2020

Real-time drinking water quality detection and classification using wireless sensor network and K-nearest neighbor algorithm

Kindie, Abebe

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

FACULTY OF COMPUTING

Real-time drinking water quality detection and classification using wireless sensor network and K-nearest neighbor algorithm

BY Abebe Kindie

BAHIR DAR, ETHIOPIA

August 20, 2020

Real-time drinking water quality detection and classification using wireless sensor network and K-nearest neighbor algorithm

Abebe Kindie Awuraries

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology, BDU in partial fulfillment of the requirements for the degree of

Master of Science in the Computer Science in the Computing Faculty.

Advisor Name: **Mekuanint Agegnehu (PhD)**

BAHIR DAR, ETHIOPIA

August 20, 2020

DECLARATION

I, the undersigned, announce that this proposition includes my own particular work. Inconsistence with globally acknowledged practices, I have recognized and refereed all materials utilized as a part of this work.

Name of the student: Abebe Kindie Awuraries Signature ________

Date of submission: 14/12/2012 E.C

Place: Bahir Dar

This thesis has been submitted for examination with my approval as a university advisor.

Advisor Name: Mekuanint Agegnehu(PhD)

Advisor's Signature: ________

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND GRADUATE STUDIES

FACULTY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student Abebe Kindie Awuraris Signature ________ Date 17/08/2020. As members of the board of examiners, we examined this thesis entitled "Real-time drinking water quality detection and classification using wireless sensor network and K-nearest neighbor algorithm" by Abebe Kinde Awuraris. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of science in "Computer Science".

Board of Examiners

	Signature	Date
Name of Advisor Mekuanent Agegnehu (PhD)	________	19/08/2020
Name of External examiner Sosina Mengistu (PhD)	________	15/08/2020
Name of Internal Examiner Esubalew Alemneh(PhD)	________	19 Aug 2020
Name of Chairperson Abreham Debasu (Ass. Professor)	________	12/12/12
Name of Chair Holder Mr. Haileyesus Amsaya	________	19/08/2020
Name of Faculty Dean Belete Biazen (Ass. Professor)	________	12/12/12

To all, who with me during my challenge

Acknowledgements

First, I would like to thank GOD for helping me to write this thesis work. Next, I would like to thanks my honorable advisor **Dr. Mekuanint Agegnehu** who contributed in giving his continuous advice and comments in this study.

I also thanks everybody who helps me on the development of this thesis work, especially my classmates and friends. Finally, I would like to say thank my family for who wishes my life success.

Table of Contents

List of figures

List of tables

Abstract

Water is the most vital resource for living things, especially for human beings. In order to be healthy, everyone must have taken good-quality water in his/her entire life. However, checking the quality of drinking water is challenging task. The traditional way of checking drinking water quality methods involves manual sampling of waters followed by laboratory analysis. This process may be difficult to cover a large geographical area and needs a huge amount of budget for running this process.

This research aims to monitor the quality of drinking water based on wireless sensor network (WSN) and machine learning algorithms. We have used water temperature, PH and turbidity sensors to collect real-time data from different water contents. These sensors sense different parameter of waters and transmit the sensed data to the base station using Wi-Fi module.

For detecting and classifying the abnormal behavior of the drinking water, we used k nearest neighbor (KNN) supervised machine learning classification algorithm on the collected data. The reason of using KNN classification is that its appropriateness for having small amount of data sets and suitability for having more than two class datasets.

Evaluation is performed on k nearest neighbor classification algorithms using accuracy as evaluation metrics. In this research the KNN classifiers classifies the data with the accuracy of 94%. The results show that the KNN algorithm is having better performance on wireless sensor data with multi labeled datasets compared to SVM classifier.

Key-Words: KNN, WSN, SVM, machine learning, classifications, water quality, sensors.

CHAPTER ONE

INTRODUCTION

1. Introductions

Ethiopia is one of the water richest country among Africa continent, which is the source of Blue Nile and many other lakes, ponds, reveres, and lakes. Even if our country is having much amount of waters especially underground water for drinking purpose, the way of treating and checking the quality of these waters is a tedious work, because of the reason that waters quality is measured in both physical as well as chemical properties. According to the American nonprofit developmental aid organization stated that 61 million Ethiopians peoples lack access to safe water, which indicates that more than half of the country peoples faced with accessing of good water quality. Those 61 million Ethiopians do not have any mechanism to check whether the given drinking water is good or poor quality unless using a traditional manual system. Traditional methods of checking water quality mainly involves on the manual collection of water sample at different locations, and then followed by laboratory analytical techniques in order to check the water is drinkable or not. Such approaches take longer time and no longer to be considering being efficient (venkatesh et al, 2017). There is need to improve existing system for monitoring water bodies especially drinkable waters body, given that laboratory methods are too slow to develop an operational response and does not provide a level of public health protection in real time (Venkateswaran et al, 2017).

In order to solve the stated problem we need have to some mechanism that will collect data from the water body continuously and then make a decision based on collected data that help us to notify about the status of drinking water quality. According to (Jayti bhatt et al, 2016), in the past decade, all human being life changed because of having worldwide telecommunication which is called in general Internet. The internet of things has been heralded as one of the major development to be realized throughout the internet collection of technologies. The WSN(wireless sensor network) represents a concept in which, network devices have the ability to collect and sense data from the world or physical environment, and then they share their data across the internet where that data can be utilized and processed for various purposes (Jayti bhatt et al, 2016).

Now days the combination of wireless sensor network and Internet of Things (IoT) are "a worldwide infrastructure for the data society, empowering propelled benefits by connecting (physical and virtual) devices in light of existing and advancing interoperable data and communication technologies" (Ray, 2018). A huge amount of data are emerging day by day to be part of the IoT infrastructure and the number of information connected to IoT is expected to reach 50 billion by 2020 . This would produce huge amounts of critical and real time information for the society in order to automate and simplify their day to day activity especially for weather monitoring , smart city, smart agriculture, smart industry and so on.

In general, the basic aim of IoT and WSN are to make our nature around us smarter, simpler, by giving real time, and relevant information's. This would upgrade our capacity to deal with our nation, cities, streets, health, homes, forests and much more. By having those large amounts of data we will have better crowd and traffic management, emergency predictions, better prediction of accidents and crimes, environmental monitoring etc.

The data's from the WSN gadgets will be gathered and used to control complex conditions around us. Simply collecting data from the WSN device is not enough unless we apply some mechanism on the data that can convert in to good knowledge and wisdom. Therefore, we need have to some pattern recognition mechanism, pattern recognition focuses on recognizing regularities and similarities across different set of data. Machine Learning (ML) is a class of methods in data analysis that learns patterns and hidden insights in the data without being explicitly programmed for it. Thanks to the better and more powerful computing devices, there is an increasing trend to apply ML in various cases recently, such as equipment failure detection, pattern and image recognition, email spam filtering, and fraud detection (G. Arroyo & B. G. Zapirain, 2011).

ML algorithms can also perform predictive analysis. Hence, ML methods are useful to analyze situations where theoretical knowledge is still not enough for the analysis. Machine learning approach is useful for wireless sensor network device data for better predictions.

In this research, we used water temperature sensors, pH sensors and turbidity sensors are used for sensing the behavior of the water (the detail description of those sensors are discussed on the chapter three). The above sensors can be directly connected to the base station through wireless network then we applied machine-learning technique for classifying the quality of the water based on the data collected by the above sensors. For classification of drinking water quality, we used K-nearest neighbor (KNN) supervised machine learning classification technique.

The KNN classification algorithm is a machine learning technique, which is theoretically mature with low complexity. The idea of KNN algorithm is by calculating the distance between a point A and all other points, remove the K points with the nearest point, and then count the K points, which belong to the classification. The nearest neighbor refers to the single or multidimensional feature vector that is used to describe the sample on the closest, and the closest criteria can be the Euclidean distance of the feature vector (WenchaoLi et al, 2014).

In this study water quality classes are three in which drinking water quality may fall under poor, excellent and normal classes. After classifying, the given water based on KNN classifier, the user can be notified by email notification mechanism. In general, the main objective of this research is to find out the class of drinking water (excellent, poor or normal) by taking temperature, turbidity and pH sensor values.

1.1 Motivation of the study

Water is one of the most important resources in the world. It has direct impact on the daily life of humankind and sustainable development of society (Maher, 2018). Monitoring the quality of the water has become imperative area nowadays. Drinking water can be easily contaminated by human actions such as industrial or waste pollution. This contaminated drinkable water can cause illness or death of human being. Water quality has to obey to strict regulations of world health organization and can be monitored constantly.

The WSN brings new possibilities for water quality control and monitoring way. In expensive connected devices can be deployed in key points such as water container, valves or needed point. These can constantly monitor water parameters and communicate to a system that would alert users as soon as a problem detected or the water is contaminated.

Our motivation is to design a model that can detect the abnormality of drinking water and then classify as poor, good or excellent using wireless sensor and machine learning technique.

1.2 Statement of the problem

For a given country, healthy people and society play an important role for sustainable development. In most country especially in Ethiopia, most of the peoples are died by the cause of water-borne diseases like Diarrhea, Giardia , Typhoid Fever, Cholera, Hepatitis A and others. All the above diseases are caused by drinking low-quality water (Manayawukal, 2019). Now a day's most of the time water quality can be verified by using manual sampling of drinkable waters within the help of laboratory equipment's. In this approach, the result is directly affected by the sampling technique, sampling time, laboratory equipment's and so on. The quality of the water may be good when the sampling is conducting in the laboratory and the samples are taken with in time intervals, we need to have some mechanism that can monitor the quality of water continuously in real time manner unless we will drink poor quality waters that can cause ill or death.

The traditional way of checking drinking water quality is not efficient in terms of human resource, cost and safety. Most of the recent works are not intelligent, in which they are not designed in knowledge based decision making system as stated in related work. In this thesis, we are trying to address the above problem by combining wireless sensor networked devises with machine learning approaches to correctly classify the class of drinking water as poor, normal or excellent water. Our approach monitors waters quality in real time and continuously by deploying different water quality sensors on the water body.

Based on the problem specified above this research answers the following question.

- Can machine-learning methods used to automate the detection of abnormal events in drinking water parameter sensed by wireless senor networked device?
- What is the performance of KNN classifications on water quality monitoring system?

1.3 Objectives of the study

1.3.1 General objective

The general objectives of this research is, building a real-time drinking water quality detection and classification using wireless sensor network by applying supervised machine learning method.

1.3.2 Specific objectives

The specific objectives of this research are listed below

- To identify important parameters for water quality detection.
- To configure water quality measurement sensors to collect real-time data.
- To design KNN-based water quality classification model
- To evaluate the performance of the water quality detection and classification method

1.4 Technique of thesis study

In the research methodology, specific procedures or techniques used to identify select, process and analyze information about the research topic presented as below in figure 1.

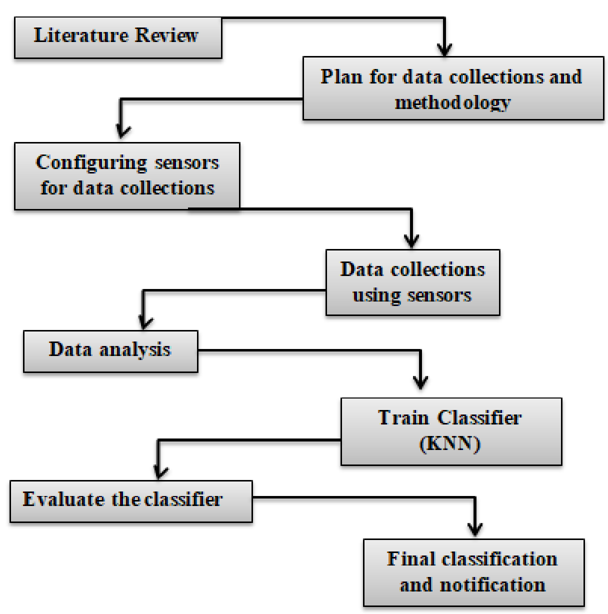


Figure 1: Flowchart of the research methodology

1.5 Data Collection

The data collection procedure for this research is done by deploying different water quality sensors. The sensors used in this thesis are water PH, Turbidity and Temperature in different water contents in controlled laboratory.

1.6 Experiment Tools

In order to achieve the research objectives different hardware and software tools are required.

1.6.1 Hardware requirement

Table 1: Hardware requirement

No	Material name	Purpose
1.	Arduino mega kit	Used for the IoT development board for getting data from the sensor
2.	Temperature sensor	Used to sense the water temperature
3.	Turbidity sensor	Used to sense the water Turbidity or to check muddiness of the water
4.	pH sensor	Used to sense the alkalinity of the water weather it acidic or basic type
5.	ESP32	This chip is used to send data to cloud using Wi-Fi communication technology
6.	Battery 9V	Used as a power supply for the system
7.	Laptop	Used for analyzing sensors data by using the appropriate tool for efficient classifications
8.	Others supporting	Resister, jumper wire, LED, Bread Board, soldering tool …

1.6.2 Software Requirement

Table 2: Software requirement

No	Software name	Purpose
1.	C language	I used C language for programming/ integrating our sensors with the development board
2.	MATLAB	Used to analyze and simulate the sensor's data on the computer
3.	Thing speak	Used as a server for storing sensors data in the cloud to access everywhere
4.	MS office	For document preparations and presentations tool

1.7 Performance Evaluation

For a given algorithm or technique there must be a way in which that algorithm is good or not in terms of time and cost as well as correctness. In our case, we selected KNN classification technique to classify waters quality based on the data collected. After we have trained the proposed experimental system by using the training data set, finally we tested systems performance by applying performance evaluation testing mechanisms, which is accuracy. We used SVM (Support Vector Machine) classification algorithm for comparing the result with the KNN model. We compare the result of both classifier using accuracy as a performance measurement matrix.

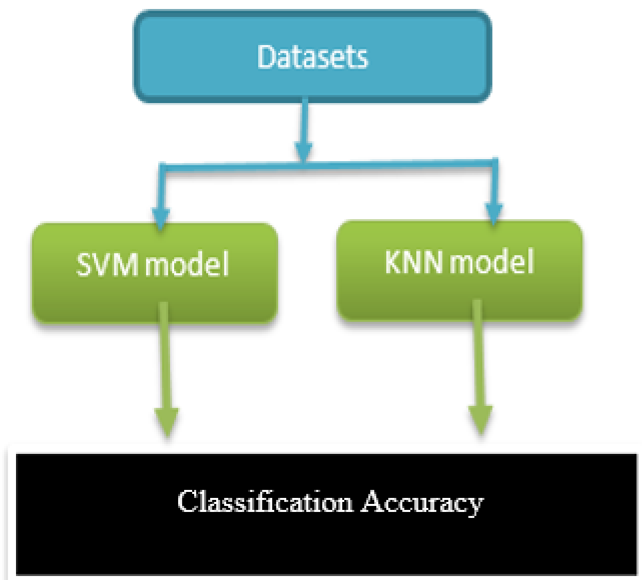


Figure 2: Performance evaluation technique

1.8 Scope and Limitations

The scope of this research relies on drinking water quality monitoring system. For monitoring, we have used temperature sensors, to sense the water temperature, pH sensor that is the basic determinant of water quality and used to sense the alkalinity of the water and turbidity sensor, which used to sense waters transparent behavior or neatness in physical view coordinating with Arduino board. For communication, we used Wi-Fi module for transferring data to the cloud. After getting of those data, we have used the k-nearest neighbor algorithm for classifying waters as poor, excellent or normal water. The limitation of this study is that we did not get more water quality sensors (minerals sensors) and large datasets for training due to time as well as budget constraints.

1.9 The significance of the thesis

Water is one of the most essential natural resources for the existence and survival of the entire life on our planet. Based on the quality of water, it can be used for different purposes like drinking, washing or irrigation. Plants and animals also depend on water for their basic survival. In short, almost all living things need a large quantity and good quality of water for existence (WenchaoLi et al, 2014). In this research drinking water is our target in which we are able to know whether given water is contaminated or not in real time and continuous monitoring approach. For checking the behavior of the water, it is better to use current technological advanced equipment, which is different water sensors that can sense continuously and pass the data to the cloud using wireless communication technology in real time. We can analyze whether the water is advisable for drinking purpose or not based on the data that we have collected in the base stations by applying machine learning algorithm. If one person knows about a certain drinking water quality, he/she will be sanctioned from drinking that contaminated water. In general, he/she is free from the water born diseases in which a killer diseases. So the advantage of this research is related to saving humans life rather than easy and safety way of living style.

