataset was collected using an IoT device from different water sources like lakes, ponds, tap water etc. The dataset contains a total of 296672 data which initially have 4 columns; Time, Turbidity, TDS and Category. We have three classes (Clean, Polluted and Moderate Polluted) in our dataset. The equation for turbidity is y = -11.25x+506.67 to convert voltage to NTU values.

Table 2. Dataset Overview

Time	Source	Turbidity	TDS	Pollution Level
21:15:47	Rain	290.47	180.66	Clean
19:50:20	Tap water	212.25	1171.88	Polluted
21:12:12	Lake water	229.09	927.73	Moderate Polluted

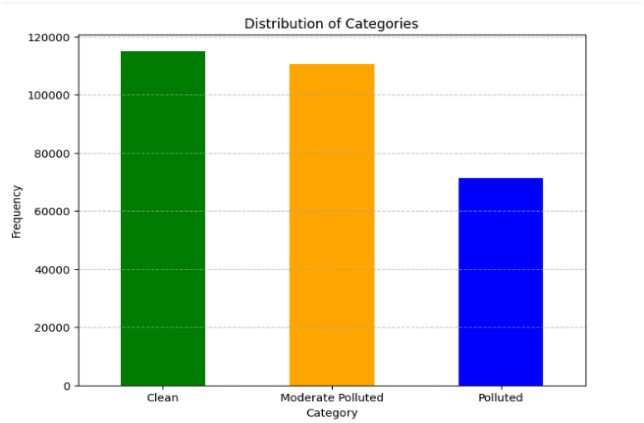


Figure 9 : Data Distribution

Preprocessing Steps

There are 296672 rows × 4 columns in our dataset. In this step we have applied many preprocessing steps into our dataset to change the data set as required for project requirements.

1. **Null value elimination (Row Wise) :** It refers to the process of removing an entire row from a dataset where one or more values are missing. This ensures that each row in the dataset contains complete information [21]. In our dataset there are 74 null values found row wise. We have applied the dropna() function to eliminate the null values.

2. **Label Encoding** : In this technique categorical data can be converted into numerical data [22]. In this process we have changed our label column as it was labeled in three classes named Clean , Moderate, and polluted. By importing LabelEncoder we have changed our label column values in numerical data which are 0,1,2 .

3. **SMOTE :** This technique is applied to imbalanced data so that the dataset helps to give a balanced class distribution [23] .We have applied the SMOTE technique to our label column to balance the data set . After applying SMOTE our imbalanced dataset got balanced.We had one class ‘clean’ that had the highest values among other classes. After applying SMOTE the other classes increased the same amount of value as the ‘clean’ class.

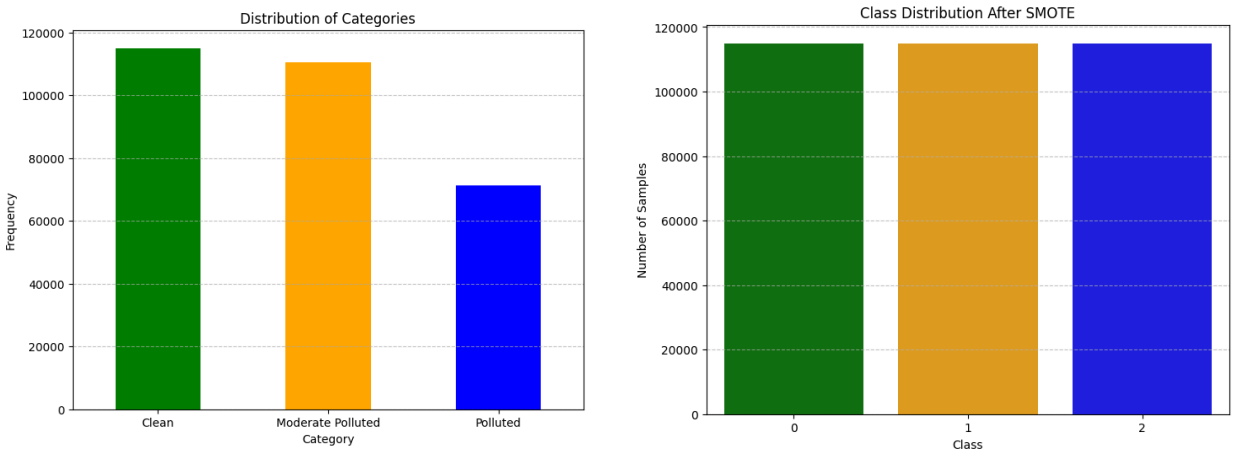


Figure 10: Dataset before and after applying Label encoding and SMOTE function

4. **Scaling (Standard Scaler) :** It is a class from sklearn.preprocessing module and creates an instance which can be used to standardize features by subtracting the mean and dividing the standard deviation [24].

5. **Splitting :** To test the model we have to first divide the feature set into train set and test set here we have also applied this process. We have split the dataset into training , testing and validation. Training 80% , Testing 10% and validation 10% .

