International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 1240–1248

Water Quality Monitoring for Disease Prediction using Machine Learning

Prajakta Patil [1], Sukanya More[2], Atharv Deshpande [3], Harshal Todkar [4],Sanjeev Wagh[5]

[1,2,3,4] *Department of Information Technology,*
Government College of Engineering Karad.

[5]*Head, Department of Information Technology,*
Government College of Engineering, Karad.

Abstract

Access to pure drinking water and sanitation has been marked as a fundamental human right as „The Human Right to Water and Sanitation" by the United Nations General Assembly on 28 July 2010. Water related diseases are the primary cause of diseases and deaths around the world with more than 3.4 million deaths per year. Lack of monitoring of water sources and inability to anticipate the proliferation of waterborne diseases are found at the root of these deaths. There has been a compelling need for disease prediction based on water quality. The present study was focused on monitoring of water quality parameters and using these parameters to predict probable waterborne diseases. The main objective of study was to apply machine learning techniques to water quality data in order to make predictions about waterborne diseases. The work involved collecting observations of some of the water quality parameters by leveraging the Internet of Things (IOT). The detailed data, involving observations of all the necessary parameters, was collected from the West Bengal Pollution Control Board's Water Quality Information System. Gradient Boosting Classifier was trained and tested on collected data. The accuracy of result was found to be 0.92 and 0.95 on cross-validation and hold-out data, respectively. Once trained, the model started making predictions based on primary data. The predicted diseases were conveyed in the form of alerts using Push bullet service. The study thus proposed usability of water quality parameters in early prediction of water related diseases.

Keywords: Internet of Things (IOT), Cloud, Machine Learning (ML), Gradient Boosting Classifier.

1. Introduction

Waterborne diseases are one of the leading causes behind human deaths across world. According to World Health Organization (WHO), waterborne Diarrheal diseases are accountable for 2 million deaths per year, mostly occurring in children under the age of 5 [5]. In India annually about 37.7 million people are impacted by waterborne diseases, out of which 1.5 million kids die of Diarrheal diseases alone and loss of 73 million working days leads to an fiscal loss of $600 million every year. The varieties of water-borne disease are caused by water contaminated with viruses, bacteria, metals, toxins and other chemical contaminants.

Internet of Things (IOT) refers to extending human to computer and computer to computer communication to computers embedded in day to day things imparting them unique identity and ability to transfer data over network. A wireless sensor network (WSN) entails numerous physically dispersed sensor nodes continuously recording physical stimuli and collecting data at a central base station. Sensors are useful in sensing physical phenomenon and representing it in digital form making it easier to process, store and act upon them [1]. With the help of Internet of Things (IoT) technology, the conventional

ISSN: 2233-7857 IJFGCN

process of water quality monitoring by laboratory testing can be automated to get the results remotely and in real time [2].

Machine Learning is a division of Artificial Intelligence (AI) that allows machines to acquire insights from data, recognize patterns within it and make decisions with less or no human intervention. Machine Learning has been proved to be helpful in automation of disease diagnosis. Considering the catastrophic impact of waterborne diseases on human life, there is need for leveraging machine learning in early prediction of such diseases. Waterborne disease outbreaks are often reported in areas with poor water quality [3]. Gastrointestinal diseases like cholera, typhoid, etc. arise due to microbial contaminants. On the other hand, water related diseases like diarrhea, alkalosis, fluorosis, bladder cancer and kidney diseases rise out of suspended metals and chemical contaminants in water. Changes in water pH levels can also significantly affect human health. Numerous efforts have been applied by analysts to create or utilize huge information investigation models and machine learning models for precise water quality evaluation [4].