1.10 Organization of the thesis

The rest of this thesis is organized as follows. In Chapter 2, literature review and available researches conducted by different scholars on the area of water quality control mechanisms using wireless sensor network, their approaches, and findings are discussed. Chapter 3 presents research methodology that includes Materials, methods, architecture, algorithms and related concepts are clearly explained. Result and conclusions are presented in Chapter 4 and 5 respectively.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

2. Introduction

This chapter discusses basic overview of drinking water quality controlling mechanisms. On this study wireless sensor network, machine learning classification technique and drinking water quality monitoring approach are discussed in detail. Finally, related works that has been worked in previous are presented.

2.1 Machine Learning and Wireless Sensor Networks

The proposed study lies on drinking water quality detection and classification using machine learning and wireless sensor network, we need to discuss on those two basic area for better understanding of the importance of ML and WSN. These two techniques are used for the tools of analyzing and collecting big data technology. Actually, machine learning was introduced in the late 1950's as a tools and techniques for developing artificial intelligence (AI) concepts (Ayodele, 2010.). Now a day's machine learning is familiar in the field of computer science in which computing devices are thinking like human beings. We can apply machine learning in different field of study to solve real facing problems. Using machine-learning algorithm we can mine meaning full information from unorganized big data. For mining knowledge from unorganized data, experts are using different machine learning algorithms and techniques. The algorithms and techniques used in machine learning comes from different fields of study including computer science, mathematics, statistics and neuroscience.

According to (Mostafa et al, 2012) most machine learning algorithms fall into the categories of supervised, unsupervised and reinforcement learning. In the first of machine learning algorithm, the machine learning task can be done by pre trained data. On this case, first the machine will learn by observing the input-output relations ship between different instances. Based on the observation, the machine will respond for new coming instances. In contrast to the first supervised learning, unsupervised learning algorithms are not provided with input-output relation rather we only feed inputs and the machine will respond the output by observing the input instance. The goal of an unsupervised learning algorithm is to classify the sample sets to different groups by investigating the similarity between the input samples (Mostafa et al, 2012).

The third category of machine learning technique is reinforcement-learning algorithms, in which the machine will learn by observing the physical environment without feeding both input and output samples of instance.

The following two basic explanation capture the essence of machine learning.

a. The development of computer models and algorithms for learning processes that provide solutions to the problem of knowledge mining and enhance the performance of developed systems (Duffy, 1997).
b. The adoption of computational methods for improving machine performance by detecting and describing consistencies and patterns in training datasets (P. Langley et al, 1995).

By combining the above two broad concept we can solve real facing problems using machine learning approach and WSN

According to (Mohammad Abu et al, 2014), sensor network designers must characterize machine learning as a collection of tools and algorithms that are used to create classifications and prediction models. However, machine-learning experts recognize it as a multi-disciplinary field of study with very large idea and patterns. Understanding such patterns will be beneficial to those who wish to apply machine learning to WSNs, in our case water quality wireless sensor network. By Applying to numerous wireless sensor networks applications, machine-learning algorithms and its approach provides various flexibility advantages. On this section, we provide some of the theoretical concepts and strategies of adopting machine learning in the context of wireless sensor networks. The importance of machine learning on the area of wireless sensor network can be grouped in to two main broad concepts.

On the first case, wireless sensor networks are widely designed for responding the dynamic nature of the world. For monitoring a given environment, we have to design systems that can respond with the environment variations. Sensor devices by itself cannot respond dynamic nature of the world unless we design some algorithm or technique on the devices. When devices are fabricated, they are designed for a specific purpose unless we re-programmed it again. However, using machine-learning approach sensors devices can respond for all dynamic nature of the environment. By Appling machine learning algorithm, wireless sensor network can also adopt changes and used in very harsh or poor environment where the wired network cannot be deployed.

On the second case, the availability of wireless sensor are essential for big data technology. According to (Beom et al, 2019) data collection is the first step of building a big data system. Among the many available data-generating sources, wireless sensor networks are receiving considerable research attention with regard to environmental monitoring. A WSN consists of a large number of sensor nodes that monitor and record the physical conditions of an environment, and the sensor data are collected. When the availability of the data increase, it is difficult for human beings to change that big data to a knowledge or information's. Machine learning play an important role for analyzing and mining knowledge from big data.

In our case, we use machine-learning approach used to classify quality of water based on the parameter that we have collected from the water using water quality wireless sensors devices.

2.2 Machine Learning Classification Technique

Classification is the main and basic technology in data mining process. It has a wide range of applications in business, decision-making, management, scientific research and other fields. At present, the main classification techniques include decision tree, SVM classification, KNN classification, artificial neural network and so on (Jingwen et al, 2018).

In the following section, we will review some of the state art in machine learning classifications technique are presented which includes support vector machines, artificial neural networks and k-nearest neighbors according to(Shkurin, 2015)

2.2.1 Support Vector Machine (SVM)

Support vector machine is one of the basic supervised machine learning classification algorithms. The bases of this algorithm refers to the family of linear models. The model is trained by transferring of the original vector in the space of higher dimension and search for dividing hyperplane with the maximum gap in this space. Two parallel hyperplanes are constructed on both sides of the hyperplane separating classes. The separating hyperplane is a hyperplane that maximizes the distance to two parallel hyperplanes, which means that the border of the two classes in linear kernel space. These classification algorithms mostly applied on binary classification cases rather than multi-class dataset. For our study, we do have three classes to group the abnormal behavior of the water. We used this classification algorithm in order to compare with the k nearest neighbor algorithms. On this thesis, we used linear kernel with one to all fashion due to the reason that we have three classes rather than binary class.

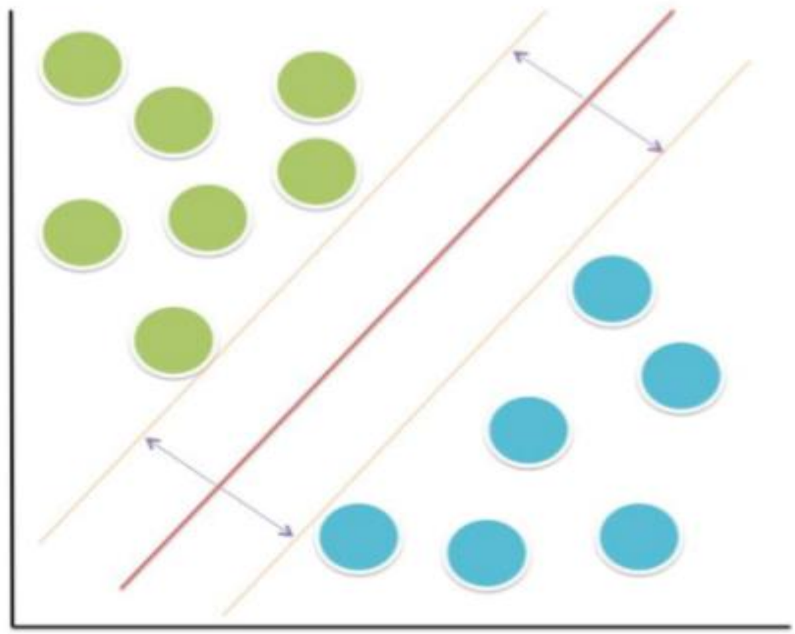


Figure 3: General graphical scheme of SVM model (adopted from Aleksei Shkurin, 2015)

2.2.2 Artificial Neural Network (ANN)

The artificial neural networks (ANN) are one of the most popular machine learning models nowadays, with a huge variety of possible applications, including regression, classification, introduced in 1943 by neurophysiologist Warren McCulloch and mathematician Walter Pitts (Shkurin, 2015). The basic of this model is in building several layers that are made up of a number of interconnected nodes, containing the activation function. The training set is presented to a model through input layer; one or more hidden layers perform processing by the system of weighted connections, taking each of the inputs for calculation, and finally output layer gives the fitted function.

The artificial neural networks needs a large amount of data for training the models , therefor then ANN classification techniques are not appropriate choice for us because of we do have small amount of datasets for training. The General graphical scheme of ANN model seems like below

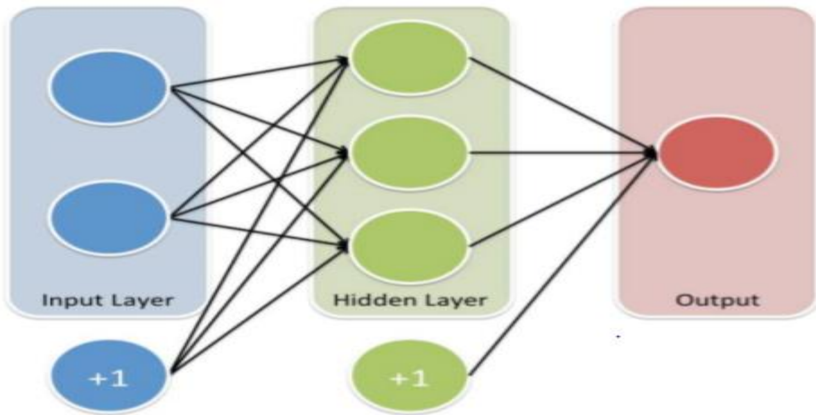


Figure 4: General graphical scheme of ANN model (adopted from Aleksei Shkurin ,2015)

2.2.3 KNN (K nearest neighbor classification).

The k-nearest neighbor algorithm is a technique for classifying objects based on closest training examples in the feature space. Not only this, KNN is also a sort of instance-based learning, or lazy learning where the function is only approximated locally and all computation is conceded until classification. The KNN classification technique can likewise be adjusted for use in assessing consistent variables. Such implementation utilize an inverse distance weighted average of the K-nearest multi-variate neighbors (Chin Heng et al, 2012). This calculation functions as follows:

Mathematical representation of the algorithm is given below:

1. Let's firstly define the distance function d(xi,xj), the distance function is Euclidean distance measuring as

$$DE(xi,xj)=\sum_{i=1}^{n}\sqrt{x_i^2-x_j^2}$$ -- (eq 2.1)

2. The result y of the model is calculated by majority voting of k neighbors with the lowest result of distance function

$$y=\frac{1}{k}\sum_{i=1}^{k}yi$$ --- (eq 2.2)

Where yi is the i-th case of the examples sample..

A principle favorable position of the KNN algorithm is that it works well with multi-modal classes in light of the fact that its decision is depend on a small neighborhood of similar target (george s, 2003). Subsequently, regardless of the fact that the target class is multi-modal, the algorithm can in any case lead to great precision.

On this research KNN classification is selected because of the reason KNN classification is a simple, effective, nonparametric method, has been widely used in text classification, pattern recognition, image and spatial classification and other fields. The basic advantage of KNN algorithm are the following

- Ease to interpret the output of the classifications.
- Calculation/execution time is slow even if it depends on the dataset size.
- Predictive power of KNN is good as illustrated in chapter 4.
- Robust to noise in the input data.
- Very easy to implement for multi-class problem.
- Simple to understand and implement.

2.3 Water quality monitoring technique

Good quality water is one of the most important natural resources in order to sustain our life existence and the quality of drinking water plays a very important role in the well-being and healthiness of living things especially for human beings. Water is used in various activities, such as daily food consumption, agriculture and travel, which may affect water quality (Venkateswaran et al, 2017). As the water is very essential for human beings, we need have to some mechanism to monitor the health of drinkable waters. In (Chapman, 1996) monitoring is defined as the collection of information at set locations which is deployed and at regular or timely intervals in order to provide data which may be used to define current conditions, establish trends or in general the status of the waters in real time. According to (Soundarya et al, 2017) , water quality monitoring system can be grouped in to two based on the way we are treating and accessing behavior of the waters.

- Traditional way of water quality monitoring approach
- Technology based water quality monitoring approach

Each monitoring technique details are described as bellow.

2.3.1 Traditional way of water quality monitoring approach

In the traditional way of water quality monitoring system, there are different water instruments, which used to monitor the quality of water. This traditional way of checking water quality, not enough to measure water quality and identify any drastic changes in it. Since the measurement is based on sampling technique the experiment is done with some period of intervals, it is difficult to monitor the water quality through 24/7 days.

As the author states that, there are three major steps to execute traditional water quality monitoring way. These three steps include different experts at different levels of the process.

The major three steps are as follows:

- Water sampling.
- Testing the samples.
- Investigative analysis.

In the first stage that is water-sampling phase, water samples will be collected in large mass using various tools in different place. These water samples are then examined in the laboratories by the help of laboratory equipment's. Water sampling and analysis are only performed by ISO-certified laboratories. Unreliable results enhance issues concerning pollution when a corrective response cannot be performed within time. Professional expert technicians can conduct the process of sampling and monitoring technique. Further to sampling, Testing is carried out. In water quality testing procedures can been classified into "Physical, Chemical, bacteriological and microscopic" categories.

Physical tests: These indicate properties that are detectable by our senses or physically we can sense the properties of the water by our eye. Some of the properties that we can sense are Color of the water may be caused by the presence of minerals or by the presence of vegetable origin such as algae and weed, neatness of the water, total solids, dissolved solids, taste and odor of the water.

Chemical tests: In these test the amount of mineral and organic substance that affect the water quality will be tested. These tests determine the parameters in water like "pH (power of hydrogen), hardness, and presence of a selected group of chemical parameters; biocides, highly toxic chemicals and so on".

Bacteriological tests: This test shows the presence of bacteria, a characteristic of faecal pollution. These tests examine to identify the presence of microbial pathogen in the water that might occur with contamination. The presence of such organisms indicates the presence of faecal material and thus of intestinal pathogens.

Finally, after the above technique have been conducted the tested water samples are then thoroughly monitored and observed by professional expert who can read through the lines of the resulted report. They then make an investigative analysis with a parallel consideration of the historical records of the previous water tests. Any similarity of the currently extracted results to the previous records will give way to an intense deliberation for prediction of any unknown changes or hazards to the water quality.

In general, this process takes much amount of time in order to do laboratory test and analyses based on the data collected in the lab.

2.3.2 Technology based water quality monitoring approach

On this approach drinking water, quality can be monitored using different technological advanced devices. As we mentioned in the above we need to improve existing system for monitoring water bodies, given that laboratory methods are too slow to develop an operational response and does not provide a level of public health protection in real time manner. The following literatures are some of the papers that are dedicated to solve the problem using technology.

In the paper (Yungze Li et al, 2017), a low-cost measurement unit of water quality is designed using a waterproof Unmanned Aerial Vehicle (UAV) that increases the reliability and flexibility of previous sensor networks. Another research (M Tarique & H. Khalee, 2016) uses integrated sensors, transmitters, receivers, myRIO microcontroller, and IEEE 802.11 Wi-Fi technology to generate water quality data including pH, conductivity, and temperature. The real-time data is sent wirelessly to a local control unit for analyzing, recording and displaying. The system also sends alarm messages automatically to a remote management center when water quality falls below the required standard. One IoT based system (Sankaranarayanan, 2017) employs PH sensor and total dissolved solids (TDS) meter for measuring the water quality parameters pertaining to hydrogen ion and total dissolved solvents. In addition, machine learning algorithm K-Means clustering been employed for predicting the quality of water based on trained data set from different water samples.

The technology based water quality checking system is implemented using devices like Arduino Uno and Raspberry Pi3. By using different technology that is mentioned earlier, we can automate the water quality checking in real time and continuously.

In (Cheng-Liang Lai, July 2011) started with the problem of the high increase in global industrial product, from rural to urban drift and the over-utilization of land and sea resources, the quality of water available to people has deteriorated greatly. The high use of fertilizers in farms and other chemicals in sectors such as mining and construction have contributed immensely to the overall reduction of water quality globally in order to solve the problem using technology. One of the researches, Dempster-Shafer (D-S) (Dibo,H et al, 2013) method employed for checking the quality of the drinking water. Autoregressive model is employed for checking the water quality parameters. The Autoregressive model has been employed for estimating the water quality parameters using automated (live) water quality sensors. Finally, the Dempster-Shafer fusion method searches for anomalous probabilities of the residuals and uses the result of that search to determine whether the current water quality is normal or abnormal.