Model Implementation:

In our large dataset we have used five machine learning algorithms like Gradient Boosting, Random Forest , Naive Bayes,SVM , Decision Tree to classify or predict the category of the water quality

perfectly. After dividing the data set into a training and testing set, we have implemented these models to get best accuracy.

GradientBoosting : Gradient boosting is a collective learning strategy that promotes robustness by combining the predictive strength of many base estimations. In order to generate accurate prediction from massive quantities of data, Gradient boosting algorithm techniques are implemented [25] .

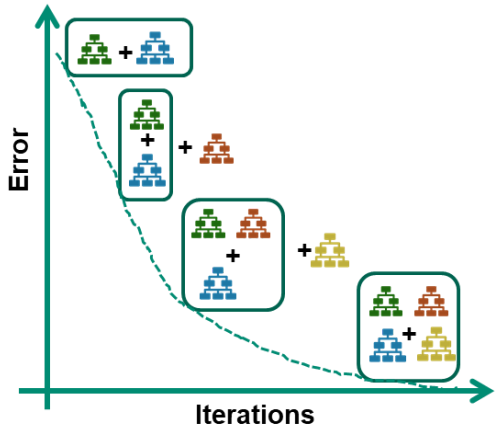


Figure 11 : Gradient Boosting

Random Forest : This technique employs ensemble learning to perform classification and regression.It generates a collection of decision trees using a bagging approach and a random subset of data.The final decision trees are constructed by combining the outputs of every decision tree in the random forest. [26]

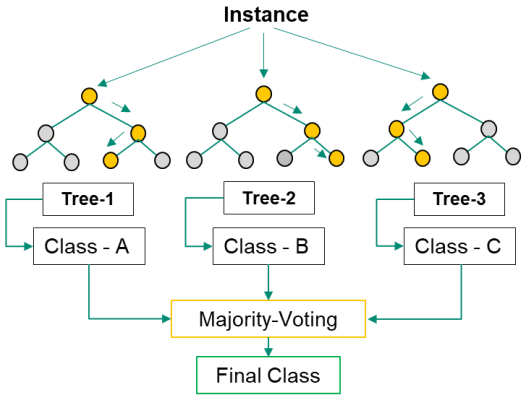


Figure 12 : Random Forest

Naive Bayes : The Naive Bayes algorithm is a supervised learning method that applies Bayes theorem to solve classification issues. It is mostly used in text classification, with a high dimensional training set.It is a probabilistic classifier , which implies that it makes predictions based on the likelihood of an item.[27].

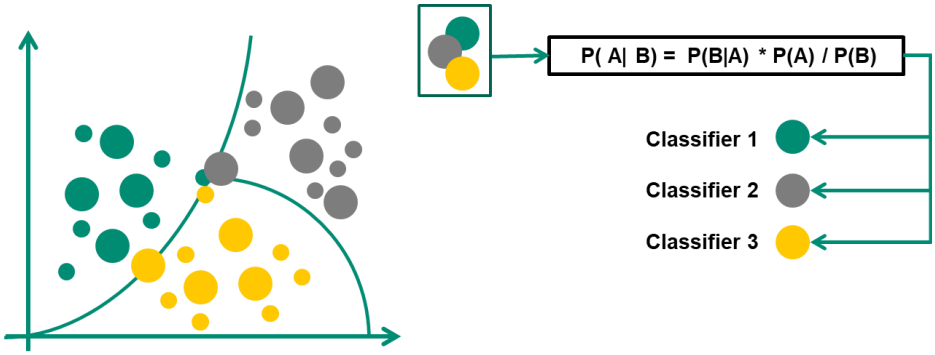


Figure 13: Naive Bayes

Support Vector Machine(SVM) : It is a supervised machine learning algorithm that is used in classification and regression tasks.Support Vector Machines determine the hyperplane that optimally divides data points into classes while maximizing the margin between classes , with support vectors controlling the hyperlane's position. SVM are successful for both linearly and nonlinearly separable data , with kernel functions [28] .

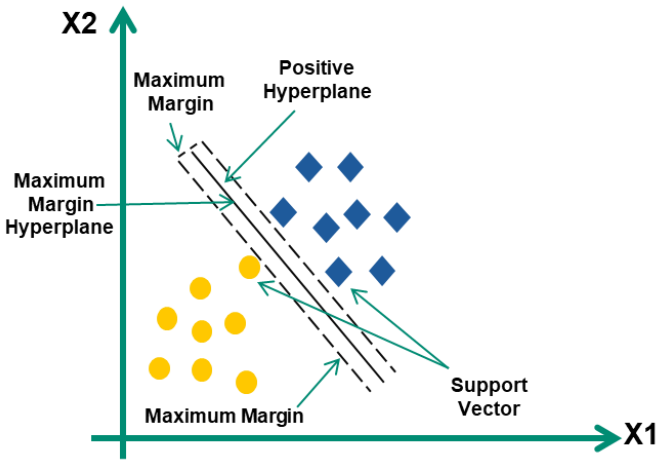


Figure 14 : SVM

Decision Tree : The decision tree approach approximates a discrete valued target function using a decision tree representation.A decision tree sorts instances from root to leaf nodes based on feature values.Each nodes indicated a choice on an attribute of the instance, whereas each branch provides a potential value for that feature [29].

