Estimating Water Quality using Internet of Things and Machine Learning

By

Umair Ahmed

MSCS600000170895

Supervisor

Dr. Rafia Mumtaz

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters in Computer Science (MS CS)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(February 2019)

Approval

It is certified that the contents and form of the thesis entitled “Estimating Water Quality using Internet of Things and Machine Learning" submitted by Umair Ahmed have been found satisfactory for the requirement of the degree.

Advisor: Dr. Rafia Mumtaz

Signature: _____

Date: _____

Member 1: Dr. Muazzam Khattak

Signature: _____

Date: _____

Member 2: Dr. Asad Ali Shah

Signature: _____

Date: _____

Member 3: Ms Hirra Anwar

Signature: _____

Date: _____

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr. **Umair Ahmed**, (Registration No **00000170895**), of **SEECS** has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: Dr. Rafia Mumtaz

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Abstract

Quality of water plays an important role in all aspects of our lives and it has been deteriorating at an alarming rate due to pollution, deeming its quick, inexpensive and accurate detection vital. Conventional methods to calculate water quality are lengthy, expensive and inefficient. This thesis reviews the conventional lab analysis methods of determining water quality to gain insight into the problem, state of the art machine learning methodologies and role of IoT in determining water quality more efficiently. Also, this thesis proposes a method to detect and predict water quality in real time, respectively, using IoT and machine learning. This thesis explores several machine learning algorithms and predicts water quality using minimal and easily attainable water quality parameters i.e. Temperature, pH, Turbidity and Total dissolved solids. Logistic Regression algorithm yields the most accurate results with accuracy up to 77.98% without TDS and accuracy up to 84.01% with TDS.

Keywords: ANN, IoT, Machine learning, Real time monitoring, Smart City, Water quality.

Dedication

Dedicated to my parents, siblings and mentors without whose continuous support and guidance this research wouldn’t have materialized as swiftly.

Certificate of Originality

I hereby declare that this submission is an original work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's de-sign and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: Umair Ahmed

Signature:

Acknowledgment

I am thankful to Allah the almighty to have directed me throughout this research at each and every step and for every new idea which you incepted in my mind to improve it. Indeed, I could have done very little if it weren't for your guidance and direction. Whoever helped me throughout the course of my research, whether my family or anyone else was your will, so indeed none be worthy of praise except you.

I am greatly thankful to my beloved parents for raising me when I wasn't capable of walking and for continued unconditional support throughout in each and every department of my life.

Also, I would like to express special thanks to my supervisor Dr. Rafia mumtaz for her support and help throughout my research and encouraging me.

I would also like to pay special thanks to my co-advisors and committee members for their cooperation and tremendous support. Every time I was stuck in some problem, they would come up with the solution. Without their help I would not have been able to complete my thesis effectively.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my parents, siblings and mentors without whose continuous support and guidance this research wouldn't have materialized as swiftly.

Table of Contents

1. Introduction

Water is the most important of sources quite vital to sustain life on earth and its quality has been deteriorating at an alarming rate. The rate of its quick consumption makes real time water quality detection necessary so that its quality is verified as its released for consumption. For real time detection of water quality, we need to employ Internet of things and machine learning. Local research concerning real time water quality detection is near to nonexistent which is the primary motivation behind this research.

This research is mainly directed towards estimating and monitoring water quality using Internet of things and machine learning methodologies. Quality of water is determined by a number of parameters like PH, turbidity, temperature, total suspended solids, fecal coliform etc. but what definitively and singularly defines water quality is water quality index WQI which in turn is calculated through water quality parameters. Different countries have different methods of calculating water quality index. After the water quality parameters are acquired through time consuming, expensive and at times inaccurate lab analysis, these methods, using acquired parameter readings, are used to calculate water quality index. Since its a cumbersome and expensive process, this research proposes an IoT system using sensors and employing different machine learning methodologies to monitor water quality parameters, study their trends and predicting water quality index & class. For that, it is quite imperative we gain an insight into the water quality parameters and other water quality systems deployed around the world.

1.1. Water Quality Parameters

The quality of water is determined by a number of parameters but a singular measure that represents water quality is water quality index(WQI) which is measured differently in different countries (Gazzaz et al. 2012). Most countries have their own standards through which they calculate WQI. WQI is measured through measuring different water quality parameters and some of the most common and effective of those parameters are briefly defined below:

- **PH:** PH of water specifies how acidic or alkaline the water is. Acidic range lies between 0 and 6 while alkaline range lies between 8 and 14. 6.5-8.5 is the most acceptable range of pH. It is highly correlated with other parameters (EPA 2001; Verma & Singh 2012).

- **Turbidity:** Turbidity of water is the measurement of non filterable, divided solids in the water. It is mostly measured in nephelometric turbidity units (NTUs). It may also interfere with the treatment of water (EPA 2001).
- **Temperature:** Temperature is one of the most important parameters which has considerable amount of effect on aquatic life. It also affects gas transfer rates and amount of dissolved oxygen. It may alter the form of some of the elements or their concentration. It is mostly measured in Celsius (EPA 2001; Verma & Singh 2012).
- **Chloride:** It is naturally present in the water and while its excess is not harmful to humans normally but water's taste grows towards the saltier range if it increases more than 250 mg/l and may be harmful for agricultural activities (EPA 2001).
- **Electrical Conductivity:** It indicates water's potential to conduct electric current. It is not directly useful in terms of water quality, but it helps more in terms of water's ionic content which in turn determines hardness, alkalinity and some of the dissolved solids. Conductivity varies with the water source. Also, it is correlated with temperature. (EPA 2001).
- **Dissolved Oxygen:** It indicates oxygen's solubility in water. Water mostly absorbs oxygen from atmosphere or produces it through photosynthesis. It is inversely correlated with temperature. It is quite important for aquatic life particularly fish (EPA 2001; Verma & Singh 2012).
- **Total Solids:** It is the amount of suspended and dissolved solids in the water. It indicates the remains in the water such as sulfur, phosphorus, calcium etc. It is measured in mg/l (EPA 2001).
- **Total Suspended Solids:** It is the amount of remains of inorganic and organic solid material suspended in the water. Increase in TSS makes water prone to high absorption of light which increases the water temperature and in turn decreases water's capability to hold oxygen.It highly affects aquatic life. It is quantified in mg/l (EPA 2001; Verma & Singh 2012).
- **Total Dissolved Solids:** It is the amount of remains of inorganic and organic soluble solids in the water. It is highly correlated with salinity and its increase makes the water saline. It is measured in mg/l (EPA 2001).
- **Biological Oxygen Demand:** It is the amount of oxygen consumed by biological activities in the water particularly protozoa and bacteria. If BOD level is quite high and surpasses DO then Other organisms die due to shortage of oxygen. It is quite an important factor indicating water quality. It is measured in mg/l (EPA 2001; Verma & Singh 2012).
- **Chemical oxygen demand:** It is the amount of oxygen consumed during breaking down of organic material and during oxidation of present inorganic material. As like

BOD, it is also an important factor representing status of the water quality. It is measured in mg/l (EPA 2001; Verma & Singh 2012).

- **Faecal Coliform:** Faecal coliforms are bacterias that are found in human and animal waste and mostly originate in intestines of warm blooded species. They indicate possible fecal contamination of water (EPA 2001).
- **Total Coliform:** Total coliforms consist of faecal coliforms and other types of similar non faecal bacteria mostly found in soil. Total coliforms reflect possible presence of pathogenic microorganisms (EPA 2001).

1.2. Water Quality Detection Systems

Post 9/11, authorities all around the world had become more cautious of all the resources that could be intentionally polluted to stir up chaos amongst the masses. Water was one of those resources. It eventually brought up the need to have real time monitoring and contamination detection systems in place. Eventually many such systems were proposed and deployed. Most of those systems were more focused on contamination event detection. One of the first such systems, **Canary**, was built by Sandia National laboratories and was funded by EPA's National Homeland Security Research Center. It is currently deployed at Greater Cincinnati Water Works (GCWW). It provides several open source components, major of them being online water quality monitoring and contamination event detection. It employs multiple direct and surrogate sensors to transmit continuous data to SCADA. It has an API which allows you to update its default algorithms. Its also rest service friendly, allows for XML input and output. It supersedes other systems in certain major aspects, It provides total algorithms transparency, it has capability to directly integrate operational data into its event detection component, it provides capability to have centralized processing on a single computing system and supports sensors from multiple sensors. Another one of such systems is **OptiEDS** by Elad Salomons. It helps detect anomalous water quality conditions in real time. Its used by various clients namely Air Liquide, Mekorot, InSyst etc. It, also, is capable of water quality monitoring and water contamination detection in real time. Next to it is **Bluebox** which is capable of identifying normal behavior, Identify normal but unusual behavior and Identify a parameter that is causing the abnormal behavior. It works even if some of the parameters are missing. It initially performs normalization, calculates points distance amongst the parameters in each data points

and plots the frequency curves of the distances to visualize. But it is quite expensive amounting up to 92,500$ and provides only proprietary transparency of algorithms. Moreover, it doesn't have the capability to directly integrate operational data into event detection. Another system, **Event monitor** was created by Hach Company which had Heuristic ability to learn events, automatically tune itself and define what constitutes an abnormality in the system. It too is quite expensive and doesn't provide transparency in terms of algorithms, neither does it allow operational data to directly integrate into event detection. Last of these mentioned systems, **Ana::tool** is another EDS system which falls under the umbrella of a vast system moni::tool introduced by s::can in 2010 which also includes a user interface reflecting dashboard to reflect real time water quality parameters. . Except Canary most of these systems are way too expensive costing upto 90,000$ with their default settings (Canary 2010; EPA 2013).

Parameter	Hach, guardian Blue	S::can con::stat	Canary
Cost: Event Detection Software (10 Stations)	$92,500	$60,000	$0.00
Cost: Required Computing Hardware (10 Stations)	$0.00 (included)	$0.00 (included)	$3500
Cost: Total	$92,500	$60,000	$3500
Algorithm Transparency	Proprietary	Proprietary	Fully transparent
Direct Integration of Operational Data into Event Detection	No	No	Yes
Centralized processing on a single computing platform	No	No	Yes
Ability to work with sensors from multiple vendors	Custom Request	Yes	Yes

Table 1.1

This research is divided into 5 chapters, in chapter 1, we introduce the water quality problem, water quality parameters and existing water quality systems deployed around the world, in chapter 2 we briefly navigate through the extensive literature review to gain insight into the water quality problem from all perspectives, in chapter 3 we explore into

our prototype IoT system for real time monitoring, explore PCRWR dataset, perform data analysis, find outliers through box plot analysis and replace them with threshold values, introduce the water quality index & water quality class, perform z-score and q-value normalization and explore different machine learning methodologies used to estimate water quality, in chapter 4 we briefly discuss results using different algorithms, in chapter 5, we conclude the thesis & discuss possible future works and propose a system combining the state of the art machine learning methodologies and IoT hardware in turn providing real time monitoring, visualizations and artificial intelligence to make informed decisions.

2. Literature Review

In order to proceed with our research we did a literature survey of a number of local and international research papers as summarized in Table 1.2. We have categorized our surveyed papers into three categories: Research concerning Manual calculation and lab analysis to gain an insight in the basic problem statement, Research concerning different machine learning methodologies employed to learn trends of water quality parameters & predicting water quality and Research concerning IoT systems employed for water quality monitoring and prediction in real time.

2.1. Use of Statistical Analysis

Research concerning manual calculation and lab analysis on some samples provide us the insight to the basic problem at hand. Daud et al. (2017), in one of such research study, have collected various water samples across all the provinces of Pakistan. Different samples were tested for different parameters and were compared against NEQS and WHO standards. Majority of the samples had presence of Total coliform, Fecal coliform, E. Coli primarily due to mixing of sewerage water and secondarily due to industrial wastes. They recommended installation & maintenance of treatment plants and to ensure enforcement of NEQS. Alamgir et al. (2015) have collected 46 piped water samples across different places of orangi town, Karachi and tested it for bacteriological and physico-chemical analyses using Standard Methods for the Examination of Water and Wastewater. They have used WHO and National Standard for Drinking Water Quality (NSDWQ) standards to compare their results against. They calculated Mean, median, minimum, maximum, standard deviation, quartile range and standard error for each of the parameters and found physio-chemical parameters to be well in limits except sulphates but bacteriological parameters such as total fecal coliform and total coliform counts were critically high reflecting poor hygienic and sanitation conditions. They have recommended continuous monitoring of water and revamping of sewerage systems.

Ejaz et al. (2010), in their study, have conducted their research on the dataset of river ravi by sampling its data for 3 years, from Jan 2005 to Mar 2007, from 14 sampling stations. They have tested for 12 parameters namely BOD, DO, COD, suspended solids, phosphorus, chloride, sodium, total nitrogen, nitrate, nitrite, oil & grease and total coliforms. The have used Standard Methods for the Examination of Water and Wastewater (1991 USA) for testing the aforementioned parameters and have used

NEQS (National Environmental Quality Standards of Pakistan) to compare their parameter readings with. They have used expensive lab analysis which is their major limitation from our research's prospective. Also, they have recommended to install more treatment plants and ensure enforcement of NEQS for better water quality.

Batabyal & Chakraborty (2015) have conducted their research in Kanksa-Panagarh area situated in West Bengal. They have collected samples from 98 tube wells during November to December 2011 for the post-monsoon period and during May to June 2012 for the pre-monsoon period. They have tested them for 13 parameters namely pH, TDS, total hardness, HCO3, Cl, SO4, NO3, F, Ca, Mg, Fe, Mn, and Zn against WHO (1993) and Indian (BIS, 1991) standards. They have done correlation analysis amongst the parameters to probe their correlation. Also, they have calculated water quality index (WQI) using the attained parameters, demonstrating, in detail, the Indian method to manually calculate the WQI.

