Smart Urban Water Quality Prediction System Using Machine Learning

Bharath Singh J¹, Nirmitha S², Kaviya S S³

¹Assistant Professor, Department of Electronics and Instrumentation Engineering, R.M.K. Engineering College, Tamil Nadu.
²&³UG Student, Department of Electronics and Instrumentation Engineering, R.M.K. Engineering College, Tamil Nadu

E.mail: jbs.eie@rmkec.ac.in

Abstract. The World Health Organization says that every year more than 3.4 million people die as a result of water-related diseases. Quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. The project aims to design a water quality prediction system using Machine learning based on the water standards suggested by BIS to prevent deaths due to water related diseases. The quality is predicted based on parameters such as pH, Temperature, TDS, Turbidity and Conductivity value. The dataset is preprocessed and split into test and training data. The data is fed into regression algorithm and been evaluated. Sensors that can measure the water parameters are also been implemented. A webpage interfaced with the Machine Learning model is created to upload sensor values and the corresponding water quality is predicted. This project can be used in urban areas to predict the quality of the drinking water thereby preventing the spread of diseases such as dysentery, typhoid and cholera due to consumption of contaminated water.

1. Introduction
Water plays a vital role in everyone’s life and is observed everywhere and in every form [1]. In Today’s world, due to climatic changes and pollution the water quality is been affected in urban areas and various experiments are done to test the quality of water [2]. Due to poor water quality, risk occurs in the industrial areas which damage the whole environment and causes an economical loss [3]. The root cause for many diseases such as typhoid, diarrhea, cholera is due to usage of contaminated water caused by increased industrialization and urbanization in India. [4]. According to reports form WHO, it is estimated that about 77 million people affected by contaminated water in India and 21% of diseases are caused due to it. [5] Due to insufficient rainfall and drying up of main reservoirs that supplies water, India faces water crisis frequently, hence making water one of the most precious and limited land resources. Many Organizations including WHO and BIS has framed standards for water parameters that can be used to efficiently analyze the quality of water. For checking the quality of water, conventionally it required to collect water samples and send it to the lab for testing which is a tedious process. [6] With IoT and Machine Learning algorithms it is easy to obtain the sensor values form a water sample, monitor and predict the quality of water at the comfort of our home. IOT is a buzzing technology that allows sensors to transfer data between them or to the cloud without the intervention of humans. [5]. Water quality index of the water, which helps in determining the quality of water, can be predicted by the extensive use of machine learning regression algorithm.

2. Literature Survey
K Spandana and V R Seshagiri Rao has brought up the idea to keep track of the water quality based on internet connection produced and has proposed that they have used Nodemcu-wifi module and sensors to
collect the real time values, they have also set the conditions in order to generate the alarms when the sensors are in peculiar conditions [15]. V Vaishnavind M A Gaikwad has brought up idea to record the water quality based on Wi-Fi connection produced. The values are viewed through the LCD display. They have also installed the Blynk app, such that with proper wifi-connection the values are displayed in the app [16]. P Suganya and S Hariharan has produced idea to keep an eye on water quality using the Blynk app based on wireless network connection produced, where various sensors are placed in order to measure real time values and uses Nodemcu for displaying the values through the Blynk App and server such that if the values are below standard values then necessary actions are performed [17].

S Gokulanathan, et.al. has come up with ideas to periodically track the water parameters using Arduino and GSM. pH and temperature sensors are used to measure real time values and the values are viewed through the LCD display. Through the GSM module values and the type of water such as acidic or base are been displayed through the PC and Smart Phone. The alarms are also been generated if there is any variations in the values [18]. Aziz, M Saroda and E Rohadu has brought up with ideas to observe the various water parameters of a river, where sensors are been implemented for measuring real time values. They have also implemented Nodemcu in order to display the values through firebase cloud page and in the Android application. Therefore calibration of sensors is required when the results are not accurate [19]. T Kalavathi Devi and P Saktihivel has come up with ideas to record and examine the water grade, which is used for identifying the secured drinking water using the internet connection produced. In this paper they have proposed to use various sensors for measuring the water parameters. Hence the microcontroller is programmed in such a way to display the values in LCD and through GSM Module the values are also displayed in Android application. The values are also been stored in cloud. In the Android page the values are shown one by one and determines the nature of water for each sensor page [20].