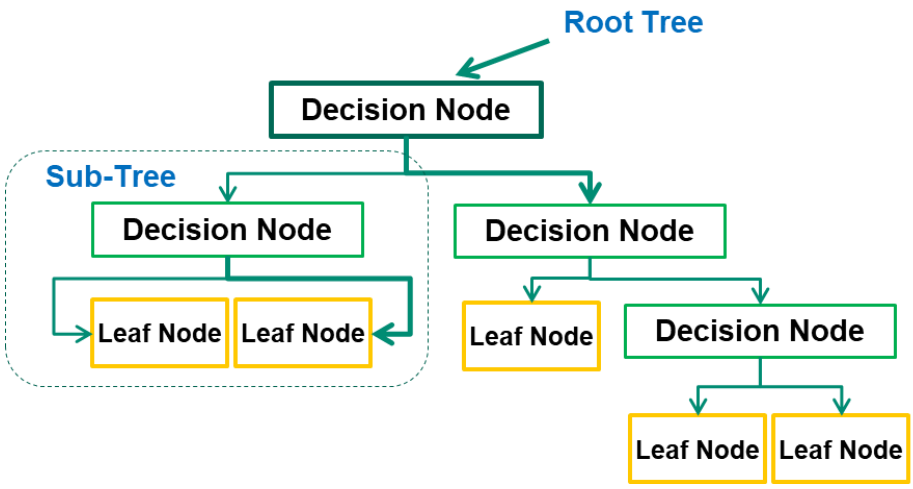


Figure 15: Decision Tree

Result Discussion:

We have used evaluation metrics to determine the effectiveness of machine learning models . These are qualifiers that determine if the output is valid. They consist of two parts: True/False and Positive/Negative.The first portion represents the output's accuracy, while the second part might represent the classification's label value.

Accuracy : The calculation of accuracy involves dividing the entire number of data sets (P + N) by the sum of two accurate predictions (TP + TN). The accuracy ranges from 0.00 to 1.0 at its worst [30] .

$$\text{Accuracy} = (TP + TN) \div (P + N) \quad (1)$$

Specificity: The specificity of a classifier is the ratio of how much was accurately labeled as negative to how much was truly negative. It is used in places with a high priority for negative categorization [31].

$$\text{Specificity} = TN \div (FP + TN) \quad (2)$$

Precision: The precision metric determines the quality of positive predictions by measuring their correctness. It is the number of true positive outcomes divided by the sum of true positive and false positive predictions [32].

$$\text{Precision} = TP \div (TP + FP) \quad (3)$$

Recall : Recall, also called sensitivity, measures the model's ability to detect positive events correctly. It is the percentage of accurately predicted positive events out of all actual positive events[33].

$$\text{Recall} = TP \div (TP + FN) \quad (4)$$

F1 Score : The F1 score or F-measure is described as the harmonic mean of the precision and recall of a classification model. The two metrics contribute equally to the score, ensuring that the F1 metric correctly indicates the reliability of a model. The F1 score is a fundamental metric for evaluating classification models and understanding how this metric works is crucial to ensure your classification model's performance [34].

$$\text{F1 Score} = 2 \times [(\text{Precision} \times \text{Recall}) \div (\text{Precision} + \text{Recall})] \quad (5)$$

To sum up, these measurements have been crucial in helping us get good results from our evaluation procedure. They have given us insightful information on how well our machine learning models are doing, allowing us to make wise choices and optimize for superior outcomes.

The evaluation of our research is shown in Table 3. The table showcases the effectiveness of different algorithms in handling the dataset, providing insights into their strengths and weaknesses.

Table 3. Evaluation Metrics of Models

Algorithms	Accuracy	Precision	Recall	F1 score	Specificity
GradientBoosting	99.99%	1.00	1.00	1.00	1.0
Random Forest	99.99%	1.00	1.00	1.00	1.0
Decision Tree	99.99%	1.00	1.00	1.00	1.0
SVM	99.47%	0.99	0.99	0.99	1.0
Naive Bayes	97.76%	0.98	0.98	0.98	0.99

Table 3. shows that all the algorithms performed quite well with high accuracy and precision. GradientBoosting algorithms achieved an accuracy of 99.99% with perfect precision,recall,F1 score and specificity of 1.0 .Similar to Gradient Boosting here Random forest and Decision Tree algorithms achieved an accuracy of 99.99% and they have perfect precision , recall , F1 score and specificity of 1.0. The SVM algorithm achieved 99.47% accuracy with the perfect precision , recall and F1 score of 0.99 and specificity of 1.0. On the other hand, the Naive Bayes algorithm

achieved 97.76% accuracy, and it has precision , recall , F1 score of 0.98 and specificity of 0.99 indicating slightly lower performance compared to other algorithms.

Result Analysis

In this part , the results obtained from all algorithms are analyzed along with different parameters like accuracy , jaccard score, cross validation score and confusion matrix performances.