The proposed thesis work is categorized under the technological based drinking water quality checking mechanisms. This research aims to develop technological way of checking drinking water quality using wireless sensor network and machine learning approaches. The technology used in this study uses different water quality sensors, development board and implementation tools like MATHLAB and ThingSpeak platform. The final output is to notify the users about the status of the water in real time event.

2.4 Related work

In (Unnikrishna Menon et al, 2012) introduces a river water quality monitoring system based on WSN. The main objective of this work is to automate monitoring of rivers water quality in a remote site with continuous data transmission using ZigBee wireless communication modules. The proposed work is mainly designed on receiving water PH value. The author stated that, water PH value is the basic parameter to determine the quality of water inside rivers.

The authors present the system architecture for a Wireless sensor Network for river water quality monitoring technique. The author also designed schematic circuit diagram for amplifying the PH value of river waters and level of shifting. They used Peripheral Interface Controller (PIC) micro-controller for receiving data from water PH sensor using serial communications. After they received the value, they can send the sensed data using ZigBee communications. Their system can provides an energy efficient and low cost sensor node platform for monitoring water quality through the use of inexpensive, low power devices for the hardware design.

The way they are looking the problem and addressing the solution is good. Their main drawback is unable to integrate others water quality sensors for a better outcome even if they are stated as future work.

In (Arun Pandi T et al, 2018) proposed real time water quality monitoring system using multiple water quality sensors. Their system is designed in low cost and can able to respond dynamic environment in the water. The abnormal condition of the water are detected remotely in online-based technology. As they are presented, their proposed system is a low-cost system for real-time monitoring of water quality in IoT environments. They can monitor the physical and chemical propoeries of the water using different water quality p sensors. They used water temperature, turbidity, PH, Conductivity sensor to sense the parameters of the water. Those water sensor devices are low cost, and efficient for data collections. The collected data values from the sensors can be processed by the main micro-controller. They used Arduino mega 2560 for data processing, analyzing and sending to the cloud using Wi-Fi communications technologies. This can implement for suitable environment monitoring, ecosystem monitoring, etc. and the data can be viewed anywhere in the world. Even if the paper is well organized and solves the problem as stated but they have not use any machine learning approach to make the system more intelligent.

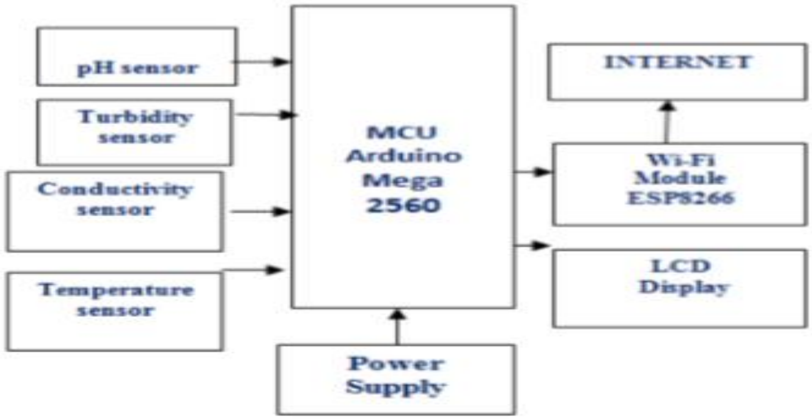


Figure 5:Multi-sensor based water quality monitoring in IoT environment framework In (Arun Pandi T et al, 2018).

In (Aaina Venkateswaran et al, 2017) they proposed a method to monitor the quality of water using the concept of internet of things. They designed to develop a low cost based real time monitoring way of water quality. They used different water quality sensors like temperature, Turbdity, PH,water conductivity and desolved oxigen sensors device. The main objective of this paper is to visualize the sensors value in Arduino terminals locally and finally users of the water can see the status of the water using their mobile by installing Blynk applications. Their system does not designed as more intelligent to advice humans whether the given water is contaminated or not. In general, the proposed work have not any decision-making system by itself unless the user can predict the quality of the water.

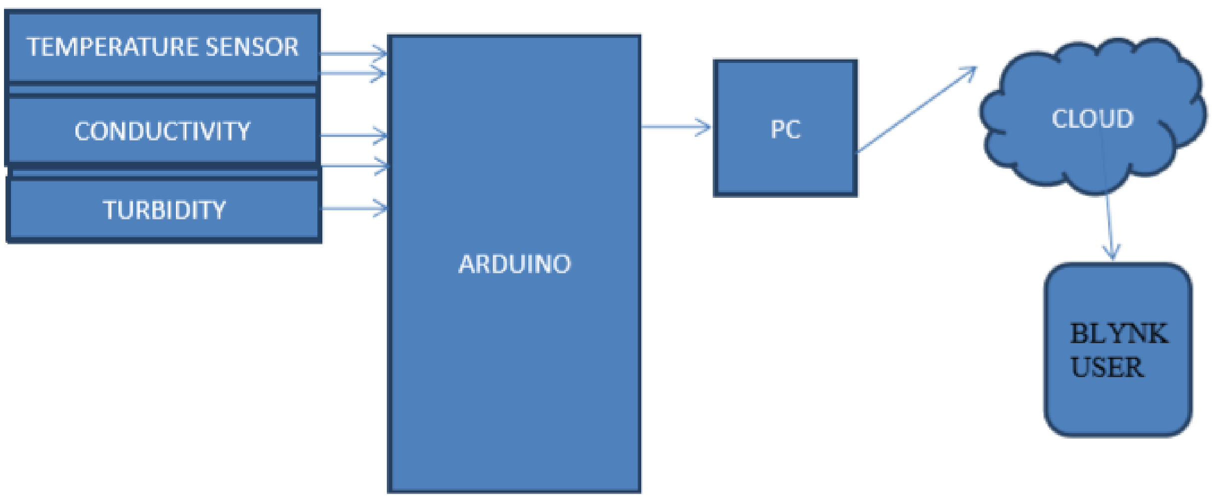


Figure 6: An IoT Based System for Water Quality monitoring system architecture

In (Vani K S and Rajesh, 2017) they have developed a system where the patient can be monitored remotely using wireless sensor devices like air quality, body temperature and pulse rate sensors. The sensed data can be stored on the cloud storage using Node MCU technology. Their cloud storages is ThingSpeak. On the ThingSpeak they applied Fuzzy logic algorithm to predict whether the patient have diseases or not. In addition to sending sensors data to cloud storage, they can also visualize the data locally using Liquide crystal display (LCD). The authors stated that after the data is collected, the data will be downloaded manually and then the inference rule can be applied locally. This system cannot monitor the patient independently without the involvement of humans.

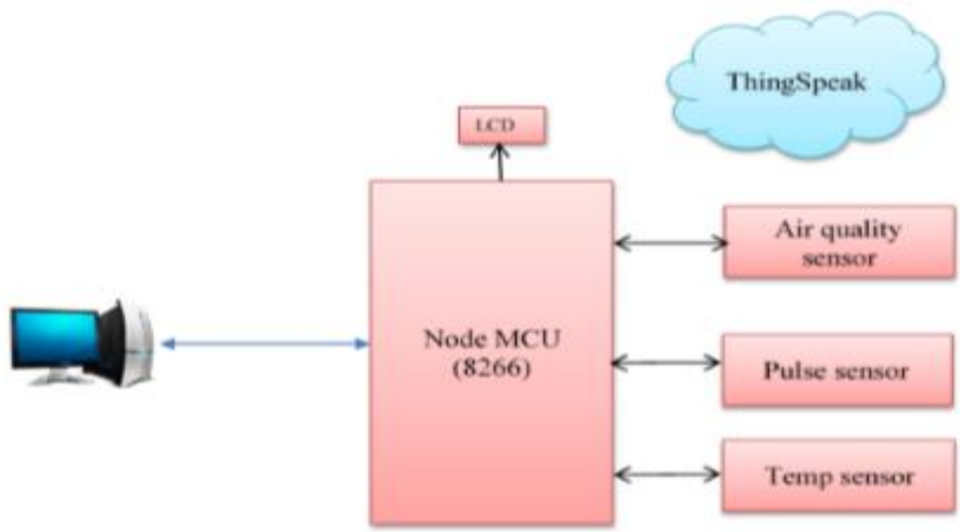


Figure 7: IoT based Health Monitoring using Fuzzy logic framework.

In (Yuthika Shekhar et al, 2017) the author's tries to develop an Intelligent IoT based Automated Irrigation system using temperature and soil moisture sensor data. The irrigation process is fully automated using wireless sensor network and machine learning approach. They used micro-controller to collect the data from the sensors using serial communications. In addition to those sensors, the authors also used water pump device as an actuator if the soil needs water based on the decision of the systems. They used KNN classification machine learning algorithm for classifying soil conditions. The different type of soil conditions are Dry, little Dry, little wet and Wet. The K-NN Machine learning algorithm used in the study, for predicting the soil condition based on soil moisture and temperature level. The predicted output is then used for sending the control signal via the serial communication to Arduino for controlling water pump for watering the field accordingly. Finally, the history of the data are stored on the cloud and the farmers can get information about their irrigations farming area using their mobile phones. In this paper the authors does not mention about the KNN classifier accuracy performance on the irrigation context.

In (Umale, 2018) ,an IoT based health monitoring way using Machine Learning was proposed for monitoring remotely located patients. The authors provide an interface among doctors, nurses and concerned relatives of patients. The main objective of this paper is to remotely monitor the health status of patients. The proposed work also can do simultaneously an alert notification if the heart rate of a given patient goes in abnormal conditions and provides data analysis using supervised machine learning models such as KNN, SVM, Naïve-Bayes and J48 classifier.

The proposed system can also sends alert notifications when abnormal condition is happen on the patient side. The data collected from the patients are Electrocardiogram (ECG), Oxygen Saturation (SpO2), Heart Rate (HR), Photo plethysmography (PPG), Blood Glucose (BG), Respiratory Rate (RR), and Blood Pressure (BP). This data can be stored on the website. After the data has been collected, prediction can done using supervised machine learning classification algorithm . In this paper, the authors states that the proposed models scores accuracy of 92% using J48 classification techniques.

In this (Mingfei Zhang et al, 2019) paper, the author presents a system framework by using wireless sensor network for the real-time monitoring on the water quality in aquaculture. They have designed the structure of the wireless sensor network to collect and continuously transmit data to the monitoring site.

The monitoring sit can represent the monitoring hardware and data visualization, and analyze the data with expert knowledge to implement the auto control. The proposed work also can send warning notification (Via SMS or graphical device twinkle) messages to the intended users when abnormal events are detected. The drawback of this paper is unable to clearly state the decision-making system. They did not state about the expert knowledge to classify the water quality based on the water quality parameters.

We have identified the gaps of those research and we proposed a KNN based model water quality checking system by applying machine-learning technique and we have got relevant different from other work.

CHAPTER THREE

RESEARCH METHODOLOGY

3. Introduction

Our research focuses on designing of wireless sensor-based drinking water quality monitoring using machine learning algorithms. In order to do this research, we have followed different procedures, from the data collection up to notifying the intended person about the quality status of the drinking water. For the success of this research, we have used controlled experiment room for varying the parameter of the water. On the experiment room, we have three different water quality sensors namely water PH, Turbidity and Temperature sensors. Using these sensors, we can get waters PH, Turbidity and Temperature parameters. According to (Kofi sarpong et al, 2017) these three drinking water parameters are basic characteristics to identify the quality of drinking waters. The following table shows that WHO drinking water quality parameter standards with definitions.

Table 3: WHO drinking water quality measurement parameters

Measurement Parameter	Definition	WHO Standard (Drinking Water)	References
pH	Effective hydrogen-ion concentration (i.e., pH = − log[H+])	7–8.5	(WHO, 2020) (R. Yue et al, 2011)
Turbidity	Amount of solid matter (particles or colloids) suspended in water that obstruct light transmission	1–5 NTU	(WHO, 2020) (R. Yue et al, 2011)
Dissolved oxygen	Amount of DO	5–6mg/l	(R. Yue et al, 2011)
Residual chlorine detection (RCD	Amount of chlorine (residual after chlorine-based water disinfection)	2–3mg/l	(WHO, 2020)
Conductivity (also Salinity)	Ability of an aqueous solution to transfer an electrical current	25◦C	(Wagner et al, 2006)
Temperature	Temperature impacts DO content	Drinking water supply (15◦C)	(WHO, 2020)
Fluoride	Salts that form when fluorine combines with minerals in soil or rocks	4mg/l	(WHO, 2020)
Calcium hardness	Amount of calcium salts (reacts with most detergents and can reduce the effectiveness of the cleaning process)	75–100mg/l	(WHO, 2020)

Total dissolved solids (TDS)	Amount of inorganic salts and small organic matter	600–1,000mg/l	(WHO, 2020)
Magnesium hardness	Amount of magnesium salts (causes an undesirable taste and stains laundry)	50–100mg/l	(WHO, 2020)
Manganese	Mineral that naturally occurs in rocks and soil and is a normal constituent of the human diet	<0.1mg/l	(WHO, 2020)
Sodium	Essential mineral that is commonly found in the form of sodium chloride (salt)	~200mg/l	(WHO, 2020)
Hydrogen sulfide	Formed by sulfur and sulfate-reducing bacteria that can occur naturally in water	0.05–0.1mg/l	(WHO, 2020)
Oxidation reduction potential (ORP)	Capacity to either release or accept electrons from chemical reactions; influences the life span of bacteria in water	650–700mV	(WHO, 2020)

Therefore, this thesis work relays on the water PH, Turbidity and Temperature drinking water quality parameters.

3.1 Research Design

This section elaborates the design techniques that we have applied. In this research, the design technique that we have applied is experimental research design. On the experiment drinking water quality sensors are configured to sense the properties of the waters. The sensors send the sensed data to our database (cloud) and the data will be further processed. Finally, the anomaly detection and classification can be done. The detail architecture of this research is depicted in section 3.2 bellow.

3.2 Proposed Architecture

This section provides an insight about the designing of the proposed model of this study. The designing involves data collection using wireless sensor networked devices and the implementation involves ML classifications algorithms used. The architecture of the proposed system is being displayed as shown in the Figure 9. The architecture begins with data collection using water PH, Turbidity and Temperature sensors. These three devices then transmit the data to the base station, which is microcontroller. This data is then transmitted to the cloud using Wi-Fi module as a communication tool. Data analysis and classification is done on Thingspeak cloud storage. According to the official definition of ThingSpeak, ThingSpeak™ is an IoT analytics platform service that allows you to aggregate, visualize and analyze live data streams in the cloud. ThingSpeak provides instant visualizations of data posted by your devices to ThingSpeak. With the ability to execute MATLAB® code in ThingSpeak you can perform online analysis and processing of the data as it comes in. ThingSpeak is often used for prototyping and proof of concept IoT systems that require analytics.

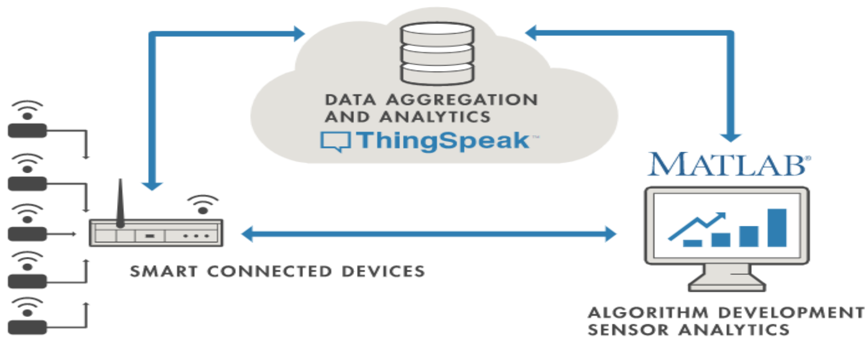


Figure 8: ThingSpeak system Architecture

In our research, water quality data can be analyzed on ThingSpeak for classification and the proposed architecture of the water quality detection and classification system is shown below.