In this regard, the main motivation in this study is to propose the method using machine learning algorithm for the efficient prediction of water related disease based on water quality parameters like pH, Temperature, Turbidity, Total Dissolved Solids (TDS), Dissolved Oxygen, etc. The system involves sensors like pH sensor, TDS sensor, Turbidity sensor, temperature sensor and conductivity sensor to continuously monitor water conditions. The monitoring system obtains and stores real time water quality data. The cloud technology provides for real time storage and access of collected data. The data collected from West Bengal Pollution Control Board's water quality information system served the purpose to train and test a Gradient Boosting Classifier, a machine learning classification model. The trained model served to predict probable waterborne disease. The prediction was raised as an alert using Push bullet API. The study in this paper highlights on the applicability of machine learning in early prediction of waterborne diseases. Continuous monitoring of water sources and studying from the trends in water quality and disease susceptibility can be more effective than disease diagnosis and cure.

2. Literature Review

[1] A Quick Survey on Wireless Sensor Networks

Manisha Bhende, Sanjeev J. Wagh and Amruta Utpat in "A Quick Survey on Wireless Sensor Networks" highlight importance of sensors and wireless sensor networks in monitoring physical world in real time. WSN is a technology which provides low-cost and flexible way to obtain data from the places that are not easily reachable, and that are perilous to humans. Wireless Sensor Networks (WSNs) have many practical, potential, and useful applications like Home Control, Agricultural, Traffic Control, Medical, and Military. The paper highlights Wireless Sensor Networks (WSN) as an effective technology to capture physical world phenomenon, convert them into digital form, store them and act upon them. At the same time it raises concerns about the efficiency in power consumption of WSNs [1].

[2] Water Quality Monitoring System using IoT

Ankita Taware , Vrunda Ghate , Minal Gaingade, Mayuri Bhandvalkarand, Dr. Sanjeev Wagh in "Water Quality Monitoring System using IoT" proposed a low-cost solution to analyze the water quality in real time by leveraging Internet of Things (IoT). They used pH sensor, Turbidity sensor, Conductivity sensor to analyze water quality in real time and collected the data on cloud using NodeMCU. Their effort was to automate the process of water quality monitoring. The advantage of this system is that it informs remotely and wirelessly about the water quality based on threshold values of measured water quality parameters. However, it fails to utilize the data collected on cloud as it was not used to learn from patterns in it and estimate water quality [2].

ISSN: 2233-7857 IJFGCN

[3] Analyzing water-borne diseases susceptibility in Kolkata Municipal Corporation using WQI and GIS based Kriging interpolation

Authors Sk Ajim Ali and Ateeque Ahmad have performed analysis of chemical parameters of ground water to come up with water quality index and map waterborne diseases, liability areas in Kolkata Municipal Corporation, India. The goal of their study was to demonstrate the importance of geographical information system based Geo statistical technique for identifying waterborne diseases sensitive areas by using ground water parameters, water quality index and water-borne disease reports into consideration. They demonstrated use of Kriging interpolation technique to generate liability map of water-borne diseases by plotting water parameters, water quality index and water-borne disease reports. The result of the study was the identification of weaker zones concerning water quality. Their proposed system is gainful since it acquires vital insights from water quality data. On the other hand, their study was limited to ground water quality in Kolkata region only and no other water body was taken into consideration [3].

[4] Prediction of water quality and smart water quality monitoring system

The authors Karthick T., Gayatri and Tarunjyot proposed a system to monitor and predict quality of water using sensors like pH and Dissolved oxygen sensor, pH sensor, turbidity sensor and temperature sensor. The system is proposed to overcome the conventional approach which includes manual collections and assessment of raw data. This is a reconfigurable shrewd sensor gadget that coordinates collecting information, operating and transmitting to remote location. They used K-Means clustering algorithm to train a water quality prediction model on collected data. Re-configurability and power efficiency are the advantages of this system. However in order to attain power efficiency the sampling was done at longer intervals [4].

3. Proposed System

A. Objectives

Following are the objectives of proposed system:

1. To produce an innovative, intelligent, autonomous data collection system to predict waterborne diseases from water quality parameters, as there does not appear to be any current system in place that fulfills all the requirements.
2. To develop knowledge platform to support disease prediction based on water quality.
3. To check water quality, predict disease based on it and inform people about possible disease so as to take appropriate preventive measures.