2.2. Use of Machine Learning

Research concerning machine learning methodologies help us understand the application of machine learning in Water quality prediction and trend analysis. Uferah et al. (2018) have used IoT for real time monitoring & different machine learning methodologies to predict water quality. They have used ATMega328 microcontroller & PH, turbidity & temperature sensors in their IoT module for real time monitoring. As far as analytics module is concerned, they have conducted their research on a dataset collected from 11 different sources of Pakistan. They have analysed the data using multiple machine learning algorithms to predict the quality of water namely SVM, KNN, ANN & Deep neural networks. Deep Neural Networks yield the highest accuracy of 93% while the close second is SVM with accuracy of 91%. They have used accuracy, precision & recall for performance evaluation.

Sakizadeh (2016) has conducted his research on the dataset of 47 wells and spring (2006-2013) acquired from Ministry of Iran. His study takes 16 water quality parameters into consideration. He has used the method proposed by Horton (1965) to calculate WQI. He has employed three methodologies: ANN with early stopping, ANN with ensemble averaging, ANN with Bayesian Regularization. He calculated the correlation coefficients between the predicted and observed values of WQI to be 0.94 and 0.77 and concluded that ANN with Bayesian Regularization generalizes the dataset better than others. But his model is prone to overfitting because it has less samples so the study has to focus on efficient generalization.

Abyaneh (2014) has predicted two prominent and not easily acquired water quality parameters, Biochemical oxygen demand (BOD) and chemical oxygen demand (COD) using multivariate linear regression and artificial neural networks. BOD and COD are predicted using easily attainable parameters pH, temperature (T), total suspended (TS) and total suspended solid (TSS). This study has been conducted on the data acquired from from Ekbatan wastewater treatment plant, Iran. To validate the model two prominent evaluation criterias were used root mean square error (RMSE) and coefficient of correlation (r). As evident in the results ANN performed better than MLR in predicting BOD and COD. Using ANN with minimal input parameters the evaluation metric of BOD was RMSE = 25.1 mg/L, r = 0.83 and for prediction of COD was RMSE = 49.4 mg/L, r = 0.81. It was established that the both models predicted BOD better than COD and PH had the most effect on the predictions.

Zhang et al. (2014) have proposed a system to monitor water quality online and employed machine learning algorithms to help users make educated decisions. Continuous data from different sites is gathered in data repository for monitoring and machine learning algorithms like pixel-based adaptive segmenter and bag of words approach are used on that data to aid user to make informed decisions. They have conducted their study on Dublin bay and they have used YSI 6600EDS for continuous monitoring of turbidity, optical dissolved oxygen, temperature, conductivity and depth. They have modified and used pixel-based adaptive segmenter from image processing domain to detect anomalous events from continuous data stream. Once anomalous events are detected then they extract features of those anomalous events and cluster those events to help in decision making.

Ali & Qamar (2013), in their research, have mapped this problem to the machine learning domain. They have conducted their research on Rawal watershed, situated in Islamabad. They collected 663 water samples from 13 different stations and tested them for Appearance, Temperature, Turbidity, pH, Alkalinity, Hardness as CaCO3, Conductance, Calcium, Total Dissolved Solids, Chlorides, Nitratres and Fecal Coliforms against WHO standards. They initially preprocessed data, filled out the missing values by attribute mean and replaced the outliers by attribute median. Followed by a correlation analysis to draw out the correlation amongst the parameters. They have employed regression models to check seasonal water quality trends (monthly and quarterly) and since there was no water quality index(WQI) in the data, they have employed unsupervised learning: Average Linkage (Within Groups) method of Hierarchical Clustering using Euclidean distance to categorize water quality. In results, they found higher values of fecal coliforms were found in the months of March, June,

July, and October. But their model had a clear limitation since no other parameters except fecal coliforms and turbidity were out of standard limits in the data set so it was a little biased and ensured accuracy mostly on turbidity and fecal coliform.

Gazzaz et al. (2012) have conducted their research on 255 samples of Kinta river Malaysia, obtained by their Department of Environment. Their dataset comprises of 9180 datapoints derived from measurements on those samples. They acquired 30 parameters from those samples and reduced them to 23 through Principal Factor Analysis(PFA). Next to it, they initially calculated the WQI manually using Malaysian WQI method and then using Artificial Neural Network with a setting of 23-34-1 to train their model. They partitioned their dataset into 3 parts, 80% for training, 10% for validation and 10% for testing. The aforementioned setting explained 99.5% of the predictions and variations of the data accurately. The only drawback to proposed approach was to have a large dataset in order to achieve satisfactory accuracy.

Verma and Singh (2012), in their study, have acquired 73 datasets from Jharia coalfield situated in Jharkhand, India. They have used 58 of those datasets for training and 15 for test. They have used three-layer feed-forward back propagation neural network and trained it for 1000 epochs. Their model takes in six inputs temperature, pH, TS, TSS, DO and oil & grease and produces two outputs BOD and COD. Their results reflect RMSE values for the BOD and COD to be 0.114 and 9.83 % and corresponding coefficients of correlation to be 0.976 and 0.981 and also, conclude that ANN with Bayesian Regularization generalizes best. One of their major limitations with respect to our study is that they don't actually predict WQI but estimate BOD and COD which might add to the error if we are to use them to predict WQI.

Mahapatra et al. (2011) have proposed to use fuzzy system to predict water quality index (WQI) of water. Since there is a certain uncertainty when you are working with crisp inputs in terms of water quality parameters. Normally, conventional simple fuzzy systems like Mamdani and Takagi, Sugeno and Kang would work but as it gets complex their efficiency is highly affected. So they have proposed a cascaded fuzzy system which works better with complex problems. The proposed fuzzy system takes multiple inputs and gives out multiple outputs by using multiple fuzzy sub systems. They have validated their system on data collected from CPCB (Central Pollution Control Board of India). They have used data of 6 indian rivers and estimated WQI using three water quality criterias (Indian, Malaysia and USA). Also, they have used six parameters for their case study, namely pH, Biological Oxygen Demand, Dissolved Oxygen, Fecal Coliform, Electric Conductivity, Ammonical Nitrogen and Temperature. As mentioned above, they

have used three fuzzy subsystems, each for different water quality criteria. Evidently, predictions of the system are quite close to the actual WQIs of each criteria making proposed system more fit, to the problem at hand, than conventional fuzzy systems.

Bucak & Karlik (2011) have emphasized on the importance of real time contamination detection of water. Their research is mostly focused on intentional contamination of water. They have used Cerebellar Model Articulation Controller Artificial Neural Network (CMACANN) for contamination detection because of its evident fast learning capabilities. They have monitored 5 parameters: pH, conductivity, chlorine residual, turbidity, and Total Organic Carbon in their system. To validate their model they intentionally introduce certain contaminants in the water: sodium cyanide, sodium arsenate, sodium fluoroacetate, parathion, Cryptosporidium parvum oocyts, and a surrogate of Bacillus anthracis spores. Their model then detects the effects that the contaminants have and classify it as an anomaly. Their proposed model works far better than conventional multilayer perceptron with back propagation (MLP with BP) . MLP achieves an accuracy of 98% after 1000 iterations while the proposed model achieves 100% accuracy with far less iterations.

Yan et al. (2010) have used adaptive neuro fuzzy inference system (ANFIS) to predict the water quality status and compared its results with the convention artificial neural networks and found it to be more efficient than it. They have used the dataset of major river basins of China obtained from CNEMC consisting of 845 observation samples. They have selected three parameters for their classification model, namely dissolved oxygen (DO), chemical oxygen demand (COD) and ammonia-nitrogen (NH3-N). The used model ANFIS combines the two algorithms ANN and fuzzy logic to map the water quality problem in an efficient manner. Fuzzy logic works in terms of IF-THEN rules which makes it easier to interpret and map but generation of those rules and its consequences requires expert knowledge which makes fuzzy logic unsuitable to our problem. While ANN comes with a certain adaptability which enables ANFIS to combine the power of both algorithms. ANN allows ANFIS to learn and construct rules of fuzzy logic as according to the problem at hand which, in results, have turned out to be more efficient than either of the models and classified 89.59% of the data correctly.

Rankovic et al (2010) have conducted their study on Gruza reservoir, Serbia. They have acquired 180 data samples by monthly sampling for 3 years (2000 - 2003) though monitoring. They have used 152 of those data samples for training and 28 for test. They have considered pH, temperature, chloride, total phosphate, nitrites, nitrates, ammonia, iron, manganese and electrical conductivity as their input and Dissolved Oxygen (DO)

to be the predicted parameter. They have used Feedforward neural network (FNN) model to predict the dissolved oxygen. Levenberg–Marquardt algorithm is used to train the FNN and they have established that 15 hidden neurons give the optimal results. Results of FNN models have been compared with the measured data on the basis of correlation coefficient (r), mean absolute error (MAE) and mean square error (MSE). The limitation of this work is that they are predicting DO instead of WQI which, from our research topic's perspective, might add to the error if we are to predict WQI using the predicted DO. Also, the model needs to be updated every now and then with real values to reflect the environmental changes.

Najah et al. (2009) have used artificial neural networks to predict three water quality parameters, total dissolved solids (TDS), electrical conductivity, turbidity. They have conducted their study on two monitoring stations, Johor River and Sayong River situated in Malaysia. They have employed separate methodology for each parameter and each monitoring station. For TDS, they have used back propagation with two hidden layers and bayesian regularization but with distinct transfer functions for each monitoring station. They have predicted TDS using EC since they are highly correlated as evident in their results. They have employed the same methodology for EC and predicted it using TDS given their correlation. For turbidity, they have employed feed-forward neural network using back propagation with single hidden layer and back propagation. Distinct function for each monitoring station was used. They have predicted turbidity using total suspended solids (TSS) since they are highly correlated. The selected models imitated each water quality parameter quite efficiently with minimal prediction error.

Rene & Saidutta (2008) have used regression analysis and artificial neural networks to predict Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) using other water quality parameters, namely, Total Organic Carbon (TOC), Total Suspended Solids (TSS), Total Dissolved Solids (TDS), Phenol concentration, Ammoniacal Nitrogen (AMN) and Kjeldahl's Nitrogen (KJN). They have employed regression analysis to find correlation of TOC with BOD & COD. After regression analysis they have run 12 different models of ANN using different combinations of aforementioned water quality parameters to predict BOD and COD. They have used Average Relative Error (ARE) to find accuracy of the model. Model having 7 hidden neurons in hidden layer and training count of 5000 with TOC, Phenol, TSS and AMN predicted the BOD most effectively with ARE = 11.6614%. Model having 8 hidden neurons and training count of 1500 with TOC, Phenol and TDS as input predicted COD better having ARE = 6.9729%. Model with 6 hidden neurons and training count of 5000 with TOC, Phenol, TSS and TDS as input performed effectively for both BOD and COD

with ARE for BOD = 8.201 % and ARE for COD = 11.0835 %. The empirical relations formed amongst parameters are quite reliable and bring versatility in the domain.

2.3. Use of Internet of Things (IoT)

Research concerning IoT systems make us familiar with the required hardware for water quality monitoring and prediction in real time. Geetha & Gouthami (2017) provide us with such basic framework of an IoT system. The have proposed a generic IoT system for real time water quality monitoring. It comprises of sensors which read parameter readings, then those parameter readings are transmitted to a controller through wireless communication devices attached with sensors and then controllers, through some wireless communication technology store those sensor readings to a data storage which are reflected in some customized application. They then implemented an instance of the generalized IoT system. They used four parameter sensors for it namely conductivity, turbidity, water level and pH. For connectivity they used TI CC3200 which is a single chip microcontroller with in-built Wi-Fi module and ARM Cortex M4 core, which can be connected to the nearest Wi-Fi hotspot for internet connectivity and in turn move the data to cloud or storage and then some application could be using that data storage to reflect the readings.. If sensors were not connected directly to the controller they could be connected using LoRa sensors.

Encinas et al. (2017) have presented a prototype for water quality monitoring of ponds. They have used temperature, PH and dissolved oxygen sensors, an arduino module and zigbee transmitters and receivers. When it comes to software they have used MySql database, SOAP web services and applications developed on C# and android. C# application allows to send request for particular sensor readings through arduino and multiplexer. Once requested, particular sensors take readings and send them back to the computer through zigbee transmitter and the readings are then received by a zigbee receiver attached with computer. The received reading are then saved in the local database and sent to the cloud through the web service and eventually are visualized in the android application. Artificial intelligence is not used in the system but it does set the base for artificial intelligence to be used in future for effective real time decision making.

Raju & Varma (2017) have proposed a real time monitoring system for aqua farmers which allows farmer to be apprised of the anomalous event if the pond is contaminated. They have used Raspberry pi3 with built in Wifi module, a solar panel and a sensor node comprising of several sensors: Dissolved Oxygen, Ammonia, pH, Temperature, Salt,

Nitrate and Carbonates, mounted on it. It continuously monitors and stores sensor data and generates an alert for the farmer if any of the data deviates from the allowed ranges. There is mobile application for the farmer in which he can monitor the sensor data in real time and access historical data. Vijaia & Sivakumar (2016) have proposed a general framework for IoT system for real time water quality monitoring, demand forecasting and anomaly detection. For IoT system they have considered the parameters: Turbidity, chlorine, ORP, Nitrates, pH, conductivity and temperature and used their sensors. For connectivity they have proposed several options 3g, Bluetooth, Zigbee etc. These components when connected make a centralized system requiring a steady power supply to keep the system online. They have proposed two other components Demand forecasting and Anomaly detection. For anomaly detection they have used ANN and fuzzy systems. Their proposed system is quite a general one and no dataset has been used to test it.