3. Proposed Methodology

This section provides the proposed optimal, cost efficient method for monitoring the water quality. The existing system methodology disadvantages are listed and the overcoming is explained in the proposed method. The proposed method is developed as hardware and the simulations are done in software and the results are validated.

3.1. Existing Model

Water Quality checking was first done by traditional methods which was a very complex as well as a lengthy process. Sathish Pasika have produced idea to keep record of water quality parameters using the network connection. He used various types of sensors to measure the values. Here in order to access internet Nodemcu Wi-Fi module has been used, such that values are sent to the cloud page. Thus the values are graphically presented through the ThingSpeak page and displayed in the Arduino Serial monitoring page. [21]. The existing system consist of various sensors to monitor and upload the values to the webpage, but the problem is that in the absence of a mechanism to predict the water quality it might be difficult for the user to know about the actual quality of the water. Thus in order to overcome this problem we have built a Machine Learning model and have integrated it with the webpage where the sensor values are displayed, So that the user can instantly know about the quality of water.

3.2. Proposed Model

![Block diagram of proposed model](image)

**Figure 1:** Block diagram of proposed model
The existing system consists of various sensors to monitor and upload the values to the webpage, but the problem is that in the absence of a mechanism to predict the water quality it might be difficult for the user to know about the actual quality of the water. Thus in order to overcome this problem we have built a Machine Learning model and have integrated it with the webpage where the sensor values are displayed, so that the user can instantly know about the quality of water.

3.3. IoT Implementation
This model is used to monitor the values. The sensors are connected to the Arduino and programmed in such a way to monitor the values. Nodemcu is used for Wi-Fi connectivity. Nodemcu and Arduino are serially communicates and data are sent to the “Thingspeak server” webpage using nodemcu Wi-Fi module. The sensors used are pH for measuring amount of hydrogen, TDS for measuring amount of solids, turbidity to measure the clearness level and conductivity for measuring conductivity level and temperature for temperature compensation. Important Sensors- Calibrated and tested in laboratories.

3.4. Hardware Setup and Circuit diagram
This section provides the hardware setup model and the circuit diagram of the proposed model. The sensors are interfaced with the controller and the necessary experiments are conducted. A detailed circuit diagram is represented with the sensors input and the web interface at the output.

4. Creation of Machine Learning Model
This model is used to predict the quality of the water. Data are been collected through sensors and saved in excel sheet. The method called Data preprocessing is to clean the given raw data. Cleaning provides the accurate data. Import and export of these data can be done anytime and anywhere. The following steps are to be followed. First the necessary libraries such as pandas, SCI-KIT learn are imported. And the important second step will be, the dataset are been converted to .CSV format and been uploaded. The dataset is checked for hidden values and if present it is filled with mean of the column.

4.1. Water Quality Index Calculation
Water quality index is a numerical value that represents the overall quality of the water based on its individual parameters. Based on the quality index the usability of a water sample can be determined. Before calculating the water index, it is necessary to choose standard values for the parameters based on WHO/BIS/ICMR prescriptions. There are various methods to calculate Water quality index and the method we use is Weighted Arithmetic water index method. The parameters we chose for analysis are pH, total dissolved salts, turbidity and electrical conductivity. The parameters and their standard values are given in Table 1.

| Parameter          | Standard (Prescribed by BIS / ICMR) |
|--------------------|-------------------------------------|
| pH                 | 6.5-8.5 (BIS)                       |
| TDS (mg/l)         | ≤500 (BIS)                          |
| Turbidity (NTU)    | ≤5 (BIS)                            |
| Electrical Conductivity (µS/cm) | ≤300 (ICMR) |
The formula for calculating the WQI is as follows \[22\] \[23\] \[24\]

\[ WQI = \frac{\sum WnQn}{\sum Wn} \]

\( Wn \) is the unit weight for all the parameters where \( Wn = K/Sn \). \( Sn \) is the standard values for the parameters and ‘\( K \)’ is a proportionality constant calculated by the formula \( K=1/\sum Sn \).