B. System Overview

Dataset

The system has a primary and a secondary data source. Collection of data about variety of water related diseases will require study and surveys to be conducted in different weather conditions and at least throughout a year. Since this is time consuming, we used secondary data source to train machine learning classifier model. The water parameter data was collected from the West Bengal Pollution Control Board's Water Quality Information System [6]. We preferred this data for machine learning task due to its authenticate and accurate nature. The primary data refers to data collected from proposed sensor network for water quality monitoring. This data is utilized to make predictions using the earlier trained model.

System Architecture:

The proposed work put forth an embedded system for water quality monitoring and disease prediction. The system comprises of hardware components like sensors, microcontroller units (MCUs) and display, etc. interfaced with each other. The software components of the system are deployed on the MCU. Fig 4.1 gives the pictorial representation of the projected system.

ISSN: 2233-7857 IJFGCN

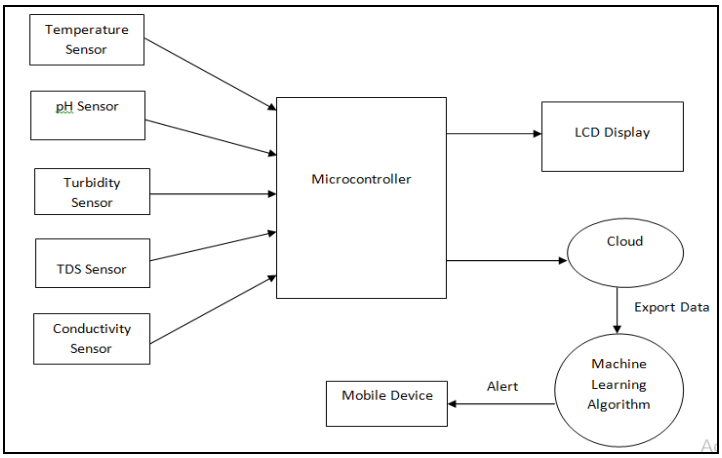


Fig 1: System Architecture

The proposed system for water quality monitoring for disease prediction involves Temperature sensor, pH sensor, Turbidity sensor, TDS sensor and conductivity sensor to continuously assess water quality. All the sensors are interfaced with central microcontroller (NodeMCU, Arduino) which performs the task to fetch sensed data from sensors. Collected data is transferred to cloud server for storage. At the same time it is displayed on a scrolling display interfaced with system for visual cue. The data is exported from cloud storage and fed to Machine Learning Classifier for the disease prediction task. The Machine Learning classifier is previously trained using secondary data on water quality parameters. Finally the predicted disease is conveyed as an alert to concerned mobile devices in the form of push notifications.

C. Technical Descriptions

DS18B20 Temperature Sensor

The DS18B20 underwater temperature sensor delivers 9 - 12-bit temperature sense output which indicates the temperature of the water. The DS18B20 links with a central processor over a 1-Wire bus .That is it requires only a single data line and ground for interconnection with a central microprocessor. It can also derive power straight from the data line also called as “parasite power”, which removes the necessity for a peripheral power source. The sensor operates within a voltage range of 3V to 5V. Every such sensor has distinct 64-bit serial code, thus many DS18B20s can work on the same one-wire bus. So we can use a single micro-controller for many DS18B20 as per the requirement of the system. The features of DS18B20 also include programmable alarm options, 12 bits of precision, usable from -55 to 125°C (67°F to +257°F).

E201-C-9 pH sensor

A pH Meter or a pH probe refers to a systematic tool that monitors the hydrogen-ion concentration of a solution, representing whether it is acidic or alkaline. It captures the variance in electrical potential among a pH electrode and a reference electrode. The pH meter needs to be calibrated properly in order to work efficiently and accurately. Calibration has to be done each time you start to use them. Once calibrated the meter precisely transforms voltage input into pH values. Standard buffer solutions , distilled water for example, are used for this process.