Birje et al. (2016), in their paper, have proposed a system to monitor two of the most descriptive water quality parameters which particularly determine whether water is safe for aquatic life. They have used PH sensor along with PH meter and turbidity sensor and their readings are sent to analog to digital converter(ADC) which in turn sends digital readings to 16F887A PIC microcontroller which shows its results on LCD. Their work is extensible to use GSM to communicate the water quality. Cloete et al. (2016), in their research have designed a sensor node which consists of temperature, conductivity, pH, ORP and flow sensors. Since sensors available in the market are expensive, they have implemented the sensor designs themselves in turn making the system cost effective. The signals generated from the sensors then go through conditioning in order to be able to interface with microcontroller. Apart from the sensor node their proposed system makes use of Zigbee module to receive and transmit measurements ahead and a microcontroller to notify the measurements. All the measurements are then shown on an LCD in front of their respective labels and a buzzer goes off if any of the measurement goes out of its allowable limits.

Wong & Kerkez (2016) have emphasized on the importance of using real time data along with historical data and on flexibility that comes with using web services along with IoT platforms. Since some of the water quality constituents are difficult to measure or their sensors are too expensive for it to be cost effective, they have used adaptive sampling of water for them along with easily available sensors. Adaptive sampling, instead of sampling after a predefined intervals, adapts and samples only when event of interest occurs i.e: flood and minimizes the number of samples to be taken. They have used Neomote wireless sensing platform which consists of ARM-Cortex M3 microprocessor

for computing and Xively IoT platform, which acted as the interface for the services. The sensor node was connected with automated sampler (ISCO 3700) which had a 24 bottle capacity. To emphasize on the flexibility of the web services, they have used three web services each of it developed in different programming language. One of them was used to receive commands for sampling and transmitting data, second service had adaptive sampling algorithm in it and sent the sampling commands to the first service and third web service helped to interface with IoT platform as to access historical data and communicate with sensors. Data transfer among these services was in the form of JSON due to convention but it also supports the XML format.

Perumal et al. (2015) have come up with a prototype for measuring water level in real time as to be apprised of events like floods and generate an alarm to authorities and on social networks. They have used ultrasonic sensors, a wireless gateway, ATmega328P controller and a cloud server. After every interval, ultrasonic sensors determine the distance between water level and sensor by sending a wave and estimating water level by its reflection. Once determined, the water level reading is transmitted to the cloud server through wireless gateway, where it is stored on a database. If the water level crosses a certain predefined threshold, an alarm is generated to alert the authorities or to broadcast it on social networks like twitter. Also, water level data stored on cloud server is visualized through a web application as to learn trends and perform decision making. Vijayakumar & Ramya (2015) have proposed an online water quality monitoring system. They have used five parameter sensors namely temperature, pH, turbidity, conductivity and dissolved oxygen. For IoT connectivity they have used , Raspberry Pi B+, IoT module USR-WIFI232-X-V4.4 which transmit the data to the cloud through the gateway. The proposed system provides water quality monitoring and is suggested to be installed in different locations of a pond to collect real time water quality data.

Cao et al.(2014) have proposed an inexpensive, easy to setup wireless network to monitor water quality using ISFET microsensors and mobile communication. Microsensors are deployed on the site to measure important water quality parameters and send the measurements to Sensing end device (ED) nodes attached with sensors. ED nodes then transmit the measurements to Sensing access point (AD) node which is connected to database server, where the sensor data is stored for future use and visualization. They use mobile networks for communication between ED nodes and AD nodes. The system was programmed to collect sensor data automatically after every two hours. To experiment with their proposed system they used two microsensors, PH and temperature.

Rasin & Abdullah (2012) have proposed a cost effective online water quality monitoring system using wireless sensor network (WSN). Their system contains two modules, Wireless node and base monitoring station. Wireless node consists of a sensor unit and a microprocessor and is powered by a 9V battery. They use ZMN2405HP Zigbee module which consists the CC2430 transceiver IC. They have used the inexpensive PH, temperature and turbidity sensors. the readings of the sensors go through signal conditioning as to determine their validity. Once conditioned, wireless sensor node sends the readings to the base monitoring station through the transceiver. The other Zigbee module consisting of transceiver at the base monitoring station receives the readings and sends to the computer using RS 232 protocol. The received data is then visualized on a custom GUI developed in C++. Wang et al. (2009) have proposed a low cost, low power, long-distance supervisory system based on the wireless sensor network (WSN) for aquaculture. Their proposed system contains two modules, coordinator and sensor node. Coordinator is composed of zigbee based wireless communication module, which uses CC2430 chip with RF transceiver and an analog-to-digital converter (ADC), and a GPRS module to transmit the data to the monitoring computer which stores the data and visualizes it. Sensor node contains the sensors which read the water quality parameters and applies signal conditioning on the readings to ready them to be digitized. After signal conditioning they are sent to the coordinator where they are digitized and processed ahead. Also, the system is modelled to consume low battery, as it goes into sleep mode when there is no request for data to be read.

In this chapter, we have reviewed the problems statement from multiple perspectives i.e. Statistical Analysis, Machine Learning and IoT. Most of the research pertaining to our problem is international and not local which motivated this study since environmental factors vary geographically and affect differently in different geographical locations.

Paper	Methodology	Limitations	Dataset	Parameters	Results	Hardware
Uferah et al. (2018)	Monitoring using sensors and classifying water quality using DNN, NN, SVM & KNN	Classifies water quality only into two categories i.e. good or poor. Doesn't use WQI.	Dataset of 667 samples collected from PCRWR.	PH, Turbidity Temperatur e	Accuracy: DNN 93% SVM 91%	ATMega328, LCD & Parameter sensors

					NN 86% kNN 76%	
Geetha & Gouthami (2017)	Monitoring using sensors and cloud infrastructure	Only monitoring, no prediction	N/A	Conductivity Turbidity Water Level pH	N/A	TI CC3200 controller & parameter sensors
Daud et al. (2017)	General Review of water quality across provinces	Only manual lab analysis	Manual samples across Pakistan	Total coliform Fecal coliform E. Coli	Excessive Total coliform due to sewerage	N/A
Encinas et al. (2017)	Water quality monitoring using sensors and SOAP web services	Only monitoring, no prediction	N/A	Temperature PH DO	N/A	Parameter sensors, arduino module and zigbee transcievers
Raju & Varma (2017)	Real time monitoring system and mobile application for aqua farmers to be apprised of contamination	Just provides monitoring, doesn't process data for trends	N/A	DO, Ammonia, pH, Temperature, Salt, Nitrate and Carbonates	N/A	Raspberry pi3 with built in Wifi module, a solar panel and a sensor node
Wong & Kerkez (2016)	Adaptive sampling of water using adaptive sampling algorithm, Xively IoT platform & webservices to monitor water quality.	It doesn't actually monitor water quality in real time but through sampling, also just provides monitoring. No predictive analysis	N/A	N/A	N/A	Neomote wireless sensing platform: ARM-Cortex M3 microprocessor & Xively IoT platform. Also, automated sampler (ISCO 3700 with 24 bottle capacity)
Vijaia & Sivakumar (2016)	Artificial Neural Network (ANN) and fuzzy systems	Proposes a generic IoT system without any dataset & results	N/A	Turbidity chlorine ORP Nitrates pH conductivity temperature	N/A	Sensors, connectivity: 3G, Bluetooth, & Zigbee
Sakizadeh (2016)	ANN with early stopping, ANN with ensemble averaging & ANN with Bayesian Regularization	prone to overfitting with less samples	47 wells and springs (2006-2013) from Ministry of Iran	16 groundwater quality variables. To calculate mentioned WQI	Bayesian regularization. WQI cor: 0.94 and 0.77	N/A
Alamgir et al. (2015)	Bacteriological and physico-chemical analyses	Only manual lab analysis	Forty six samples of piped water in Orangi	pH TSS TDS Turbidity	well within limits except sulphates and total	N/A

			town 2014	TCC TFC TFS	fecal coliform	
Batabyal & Chakraborty (2015)	Calculate WQI using manual indian method	Manual calculations	98 tube wells	pH, TDS, Total hardness, HCO3, Cl, SO4, NO3, F, Ca, Mg, Fe, Mn, and Zn	poor quality was attributed to high contents of TDS, NO 3 and Cl	N/A
Vijayakumar & Ramya (2015)	Monitoring employing IoT through sensors and cloud	Just provides monitoring.	N/A	Temperature PH Turbidity Conductivity DO	N/A	Sensors, Raspberry Pi B+, IoT module USR-WIFI232-X-V4.4)
Abyaneh (2014)	multivariate linear regression (MLR), Artificial neural networks (ANN), RMSE, r	Just predicts BOD which doesn't wholly reflect water quality	Data acquired from Ekbatan wastewater treatment plant, Iran	pH temperature total suspended (TS) total suspended solid (TSS)	Both models predicted BOD better than COD and PH had the most effect on the prediction	N/A
Zhang et al. (2014)	continuous monitoring, pixel-based adaptive segmenter and bag of words	Doesnt actually predict water quality, just clusters possible anomalous events	Dublin bay	turbidity, optical dissolved oxygen, temperature, conductivity and depth	N/A	YSI 6600EDS
Ali & Qamar (2013)	Preprocessing : Attribute mean, Regression models, Hierarchical clustering	Biased dataset: No other parameters except fecal coliforms and turbidity were out of standard limits	13 different stations, 2009 to 2012, 663 water samples	Appearance temperature turbidity Ph alkilinity hardness as CaCO3 conductance Calcium TDS Chlorides	High fecal coliforms were found in the months of March, June, July, and October	N/A

				Nitrates Fecal Coliform		
Verma and Singh (2012)	ANN with Bayesian Regularization: 1000 epochs	Doesn't actually calculate WQI but predicts BOD and COD	73 datasets, 58 train and 15 test	Six inputs (temp, pH, TS, TSS, DO and oil and grease) and two outputs (BOD and COD)	(RMSE) values for BOD and COD are 0.114 and 9.83 % & correlation is 0.976 and 0.981	N/A
Gazzaz et al. (2012)	Artificial Neural Network , 23-34-1	Must have huge datasets.	9180 data points, 255 samples	30 Parameters reduced to 23 through PFA	Predictions explain almost 99.5% of the variations	N/A
Rankovic et al (2010)	FNN. Levenberg–Marquardt algorithm is used to train the FNN. 15 hidden neurons.	Prone to overfitting. Doesn't actually calculate WQI but predicts DO which might result in error ahead	180 data samples, 152 train, 28 test	pH, temperature , chloride, total phosphate, nitrites, nitrates, ammonia, iron, manganese and electrical conductivity	correlation coefficient (*r*), mean absolute error (MAE) and mean square error (MSE) indicate accurate results	N/A

3. Methodology

In methodology, different aspects of water quality have been explored. Initially water quality index and water quality class is discussed which is the building block of the research. Following it, dataset has been described and preprocessed. After preprocessing, machine learning methods are employed to estimate aforementioned water quality index. Eventually, the proposed IoT system, built on top of machine learning module, is discussed in detail. The most basic of building blocks of this research is water quality index which is the definitive and singularly quantifiable measure to define water quality and it is discussed below.

3.1. Water Quality Index

Water quality index is the singular measure which indicates the quality of water and it is calculated using various parameters that are truly reflective of water's quality. Once WQI is calculated, we use it to define Water Quality Class.

Water Quality Index (WQI):

To conventionally calculate WQI 9 Water quality parameters are used but if we don't have all 9 parameters we could still estimate water quality index with atleast 6 of defined parameters. We have 6 of those parameters, namely Fecal coliform, pH, temperature, nitrates, turbidity and total dissolved solids in our data set. Using these parameters and their assigned weightages we have calculated WQI as reflected in the equation below(Thukral, Bhardwaj & kaur 2015; Srivasstava & Kumar 2013).

$$\boldsymbol{wqi} = \frac{\sum Q - value * Weighing\ Factor}{\sum Weighing\ Factors} \ (\boldsymbol{Eq.1})$$

Weighing Factor	**Weight**
pH	0.11
Temperature	0.10
Turbidity	0.08
Total Dissolved Values	0.07
Nitrates	0.10
Fecal Coliform	0.16

Water Quality Class (WQC):

Once we estimated the water quality index, we defined the water quality class using WQI, to use in classification algorithms (Thukral, Bhardwaj & kaur 2015; Srivasstava & Kumar 2013).

Water Quality Index Range	Class
0 – 25	**Very bad**
25 – 50	**Bad**
50 – 70	**Medium**
70 – 90	**Good**
75 – 100	**Excellent**

3.2. Data Analysis & Preprocessing

The most important and initial part of any research is a definitive data which defines and drives the research. We have collected our dataset of Rawal water shed from PCRWR. Which doesn't contain all the parameters but contains most definitive parameters which estimate water quality well. In this part we describe the data, detect & remove outliers though Box plot Analysis, perform q-value & z-score normalization, perform correlation analysis, do parametric analysis, select minimal features to estimate WQI and split data for machine learning algorithms.

3.2.1. Dataset Description

Dataset collected from PCRWR contains 663 samples from 13 different sources of rawal water lake collected throughout 2009 to 2012. It contains 51 samples from each source. It contains following 12 parameters as listed in table 1.2.