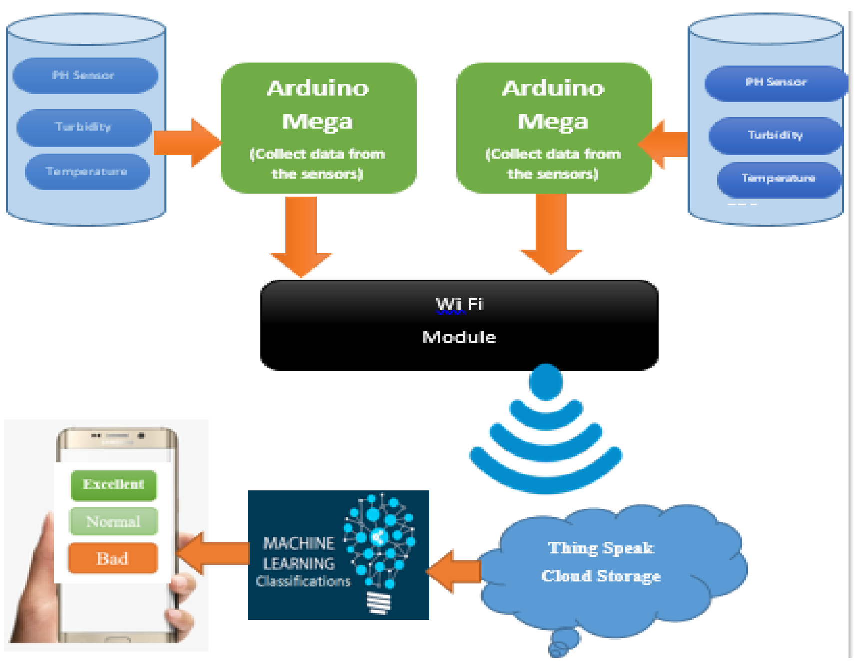


Figure 9: Architecture of the proposed systems

This research model has three major components; the first component is to collecting the data from the sensors, the second is communication to the cloud storage and the third one is to build water classification and analysis technique.

For the first phase, the sensing phase deals with the physical interaction of the devices with the different water contents. we have used three wireless sensor device that can sense the waters parameters from different water contents. The hardware device used in this research are described in section 3.3.2. For the second phase, we used Wi Fi communication protocol to send data from the intermediate device to the cloud storage. The data is sensed on every 30 seconds stored on the intermediate device and then it passed to ThingSpeak cloud storage for further analysis and classifications. Analysis and classification techniques are described on sections 3.8 in detail.

3.3 Development Tools

For the completion of this research, we used the following hardware and software tools.

3.3.1 Software Tools

We are installing the following application software for training and testing our proposed system

- Data storage – ThingSpeak
- Dataset processing –MS –excel
- Programming language- C and MATLAB

3.3.2 Hardware Tools

For the completion of this research, we used the following hardware device.

- Desktop (Processor - core I3 @ 3.7 GHz , RAM - 4 GB, Hard Drive - 500 GB , Hard Drive - 500 GB , Monitor - 16" LCD
- Sensors and devices(listed and described in section 3.5)
- Others supporting materials like breadboard, LCD(16x2), LED, resisters Buzzers

3.4 Sensors and Devices

In the following, we discussed about the devices and sensors that we have used in this research.

3.4.1 PH sensor

This is one of the most valuable water quality parameter to determine whether a given water is contaminated or not. The pH indicates the concentration of hydrogen [H] + ions present in certain solutions. The pH of a solution is the measure of the acidity or alkalinity of that solution. A sensor that measures the potential difference between two electrodes can accurately quantify it: a reference electrode (silver / silver chloride) and a glass electrode that is sensitive to hydrogen ion. This is what forms the probe. We also have to use an electronic circuit to condition the signal appropriately and we can use this sensor with a micro-controller, such as in our case Arduino mega. The pH of pure water is 7. The pH scale is a logarithmic scale whose range is from 0-14 with a neutral point being 7. Values above the neutral point indicate a basic or alkaline solution and values below seven would indicate an acidic solution. It operates on 5V power supply and it is easy to interface with Arduino. According to the world health organization's standard an excellent water PH level is exact seven which is neutrals.

To calculate the pH of an aqueous solution you need to know the concentration of the hydronium ion in moles per liter (molarity). The pH is then calculated using the expression:

$$pH = -\log [H_3O^+].$$

Device Specifications

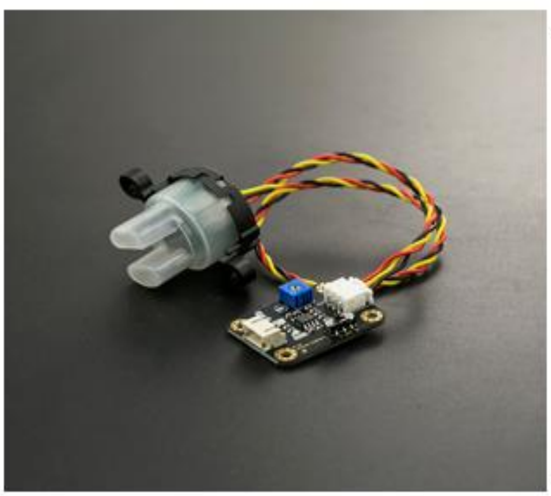
- Module Power : 5.00V
- Module Size : 43 x 32mm(1.69×1.26")
- Measuring Range :0 – 14PH
- Measuring Temperature: 0 – 60 °C
- Accuracy : ± 0.1pH (25 °C)
- Response Time : ≤ 1min
- pH Sensor with BNC Connector

Figure 10: Water PH sensor device and specifications

3.4.2 Turbidity Sensor

Turbidity is a measure of the cloudiness of water. Turbidity has indicated the degree at which the water loses its transparency. It is considered as a good measure of the quality of water. Turbidity blocks out the light needed by submerged aquatic vegetation. It also can raise surface water temperatures above normal because suspended particles near the surface facilitate the absorption of heat from sunlight. The turbidity sensor detects water quality by measuring the levels of turbidity, or the opaqueness. It uses light to detect suspended particles in water by measuring the light transmittance and scattering rate, which changes with the amount of total suspended solids (TSS) in water. As the TTS increases, the liquid turbidity level increases. Turbidity sensors are used to measure water quality in rivers and streams, wastewater and effluent measurements, control instrumentation for settling ponds, sediment transport research and laboratory measurements. This liquid sensor provides analog and digital signal output modes.

The SI unit of the turbidity is NTU (Nephelometric Turbidity Units) in which the drinking water turbidity must be between 0 and 5 NTU. Water turbidity values greater than 5 NTU become perceptible to the eye, especially in large volumes such as white containers. This effect might be increased if the water also contains colored materials such as humid acid or inorganic colored products such as iron compounds.



Device Specifications

- Operating Voltage: 5V DC
- Response Time: <500ms
- Output Method: Analog
- Analog output: 0-4.5V
- Operating Temperature: 5°C~90 °C
- Storage Temperature: -10°C~90°C
- Weight: 30g

Figure 11: Water turbidity sensor device and specifications

3.4.3 Water Temperature Sensor

Water Temperature indicates how water is hot or cold. This temperature sensor is digital type, which gives an accurate reading. The DS18B20 waterproof temperature sensor is quite versatile. It can be powered through the data line (so called "parasite" mode, which requires only 2 wires versus 3 in normal mode).

The basic reason that we selected temperature as quality parameter is that it is the cause for other contamination for example high temperature value in the water can create an environment for creation of algae and organisms inside the water.

Device Specifications

- Communicates using 1-Wire method
- Operating voltage: 3V to 5V
- Temperature Range: -55°C to +125°C
- Accuracy: ±0.5°C
- Output Resolution: 9-bit to 12-bit (programmable)
- Unique 64-bit address enables multiplexing
- Programmable alarm options

Figure 12: Water temperature sensor device and specifications

3.4.4 Arduino Mega 2560

This our main development board used for interfacing different water quality sensors with the microcontroller and sends the sensed data to Wi-Fi module (ESP32) via serial communications principles. Arduino can sense the environment by receiving input from a variety of sensors and can affect its surroundings by controlling lights, motors, and other actuators. The microcontroller on the board is programmed using the Arduino programming language (usually C programming language).

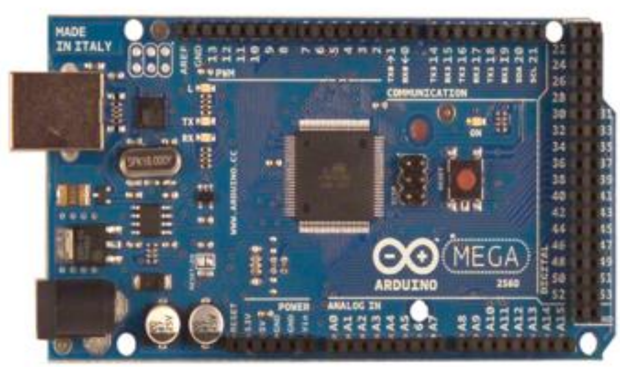
The main advantage of Arduino microcontroller are:-

Inexpensive - Arduino boards are relatively inexpensive compared to other microcontroller platforms. The least expensive version of the Arduino module can be assembled by hand, and even the pre-assembled Arduino modules cost less than $50.

Cross-platform - The Arduino software runs on Windows, Macintosh OSX, and Linux operating systems.

Simple, clear programming environment - The Arduino programming environment is easy-to-use for beginners, yet flexible enough for advanced users to take advantage of as well.

Open source and extensible software- The Arduino software is published as open source tools, available for extension by experienced programmers through C++ programming language.



Device Specifications

Microcontroller: ATmega2560

Operating Voltage: 5V

Input Voltage (recommended 7-12V

Input Voltage (limits 6-20V

Digital I/O Pins: 54

Analog Input Pins: 16

Flash Memory: 256 KB

SRAM: 8 KB

EEPROM: 4 KB

Clock Speed: 16 MHz

Figure 13: Arduino mega 2560 development board and specifications

3.4.5 ESP32 WI FI module

ESP-32 is a series of low-cost and low-power system-on-chip (SoC) microcontrollers with integrated Wi-Fi and dual-mode Bluetooth. The microcontroller is cheap with low-power consumption and a great number of pins. Evidently, with its varied features, IoT becomes easier when it comes to ESP-32. We used this module for sending the water quality sensors data to the cloud or to our database.

Device Specifications	
Number of cores	2 (dual core)
Wi-Fi	2.4 GHz up to 150 Mbits/s
Bluetooth	BLE (Bluetooth Low Energy)
Architecture	32 bits
Clock frequency	Up to 240 MHz
RAM	512 KB
Pins	30 or 36 (depends on the model)

Figure 14:ESP32 WI fi module

3.4.6 Others Supporting Materials

All the above Martials that we have listed are the main device used to do this research but without the supporting materials, we cannot proceed a head unless we use them.

The following are the basic list supporting material that we used.

Breadboard: A breadboard is a construction base for prototyping of electronics. We used this device for temporary prototyping wiring procedure of the proposed work.

Jumper wire: We used different jumper wire for connecting sensors with main board. Male-Female and Male- Male jumper wire is used for the completion of the research.

Buzzer: We used to set alert sound when abnormal condition is happen may be outliers.

Potentiometer: We used for changing the brightness of the LCD display.

LCD: This device is used to display characters in 16x2 screens.

Resister: We used for resisting current from damaging LEDs

LED: We used RGB LEDs to see the sensors are working or not.

USB Cable: We used Arduino USB cable for programming the microcontroller via computer.

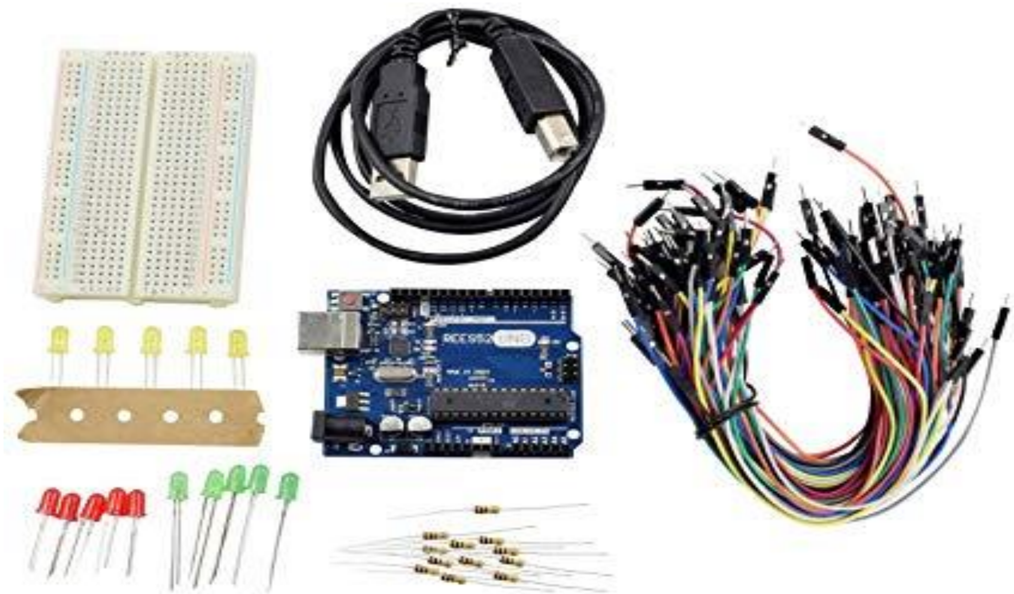


Figure 15: Others supporting materials

3.5 Schematic diagram of proposed work

The following two image describes the proposed system schematic design of real time drinking water quality detection and classification based on WSN and machine learning

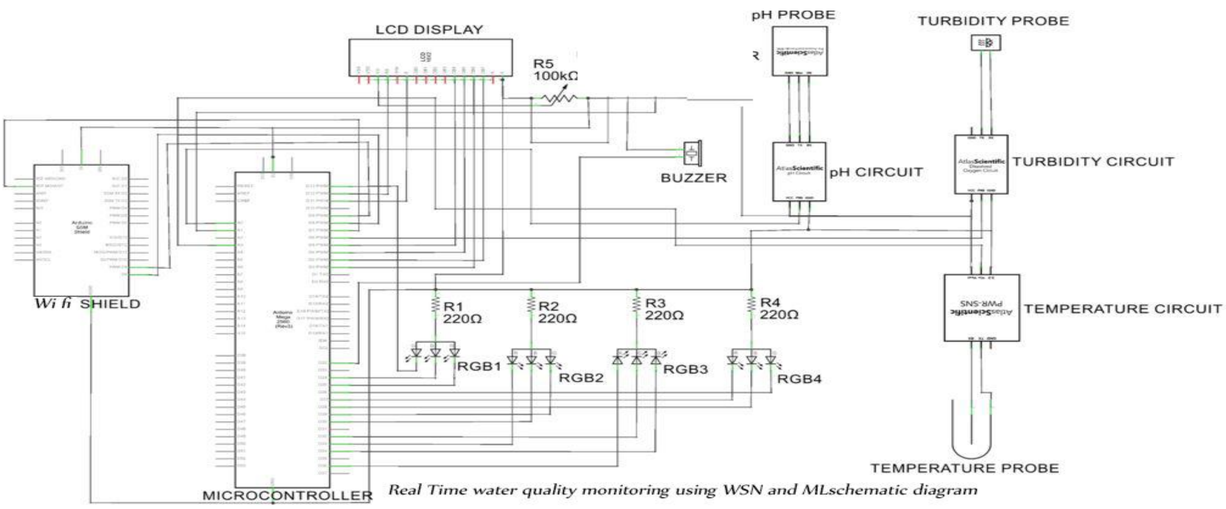


Figure 16: Detail Schematic diagram of the proposed sensors configurations with controller