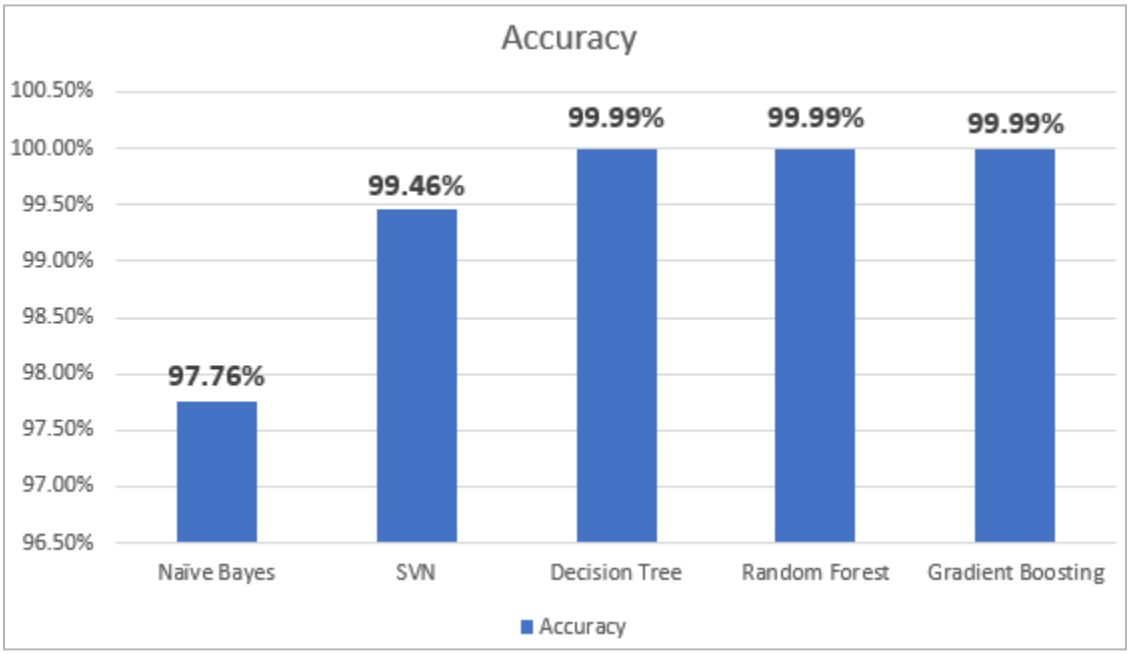


Figure 16: Accuracy Of All Algorithms

Figure 16. shows the performance of algorithms based on their accuracy scores. GradientBoosting, Random forest , Decision Tree algorithms all exhibit identical accuracy ratings of 99.99%. This consistency underscores the robustness and reliability of these algorithms in accurately predicting outcomes. The SVM algorithm showed an accuracy of 99.46%. On the other hand, the Naive Bayes algorithm, slightly trailing behind,still shows an accuracy rate of 97.76%. Overall, Figure 6. shows high performance levels of the evaluated algorithms, emphasizing their efficacy in handling the dataset.