Parameter	WHO Limits
Alkalinity	500 mg/l
Appearance	Clear
Calcium	200 mg/l
Chlorides	200 mg/l
Conductance	2000 µS/cm
Fecal Coliforms	Nil Colonies/100ml
Hardness as CaCO3	500 mg/l
Nitrite as NO^-	<1 mg/l
pH	6.5 - 8.5
Temperature	◦C

Total Dissolved Solids	1000 mg/l
Turbidity	5 NTU

Table 1.3. Parameters along with their WHO standard limits

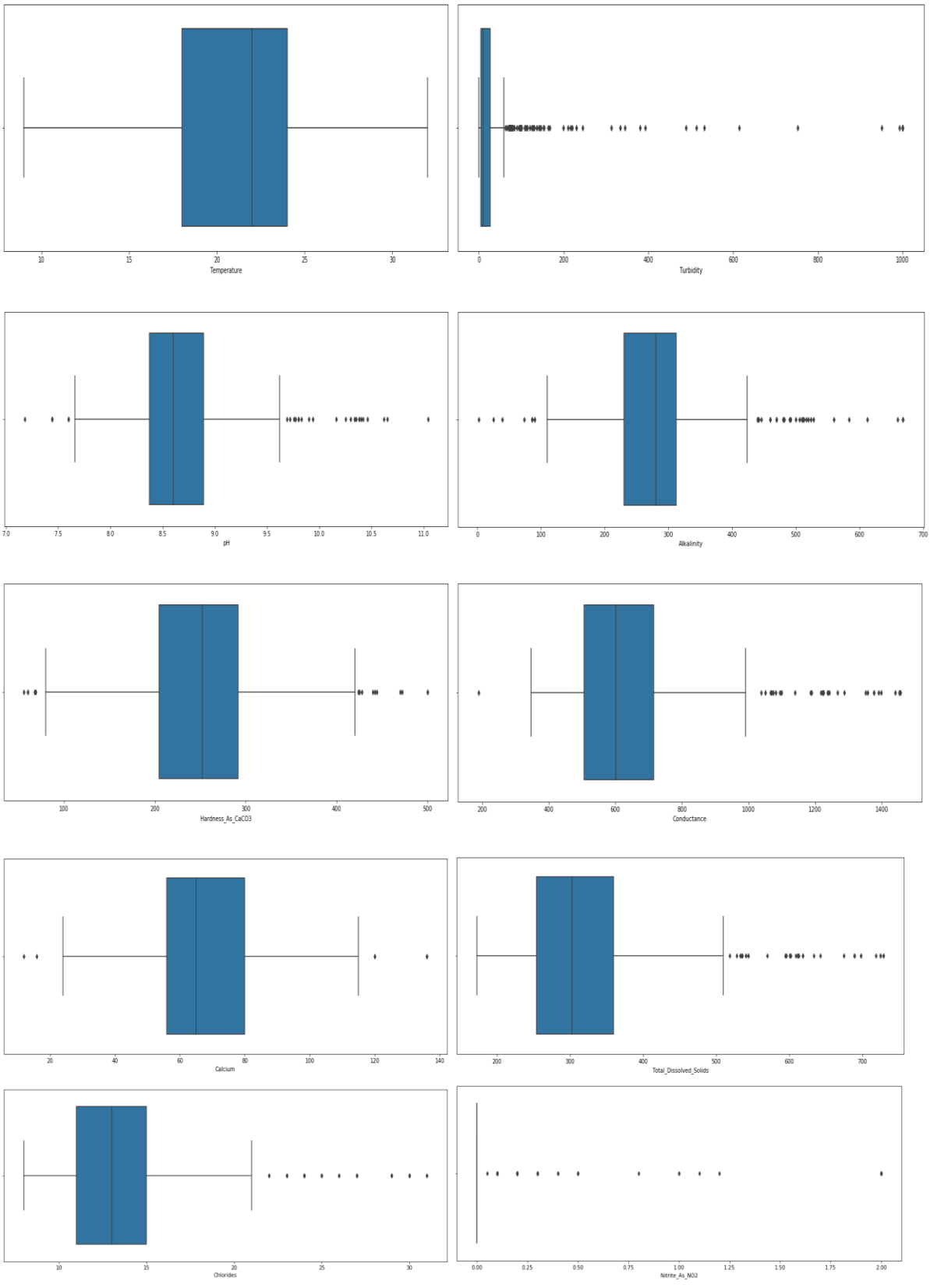
Source	Location
S1	Stream coming from Bidhawa Village
S2	Stream coming from Ghora Gali
S3	At Junction of Bidhawa Village and Ghora Gali Streams
S4	At Junction near DESTO Lab and Upstream of Chattar Park
S5	Korang River before Chattar Park near Bahria Town
S6	At Junction of Korang River and Chattar Park Streams
S7	At Downstream Chattar Park and Upstream of Bara kahu
S8	Stream coming from Shahdra
S9	Korang River before entering Rawal Lake
S10	Stream coming from Bari Imam and Diplomatic Enclave
S11	Stream coming from Quaid-e-Azam University
S12	Stream coming from Bari Imam at Noor Pur Shahan
Reservoir	Rawal Lake

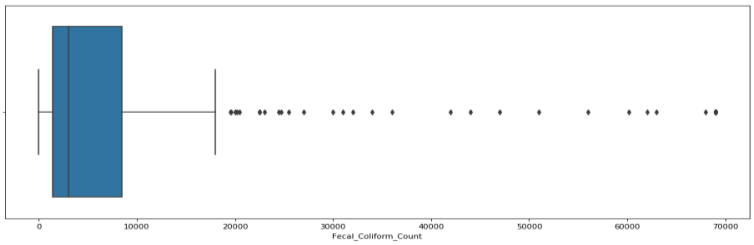
Table 1.4. Rawal Lake sources

3.2.2. Boxplot Analysis & Outlier Detection

We have chosen boxplot analysis for outlier detection since most of parameters were varying enough and were on the higher end of the values and boxplot provides insightful visualization to decide outlier detection threshold values depending upon the problem domain. Boxplot Analysis yielded that most parameters lied outside the box deeming outliers normal so we adapted an upper cap strategy to filter out outliers. We recognized the parameter values that were way off than other values and replaced them with the max threshold value. We set max threshold value as the parameter value that was just below the outlier values. For example, for turbidity we set threshold value as the sample value which just below 80 and applied it to all values above 80. We repeated the same

process with all the parameters and manually removed the outliers such as to not risk any data loss at all, given our limited dataset (Gazzaz et al. 2012).





3.2.3. q-value Normalization

Q-values normalization is used to normalize parameters, particularly water quality parameters to fit them with in the range of 0 to 100 as for easier index calculation. Following are the q value charts for 6 of the Water quality parameters. We have used them to convert these 6 parameters within the range of 0 to 100 (Thukral, Bhardwaj & kaur 2015; Srivasstava & Kumar 2013).

Fig. 1. Q- value for fecal coliforms (Q=2 for FC>100000/ 100ml)

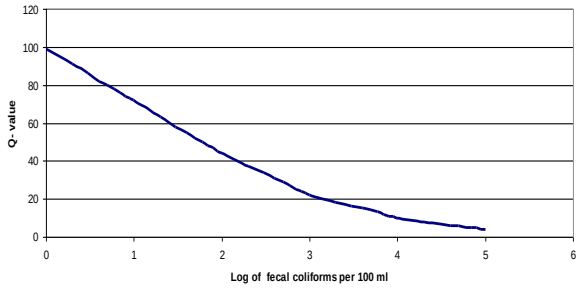


Fig. 4. Q- value for pH (Q=0 for pH<2 and pH>12)

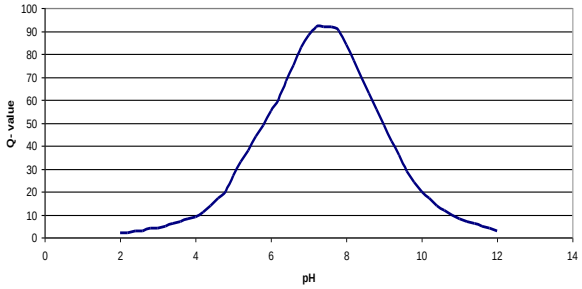


Fig. 5. Q- value for temperature change

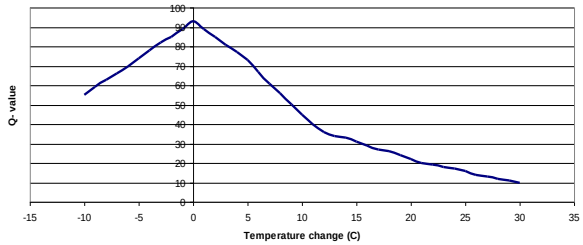


Fig. 6. Q- value for turbidity (Q=5 for NTU>100)

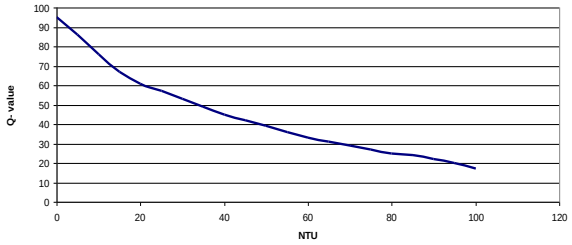


Fig. 7. Q- value for total solids (Q=20 for TS>500 mg/L)

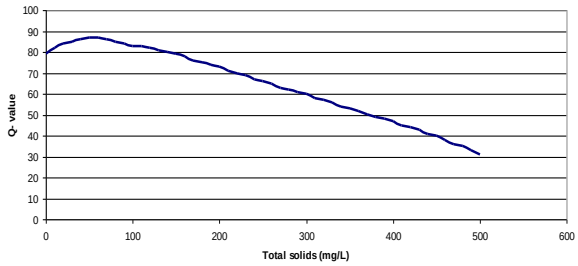
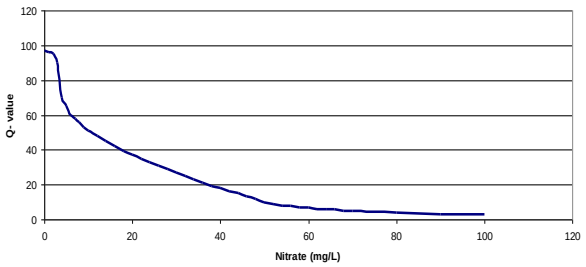


Fig. 9. Q- value for nitrate (Q=1 for nitrates>100 mg/L)



3.2.4. z-score Normalization

z-score is the conventional standardization and normalization method which represents the number of standard deviations, a raw data point is above or below the population mean. It ideally lies between -3 and +3. It normalizes all the data to the aforementioned scale to bring all the data with varying scales on the same scale.

To normalize data, using z-score, we subtract mean of the population from a raw data point and divide it by standard deviation which gives out a score ideally varying between -3 to +3 reflecting how many standard deviations a point is above or below the mean (Jayalakshmi & Santhakumaran 2011).

$$zscore = \frac{(x-\mu)}{\sigma} \quad \textbf{(Eq. 2)}$$

3.2.5. Correlation Analysis

To find the dependent variables and to predict hard to estimate variables through easily attainable parameter we performed correlation analysis to extract the relations of the parameters. We have used the most commonly used and effective correlation method known as Pearson correlation. We have applied pearson correlation on the raw values of the parameters and applied it after normalizing the values through q-value analysis as explained in the subsequent chapters.

	Temp	Turb	pH	Alk	CaCO3	Cond	Ca	TDS	Cl	NO2	FC	WQI
Temp	1.000	0.103	0.005	-0.193	-0.288	0.266	-0.150	0.274	0.293	-0.154	0.194	-0.483
Turb	0.103	1.000	-0.0886	-0.093	-0.146	0.048	-0.122	0.042	0.037	0.0002	0.037	-0.360
pH	0.005	-0.088	1.000	-0.177	-0.278	-0.065	-0.236	-0.060	-0.149	0.167	0.054	-0.423
Alk	-0.193	-0.092	-0.177	1.000	0.462	0.011	0.444	0.012	0.061	0.046	0.013	0.228
CaCO3	-0.288	-0.146	-0.278	0.462	1.000	0.068	0.637	0.060	0.135	0.078	0.016	0.370
Cond	0.266	0.048	-0.064	0.011	0.068	1.000	0.225	0.973	0.780	0.100	0.456	-0.367
Ca	-0.150	-0.122	-0.236	0.444	0.637	0.225	1.000	0.219	0.262	0.124	0.113	0.198
TDS	0.273	0.041	-0.060	0.012	0.060	0.974	0.219	1.000	0.765	0.095	0.454	-0.380
Cl	0.292	0.037	-0.149	0.061	0.135	0.780	0.262	0.765	1.000	0.036	0.353	-0.275
NO2	-0.154	0.0002	0.167	0.046	0.078	0.100	0.124	0.095	0.036	1.000	0.193	-0.139
FC	0.194	0.037	0.053	0.012	0.016	0.456	0.113	0.454	0.353	0.193	1.000	-0.418
WQI	-0.483	-0.359	-0.423	0.228	0.370	-0.367	0.198	-0.380	-0.275	-0.139	-0.418	1.000

As the correlation chart indicates:

- Alkalinity is highly correlated with hardness and calcium.
- Hardness is highly correlated with Alkalinity and calcium and loosely correlated with pH.
- Conductance is highly correlated with Total Dissolved Solids, Chlorides and Fecal coliform count and loosely correlated with Calcium and temperature.
- Calcium is highly correlated with Alkalinity and hardness while loosely correlated with TDS, chlorides, conductance and pH.
- TDS is highly correlated with conductance, chlorides and fecal coliform and loosely correlated with calcium and temperature.
- Chlorides are highly correlated with conductance and TDS and loosely correlated with temperature, calcium and fecal coliform.
- Fecal coliform are correlated with conductance and TDS and loosely correlated with chlorides

Now that we have listed down the correlation analysis observations, we find that our predicting parameter WQI is correlated with 7 parameters namely Temperature, Turbidity, pH, Hardness as CaCO3, Conductance, Total Dissolved Solids and Fecal Coliform Count. Since we have to choose minimal number of parameters to predict WQI as to lower the cost of the system. The 3 parameters whose sensors are easily available, cost the lowest and contribute distinctly to the WQI are Temperature, turbidity and pH which deems them naturally selected. The other convenient feature is Total Dissolved Solids whose sensor is also easily available and is correlated with Conductance & Fecal Coliform Count which means selecting TDS would allow us to discard other two features.