Turbidity Sensor

Turbidity relates to cloudiness of water. The working principle of a Turbidity sensor is that when light is transmitted through water, the extent of light transmitted through it is reliant on on the quantity of soil present in the water. The amount of light transmitted through water decreases with the increase in soil

ISSN: 2233-7857 IJFGCN

present in water. The turbidity sensor thus captures the quantity of transmitted light to estimate the turbidity of the water.

TDS Sensor
Total Dissolved Solids (TDS) refers to the extent of soluble solids present in water. The larger value of TDS notifies more soluble solids concentration in water and thereby the impurity of the water. Hence, the TDS value might be used as an indication of the purity of water. We have used a plug and play analog TDS sensor kit which is easy to use.

Conductivity Sensor
Conductivity is nothing but the capacity of the material to convey electric current. It can be considered as the reciprocal of the resistance. Conductivity is a significant factor of water quality. It can indicate the extent of electrolytes existing in water. We have used an elevated form of electrical conductivity meter V1 that highly increases the accuracy of measurement. It works within 3V to 5V input range, and is well-suited with 5V and 3.3V microcontroller board.

DOT Matrix Display
A P10 dot-matrix display is an electric alphanumeric display tool that displays information requiring a modest alphanumeric (and/or graphic) display device of restricted resolution. The display comprises of a dot matrix of lights or mechanical displays arranged in a rectangular formation such that by swapping on or off particular lights, text or graphics can be shown. A dot matrix regulator translates commands from a microprocessor into signals which turns on or off display features in the matrix so that the necessary presentation is formed.

4. Implementation Details

A. Feature Selection
Ten parameters chosen to train disease predictor include Ammonia, Chloride, Fluoride, Nitrate, pH, TDS, TSS, Turbidity, Zinc, and Total Coliform. These parameters are chosen depending on literature review, study of causes of various water related diseases and relevance of parameter with respect to disease outbreak [3]. Diseases included in study are Diarrhea Fluorosis, Convulsions, Bladder Cancer, Kidney Diseases, Metabolic Alkalosis, Methemoglobinemia, Dental Corrosion, and other Gastrointestinal Diseases like Cholera and Typhoid.

Table 1: Water Quality Parameters and associated Diseases

SR. NO.	PARAMETER	UNIT	MAXIMUM PERMISSIBLE LIMIT	ASSOCIATED DISEASE
1	TOTAL COLIFORM	PPM	0	GASTROINTESTINAL DISEASES LIKE CHOLERA, DIARRHEA, HEPATITIS
2	TURBIDITY	NTU	3	GASTROINTESTINAL DISEASES
3	AMMONIA	PPM	0.5	CONVULSIONS
4	CHLORIDES	PPM	250	BLADDER CANCER
5	FLUORIDES	PPM	1.5	FLUOROSIS