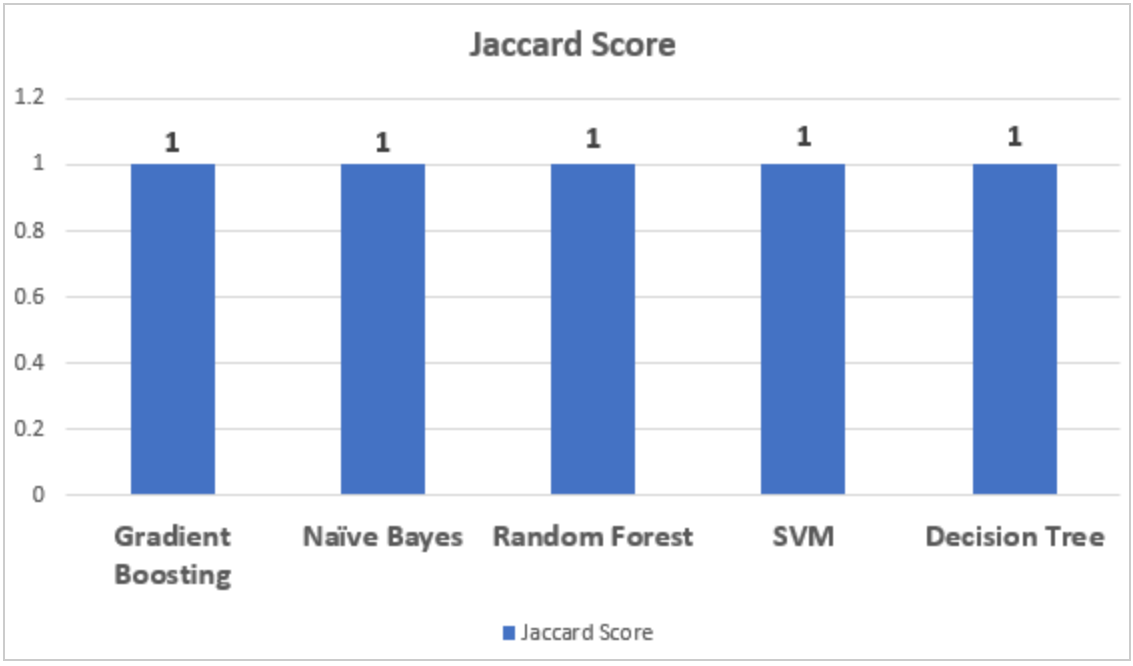


Figure 17: Jaccard Similarity Score of All Algorithms

Figure 17. shows the performances of all algorithms based on the Jaccard similarity score. This score measures the similarity between two sets by comparing the intersections of the sets of their union.It's commonly used in machine learning to evaluate similarity between predicted and actual outcomes, with a score of '1' indicating perfect similarity and '0' indicating no similarity. In figure 7. Impressively all algorithms - GradientBoosting, Random forest , Decision Tree ,SVM ,Naive Bayes achieved a perfect Jaccard score of 1.0. This uniformly emphasizes the outstanding performance of each algorithm in accurately predicting outcomes. The algorithms showed exceptional performance and reliability by showcasing their effectiveness in classification tasks with high precision and consistency.

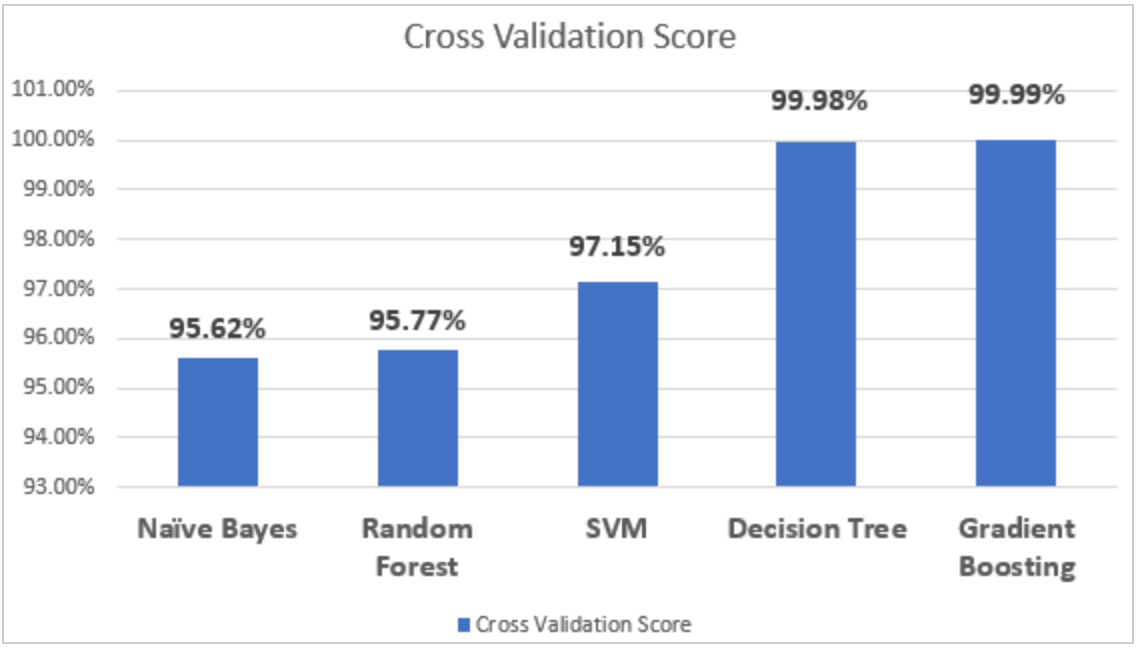


Figure 18: Cross Validation Score of All Algorithms

Figure 18 . shows the cross validation score of all algorithms. Cross validation score refers to a method to evaluate the performance of a machine learning model by repeatedly splitting the dataset into training and testing sets to ensure reliable assessment and generalization to new data. We have applied the K-fold cross-validation technique, which is used to evaluate the performance of machine learning models by splitting the dataset into k subsets. In our research here k= 5 so the model is trained and tested 5 times , using 5 different subsets for testing each time. Here GradientBoosting and Decision tree algorithms show high performance with scores of 99.99% and 99.98%, respectively. SVM follows closely with a score of 97.15% while Naive Bayes and Random Forest achieved respectable scores of 95.62% and 95.77% respectively.