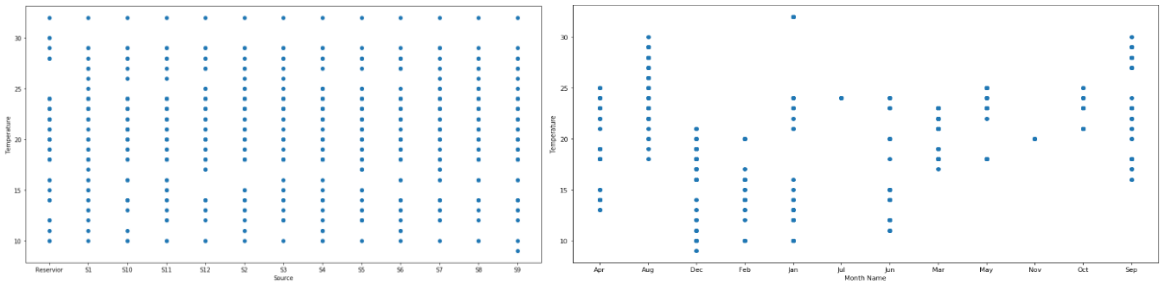
We leave the remaining inconvenient feature Hardness as CaCO3 out, since it is not that highly correlated comparatively and is not easy to acquire.

To conclude the correlation analysis, we select four features for the prediction of WQI, namely, Temperature, turbidity, pH and Total Dissolved Solids. This research would initially just consider the first three parameters given their low cost and if needed, TDS would also be included later to analyze its contribution to the accuracy.

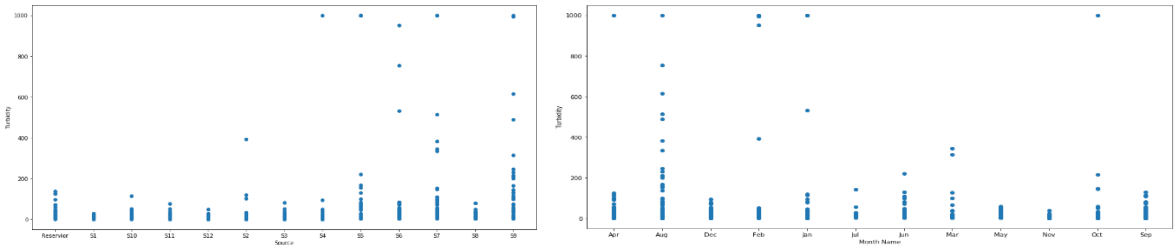
3.2.6. Parametric Analysis

To learn the data trends and gain deeper insight into the data we performed parametric analysis on each of the data feature. We studied data by rearranging each parameter month-wise and source-wise.

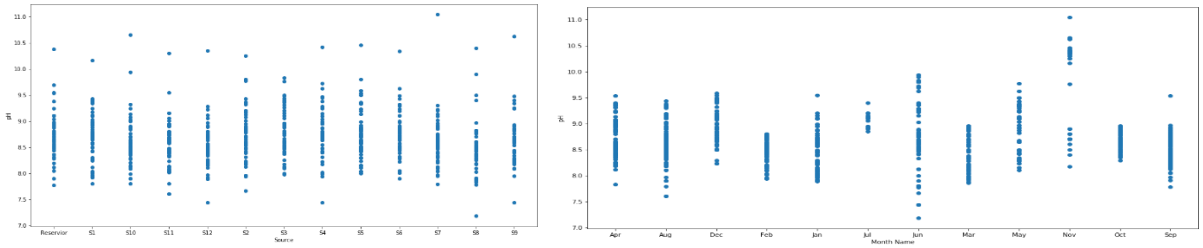
Temperature almost follows a similar distribution in all sources when studied source-wise while it shows high values in the months of August and September when studied month-wise.



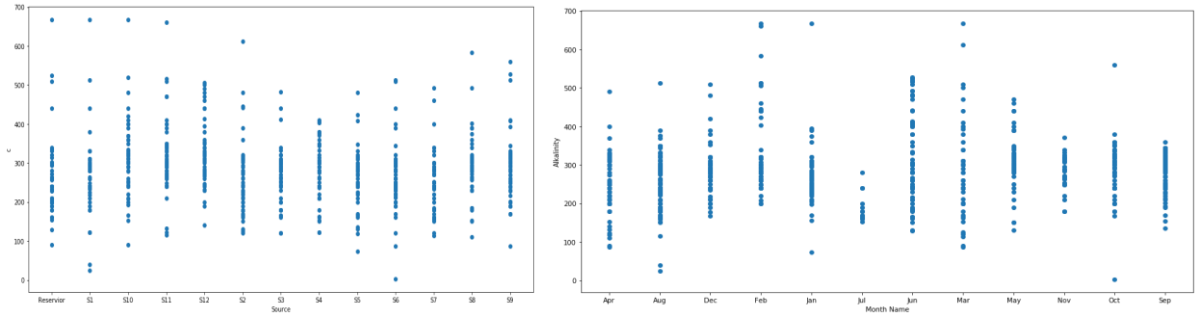
Turbidity goes high in Sources S6, S7 and S9. Also, it goes high particularly in the month of August.



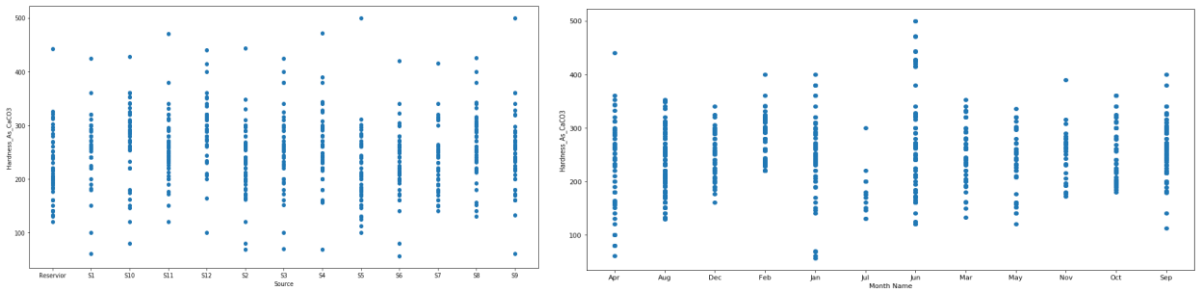
pH is on the higher end in most of the sources, atleast some of the samples are in each Source. It is particularly high in the month of June, May and November.



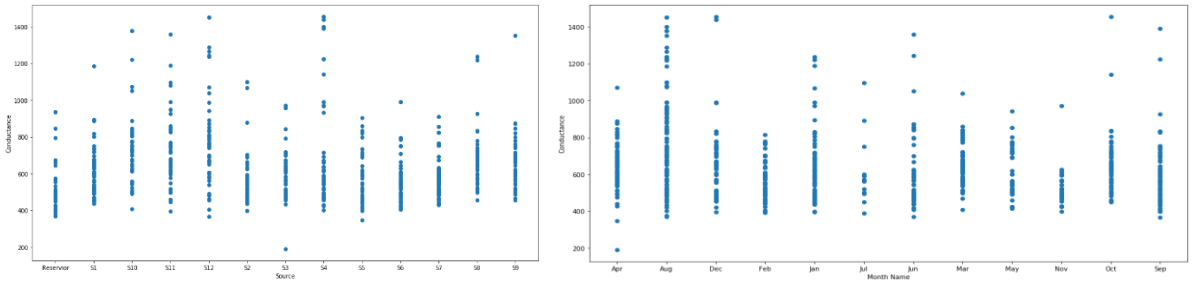
Alkalinity varies in each of the sources, but is particularly on the higher end in thee months of January, March and June.



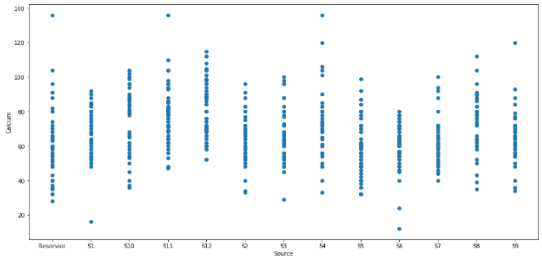
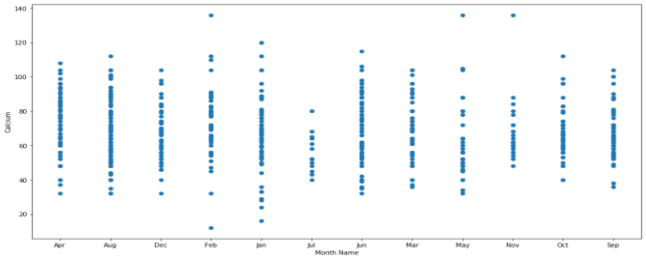
Hardness as CaCO3 is on the higher end in the month of June.



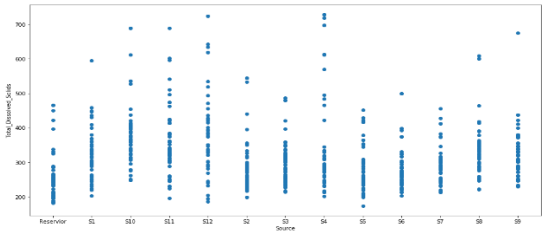
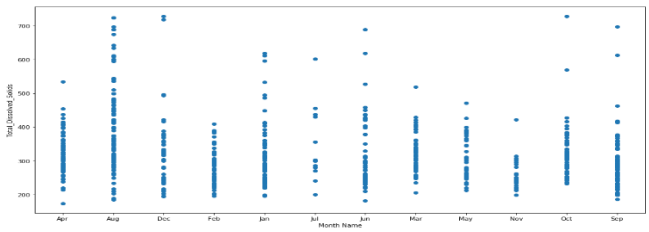
Conductance is on the higher end in the sources s4,s10,s11, s12 and in the months of August, December, January and June.



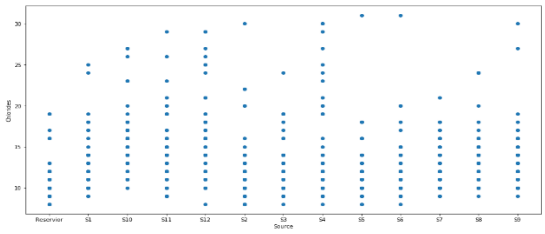
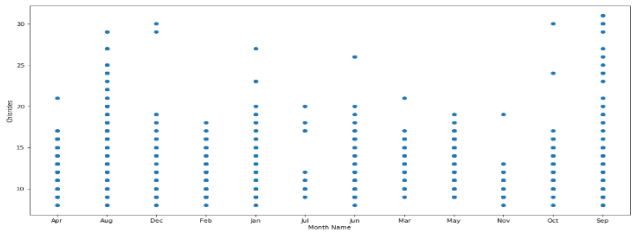
Calcium is on the higher end in source s4 and in the months of January, February, June and May.

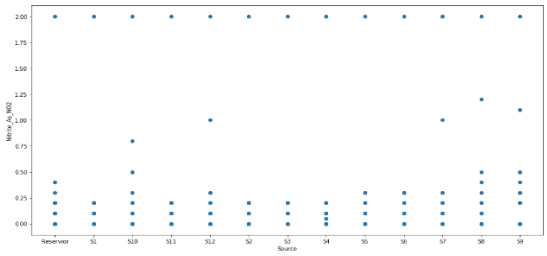
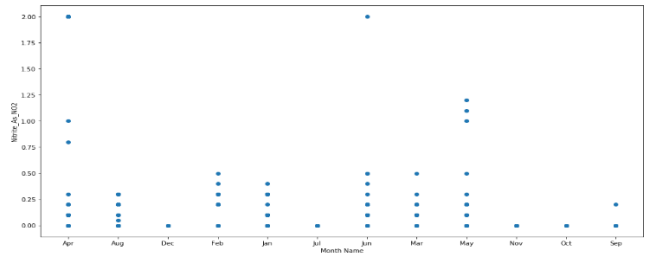
Total Dissolved Solids are on the higher end in sources s4, s10,s11 & s12 and in the months of January, June, August and December.

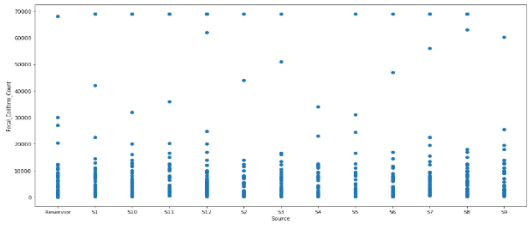
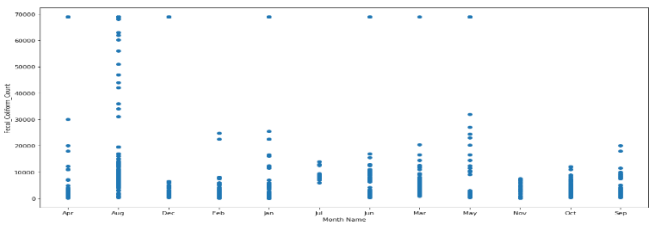
Chlorides are on the higher end in sources s4, s11 & s12 and in the months of August and December.

Nitrates are on the higher end in the months of April and May.

Fecal Coliform count are generally higher than allowed limit but are particularly in the month of August.

In similar way most of the graphs are self-explanatory and help in identifying the pattern of the parameter values.

3.2.7. Data Splitting

The last step prior to applying machine learning model is splitting the provided data as to train the model and test it with certain part of data and feed it to the accuracy measures to establish the model's performance. This research explores two data splitting techniques: Train-Test split and Cross validation.