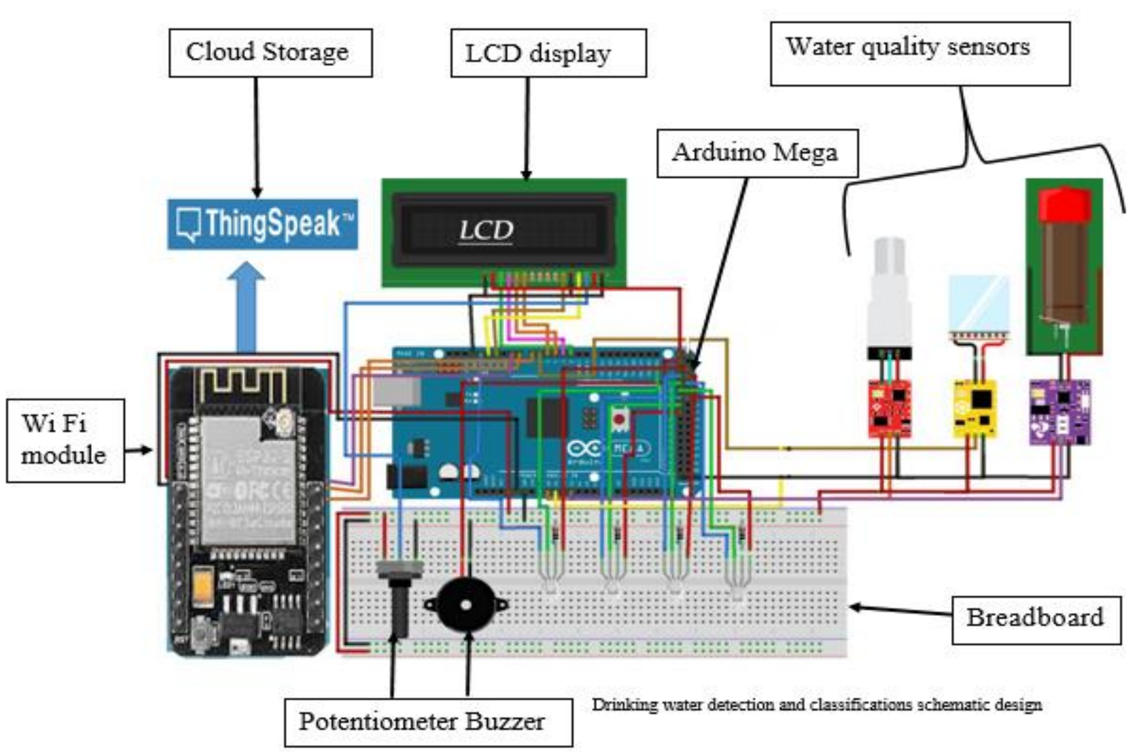


Figure 17: Schematic Diagram of proposed system

3.6 Data Collection

For this research, we used the dataset that we have collected by water quality sensors by applying the sensors in different water bodies in our laboratory rooms. In this research PH, Turbidity and temperature of the water are collected on different conditions and water content for variation of our datasets. For varying the contents of water , we used different technique and tools. In order to vary water temperature, we used portable boiler, to vary the PH content of the water we used both Acid and Base solution and we insert via syringe, and lastly for vary the turbidity content we used different dusts like soils and chock to decrease and increase the transparency of the water. The dataset that we have collected and used for this research is 1,000 datasets (1,000 PH sensor, 1,000 Temperature, 1,000 turbidity values). Due to time and budget constraint, we cannot able to collect more water quality datasets. The sensors data can be sent via wireless sensor network using Wi-Fi module to our cloud storage, which is ThingSpeak.

The following picture illustrates when we are collecting data using those three sensors by varying water parameter contents in each bottle.

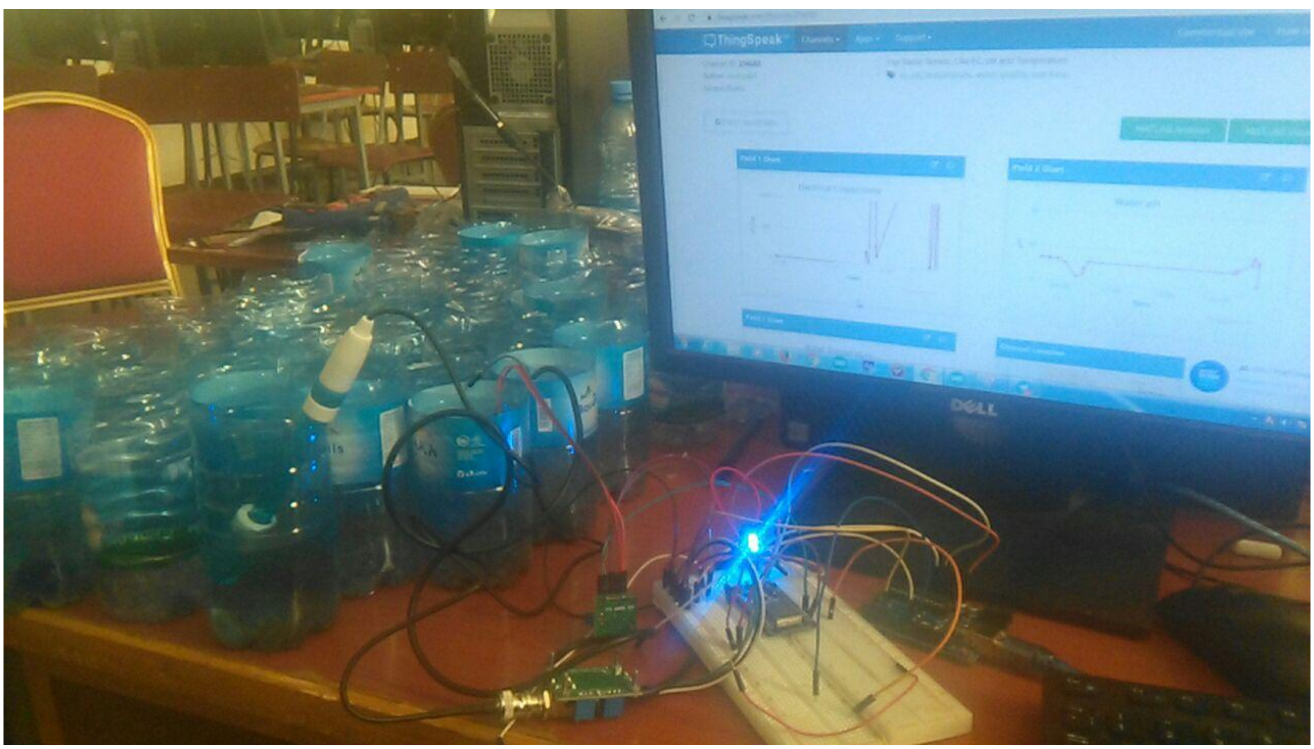


Figure 18: Data collections technique

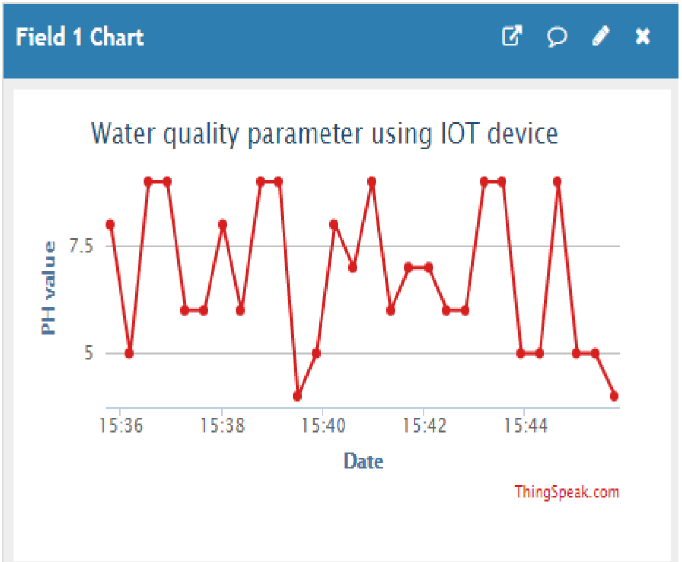


Figure 21: Water PH collected value in ThingSpeak

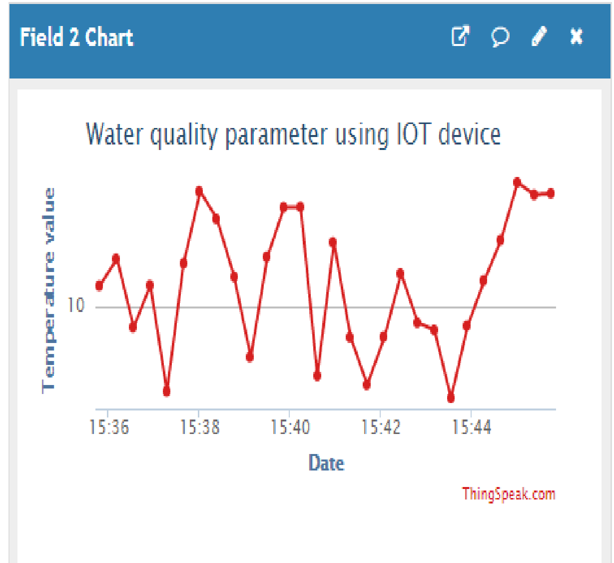


Figure 20: Water Temperature collected sample Value in °C

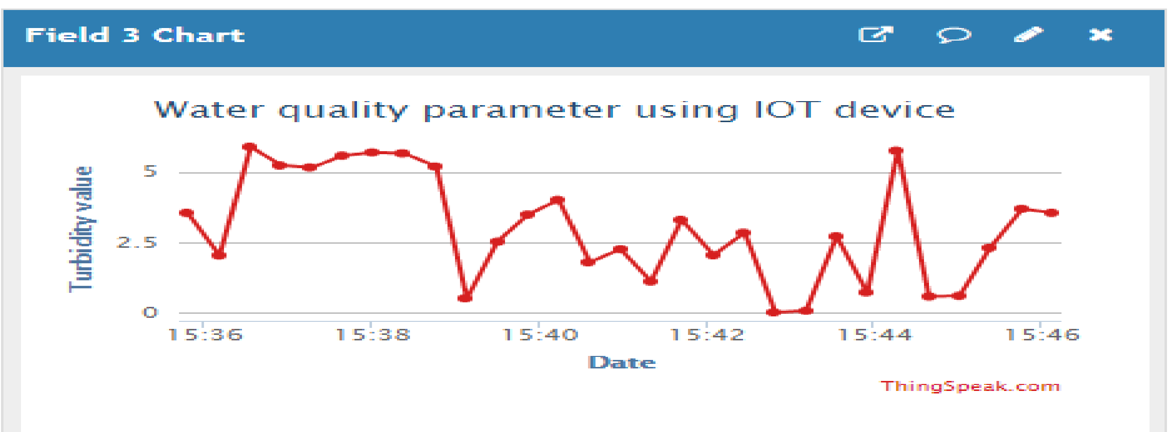


Figure 19: Water Turbidity collected value in ThingSpeak

3.7 Dataset preparations

Dataset preparation is the way in which we are arranging the datasets that is suitable for machine learning approaches. We need to have well-organized training and testing dataset in order to achieve better result. We gave label for the training and testing data set according to the world health organization’s drinking water standard. The classes of the waters are three in which represented as excellent water, normal waters and poor waters. The excellent waters are highly recommended for drinking and the parameters are sharp according to WHO standard.

The poor water class indicates that, they are not recommended for drinking purpose because of bad water contents. Good water quality classes are recommended for drinking but they are not strictly sharp the standard (i.e. PH value may be 7.5 rather 7).Some of the preprocessed training datasets that we have labeled are shown in table 3 bellow.

Table 4: sample preprocessed training datasets

	A	B	C	D
1	PH value	Temp va	Turbdity	Labels
2	6	16	5	0
3	5	15	3	0
4	9	14	3	2
5	7	18	2	1
6	7	21	1	1
7	6	12	4	0
8	7	18	2	1
9	6	13	1	0
10	5	13	3	0
11	9	16	3	2
12	8	16	2	1
13	7	19	1	1

3.8 Anomaly Detection and Classification Technique

After the data has been collected, we have applied machine-learning approach to our data that is KNN classification technique. According to (Jingwen Sun et al, 2018), KNN algorithm is simply based on the idea that objects that are 'near' each other will also have similar characteristics. Thus if we know the characteristic features of one of the objects, we can also predict it for its nearest neighbor.

It is based on the idea that any new instance can be classified by the majority vote of its 'K' neighbors, where k is a positive integer, usually K a small number. The distance between the sample to be sorted and the training sample of the pre-labeled data is calculated, and the k neighbors closest to the sample data to be sorted are found. The categories of the sample data to be classified are determined according to the category to which the neighbors belong. When we see the working principles of KNN algorithm firstly we must calculate the distance between the test data with all training dataset, secondly, we have to specify the value k (select the maximum class that is nearest or small distance). The value of k can affect the performance KNN classification algorithm. In this thesis, we test the performance of the classifier by varying the value of k from 1 to 8 and then the optimal value of k chosen for detection and classifications.

The reason that we selected the KNN algorithm is that, it is suitable for having small amount of training and testing data and it fits the data that we have. Even if, when the training data is too large, the classification output will be smooth but the execution time will be large because of the reason that KNN calculates the distance between each dataset. The other advantages of this machine learning classification algorithm is a memory-based approach which means that the classifier immediately adapts as we collect new training data. It allows the algorithm to respond quickly to changes in the input during real-time use.

3.8.1 Pseudo code of KNN classifications

1. Load the training and test data

2. Choose the value of K

3. For each point in test data:

   - find the Euclidean distance to all training data points

   - store the Euclidean distances in a list and sort it

   - choose the first k points

   - assign a class to the test point based on the majority of classes present in the chosen points

4. End

3.8.2 Euclidean Distance Functions

In order to find the Euclidian distance between two points. Let say the first point is *A* and the second point is *B* and assume each point can be represented as vectors.

$$A = (x1, x2 \ldots\ldots xm) \text{ -- (eq 3.1)}$$

$$B = (y1, y2 \ldots\ldots ym) \text{ -- (eq 3.2)}$$

To calculate the distance between point *A* and *B*, the normalized Euclidean metric is generally calculated as

$$dist(A,B) = \frac{\sqrt{\sum_{i=1}^{m}(xi-yi)}}{\sqrt{m}}$$ --(eq 3.3)

Using the above formula, we find the similarity of two water quality parameters datasets. In this work, the datasets are grouped in to two. The first categories are training and the second is testing datasets. These two datasets are preprocessed and feed in to the KNN classifier.

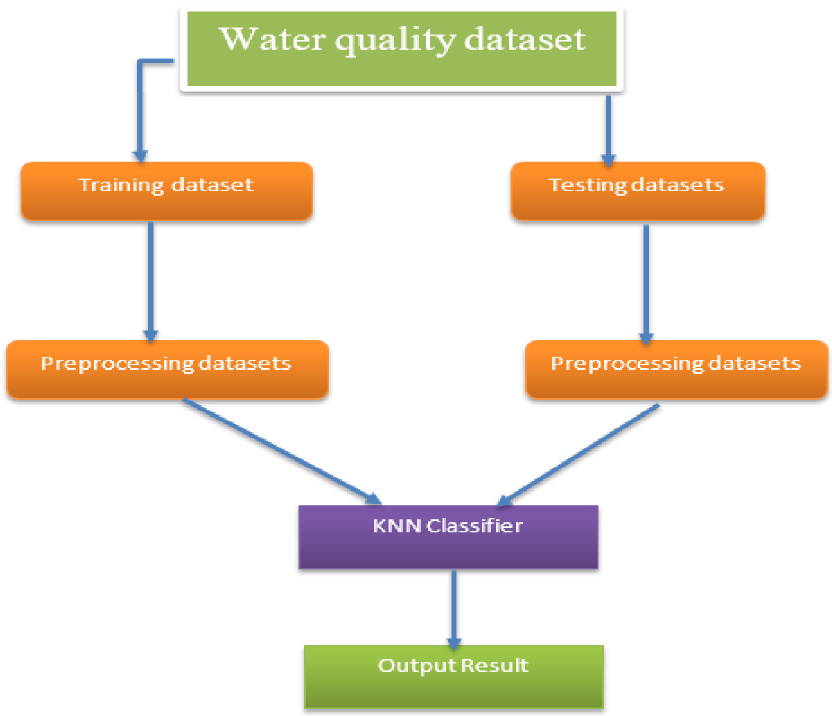


Figure 22: KNN classification model for drinking water quality datasets

The classification done after the KNN model is learn. The KNN classifier can learn from the pre labeled data, which is training datasets. The drinking water anomaly detection and classification proven by the test datasets. The KNN model detects the abnormal behavior of the water with the accuracy of 94%.