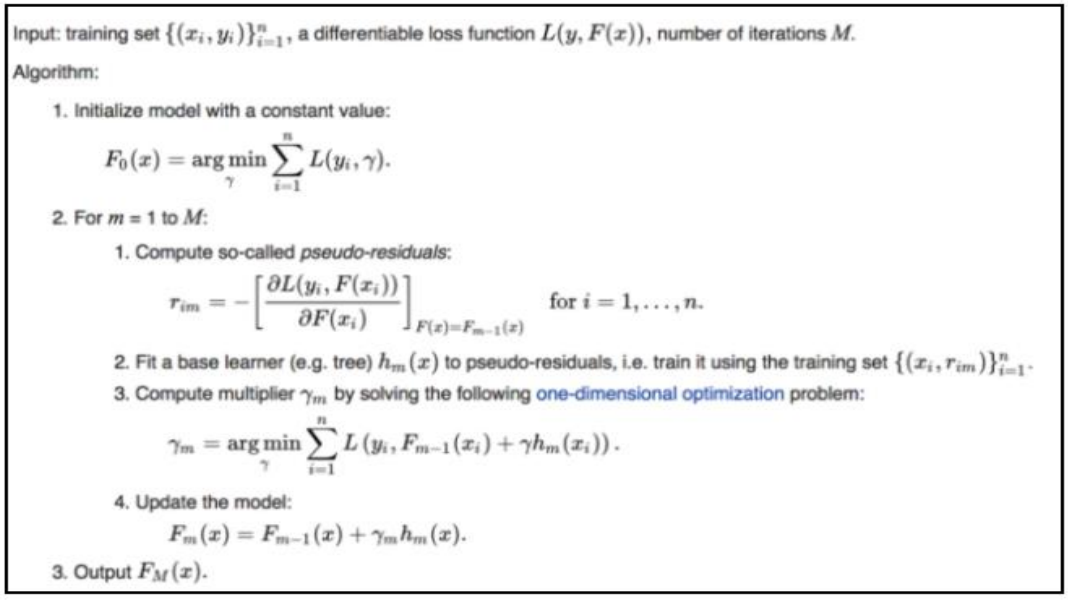
ISSN: 2233-7857 IJFGCN

6	NITRATE	PPM	1.5	METHEMOGLOBINEMIA
7	TOTAL DISSOLVED SOLIDS (TDS)	PPM	300 – 500	KIDNEY DISEASES
8	TOTAL SUSPENDED SOLIDS (TSS)	PPM	50	KIDNEY DISEASES
9	PH **I.** ACIDIC **II.** ALKALINE	-	 6.5 – 7 7 – 8.5	 DENTAL CORROSION METABOLIC ALKALOSIS
10	ZINC	PPM	5	DIARRHEA

B. Gradient Boosting Algorithm

Gradient Boosting is an ensemble based machine learning algorithm that involves combining a group of weak machine learning models in order to get a better and stronger learning model. Since it is a boosting technique the models are trained sequentially so that each subsequent model improves on the error of its precursor while optimizing a loss function at the same time. Usually for gradient boosting classification decision trees are used. Thus a gradient boosting classifier is an ensemble of a number of weak decision tree models. Gradient boosting involves following major tasks:

1. Modeling data with one of the weak models.
2. Obtaining error residual that is data which cannot be fit by the model.
3. Fitting another model focusing to fit residual data of precursor model.
4. Summing up models to get stronger learning model. Repeat steps 2 to 4 until a predetermined loss function is optimized.


Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For $m = 1$ to M:
   1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner (e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$.
   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