Train-Test Split

In Train-Test split data is divided into two subsets, training set and testing set. The model is trained on the training data set and tested on test dataset. But it comes with a risk of data not being properly split and test data leaking information into training data. This research uses splits data into 80% training data and 20% test data.

Cross validation

Cross validation splits data into k subsets and iterates over all the subsets considering k-1 subsets as training dataset and 1 subset as testing dataset. This ensures the efficient split and use of proper and definitive data for training and testing. This is generally computationally expensive given the iterations but this research uses a small dataset which most of water quality datasets are which makes cross validation more suited for this problem. This research splits data into k=6 subsets and runs cross validation.

3.3. Machine Learning Algorithms

We have used both regression and classification algorithms. Regression algorithms to estimate water quality index and classification algorithms to classify samples into previously defined water quality classes. We have used 8 regression algorithms and 10 classification algorithms. To measure accuracy of regression algorithms we have employed Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean absolute Error (MAE) and R squared error (RSE). To measure accuracy of classification algorithms we have used Precision, Recall, F1 score and accuracy. Following are the listed algorithms that we have employed in our study:

- **Linear Regression:** Multiple linear regression is a form of linear regression used when there is more than one predicting variables at play. When there are multiple input

variables we use multiple linear regression to assess the input of each variable that affects the output as reflected in the following equation (Amral et al. 2007).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \in$$

- **Polynomial Regression:** Polynomial regression is used when the relation between input and output variables is not linear and a little complex. We use higher order of variables to capture the relation of input and output variables which is not as linear. But using high order of variables carries the risk of overfitting (Ostertagovaa. 2012).

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_k x_i^k + \in_i , for\ i = 1,2, \dots , n$$

- **Random Forest:** Random forest is a model which uses multiple base models on subsets of given data and make decisions based on all the models. In Random Forest base models used are decision trees and carry all the pros of decision trees with additional efficiency of using multiple models (Liaw & Wiener. 2002).
- **Gradient Boosting Algorithm:** It is the most contemporary algorithm used in most of competitions. It uses an additive model which allow for optimization of differentiable loss function (Friedman. 2002).
- **Support Vector Machines:** Support Vector Machines are mostly used for classification but they can be used for regression as well. Suppose data points are plotted on a plane, SVMs define a hyperplane between the classes and extend the margin as to maximize the distinction between two classes which results in lesser close miscalculations (Tong. & Koller. 2001).
- **Ridge Regression:** Ridge regression works on the same principles as linear regression it just adds a certain bias to negate the effect of large variances and to void the requirement of unbiased estimators. It just penalizes the coefficients that are way off from zero. It minimizes the sum of squared residuals (Hoerl & Kennard. 1970; Zhang et al. 2015).
- **Lasso Regression:** Lasso regression works on the same principles as the ridge regression, the only difference is how they penalize their coefficients being off. Lasso penalizes the sum of absolute errors instead of sum of squared coefficients (TIBSHIRANI. 1994).
- **Elastic Net Regression:** Elastic net regression combines the best of both ridge and lasso regression. It combines the method of penalties of both methods and minimizes loss function (Zou. & Hastie. 2005).
- **Neural Net:** Neural nets are loosely based on structure of neurons. It contains multiple layers with interconnected nodes. It contains an input layer and output layer and hidden

layers between those two mandatory layers. Input layer takes in the predicting parameters and output layer shows the prediction based on the input. It iterates through each of training data point and generalizes the model by giving and updating weight on each node of each layer. The trained model then uses those weights to decide what units to activate based on the input. Neural Net is mostly used for classification but it can be used for regression as well (Gunther. & Fritsch. 2010).

- **Gaussian Naïve Bayes:** Naïve Bayes is a simple and a fast algorithm which works on the principle of Bayes theorem with the assumption that the probability of the presence of one feature is unrelated to the probability of presence of the other feature (Zhang. 2004).
- **Logistic Regression:** Logistic regression is a classification algorithm. It is based on the logistic function or the sigmoid function hence the name. It is mostly the go to algorithm in case of binary classification but in our case we use multinomial logistic regression due to Y having more than two classes (Hosmer. et al 2013).
- **Stochastic Gradient Descent:** It is an iterative optimization algorithm that minimizes the loss function iteratively as to find the global optimum. In stochastic gradient descent the sample selection is random (Bottou. 2010).
- **K nearest neighbor:** K nearest neighbor algorithm classifies by finding the given points nearest N neighbors and assigns the class of majority of n neighbors to it. In case of draw one could employ different techniques to resolve it e.g. increase n or add bias towards one class etc. K nearest neighbors is not recommended for large data as all the processing takes place while testing and it iterates through the whole training data and computes nearest neighbors each time (Beyer et al. 1999).
- **Decision Trees:** A decision tree is a simple self-explanatory algorithm which can be used for both classification and regression. Decision tree after training makes decisions based on values of all the input parameters. It used entropy to select the root variable on basis of which it looks towards the other parameters' values. It has all the parameter decisions arranged in a top to down tree and projects the decision based on different values of different parameters (Ouinlan. 1990).
- **Bagging Classifier:** Bagging classifier fits multiple base classifiers on random subsets of data and then average out their predictions to form out final prediction. It highly helps out with the variance (Breiman. 1996).

3.4. Accuracy Measures

Since we have used two class of algorithms, Regression and Classification. There are different accuracy measures for regression and classification.

For regression we have used the following accuracy measures:

Mean Absolute Error:

Mean Absolute Error is a rather simple measure of accuracy. It sums up absolute values of errors and divides it by total number of values. It gives equal weight to each error value (Willmott & Matsuura 2005).

$$\frac{\sum(\|x_{obs} - x_{pred}\|)}{n}$$

Mean Square Error:

Mean Square Error is the sum of squares of errors divided by total number of predicted values. This attributes greater weight to larger errors, this is particularly useful in the problems when there needs to be a larger weight for larger errors (Willmott & Matsuura 2005).

$$\frac{\sum(x_{obs} - x_{pred})^2}{n}$$

Root Mean Square Error:

Root Mean Squared Error is just the square root of Mean Square Error and just scales the values of MSE near to the ranges of observed values (Willmott & Matsuura 2005).

$$\sqrt{\frac{\sum(x_{obs} - x_{pred})^2}{n}}$$

R^2 Error:

R squared error, also known as coefficient of determination, determines the goodness of fit of the model. It particularly explains the amount variance of dependent variable that is explainable through independent variable. It ranges between 0 and 100. Higher values entail that independent variables highly explain the variance of the dependent variable (Menard 2000).

$$R^2 = 1 - \frac{Explained\ variation}{Total\ variation}$$

For classification we have used the following accuracy measures:

Accuracy:

Accuracy is the correct number of predictions made by the model over all the observed values (Goutte & Gaussier 2005; Sokolova, Japkowicz & Szpakowicz 2006; Shafi et al 2018).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision:

It reflects the measure of correctly classified instances of a particular positive class out of total classified instances of that class (Goutte & Gaussier 2005; Sokolova, Japkowicz & Szpakowicz 2006; Shafi et al 2018).

$$Precision = \frac{TP}{TP + FP}$$

Recall:

Recall is the proportion of instances of a particular positive class that were actually classified correctly (Goutte & Gaussier 2005; Sokolova, Japkowicz & Szpakowicz 2006; Shafi et al 2018).

$$Recall = \frac{TP}{TP + FN}$$

F1 Score:

Since precision and recall, individually, don't cover all the aspects of the accuracy so we take their harmonic mean to reflect F1 score which covers both aspects and reflects better overall accuracy measure. It ranges between 0 and 1. The higher the score the better the accuracy (Goutte & Gaussier 2005; Sokolova, Japkowicz & Szpakowicz 2006).

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{Precision + Recall}$$

In previous sections we performed data description, data analysis and outlier detection to eradicate irregularities in the data. Following that, we performed parametric analysis to learn data trends, then we performed data normalization to bring all parameters on the same scale. After all the preprocessing, we applied various machine learning algorithms and explored accuracy measures that go with them as reflected in fig 1. In next section explore into our proposed IoT system which is built on top of aforementioned machine learning module and following that, we conclusively explore into the evaluation of the applied algorithms and the system.

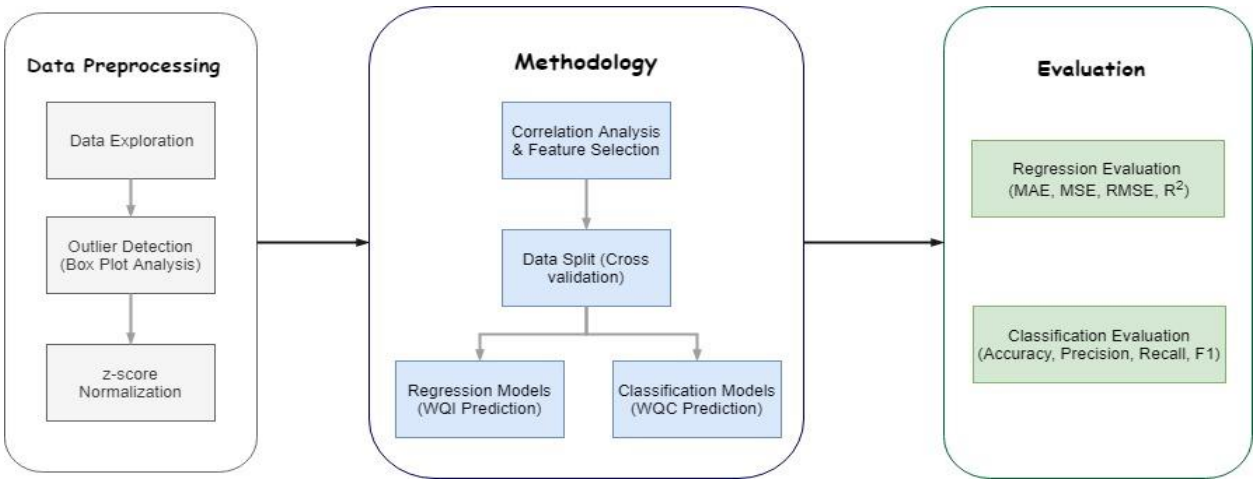


Fig 1: Methodology Flow

3.5. IoT System

Real time quality detection is preceded by live parameter monitoring. One needs the utility of monitoring water quality parameters instantly and for that we have employed IoT sensors and hardware and built a prototype system. Our prototype IoT system proposed to employ 3 parameter sensors namely, pH, Turbidity and Temperature to monitor and predict water quality as according to methodologies in the trailing sections but we eventually could use only two of those sensors, namely pH and Turbidity. Our proposed system has 4 modules, namely sensing module, actuator module, data analytics module and application module.

1. Sensing Module:

Sensing module was proposed to contain several sensors to be integrated in the system but due to lack of funds we could integrate only two of the parameter sensors namely pH and turbidity. These sensors are integrated with the Arduino board. For connectivity purposes we have also integrated WIFI shield to communicate readings. These parameter sensors read the parameter readings and then are transmitted after every 30 seconds to the cloud for data collection using Arduino board and Arduino WIFI shield. The outlook of sensing module is reflected in fig 2.

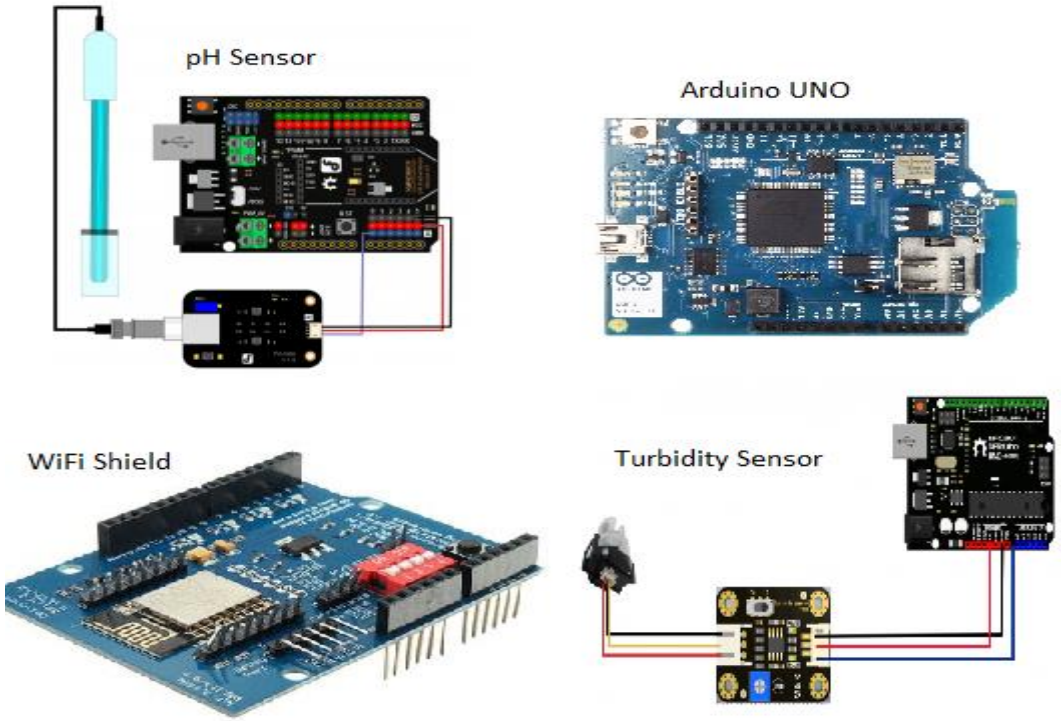


Fig 2: Sensor and Coordinator Module

2. Actuator Module:

The actuator module is responsible for managing the water flow. It does so, using a solenoid valve which makes use of magnetic field to direct water flow. It also has an attached LCD display which reflects parameter readings in real time. These readings are dispatched to the cloud at an iteration of each 30 seconds to be accessed by the analytics module for estimation and also, retrieved by an android app used by the user for real time monitoring.