\( Qn \) is the quality rating of the parameters given by the formula \( Qn = \frac{|Vn-Iv|}{|Sn-Iv|} \), where \( Vn \) represents the value observed from the sample and \( Iv \) represents the ideal value for a parameter (\( Iv=7 \) for pH and 0 for the other parameters). A sample Water Quality Index calculated for a water sample is given in Table 2.

| Parameter | Sn  | 1/Sn | \( \sum (1/Sn) \) | \( k=1/(\sum (1/Sn)) \) | \( Wn=K/Sn \) | Ideal Value | mean Concentration value (Vn) | Qn= \( Vn/Sn*100 \) | \( Wn*Qn \) |
|-----------|-----|------|-----------------|-----------------|-----------------|-------------|------------------------|-----------------|-------------|
| pH        | 8.5 | 0.117| 0.32298         | 3.096163        | 0.364254        | 7           | 7.59                   | 39              | 14.20593    |
| TDS       | 500 | 0.002| 0.32298         | 3.096163        | 0.006192        | 0           | 156.25                | 31.25           | 0.19351     |
| Turbidity | 5   | 0.2  | 0.32298         | 3.096163        | 0.619233        | 0           | 2                      | 40              | 24.76931    |
| Conductivity | 300 | 0.003| 0.32298         | 3.096163        | 0.010321        | 0           | 200                    | 66.6667        | 0.688036    |
| Sum       | 0.322| 1    | 1               | 1               | 3.096163        | 0           | 200                    | 66.6667        | 0.688036    |

The ideal water quality index is calculated and found as WQI=39.85678. The classification of water we deal with based on the water quality index is given in Table 3.

| Water Quality Index | Water Quality Class |
|---------------------|---------------------|
| <50                 | Good                |
| 50 – 75             | Partial             |
| >75                 | Poor                |

The model has a graphical option called Data visualization. It presents data in graphical format to detect patterns easily.

Figure 3: Histogram of Water Quality Parameters

The model has another option called Extraction/Vectorization. This method is to convert texts to numbers as machine learning deals with it. The data are then split to test and train dataset in the ratio 2:8 to feed and train the Model.

5. Creation of Machine Learning Model

We will be using Regression model for predicting the Water Quality Index. Various Regression models were trained and their accuracy scores were calculated as per Table 4.

| Model                   | Accuracy |
|-------------------------|----------|
| Linear Regression       | 35%      |
| KNN Regression          | 50%      |
| Gradient Boosting Regression | 94%     |
| Polynomial Regression   | 94%      |
| Random forest Regression| 97%      |

Table 4: Accuracies of Various models
The Random Forest Regressor was chosen for its accuracy. Random Forest classifier / Regressor introduced by Breiman are a supervised machine learning that consists of multiple decision trees as the base classifiers. Random Forest Classifier gives good accuracy with large datasets and also provides great estimation of missing values in a dataset when large number of values is missing. [7]. Sub sampling the training data and selecting node sets introduces randomness and replacement [8]. The algorithm to be followed has the following steps. Select N samples from the dataset through row sampling and feature sampling. These samples are training datasets to draw N decision trees from the samples, each capable of producing a prediction. Aggregate the results produced by each decision tree. Determine the final prediction by considering the majority of results produced by N TREES. [9] [10]

![Figure 4: Random Forest Regression Block Diagram](image)

The Random Forest Regressor can be imported from the SCIKIT learn package in python. WE have chosen 10 as our n_estimators value (number of decision trees). The training data was given and the model predicts the Water Quality Class.

6. Model Evaluation
After building the model, it is essential to test its accuracy using the test data. We use regression metrics such as Mean Squared error, Root Mean Square Error, Mean Absolute Error and r2_score, which can be imported form sklearn.metrics package. R2_score indicates the accuracy of the predictions. It ranges form 0-1, and a larger value indicates larger accuracy.

| Metric    | Value obtained |
|-----------|----------------|
| MAE       | 0.4478         |
| MSE       | 2.6955         |
| RMSE      | 1.6365         |
| R2_score  | 0.9745         |

After evaluating the model, it is saved using pickle with an extension .pkl

7. Application Building
The webpage consist a form that accepts the water parameters as input and sends it to the model. The values go into the machine learning model and give a prediction that will be displayed on the screen. The webpage is deployed using flask web framework.