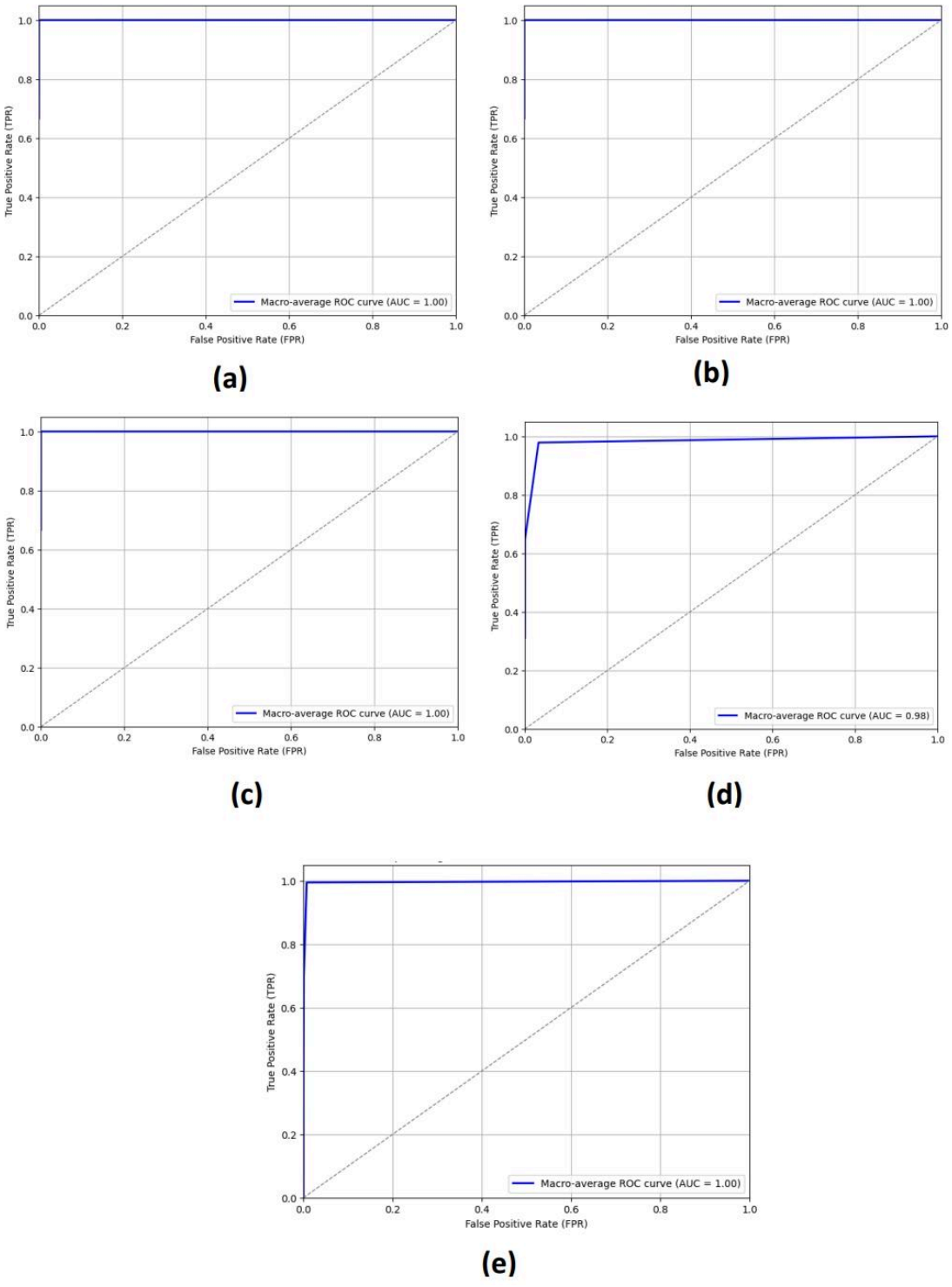


Figure 19 : ROC Curve of a) GradientBoosting, b) Random forest , c) Decision Tree , d) Naive Bayes , e) SVM classifier

Here Figure 19. shows the ROC curve of algorithms according to their macro average AUC values. Here macro average of AUC score = 1.00 is achieved for GradientBoosting, Random

forest , Decision Tree ,SVM algorithms.The curve showcases flawless discriminatory power , achieving highest possible accuracy in distinguishing between positive and negative instances across different thresholds.The Naive Bayes algorithm achieved AUC value of 0.98 which scored closed to perfect, the curve illustrates high discriminatory power as well.