   4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

Fig 2: Gradient Boosting Algorithm –Wikipedia

ISSN: 2233-7857 IJFGCN

C. DFD Diagram

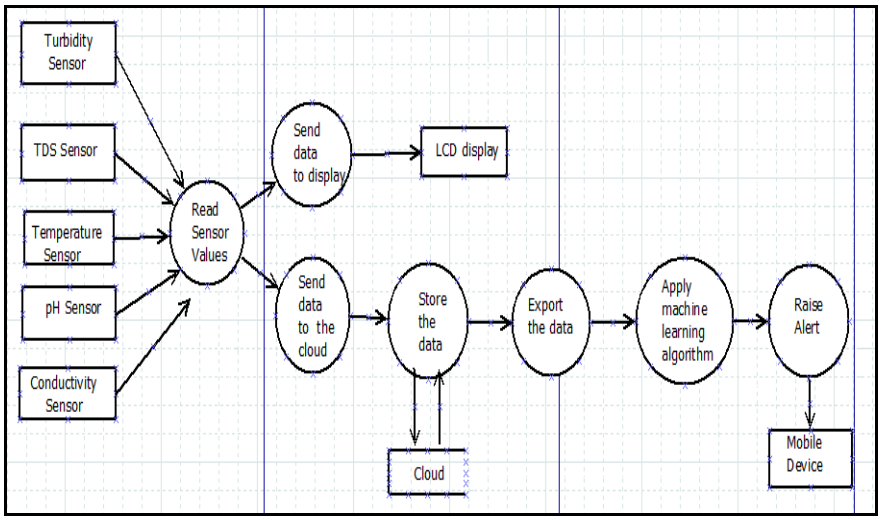


Fig 3: DFD Level-1

Above diagram shows the flow of data through Water Quality Monitoring for Disease Prediction (WQMDP) system. First of all we collect all five sensor values, turbidity, TDS, temperature, pH and conductivity. Microcontroller then forwards these values to cloud server for further processing and to LCD display. Cloud server stores the data and raise alerts after processing regarding diseases which are likely to spread.

5. Results

Waterborne diseases can have severe impacts on human health and may even lead to loss of life. One way to minimize this destruction is continuous monitoring of water quality and early prediction of such disease outbreak. Keeping this as goal a water quality monitoring and disease prediction system was developed. Sensors were deployed to capture measures of water quality parameters. It included pH sensor, temperature sensor, conductivity sensor, TDS sensor and turbidity sensor deployed at water source. Sensors were connected to microcontroller board to fetch sensed values and display them on scrolling display matrix. Collected data was stored on cloud storage leveraging wireless communication technology.

The parameters for disease prediction were analyzed and selected based on their correlation with water related diseases. A Gradient Boosting Classifier model was trained based on secondary data on water quality. An accuracy of 0.92 and 0.95 was achieved on cross-validation and testing data, respectively.

Pipeline leaderboard

Rank	Name	Algorithm	Accuracy (Opti...	F1 macro	F1 micro	F1 weigh...	Precision...	Precision...
★ 1	Pipeline 1	Gradient Boosting Classifier	0.952	0.857	0.952	0.932	0.844	0.952

ISSN: 2233-7857 IJFGCN

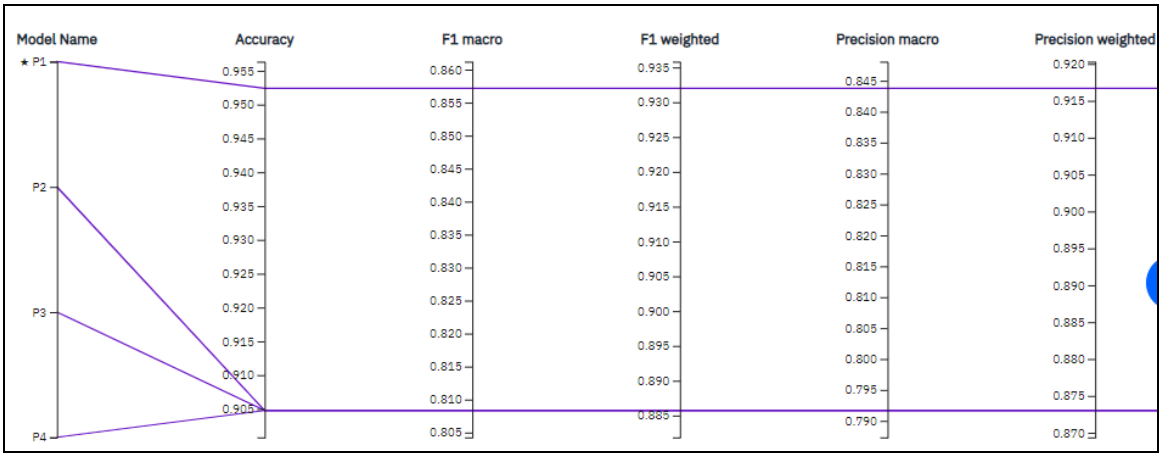


Fig 4: Performance of Gradient Boosting Classifier (*P1) on primary dataset

Fig 5 represents the classification report of trained machine learning model which informs about various performance metrics on individual prediction classes. For all disease classes, precision, recall and F1-score was calculated based on test dataset. The average of these individual values gives an accuracy of 0.95 to the classifier.