3.9 Real Time Alert Notification

After the model classifies the new data that comes from the water quality sensors, the next step is to announce the status of drinking water at real time for the intended users. If the system detects abnormal water or poor-quality water, then the developed model sends an alert notification to the user through email. For excellent and good water, the system will remain silent until it detects abnormal behavior.

The alert notification is sent to the user with the help of Thingspeak application that integrates both Math lab code analysis and email based alert notification systems on real time events. The data analysis and visualization is done using MATLAB analytics. There is also option to add various plugins that enable user to display google guage and other custom visualization and controls in private view. Various actions can be done using apps provided by the platform. ThingSpeak provides apps that allow us for an easier integration with the web services, social networks and other APIs. It include Thing-Tweet app to tweet alerts and messages, Email alert notifications, Time control App and so on. Using this feature we selected email based alert notification systems.

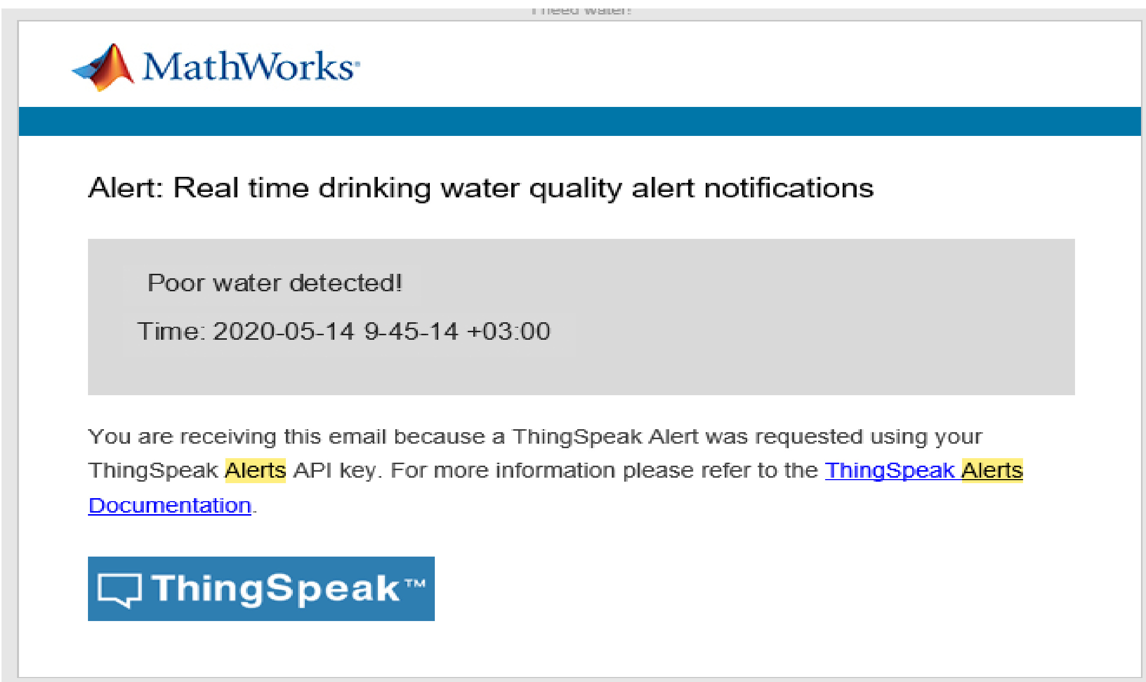


Figure 23: Real-time alert notification

3.10 Sample Implementations

Apps / MATLAB Analysis / Drinking water quality detection and classifications / Edit

Name

Drinking water quality detection and classifications

MATLAB Code

```
% Read the soil moisture channel data from the past two weeks.
% Send an email and tell the user to add water if the value
% is in the lowest 10 %.

alert_body = 'Real time drinking water alert notification';
alert_subject = 'poor water detected';
alert_api_key = 'I2CA5Y17Y9PUBRDL';
alert_url= "https://api.thingspeak.com/alerts/send";
jsonmessage = sprintf(['{"subject": "%s", "body": "%s"}'], alert_subject,alert_body);
options = weboptions("HeaderFields", {'Thingspeak-Alerts-API-Key', alert_api_key; 'Content-Typ
result = webwrite(alert_url, jsonmessage, options);
```

Figure 24: Email body Alert notification sample code

3.10.1 C sample code for sending to ThingSpeak

```
    if (client.connect(server,80))                         //   "184.106.153.149" or
    {

    String postStr = apiKey;
    postStr +="&field1=";
    postStr += String(phValue);
    postStr += "&field2=";
    postStr += String(Turbdity);
     postStr += "&field3=";
      postStr += String(Temprature);
    postStr += "\r\n\r\n";

    client.print("POST /update HTTP/1.1\n");
    client.print("Host: api.thingspeak.com\n");
    client.print("Connection: close\n");
    client.print("X-THINGSPEAKAPIKEY: "+apiKey+"\n");
    client.print("Content-Type: application/x-www-form-urlencoded\n");
    client.print("Content-Length: ");
    client.print(postStr.length());
    client.print("\n\n");
    client.print(postStr);

    delay(5000);
 }
  delay(5000);
    client.stop();
    Serial.println("Waiting...");

 }
```

CHAPTER FOUR
RESULT AND DISCUSSION

4.1 Result and Discussion

For the evaluations of the proposed algorithm, we used accuracy as a classification evaluation Metrix**.** Accuracy is the ratio of number of correct predictions to the total number of input samples. The reason of using only accuracy as performance measurement is that, we have used balanced datasets. If the datasets are balanced, accuracy is enough to evaluate the performance of the classifier. The numbers of instances in each classes are equal proportions.

In this thesis we did not used confusion matrix for the evaluation of the model because of mostly confusion matrix are applied on imbalanced detests. For the completion of the evaluation, we have 1,000 water quality balanced datasets. The expected labels of the datasets are grouped in to three class's namely excellent, normal and poor water. The data proportion of each classes are 333 i.e. 333x3 ~1000.

After the datasets are collected, we need to have train the model using the training datasets. We use the training data to fit the model and then able to classify unknown instances based on pervious experiences. For training the model, we used 60% of the datasets in which the model must learn. After training the model using the training datasets, we tested the model in order to evaluate the performance of the model. For testing of the k-nearest neighbor classifier model, we used 40% of datasets that we have collected.

By having those data we evaluate the accuracy of k-nearest neighbor classifier by varying the value of K from 1 to 8. On the first value where K is 1 the proposed model scores 93.8% based on the above testing dataset. When we vary the value of K, the accuracy of the classifier also vary due to the reason that k-nearest neighbor classifier calculates the distance between the new instance with all training datasets and then select the majority number of voted among the top K nearest instances to the new instance data.

This classification model predicts the abnormality of drinking water with the accuracy of 94% when the value of k is five.

The k-nearest neighbor classifier achieves the following result shown in table 5.

Table 5: KNN accuracy with different K value

No	Value of K	Accuracy of KNN
1	1	93.8%
2	2	87.4%
3	3	89.2%
4	4	91%
5	**5**	**94%**
6	6	93.2%
7	7	93.78%
8	8	90.8%

In order to compare the k-nearest neighbor classifier with other machine learning classification algorithms, we used support vector machine classification technique. The working principles of SVM classifier are described in chapter two. On this thesis, we used the same datasets for testing SVM classifier as we used in k-nearest classification model. Using the above datasets, we tested the SVM classifier with linear one to all fashion. In most literature, researchers are using SVM classification in order to classify binary classes. However, we can also apply SVM in one to all based technique. In one-vs-All classification technique of SVM, for the N-class instances dataset, we have to generate the N binary classifier models. In our case, there are three binary classifier. The number of class labels present in the dataset and the number of generated binary classifiers must be the same. In this thesis, we used SVM binary classifier as follow.

Classifier 1:- [poor] vs [good, excellent]

Classifier 2:- [good] vs [poor, excellent]

Classifier 3:- [excellent] vs [poor, good]

Based on the above technique the SVM classifier detects the abnormality of water with accuracy of 92.6%. We have also tested the SVM classifier in different training and testing datasets and the result shows that SVM classifier can also depends on the datasets that we have trained. Based on the data that we have observed, there are two principles that can vary the accuracy of the classifier.

1. The value of K, which belongs to KNN classifications.
2. The size of training data can affect the accuracy of both classifiers.

In the first case, when the value of the K varies the accuracy of the KNN classifications also vary.

On the second case, when we minimize the training datasets, the accuracy of the SVM classifier can classify with higher accuracy. In this thesis, we tested the SVM classifier by minimizing the original training datasets to 300 out of 600; the accuracy of the SVM classifier achieves 93.6% accuracy. According to this when the training data are small, the SVM classifier can achieve a relatively high accuracy rate. However, for the KNN classification technique, it will need more training data to reach the same level. The KNN method shows its potential to achieve a higher classification performance with the increase in training data quantity and a suitable parameter K (Fei Wang, 2017).

In general, this study shows that the k-nearest neighbor classifier can achieve good result on detecting the abnormality of drinking water by the help of wireless sensor network devices.

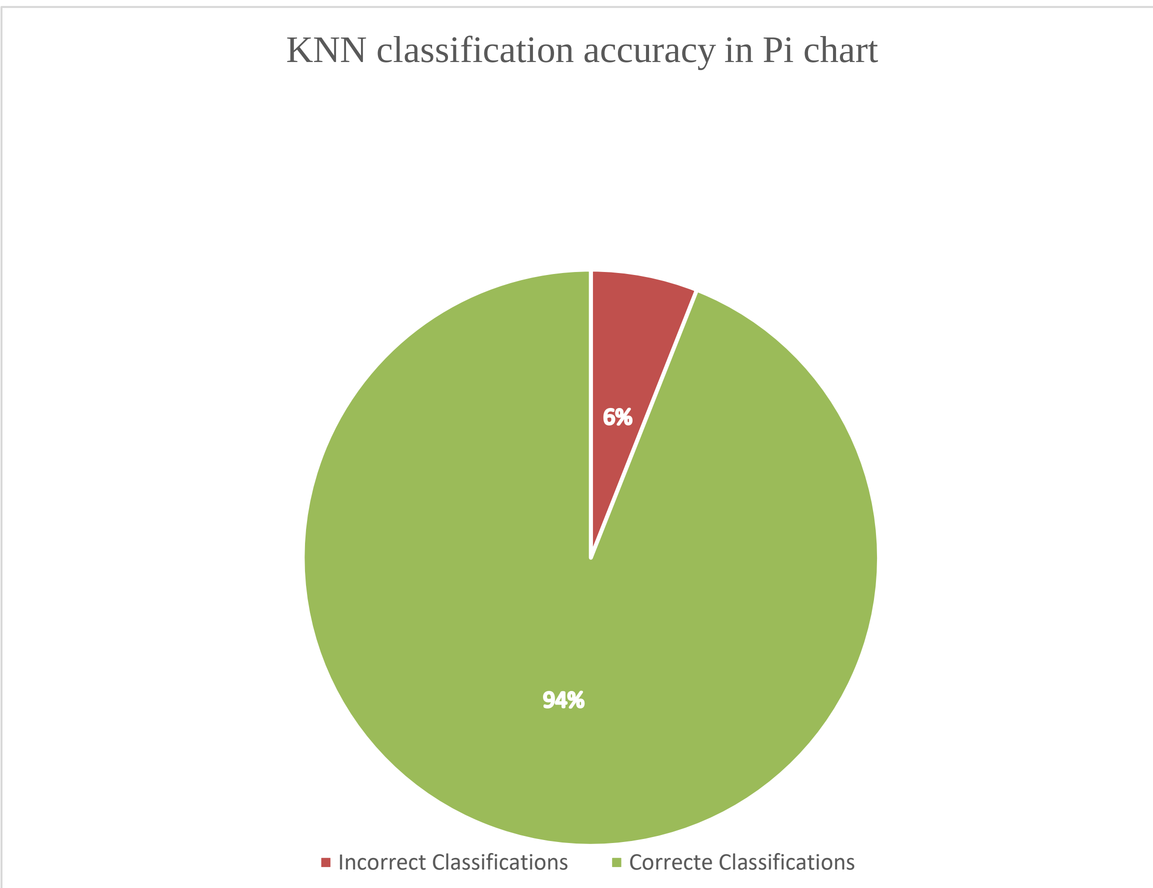


Figure 25: KNN classification accuracy in Pi chart

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

On this work a drinking water quality control mechanisms has proposed by using wireless sensor network and Machin learning approach. In this paper, we used experimental research for conducting the research. Data collection is done via deploying water quality sensors from different drinking waters contents. On this study k-nearest neighbor machine learning technique are used to classify the quality of the water. This classification algorithm can detect the contaminated water with accuracy of 94%. The KNN algorithm is affected by the value of K that we choose and the size of the training data. As we increase the value of K, our predictions become more stable due to majority voting or averaging, and thus, more likely to make more accurate predictions (up to a certain point). When the training data is increase the accuracy of the classifier also increase but the computational speed will be decreases because of the reason that KNN algorithm works by calculate the Euclidian distance between the datasets. In general, in this study we presented the way of controlling the drinking water quality by integrating machine learning and wireless sensor networked device.

5.2 Recommendation

We have tried to classify drinking water contaminated status using three sensors namely water PH, Turbidity and Temperature sensors, but water quality can be characterized by many physical and chemical properties in which we haven't use on this study ,for example amount of chloride, calcium, potassium , TDS and many more basic properties are there. Therefore, for future it is better to use more water quality sensors to achieve a better result on water quality classifications. In addition to the above, it will be better if the water quality data can be large and tested on all classification algorithms.

Reference

Unnikrishna Menon et al. (2012). Wireless Sensor Network for River Water Quality Monitoring in Indi. *ICCCNT*.

A.Lanitis, C. a. (2004). Comparing different classifiers for automatic age estimation. *IEEE* .

Aaina Venkateswaran et al. (2017). An IoT Based System for Water Quality Monitoring. *International Journal of Innovative Research in Computer and Communication Engineering*.

Aaina Venkateswaran et al. (April 2017). an IoT Based System for Water Quality Monitoring. *ijircce*.

Aaina Venkateswaran1 et al. (2017). An IoT Based System for Water Quality Monitoring. *International Journal of Innovative Research in Computer and Communication Engineering*.

Arun Pandi T et al. (2018). MultiSensor based Water Quality Monitoring in IoT Environment. *International Research Journal of Engineering and Technology (IRJET)*.

Ayodele, T. O. (2010.). in New Advances in Machine Learning. ,. In *Introduction to machine learning.* InTech.

Chapman, D. (1996). - A guide to Use of Biota, Sediments and Water in Environmental Monitoring. *Water Quality Assessments.* London, UK: F and FN Spon.

Cheng-Liang Lai, C.-L. C. (July 2011). Using image processing technology for water quality monitoring system. *IEEE*.

Chin Heng Wan & Lam Hong Lee. (2012). A Hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbour and support vector machine, Expert Systems with Applications. *elsevier journal*.

David Nettikadan et al. (2018). Smart Community Monitoring System using Thingspeak IoT Plaform . *International Journal of Applied Engineering Research*.

Dibo,H et al. (2013). Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster–Shafer method. *Measurement Science and Technology, 21*.

Duffy, A. H. (1997). *“The “what” and “how” of learning in design,”.*

Fei Wang, Z. Z. (2017). Comparative Study on KNN and SVM Based Weather Classification Models for Day Ahead Short Term Solar PV Power Forecasting. *Applid science*.

G. Arroyo & B. G. Zapirain. (2011). Blue-white veil and dark-red patch of pigment pattern recognition in dermoscopic images using machine-learning techniques. *IEEE*.

george s. (2003). Content-based audio classification and retrieval by support vector machines. *IEEE*.

hui h.et al. (2012). A new Autonomous Underwater Robotic Fish designed for water quality monitoring. *Proceedings of IEEE International Conference on Modelling, Identification and Control.* Wuhan.