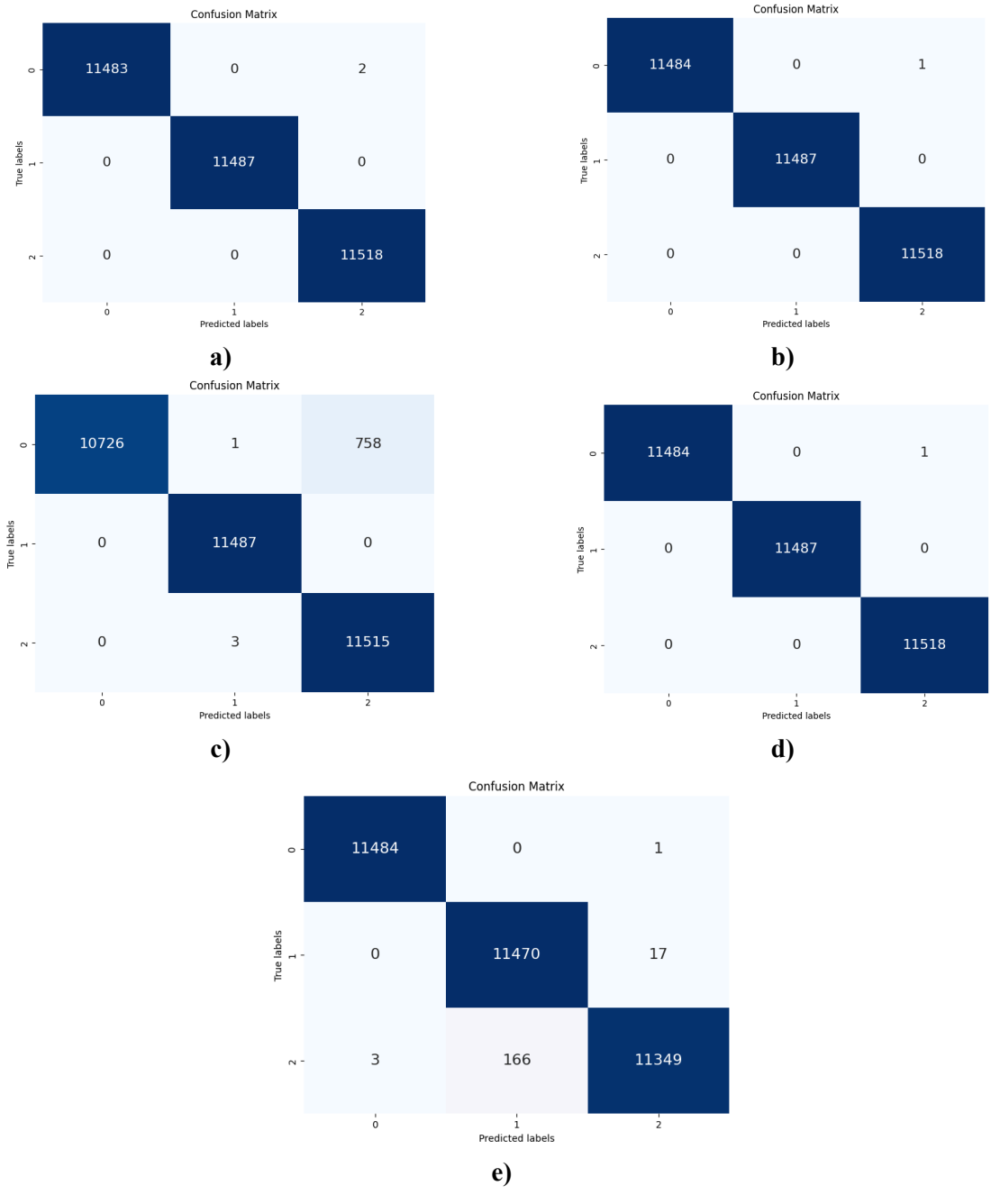


Figure 20 : Confusion Matrix of a) GradientBoosting, b) Decision Tree , c) Naive Bayes, d) Random forest, e) SVM classifier

Figure 20 shows the confusion matrix of all algorithms. Confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions made by the model compared to the actual outcomes.It consists of four components : true positive , true negative ,false positive and false negative, providing a clear picture of models performance for different classes. In our analysis the GradientBoosting algorithm classified 11483 data points correctly and there is a bit of misclassification of 2 data points. In decision tree and random forest algorithms confusion matrix the data points also 11484 data points are correctly classified and little bit of misclassification seen. The Naive Bayes and SVM algorithms showed miss classification conflicts with many data points though the classification results are satisfactory

Conclusion:

Water is the quintessential element of life.Within this context, the accurate classification of water pollution levels emerges as a critical endeavor, resonating with the imperative to safeguard this precious resource. Our research underscores the effectiveness of IoT technology and machine learning algorithms in accurately classifying water pollution levels. After applying four machine learning algorithms, we have achieved 99.99% accuracy from Gradient Boosting, Random Forest and Decision Tree. Naive Bayes algorithm attained 97.76% accuracy. All the algorithms were evaluated and showed high precision ,recall,f1-score.Achieving remarkable accuracy rates across various models highlights the reliability of these approaches for real-time assessment of water quality. In future , more factors related to water quality may be taken under consideration for analysis and more IoT driven approaches will be applied.

Reference

1. Gorde, S. P., & Jadhav, M. V. (2013). Assessment of water quality parameters: a review. J Eng Res Appl, 3(6), 2029-2035.
2. O'Donnell, D. (2022, April 27). Three Main Types of Water Quality Parameters Explained. Sensorex Liquid Analysis Technology. https://sensorex.com/three-main-types-of-water-quality-parameters-explained/
3. Serajuddin, M., Chowdhury, M. A., Haque, M. M., & Haque, M. E. (2019, January). Using turbidity to determine total suspended solids in an urban stream: a case study. In Proceedings of the 2nd International Conference on Water and Environmental Engineering, Dhaka (pp. 19-22).
4. Rusydi, A. F. (2018, February). Correlation between conductivity and total dissolved solid in various type of water: A review. In IOP conference series: earth and environmental science (Vol. 118, p. 012019). IOP publishing.