Classification Report:

	precision	recall	f1-score
Bladder Cancer	1.00	1.00	1.00
Convulsions	1.00	0.60	0.75
Dental Corrosion	1.00	1.00	1.00
Diarrhea	1.00	1.00	1.00
GIS	1.00	0.97	0.99
Kidney Diseases	0.83	1.00	0.91
Metabolic Alkalosis	1.00	1.00	1.00
Methemoglobinemia	1.00	1.00	1.00
No Disease	0.67	0.80	0.73
avg / total	0.95	0.95	0.94

Fig 5: Classification Report of Gradient Boosting Classifier

The model was used to predict possible waterborne and water related disease. The disease predicted was conveyed to mobile devices in the form of push notification. Based on accuracy obtained, it can be demonstrated that water related diseases can be effectively predicted based on water quality parameters.

ISSN: 2233-7857 IJFGCN

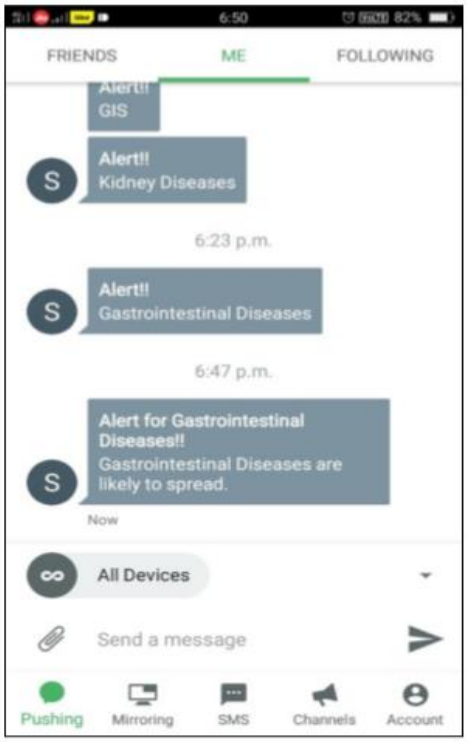


Fig 6: Push Bullet Alert about predicted disease

6. Conclusion

The water quality monitoring system for disease prediction is successfully designed and implemented. The sensors captured water quality data on continuous basis. The data was visualized using matrix display interfaced with microcontrollers. Collected data was uploaded to cloud storage using wireless networks. For disease prediction a machine learning model was trained and tested successfully with significantly good precision. The model is employed to make predictions on newer data. The alerts from the system are raised wirelessly and remotely. The study thus demonstrated importance of water quality monitoring and use of machine learning techniques in early prediction of water related diseases.

Acknowledgement

Authors are extremely thankful to Arun Patokar Sir, Deltiin India Tech Pvt. Limited for his continued support throughout the implementation of system.

References

[1] Manisha Bhende, Sanjeev J. Wagh, Amruta Utpat, “A Quick Survey On Wireless Sensor Networks”,fourth International Conference On Communication Systems And Network Technologies, 2014
[2] Ankita Taware, Vrunda Ghate, Minal Gaingade, Mayuri Bhandvalkar, Dr. Sanjeev Wagh, “Water Quality Monitoring System using IoT”, IJARP,ISSN 2456-9992,Volume 1,Issue 1,July 2017
[3] Sk Ajim Ali and Ateeque Ahmad, “Analysing Water-Borne Diseases Susceptibility in Kolkata Municipal Corporation Using WQI And GIS Based Kriging Interpolation”, GeoJournal, May 2019.
[4] Karthick T, Gayatri Dutt’s “Prediction of Water Quality and Smart Water Quality Monitoring System in IoT Environment”, Department of Information Technology. SRM University. Chennai, India.
[5] https://www.who.int/sustainabledevelopment/housing/healthrisks/waterborne-disease/en/
[6] http://www.wbpcb.gov.in/

ISSN: 2233-7857 IJFGCN