Jayti bhatt & Jignesh patoliya. (2016). iot based water quality monitoring system. *International Journal of Industrial Electronics and Electrical Engineering*.

Jingwen Sun et al. (2018). A Survey of knn Algorithm. *Information Engineering and Applied Computing*.

Jingwen Sun, et al. (2018). A Survey of kNN Algorithm. *Information Engineering and Applied Computing* .

KOFI SARPONG et al. (2017, january). *ACM digital library*. Retrieved from ACM digital library: https://dl.acm.org/doi/10.1145/3005719

M Tarique & H. Khalee. (2016). *"A Reliable Wireless System for Water Quality Monitoring.*

MAHER, D. (2018). *IoT for fresh water quality monitoring.* SWEDEN .

MANYAWKAL. (2019, 06 10). *ASSESSMENT OF WATER ACCESS AND PREVALENCE OF WATERBORNE DISEASES IN RURAL COMMUNITY: A CASE OF ENDEGAGN WOREDA, GURAGE ZONE, SNNPR, ETHIOPIA*. Retrieved from http://repository.smuc.edu.et/: http://repository.smuc.edu.et/bitstream/123456789/108/1/MANYAWKAL%20BIREDA.pdf

Mohammad Abu Alsheikh1,et al. (2014). Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. *IEEE*.

O. Chapelle et al. (2006). Semi-supervised learning. *MIT press Cambridge, 2*.

Sankaranarayanan, K. a. (2017). Intelligent IoT Based Water Quality Monitoring System. *International Journal of Applied Engineering Research*.

Shkurin, A. (2015). *WATER QUALITY ANALYSIS USING MACHINE LEARNING ALGORITHMS.*

Soundarya Pappu, et al. (2017). Intelligent IoT Based Water Quality Monitoring System. *International Journal of Applied Engineering Research , 12*.

Vaishnavi V. Daigavane and Dr. M.A Gaikwad. (Number 5 (2017)). Water Quality Monitoring System Based on IOT. *ISSN* .

Vani K S and Rajesh. (2017). IoT based Health Monitoring using Fuzzy logic. *International Journal of Computational Intelligence Research*.

venkatesh waramma b & jyoti b. (2017). drinking water quality monitoring in real-time using smart objects. *ISSN*.

WenchaoLi et al. (2014). A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network. *Hindawi Publishing Corporation*.

WenchaoLi et al. (2014). A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network. *Hindawi Publishing Corporation*.

Y. S. Abu-Mostafa et al. (2012). *learning from data.*

Yungze Li et al. (2017). Design and Development of a Water Quality Monitoring Network and System. *IEEE*.

Yuthika Shekhar et al. (2017). Intelligent IoT Based Automated Irrigation System. *International Journal of Applied Engineering Research*.