5. Sedighkia, M., Datta, B., Saeedipour, P., & Abdoli, A. (2023). Predicting Water Quality Distribution of Lakes through Linking Remote Sensing–Based Monitoring and Machine Learning Simulation. Remote Sensing, 15(13), 3302.
6. Elvin, E., & Wibowo, A. (2024). Forecasting water quality through machine learning and hyperparameter optimization. Indonesian Journal of Electrical Engineering and Computer Science.
7. Siahaan, A. P. U., Silitonga, N., Iqbal, M., Aryza, S., Fitriani, W., Ramadhan, Z., ... & Hasibuan, H. A. (2018). Arduino Uno-based water turbidity meter using LDR and LED sensors. Int. J. Eng. Technol, 7(4), 2113-2117.
8. Mohiddin, M., Bodapally, K., Siramdas, S., Shainaz, & Sriramula, S. K. (2020). A Real Time Low Cost Water Quality Progress Recording System Using Arduino Uno Board. In Advances in Decision Sciences, Image Processing, Security and Computer Vision: International Conference on Emerging Trends in Engineering (ICETE), Vol. 2 (pp. 201-209). Springer International Publishing.
9. Feng, C., Yuan, J., Sun, Y., & You, J. (2020, October). Design of Water Quality Monitoring System. In 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE) (pp. 264-267). IEEE.
10. Pantjawati, A. B., Purnomo, R. D., Mulyanti, B. U. D. I., Fenjano, L. A. Z. U. A. R. D. I., Pawinanto, R. E., & Nandiyanto, A. B. D. (2020). Water quality monitoring in Citarum River (Indonesia) using IoT (internet of thing). Journal of Engineering Science and Technology, 15(6), 3661-3672.
11. Jha, B. K., Sivasankari, G. G., & Venugopal, K. R. (2020). Cloud-based smart water quality monitoring system using IoT sensors and machine learning. International Journal of Advanced Trends in Computer Science and Engineering, 9(3).
12. Chowdury, M. S. U., Emran, T. B., Ghosh, S., Pathak, A., Alam, M. M., Absar, N., ... & Hossain, M. S. (2019). IoT based real-time river water quality monitoring system. Procedia computer science, 155, 161-168.
13. Lakshmikantha, V., Hiriyannagowda, A., Manjunath, A., Patted, A., Basavaiah, J., & Anthony, A. A. (2021). IoT based smart water quality monitoring system. Global Transitions Proceedings, 2(2), 181-186.
14. Ahmed, U., Mumtaz, R., Anwar, H., Mumtaz, S., & Qamar, A. M. (2020). Water quality monitoring: from conventional to emerging technologies. Water Supply, 20(1), 28-45.
15. Ramadhan, A. J., Ali, A. M., & Kareem, H. K. (2020). Smart water-quality monitoring system based on enabled real-time internet of things. J. Eng. Sci. Technol, 15(6), 3514-3527.
16. Pasika, S., & Gandla, S. T. (2020). Smart water quality monitoring system with cost-effective using IoT. Heliyon, 6(7).
17. Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, 102920.

18. Sujadi, Harun, and A. Bastian. "Design prototype detection tools of Porous Tree using microcontroller Arduino Uno R3 and piezoelectric sensor." *Journal of Physics: Conference Series*. Vol. 1013. No. 1. IOP Publishing, 2018.
19. Ramdane, Kaouthar. "An Arduino-based Water Quality Monitoring System using pH, Temperature, Turbidity, and TDS Sensors."
20. Rose, Lina, and X. Anitha Mary. "TDS Measurement Using Machine Learning Algorithm." *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*. IEEE, 2018.
21. Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, *25*, 319-335.
22. Etaiwi, W., & Naymat, G. (2017). The impact of applying different preprocessing steps on review spam detection. *Procedia computer science*, *113*, 273-279.
23. Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.
24. Frye, M., Mohren, J., & Schmitt, R. H. (2021). Benchmarking of data preprocessing methods for machine learning-applications in production. *Procedia CIRP*, *104*, 50-55.
25. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21
26. Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series* (Vol. 1142, p. 012012). IOP Publishing.
27. Meyfroidt, G., Güiza, F., Ramon, J., & Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, *23*(1), 127-143.
28. Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.
29. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.
30. Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599-606.
31. Intawong, K., Scuturici, M., & Miguet, S. (2013). A new pixel-based quality measure for segmentation algorithms integrating precision, recall and specificity. In *Computer Analysis of Images and Patterns: 15th International Conference, CAIP 2013, York, UK, August 27-29, 2013, Proceedings, Part I 15* (pp. 188-195). Springer Berlin Heidelberg.

32. Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38(2), 99-129.
33. Bradley, A. P., Duin, R. P. W., Paclik, P., & Landgrebe, T. C. W. (2006, August). Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 4, pp. 123-127). IEEE.
34. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015-1021). Berlin, Heidelberg: Springer Berlin Heidelberg.