7.1. Build a HTML page
The html page we built consists of a form element with various input tags to receive the input parameters from the user. There are two attributes included in the form element. One is the action attribute that species where are data is submitted when the submit button is clicked. Another attribute is method that specifies the type of HTTP request our form produces. [12]. In our case action URL would be ‘/login’ and method will be ‘post’ which provides secure communication of data than other methods. CSS is used to style our HTML for a good-looking interface.

7.2. Build a Build a Python Code
Flask is a web-framework written in python which is easy to use and comes with a built-in development server that allows us to deploy applications instantly. Flask structure must contain two folders namely, the static folder that consists of CSS and images and the templates folder that consists of HTML pages. The model is loaded into the webpage using pickle.load() function and stored in variable named

---

---
model. The main elements used in our coding would be `@app.route`, `render_template()` and `request.form()`. The `render_template()` function is used to return the HTML page we built. The `@app.route` decorator converts the returned template into a HTTP request to be displayed by the browser. [13]. The flask app is run using its built-in development server to view the page.

7.3. **Display ThingSpeak values on the webpage**

It is possible to display the sensor values stored in ThingSpeak server in our webpage. There is an provision to embed the graphs or widgets available in thing-speak in our webpage by including the `<iframe>` tag provided with every widget in ThingSpeak, to our HTML code [14].

8. **Interference and Analysis**

8.1. **Drinking Water Sample**

A drinking water sample is taken for analysis. The sensors are placed into the water sample. The website displays the quality as good.

![Figure 5: Drinking Water Sample and Website displaying Drinking Water Quality](image)

8.2. **Salt Water Sample**

A salt sample is taken for analysis. The sensors are placed into the water sample. The website displays the quality as partially good.

![Figure 6: Salt Water Sample and Website displaying Salt Water Quality](image)

8.3. **Mud Water Sample**

A mud water is taken for analysis. The sensors are placed into the water sample. The website displays the water quality as poor.

![Figure 7: Mud Water Sample and Website displaying Mud Water Quality](image)
9. **Drawbacks**

Our System is low cost and efficient but also has its drawbacks. The lifetime of the low cost sensors which we have used are less and might require the frequent purchase of new sensors. The use of high grade industrial sensor might help us to overcome this problem. Internet Connectivity and times may be a problem, since data won’t be updated.

10. **Conclusion and future scope**

Our Solution that addresses water quality problems are been discussed in our paper. We have used low cost microcontroller and affordable sensors to build the system. The Webpage integrated with ThingSpeak and Machine Learning model was built which will help the user to monitor the values as well as predicts the quality of the sample water.

Our further research is been done to improve the system that will alert the concerned official in case the water quality is poor, so that they would take necessary actions. The system will also fetch real time values from sensors and automatically predict the quality without requiring the user to manually enter the value.

11. **Acknowledgement**

The authors would like to thank the management of R.M.K. Engineering College for providing the required resources.

12. **Compliance with Ethical Standards**

12.1 **Funding**

The authors declare that they have no conflicts of interest to report regarding the present study.

12.2 **Ethical Conduct**

This chapter does not contain any studies with human participants or animals performed by any of the authors.

13. **References**

[1] Shamsul Haq, Shoukat Ara, Syed Maqbool Geelani and Asma Absar Bhatt, 2016, Assessment of surface and ground water for irrigational purposes, *International Journal of Applied And Pure Science and Agriculture*, vol.2, no. 2.

[2] Manish Kumar, Gurmeet Singh, Tushara Chaminda, Pham Van Quan and Keisuke Kuroda, 2014, Emerging Water Quality Problems in Developing Countries, *The Scientific World Journal*. vol.2014, no.158796.

[3] Roy and Ritabrata, 2018, An Introduction to water quality analysis, *International Journal for Environmental Rehabilitation and Conservation*, vol.2018, no. 3, pp. 94-100.

[4] Subodh Kumar, Hari Mohan Meena and Kavita Verma, 2017, Water Pollution in India: Its Impact on the Human Health: Causes and Remedies, *International Journal of Applied Environmental Sciences*, vol.12, no. 2, pp. 275-279.