3. Machine Learning Module:

This particularly essential module bears responsibility for learning trends out of data & predicting general fitness of the sample using available parameters. Moreover, the proposed system consists data analytics part to access real time data i.e. PH and turbidity of water in that particular instant from cloud and inform about its overall water quality. Its estimations are reflected in the Application Module below. Also, this module and its results will be explored further in the trailing sections.

4. Application Module:

Application Module is responsible for the visualization of the data on an android application. Once data is received and stored, Android application accesses the data from cloud through the service and shows visualizations. Application generates an alert if anything seems to be gravely out of limit to intimate an informed measure for decision making. It reflects the parametric information, geo location information and an informed water quality index whether the water sample is fit for use or not which is reflected in fig 3.

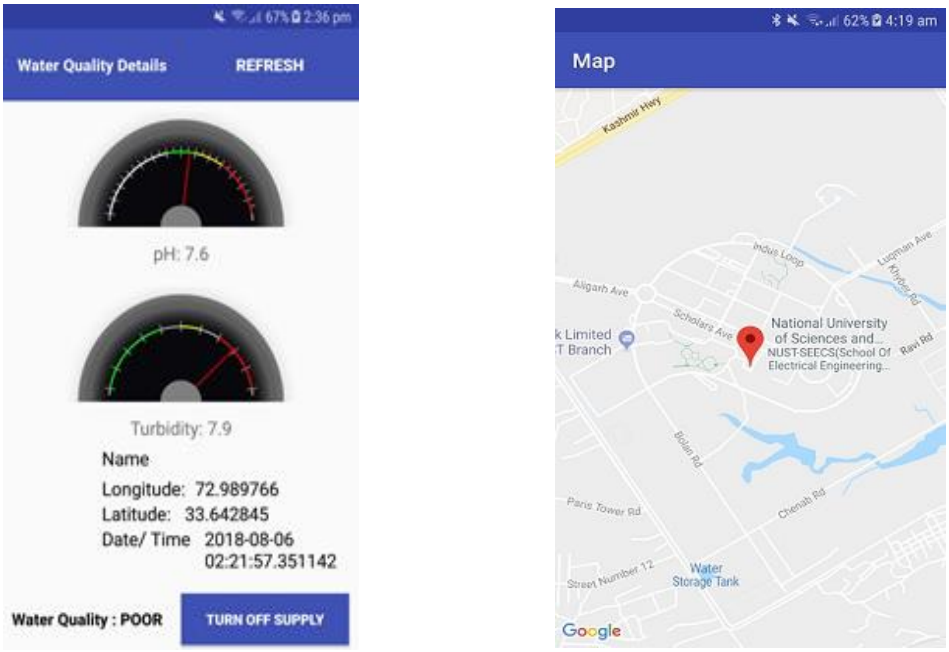


Fig 3: Android application for real time monitoring

The system as a whole is equipped to help user in surveilling water quality of a certain site instantaneously and be informed about the quality of water in any particular moment and its architecture is reflected in fig 4.

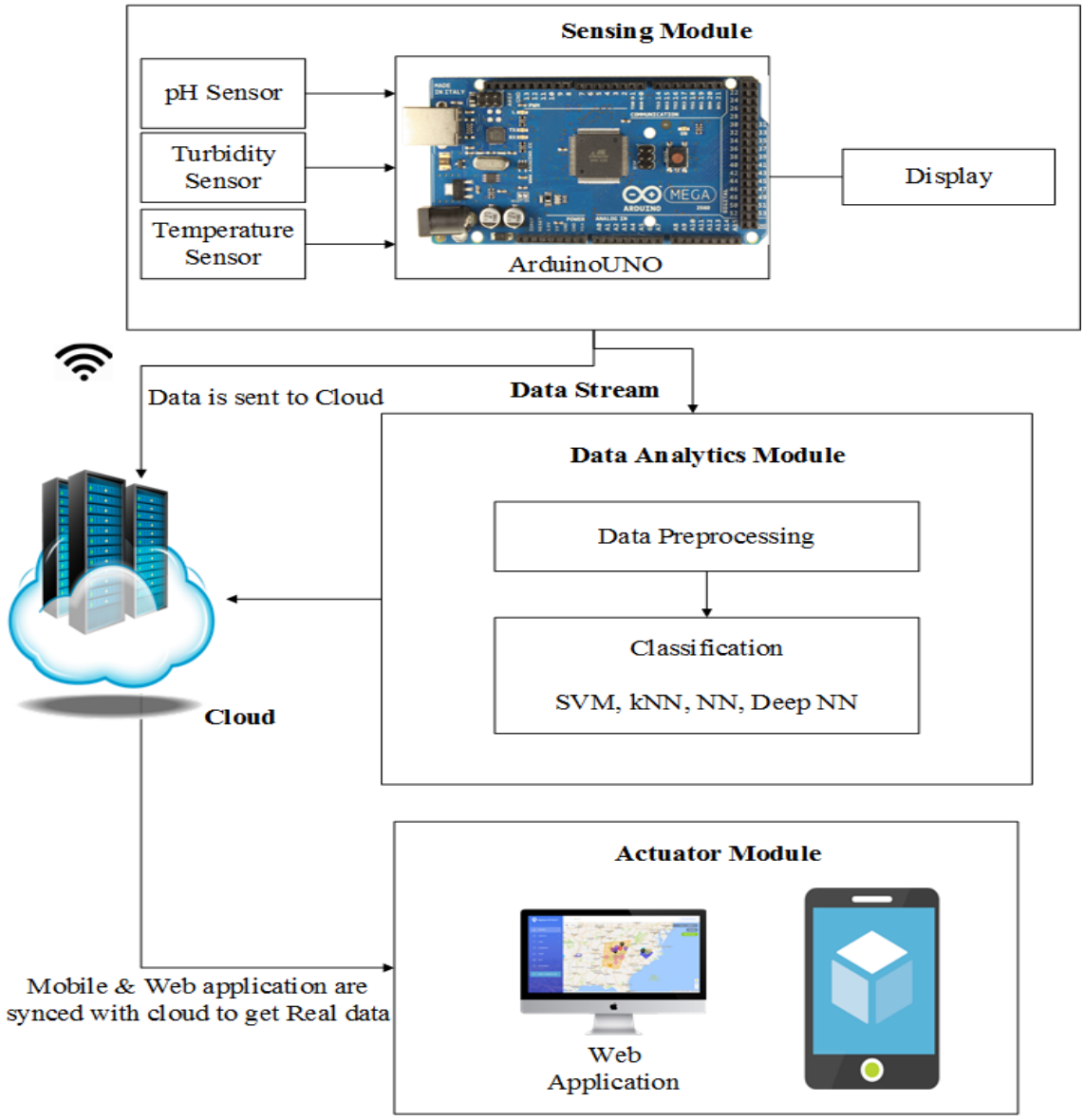


Fig 4: Synopsis of the Water Quality System

Our proposed solution, although employing minimal parameter sensors, is tangible enough to be used at any site with maximal scalability after minimal modifications. The snapshot of the system hardware is reflected in fig 5.

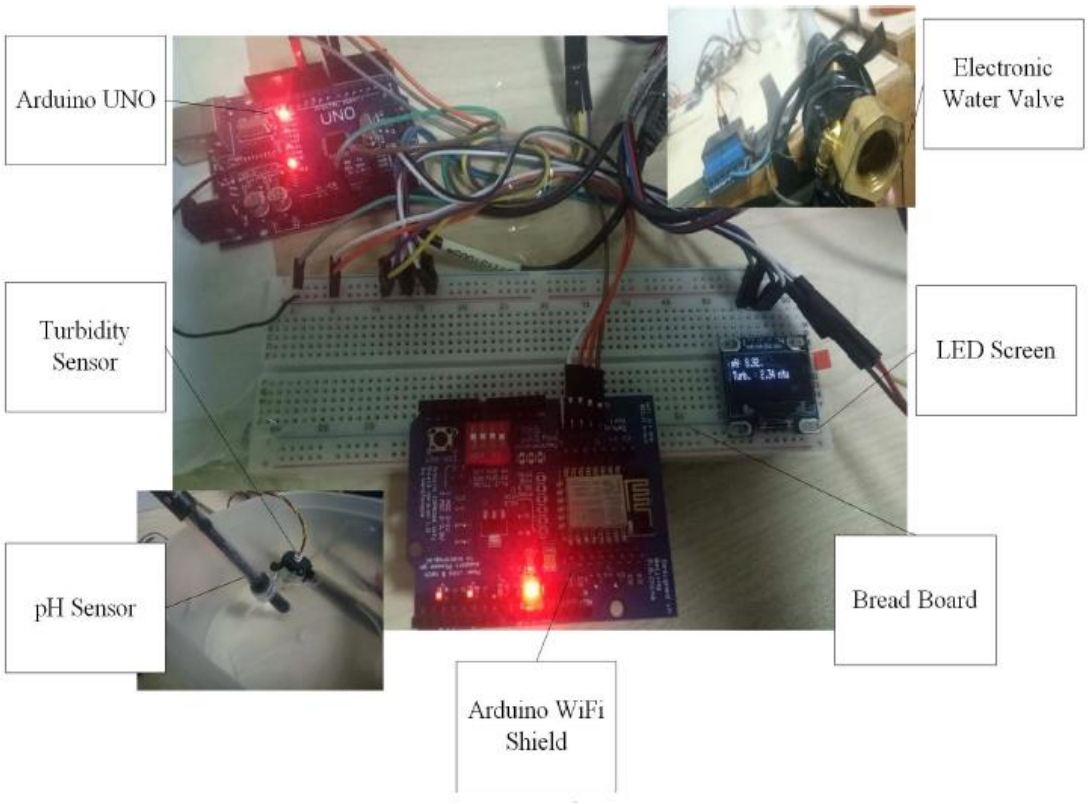


Fig 5: Hardware of Water Quality System

We collected several samples from all around Islamabad and Rawalpindi and labelled them using WHO standards as reflected in some of the samples in the table below. If it lied in between the standards it was good else it was labelled poor. The collected data validated our prototype of being able to be incorporated with machine learning component to make up Water quality detection system. Although we could not estimate water quality index given the system only employed two parameter sensors but proved to be a capable prototype to monitor water quality and predict the general goodness of the water sample.

Sr #	Long	Lat	Place	pH	Turbidity	Water Quality
S1	33.577	73.039	Harley Street, RWP	7.2	2.30	Good
S2	33.642	72.979	Mohalla Riazbad, RWP	7.1	2.40	Good
S3	33.573	73.044	Tahli Mohri, RWP	7.0	2.30	Good
S4	33.594	73.054	Saddar Metro, RWP	7.9	2.32	Good
S5	33.609	73.009	Chor chok, RWP	8.0	2.31	Good
S6	33.536	73.052	Gulshanabd, RWP	7.2	1.18	Good
S7	33.599	73.100	Chaklala, RWP	8.3	1.19	Good
S8	33.653	73.030	Ravigroup I/10-3 ISB	8.1	1.22	Good
S9	33.643	73.988	SEECS, NUST, ISB	7.6	2.24	Good
S10	33.649	73.026	I/10-3 ISB	7.0	2.32	Good

Following the methodology section, we evaluate our applied methodologies in the next section to quantify the accuracy of our methods and to validate this research

4. Results

4.1. Regression Algorithms

Since water quality parameter sensors are expensive this study aims to use minimal number of parameters with cheap sensors to predict water quality. Initially we used four parameters namely, temperature, turbidity, pH and total dissolved solids. While employing following regression algorithms we found Gradient boosting to be the most efficient of algorithms having mean absolute error of 1.93.

Algorithm	MSE	RMSE	MAE	R Squared
Linear Regression	11.1958	3.3460	2.5776	0.6678
Polynomial Regression	7.1408	2.6722	1.9070	0.7373
Random Forest	8.9845	2.9974	2.2331	0.6901
Gradient Boosting	6.7040	2.5892	1.9309	0.7549
SVM	10.0694	3.1732	2.3831	0.3821
Ridge Regression	11.1912	3.3453	2.5784	0.5189
Lasso Regression	19.4505	4.4103	3.5189	-2.7914
Elastic Net Regression	20.2484	4.4998	3.5884	-3.7857

Following that we tried to reduce more parameters and dropped total dissolved solids whose sensor is a little expensive than others. We found that Gradient boosting was still the most efficient and there was decrease in the overall error rate but the decrease was not alarming and still performed well within limits given the cost.

Algorithm	MSE	RMSE	MAE	R Squared
Linear Regression	15.0100	3.8743	3.0691	0.5546
Polynomial Regression	11.7935	3.4342	2.6280	0.5153
Random Forest	14.4142	3.7966	2.9329	0.4240
Gradient Boosting	12.6556	3.5575	2.7421	0.5029
SVM	13.1365	3.6244	2.7583	0.1930
Ridge Regression	15.0105	3.8743	3.0701	0.2486
Lasso Regression	22.0522	4.6960	3.7980	-3.3992
Elastic Net Regression	23.1448	4.8109	3.8887	-5.1591

4.2. Classification Algorithms

The same parameters were used for classification as well. Initially the same four parameters were used. We found that Logistic Regression performed better than other algorithms with accuracy of 84.01%.