Appendix

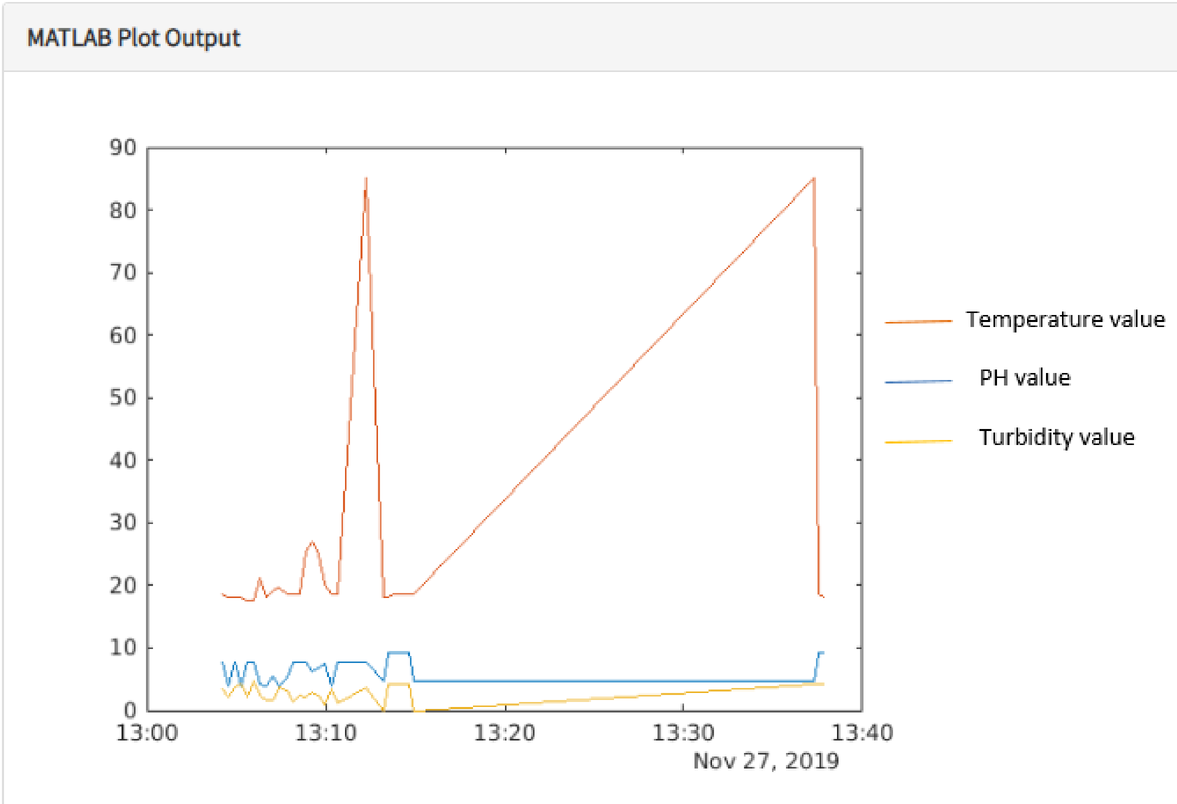


Figure 26: 2D plot of water quality sensors

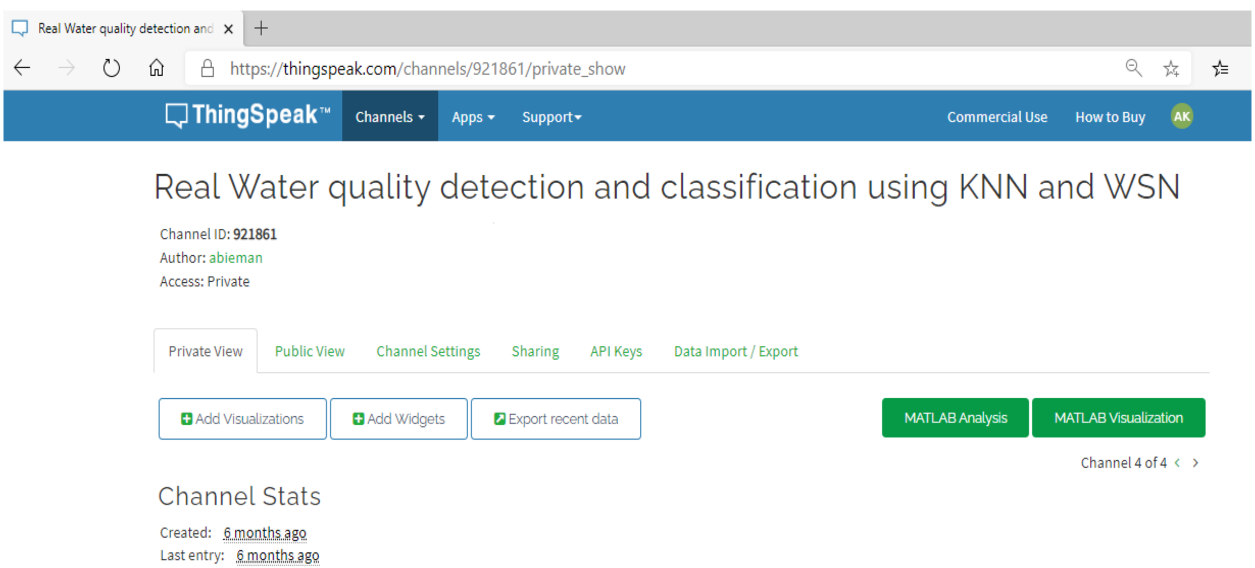


Figure 27: Channel visualizations in ThingSpeak

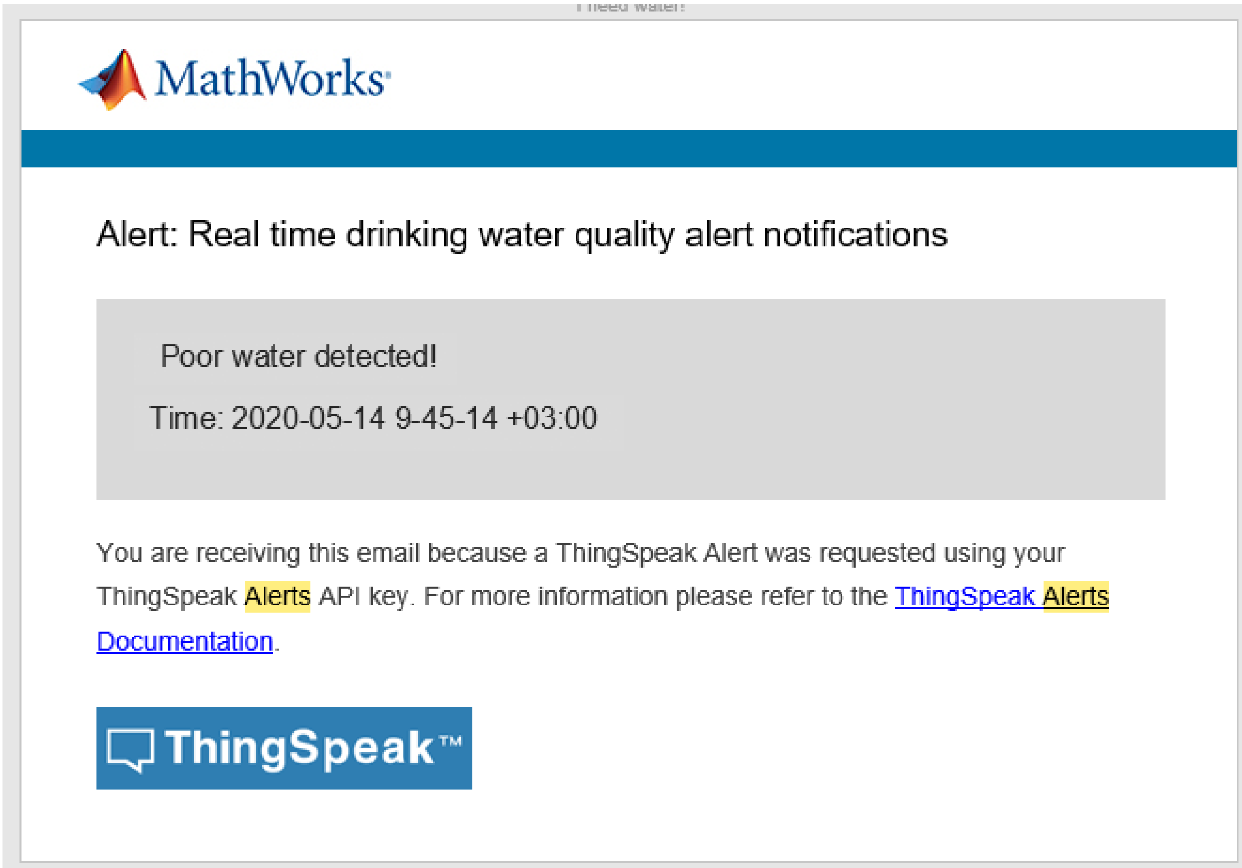


Figure 28: Alert notification using email

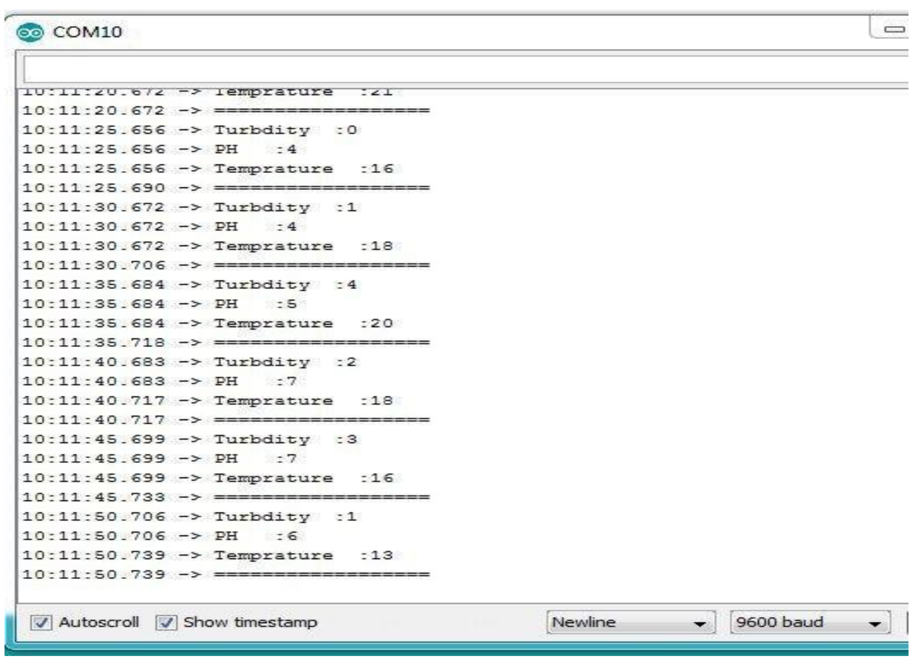


Figure 29: Sensors value when we access in Arduino

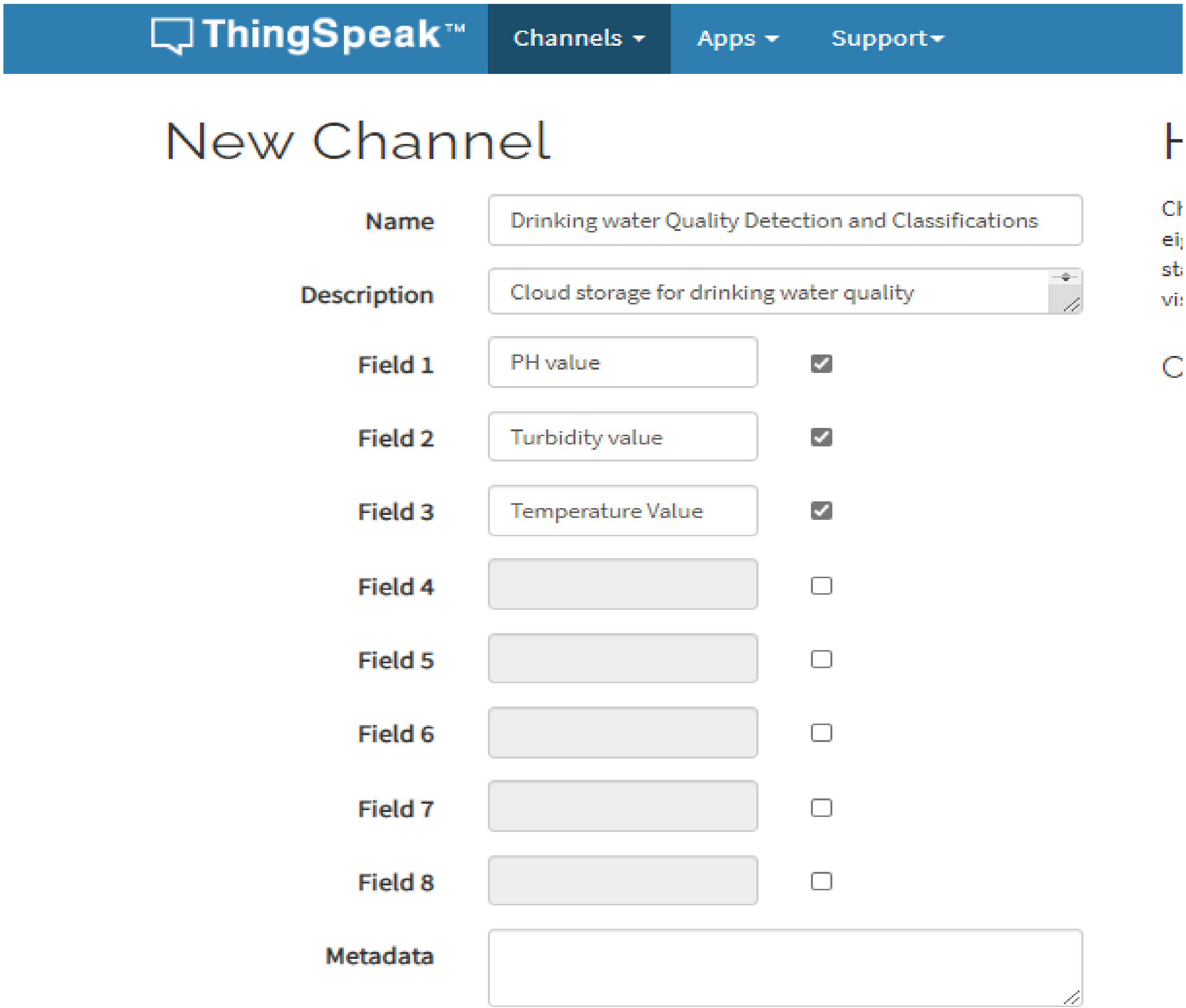


Figure 30: Channel creation for the first time in ThingSpeak

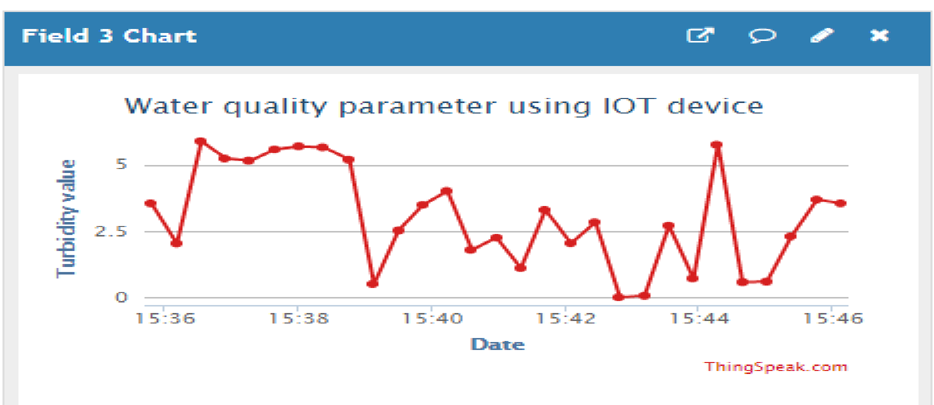


Figure 31: Water turbidity data

KNN classification implementation sample code

```
%clear all screan and perivous value
clear all;
%get the file from csv file in the given folder
get_test_file=csvread('Testing_Data.csv');
get_train_file=csvread('Training_Data.csv');
%get the content of the csv file all rows and columns for both training and
%testing data
read_test_data=get_test_file(1:500,1:3);
read_test_lable=get_test_file(1:500,4);
%initialize the correct counter and incorrect counter with valuue 0
correct_counter=0;
incorrect_counter=0;
%for the value row start form1 1 upto 500 it passess one row of test
%data,whole train data, train labele and neighbor selector 3
disp('wait')
for row=1:500

class=Knn_user_defined_function(read_test_data(row,1:3),get_train_file(1:500,
1:3),get_train_file(1:500,4),5);
    %if the returned value from the methode is simmilar with testlable that
    %is correctly classified else not
    if(class==read_test_lable(row,1))
       correct_counter=correct_counter+1;
    else
       incorrect_counter=incorrect_counter+1;
    end
end
fprintf('-----------------------------------------------------------------------------
----------- \n' );

fprintf('correctly classified instances from 500  are: %d \n'
,correct_counter);

fprintf('incorrectly classified instances from 500 are:%d \n',
incorrect_counter);
       p='%';
        accuracy=(correct_counter/500)*100;
fprintf('the accuracy of the knn classifier is %.2f %s \n',accuracy,p);
fprintf('-----------------------------------------------------------------------------
----------- \n' );

%function which is called in the main program
function
class=Knn_user_defined_function(data_to_be_tested,training_data,traininglable
,k)
%get the size of the column and row(row=500 &column=4)
[row,column]=size(training_data);
make_matrix=zeros(row,2);
%initialize the Euclidian_distance
Euclidian_distance=0;
%for i start from one up to 500, calculate the distance between the one
%row test data with all rows of training data to select the smallest one.
for i=1:row
```

```
    for col=1:column
        Euclidian_distance=Euclidian_distance+((data_to_be_tested(1,col)-
training_data(i,col))^2);
    end
   % save the distance in column one of make_matrix for all rows(500)
    make_matrix(i,1)=sqrt(Euclidian_distance);
    %assign the equivalent class or lable for the given calculated
    %distance in column 2
    make_matrix(i,2)=traininglable(i,1);
    %reset the Euclidian_distance to calculate the other row.
    Euclidian_distance=0;
end
%sort all the calculated distance based on ascending order
sortedrows=sortrows(make_matrix,1);
%initialize to zero for all classinstance to select the maximum vote
class_instance_counter_zero=0;
class_instance_counter_one=0;
class_instance_counter_two=0;
%select the number of nearest neighbors sorted values to get the class
get_the_selected_rows_class=sortedrows(1:k,2);
% from the first to the last number of nearest neighbors
for n=1:k
    %if the this row class is zero increament zeros class counter
    if(get_the_selected_rows_class(n,1)==0)
        class_instance_counter_zero=class_instance_counter_zero+1;
    %if the this row class is one increament one's class counter
    elseif(get_the_selected_rows_class(n,1)==1)
            class_instance_counter_one=class_instance_counter_one+1;
   %if the this row class is two increament two's class counter
    elseif(get_the_selected_rows_class(n,1)==2)
         class_instance_counter_two=class_instance_counter_two+1;
    end
end
%if zero class counter is greater than one and two class return o to our
methode
if(class_instance_counter_zero>=class_instance_counter_one&&class_instance_co
unter_zero>=class_instance_counter_two)
    class=0;
%if one class counter is greater than zero and two class return 1 to our
methode
elseif(class_instance_counter_one>=class_instance_counter_zero&&class_instanc
e_counter_one>=class_instance_counter_two)
    class=1;
    %if two class counter is greater than zero and one class return 2 to aur
methode
elseif(class_instance_counter_two>=class_instance_counter_zero&&class_instanc
e_counter_two>=class_instance_counter_one)
    class=2;
end

end
```

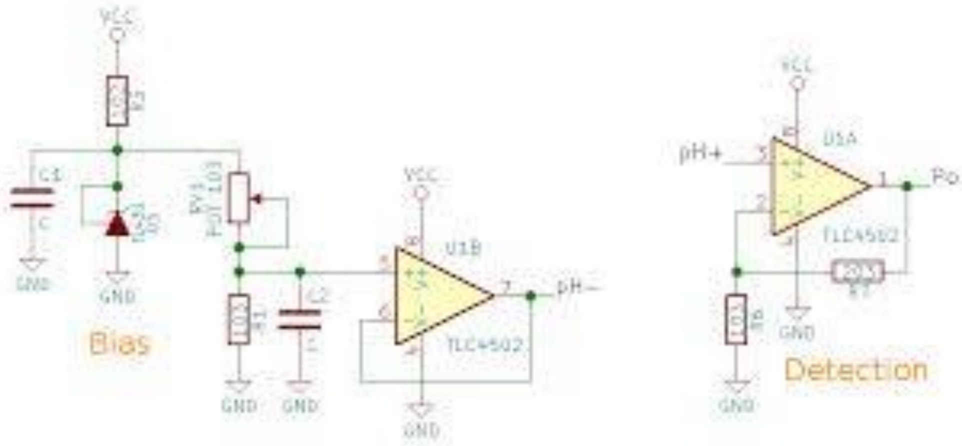


Figure 32: Water PH sensor schematic diagram

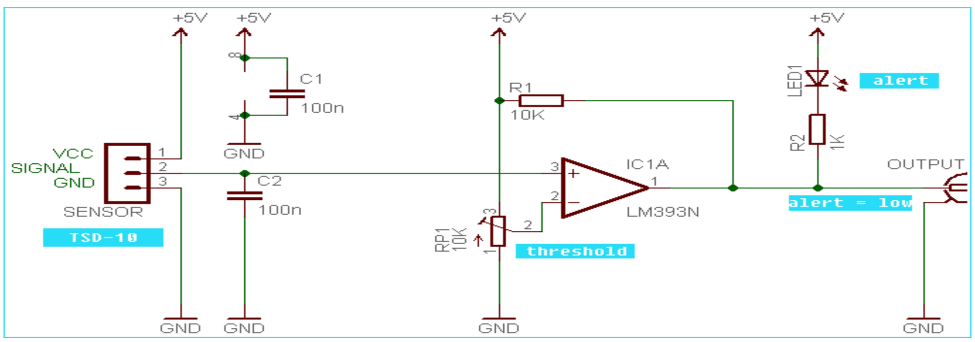


Figure 33: Water turbidity sensor schematic diagram

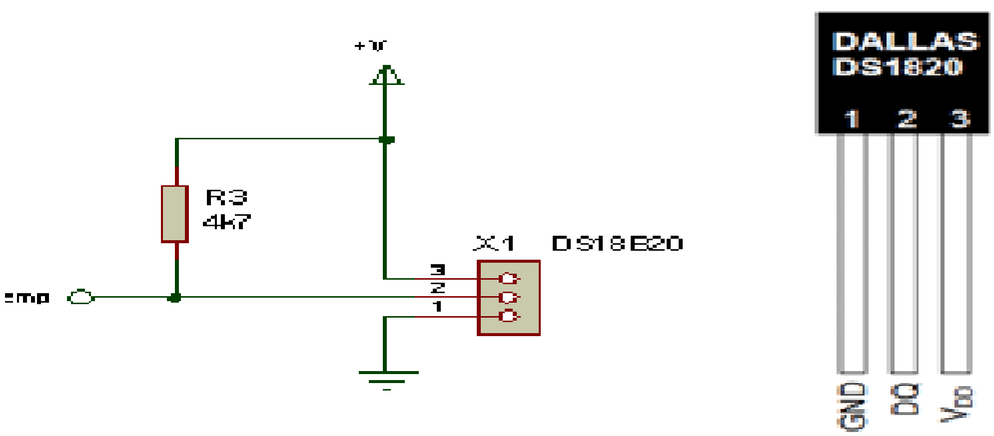


Figure 34: waterproof temperature sensor schematic diagram

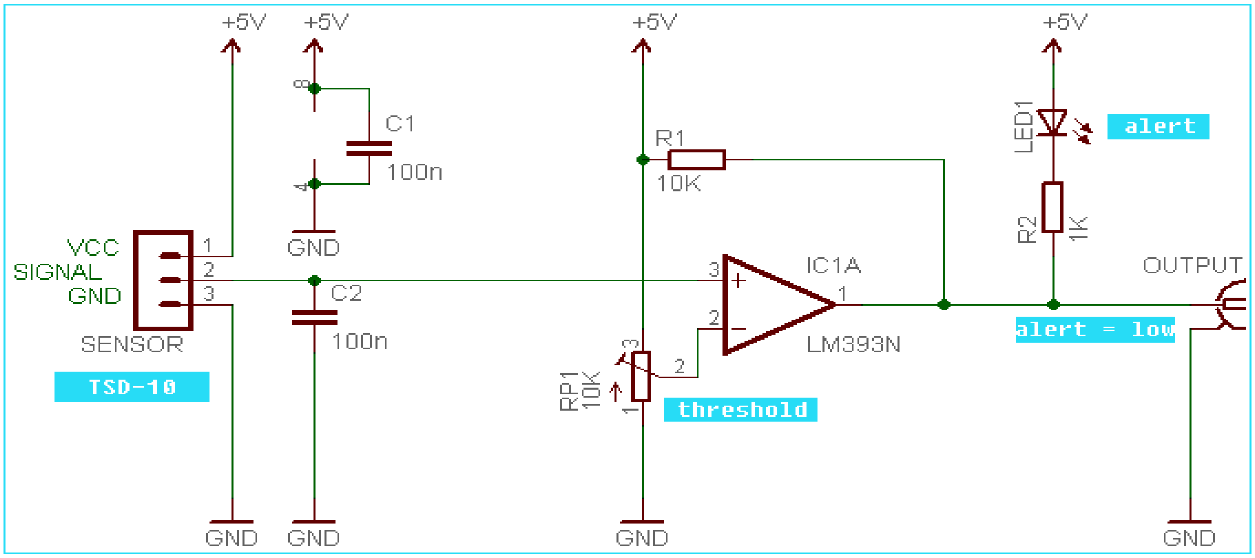


Figure 35: Water turbidity sensor schematic diagram

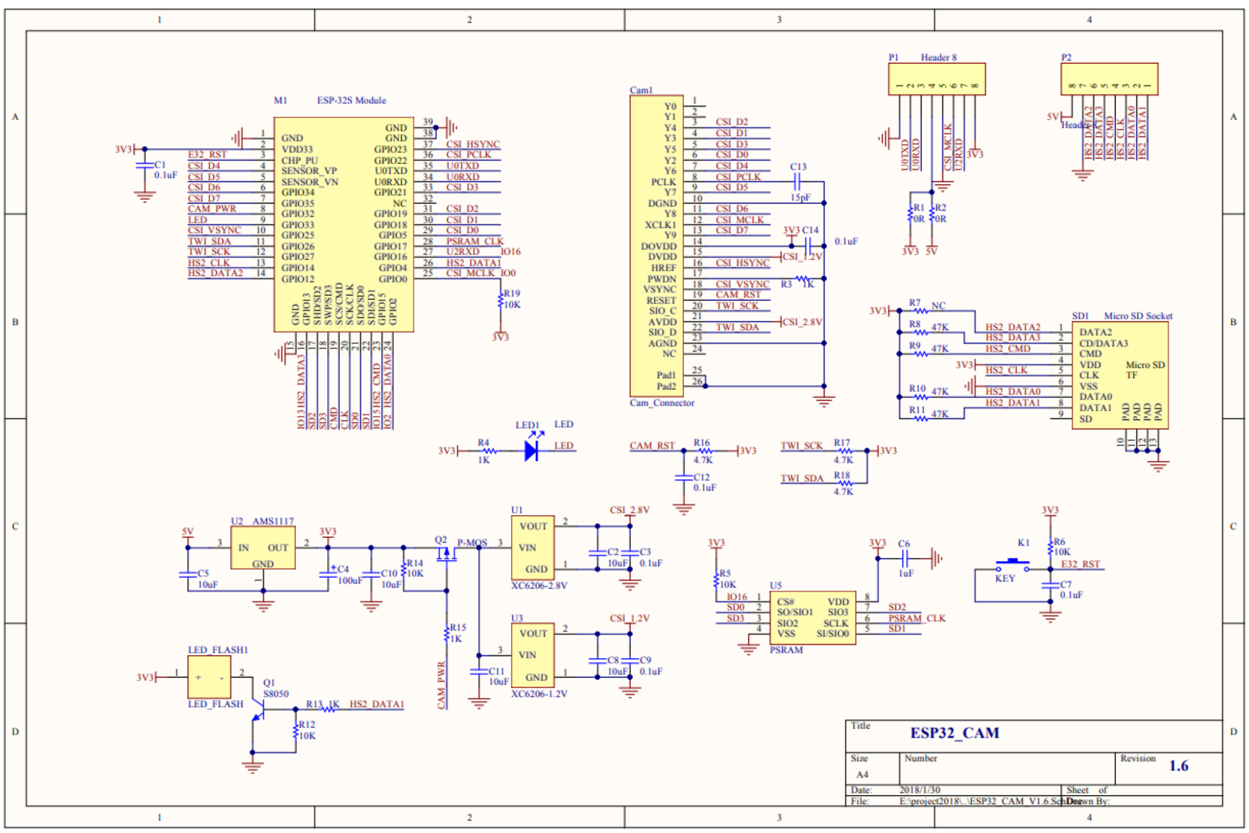


Figure 36:ESP32 WIFI module schematic diagram