[5] K Gopavanitha and S Nagaraju, 2017, A low cost system for real time water quality monitoring and controlling using IOT, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), DOI.10.1109/ICECDS.2017.8390054.

[6] Brinda Das and P C Jain, 2017, Real-time water quality monitoring system using Internet of Things, *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, DOI 10.1109/COMPTELIX.2017.8003942.

[7] Vrushali Y Kulkarni and Pradeep K Sinha, 2014, Effective Learning and Classification using Random Forest Algorithm, *International Journal of Engineering and Innovative Technology (IJEIT)*, vol.3, no. 11, pp. 45-53.

[8] Anna Bosch, Andrew Zisserman and Xavier Munoz , 2007, Image Classification using Random Forests and Ferns, *IEEE International Conference on Computer Vision*, DOI.
10.1109/ICCV.2007.4409066

[9] Hossam M Zawbaa, Maryam Hazman, Mona Abbass and Aboul Ella Hassanien, 2014, Automatic fruit classification using random forest algorithm, International Conference on Hybrid Intelligent Systems, DOI.10.1109/HIS.2014.7086191

[10] Gerard Biau, 2012, Analysis of a Random Forests Model, Journal of Machine Learning Research, vol.13, pp. 1063-1095.

[11] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A Shah and Rabia Irfan, 2019, Efficient Water Quality Prediction Using Supervised Machine Learning, MDPI, vol.11, no. 2210, pp. 1-14.

[12] Martin Breuss, Python Web Applications: Deploy Your Script as a Flask App, https://realpython.com/python-web-applications/

[13] Abdelhadi Dyouri, How to Make a Web Application Using Flask in Python 3, https://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3

[14] https://in.mathworks.com/help/thingspeak/

[15] Kancheti Spandana and V R Seshagiri Rao, 2018, Internet of Things (IOT) Based Smart Water Quality Monitoring System, International Journal of Engineering & Technology, vol.7, no. 3, pp. 259-262.

[16] V D Vaishnavi and M A Gaikwad, 2017, Water Quality Monitoring System Based on IOT, Advances in Wireless and Mobile Communications, vol.10, no. 5, pp. 1107-1116.

[17] P Suganya and S Hariharan, 2018, IOT Based Water Quality Monitoring System Using Blynk App, International Journal of Research in Advent Technology, Vol.6, no.11.

[18] S Gokulanathan, P Manivasagam, N Prabu and T Venkatesh, 2019, A GSM Based Water Quality Monitoring System using Arduino, Shanlax International Journal of Arts Science and Humanities, vol.6, no. 4, pp. 22-26.

[19] F A Aziz, M Sarosa and E Rohadi, 2020, Monitoring system water pH rate, turbidity, and temperature of river water, IOP Conf. Series: Materials Science and Engineering, vol.732 no.012106.

[20] T Kalavathi Devi and P Sakthivel, 2019, Monitoring and Analysis of Water Quality Parameters for Diagnosis of Safe Drinking Purpose using Wireless Sensor Network, International Journal of Innovative Technology and Exploring Engineering, vol.9, no.2.

[21] Sathish Pasika and Sai Teja Gandla, 2020, Smart water quality monitoring system with cost-effective using IOT, Heliyon, Vol.6, no.7.

[22] Shweta Tyagi, Bhavtosh Sharma, Prashant Singh and Rajendra Dobhal, 2013, Water Quality Assessment in Terms of Water Quality Index, American Journal of Water Resources, vol.1, no. 3, pp.34-38.

[23] D Satish Chandra, S S Asadi and M V S Raju, 2017, Estimation of Water quality index by weighted arithmetic water quality index method: a model study, International Journal of Civil Engineering and Technology (IJCIET), vol.8, no. 4, pp. 12-20.

[24] M Jeyalakshmi, M Natesan and R Udayakumar, 2019, Application of Water Quality Index (WQI) For Ground Water Quality Interpretation, International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol.8, no. 12, pp. 12-20.

[25] Zexin Lin, Weixing Wang, Huili Yin and Sheng Jiang, 2017, Design and implementation of cost effective water quality evaluation system, International Conference on Computer Network, Electronic and Automation (ICCNEA) Conference, 2017.