Algorithm	Accuracy	F1 score	Precision	Recall
MLP	0.8326	0.5599	0.5577	0.5634
Guassian Naïve Bayes	0.7783	0.5058	0.5041	0.5479
Logistic Regression	0.8401	0.5588	0.5585	0.5593
Stochastic Gradient Descent	0.8115	0.5409	0.5443	0.5407
KNN	0.6787	0.4490	0.4489	0.4494

Decision Tree	0.7798	0.5185	0.5190	0.5179
Random Forest	0.7888	0.5262	0.5317	0.5290
SVM	0.7753	0.5145	0.5140	0.5154
Gradient Boosting Classifier	0.8009	0.5332	0.5347	0.5323
Bagging Classifier	0.7662	0.5103	0.5126	0.5096

Following that we applied the same algorithms using the 3 parameters. There was again a drop in accuracy but Logistic Regression still performed better than other algorithms with an accuracy of 77.98%

Algorithm	**Accuracy**	**F1 score**	**Precision**	**Recall**
MLP	0.7466	0.4961	0.4966	0.4956
Guassian Naïve Bayes	0.7164	0.4514	0.4574	0.5205
Logistic Regression	0.7798	0.5165	0.5153	0.5199
Stochastic Gradient Descent	0.7753	0.5126	0.5112	0.5176
KNN	0.6365	0.4201	0.4201	0.4205
Random Forest	0.7511	0.5000	0.5016	0.4992
SVM	0.7406	0.4888	0.4876	0.4937
Gradient Boosting Classifier	0.7270	0.4836	0.4849	0.4828
Bagging Classifier	0.7044	0.4698	0.4725	0.4705

Apart from extensive wqi prediction we also predicted water quality class of samples gathered and tested through our IoT system. Class was assigned to good or bad depending upon the WHO standards of PH and turbidity of the samples and classified

using several classification algorithms namely SVM, NN, Deep NN and kNN and the results were validated using accuracy, precision and recall and reflected in the table below.

Algorithm	Accuracy	Precision	Recall
SVM	0.91	0.93	0.90
NN	0.86	0.87	0.85
Deep NN	0.93	0.94	0.93
kNN	0.76	0.79	0.76

As evident in the results, performance of Deep NN stood out while classifying quality of water to be good or bad and the other algorithm that came close was SVM.

In this chapter, we iterated through our study's results and established that Gradient Boosting Classifier performed better in each of the case with the highest accuracy while predicting wqi and wqc while Deep NN performed better while predicting whether a particular sample was good or bad.

5. Conclusion & Future Work

Water is one of the most essential resources for survival and its quality is determined by water quality index(WQI) which is measured through various water quality parameters depending upon the type of standard used. Conventionally, to measure water quality parameters one has to go through expensive and time-consuming lab analysis which makes timely recognition of contamination and action for it difficult. In this age of technology, we could, alternatively, employ IoT systems to monitor water in real time. A number of such systems such as CANARY are deployed at various places using IoT effectively and they prove to be an effective alternative to expensive manual lab analysis. While IoT systems are employed for real time water quality monitoring, machine learning methodologies such as artificial neural networks, support vector machines, regression, correlation analysis, hierarchical clustering etc aid to learn trends of the water quality parameters, to predict WQI and detect anomalous events like intentional contamination to enable real time contamination detection and action.

This research particularly focused on predicting water quality using minimal water quality parameters given the price of sensors. This research compared several regression algorithms to estimate water quality index and several classification algorithms to predict water quality class. Gradient Boosting regression outperformed other regression algorithms in predicting WQI while Neural Networks outperformed other classification algorithms in predicting WQC. This study also proposes a large-scale system for commercial use that incorporates the findings of this research.

Since local research & products of water quality monitoring and prediction are near to nonexistent, this research also proposes a system to monitor water quality in real time and learn and predict water quality, trends of water quality and recognize anomalous events. The proposed system consists of 5 modules namely, sensor module, coordinator module, cloud & services module, application module and machine learning module using one of the above methodologies.

The proposed system, at present, is feasible for a single site but could easily be expanded for a large multiple site system by altering the network architecture a bit. Apart from the scalability of large scale IoT system we plan to explore enhanced water quality index which uses the geographical location (latitudes and longitudes) and time of the readings as well. Since quality of water varies in different locations and on different times owing to environmental factors of the location particularly air quality, which has a high

impact on water quality. We plan to explore into an enhanced wqi considering all of the aforementioned factors and somehow integrate them in the calculation of wqi.

Bibliography:

Aamir Alamgir., Moazzam Ali Khan. & Omm-e-Hany. 2015 Public Health Quality of Drinking Water supply in Orangi Town, Karachi, Pakistan. *Bulletin of Environment, Pharmacology and Life Sciences,* 4(11), 88-94.

A.K. Thukral., Renu Bhardwaj. & Rupinder Kaur. 2005 “Water Quality Indices, Statistical Accounting of Water Resources”

Ali Najah., Ahmed Elshafie., Othman A. Karim. & Othman Jaffar. 2009 Prediction of Johor River Water Quality Parameters Using Artificial Neural Networks. *European Journal of Scientific Research*. 28(3), 422-435.

Andy Liaw. & Wiener. 2002 Classification and Regression by randomForest.

Arthur E. Hoerl & Robert W. Kennard. 1970 Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 12(1).

Asit Kumar Batabyal. & Surajit Chakraborty. 2015 Hydrogeochemistry and Water Quality Index in the Assessment of Groundwater Quality for Drinking Uses. *Water Environment Research.* 87(7), 607-617.

A. K. Verma. & T. N. Singh. 2012 Prediction of water quality from simple field parameters. *Environmental Earth Sciences*. 69(3), 821–829.

Bureau of Indian Standards *1991. Indian Standard Drinking WaterSpecification. 1st rev. Bureau of Indian Standards: New Dehli, India*.Brandon P. Wong. & Branko Kerkez. 2016 Real-time environmental sensor data: An application to water quality using web services. *Environmental Modelling & Software* 84, 505-517.

Brandon P. Wong. & Branko Kerkez. 2016 “Real-time environmental sensor data: An application to water quality using web services”

Canary Event Detection Software 2010, Sandia National Laboratories. https://www.sandia.gov/research/research_development_100_awards/_assets/documents/2010_winners/SNL_Canary_SAND2010-2228P.pdf (accessed 14 January 2019)

Cesar Encinas., Erica Ruiz., Joaquin Cortez. & Adolfo Espinoza. 2017 Design and implementation of a distributed IoT system for the monitoring of water quality in aquaculture. *Wireless Telecommunications Symposium (WTS).*

Cort J. Willmott. & Kenji Matsuura. 2005 “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”

Cyril Goutte. & Eric Gaussier. 2005 “A Probabilistic Interpretation of Precision, Recall and *F*-Score, with Implication for Evaluation”

Dian Zhang., Timothy Sullivan., Ciprian Briciu-Burghina., Kevin Murphy., Kevin McGuinness., Noel E. O’Connor., Alan Smeaton. & Fiona Regan. 2014 Detection and Classification of Anomalous Events in Water Quality Datasets Within a Smart City-Smart Bay Project. *International Journal on Advances in Intelligent Systems*. 7 (1&2), 167-178.

Environmental Protection Agency. 2001 “Parameters of Water Quality, Interpretation and Standards”. https://www.epa.ie/pubs/advice/water/quality/Water_Quality.pdf (accessed 19 November 2018)

Environmental Protection Agency. 2013 Water Quality Event Detection System Challenge: Methodology and Findings. https://www.epa.gov/sites/production/files/2015-07/documents/water_quality_event_detection_system_challenge_methodology_and_findings.pdf (accessed 19 November 2018)

E R. Rene. & M B Saidutta. 2008 Prediction of Water Quality Indices by Regression Analysis and Artificial Neural Networks. *International Journal of Environmental Research***.** 2(2), 183-188**.**

Eva Ostertagovaa. 2012 Modelling using polynomial regression. *Procedia Engineering*. 48, 500-506.

Feifei Cao., Feng Jiang., Ziqinq Liu. & Zhu Yang. 2014 Application of ISFET Microsensors with Mobile Network to Build IoT for Water Environment Monitoring. *International Conference on Intelligent Environments*

Frauke Gunther. & Stefan Fritsch. 2010 neuralnet: Training of Neural Networks. *The R Journal*. 2(1), 30-38.

Garima Srivastava. & Dr. Pradeep Kumar 2013 "Water Quality Index with Missing Parameters"

Hamid Zare Abyaneh. 2014 "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *JOURNAL OF ENVIRONMENTAL HEALTH SCIENCE & ENGINEERING*. 12(1), 40.

Han Yan., Zhihong Zou. & Huiwen Wang. 2010 Adaptive neuro fuzzy inference system for classification of water quality status. *Journal of Environmental Sciences*. 22(12), 1891-1896.

Harry Zhang. 2004 The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA.*

Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied logistic regression* (398). John Wiley & Sons.

Hui Zou. & Trevor Hastie. 2005 Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 67(2), 301-320

Ihsan Omur Bucak. & Bekir Karlik. 2011 Detection of Drinking Water Quality Using CMAC Based Artificial Neural Networks. *Ekoloji Dergisi*. 20(78), 75-81.

Jerome H. Friedman. 2002 Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 38, 367–378.

J. R. Ouinlan. 1990. Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*. 20(2), 339-346.

Kevin Beyer., Jonathan Goldstein., Raghu Ramakrishnan. and Uri Shaft. 1999. When is “nearest neighbor” meaningful?. *International conference on database theory* (217-235).

K.Raghu Sita Rama Raju. & G.Harish kumar Varma. 2017 Knowledge Based Real Time Monitoring System for Aquaculture Using IoT. *IEEE 7th International Advance Computing Conference (IACC).*

Léon Bottou. 2010 Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*. 177-186.

L. Breiman. 1996 Bagging predictors. *Machine Learning*. 24(2), 123-140.

Maqbool Ali & Ali Mustafa Qamar 2013 Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed in Pakistan. *Eighth International Conference on Digital Information Management (ICDIM 2013).*

Marina Sokolova., Nathalie Japkowicz. & Stan Szpakowicz. 2006 “Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation”

Mohamad Sakizadeh. 2016 Artificial intelligence for the prediction of water quality index in ground water systems. *Modeling Earth Systems and Environment*, 2(1), 8.

M. K. Daud., Muhammad Nafees., Shafaqat Ali., Muhammad Rizwan., Raees Ahmad Bajwa., Muhammad Bilal Shakoor., Muhammad Umair Arshad., Shahzad Ali Shahid Chatha., Farah Deeba., Waheed Murad., Ijaz Malook. & Shui Jin Zhu. 2017 Drinking Water Quality Status and Contamination in Pakistan. *BioMed Research International*. 2017.

Nabeel M. Gazzaz., Mohd Kamil Yusoff., Ahmad Zaharin Aris., Hafizan Juahir. & Mohammad Firuz. 2012 Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine Pollution Bulletin*. 64(11), 2409-2420.

Naeem Ejaz., Hashim Nisar Hashmi. & Abdur Razzaq Ghuman. 2010 Water Quality Assessment of Effluent Receiving Streams in Pakistan: A case of River Ravi. *Mehran University Research Journal of Engineering & Technology*. 30(3).

Niel Andre Cloete., Reza Malekian., & Lakshmi Nair. 2016 Design of Smart Sensors for Real-Time Water Quality Monitoring. *IEEE Access*. 4.

N Vijayakumar. & R Ramya. 2015 The Real Time Monitoring of Water Quality in IoT Environment. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).*

N. Amral., C.S. Ozveren. & D King. 2007 Short Term Load Forecasting using Multiple Linear Regression. *2007 42nd International Universities Power Engineering Conference.*

Praveen Vijaia. & Bagavathi Sivakumar P. 2016 Design of IoT Systems and Analytics in the context of Smart City initiatives in India. *Procedia Computer Science*. 92(2016), 583–588.

ROBERT TIBSHIRANI. 1994 Regression Shrinkage and Selection via the Lasso*. J. R. Statist. Soc. B*. 58(1), 267-288

Scott Menard. 2000 “Coefficients of Determination for Multiple Logistic Regression Analysis”

Simon Tong. & Daphne Koller. 2001 Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research (2001).* 45-66.

Swati V. Birje., Trupti Bedkyale., Chaitali Alwe. & Vivek Adiwarekar. 2016 Water pollution detection system using pH and turbidity sensors. *International Journal of Advanced Research in Computer and Communication Engineering*. 5(4), 530-533.

S.S. Mahapatra., Santosh Kumar Nanda. & B.K. Panigrahy. 2011 A Cascaded Fuzzy Inference System for Indian river water quality prediction. *Advances in Engineering Software.* 42(10), 787-796.

S. Geetha. & S. Gouthami. 2017, Internet of things enabled real time water quality monitoring system. *Smart Water*. 2(1).

T.Jayalakshmi. & Dr.A.Santhakumaran. 2011 “Statistical Normalization and Back Propagation for Classification”

Thinagaran Perumal., Md Nasir Sulaiman. & Leong.C.Y. 2015, Internet of Things (IoT) Enabled Water Monitoring System. *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE).*

Uferah Shafi., Rafia Mumtaz., Hirra Anwar., Ali Mustafa Qamar. & Hamza Khurshid. 2018 Surface Water Pollution Detection using Internet of Things*. 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT).*

Vesna Rankovic., Jasna Radulovic., Ivana Radojevic., Aleksandar Ostojic. & Ljiljana Comi. 2010 Neural network modeling of dissolved oxygen in the Gruza reservoir, Serbia. *Ecological Modelling*. 221(8), 1239-1244.

World Health Organization 1993. *Guideline for Drinking Water Quality, 2nd ed., vol 1. World Health Organization: Geneva, Switzerland.*

Yuchen Zhang., John Duchi. & Martin Wainwright. 2015 Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates. *Journal of Machine Learning Research*. 16(2015), 3299-3340

Zhu Wang., Qi Wang. & Xiaoqiang Hao. 2009 The Design of the Remote Water Quality Monitoring System based on WSN. *2009 5th International Conference on Wireless Communications, Networking and Mobile Computing.*

Zulhani Rasin. & Mohd Rizal Abdullah. 2009 Water Quality Monitoring System Using Zigbee Based Wireless Sensor Network. *International Journal of Engineering & Technology IJET.* 9(10).