Water quality monitoring: from conventional to emerging technologies

Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Sadaf Mumtaz and Ali Mustafa Qamar

ABSTRACT

The rapid urbanization and industrial development have resulted in water contamination and water quality deterioration at an alarming rate, deeming its quick, inexpensive and accurate detection imperative. Conventional methods to measure water quality are lengthy, expensive and inefficient, including the manual analysis process carried out in a laboratory. The research work in this paper focuses on the problem from various perspectives, including the traditional methods of determining water quality to gain insight into the problem and the analysis of state-of-the-art technologies, including Internet of Things (IoT) and machine learning techniques to address water quality. After analyzing the currently available solutions, this paper proposes an IoT-based low-cost system employing machine learning techniques to monitor water quality in real time, analyze water quality trends and detect anomalous events such as intentional contamination of water.

Key words | artificial intelligence (AI) techniques, internet of things, machine learning, real-time monitoring, smart city, water quality

Umair Ahmed
Rafia Mumtaz (corresponding author)
Hirra Anwar
Ali Mustafa Qamar
School of Electrical Engineering and Computer Science (SEECS),
National University of Sciences and Technology (NUST),
Islamabad,
Pakistan
E-mail: *rafia.mumtaz@seecs.edu.pk*

Sadaf Mumtaz
HITEC-Institute of Medical Sciences,
National University of Medical Sciences,
Rawalpindi,
Pakistan

INTRODUCTION

Water plays an important role in all aspects of our lives and its quality is deteriorating with ever-increasing pollution due to urbanization, industrialization and population growth. To sustain quality of life, it is imperative to detect water pollutants causing contamination. Typically, the detection of water quality is time-consuming, and a cumbersome task requiring manual laboratory analysis and statistical inferences (Gazzaz *et al.* 2012). There are several systems developed around the world that monitor and detect water pollution in real time. However, such research in Pakistan is limited. Although Pakistan has limited laboratory facilities through which more than 200 water quality parameters can be analyzed, laboratory analysis itself is time consuming and does not offer real-time detection of deteriorating water quality. In addition to this, Pakistan has no web portal for data visualization and for public viewing, nor has there been any extensive local research in this direction. These are the main factors that are the primary motivation behind this research to increase the national impact.

The quality of water is affected by several parameters that are biological, chemical and physical in nature. There is no single parameter that defines water quality completely, due to which the Water Quality Index (WQI) was developed to measure water quality. The computation of WQI is a lengthy process, which is why there is a need for an alternative method to simplify the WQI calculation process. Additionally, there are certain water quality parameters that are more expensive to attain than others. As of today, Internet of Things (IoT) and machine learning are two promising technologies that can be employed as an alternative to solve the aforementioned water quality problems (Gazzaz *et al.* 2012).

doi: 10.2166/ws.2019.144

In view of the above, our research is directed towards analyzing different methods to estimate and monitor water quality using IoT and machine learning. As indicated before, the quality of water is determined by several parameters, but what solely defines water quality is WQI. Different countries have different methods for calculating WQI, but all of them are computationally expensive (Gazzaz *et al.* 2012). Towards such ends, we propose an IoT-based system that can monitor water quality parameters in real time, identify quality trends and predict water quality using machine learning methodologies.

This paper is organized into five sections as follows: the first section defines common water quality parameters and their role in determining the status of water quality. The second section discusses existing systems across the world with a comparative analysis. The third section highlights research conducted in the domain of water quality using manual laboratory analysis to gain more insight into the problem, machine learning algorithms employed in the domain, and IoT systems employed for water quality monitoring. The fourth section outlines our proposed system based on IoT and machine learning to provide real-time water quality monitoring. The last section concludes the paper along with giving some future directions.

Water quality parameters

WQI is measured through different water quality parameters. The commonly used parameters (EPA 2001; Verma & Singh 2012) are briefly discussed below and their relations are defined in Table 1:

- **pH:** pH of water specifies how acidic or alkaline the water is. The acidic range lies between 0 and 6 while the alkaline range lies between 8 and 14. 6.5–8.5 is the most acceptable range of pH. It is measured through electrometry and pH electrodes. It is significantly correlated with electrical conductivity, total hardness, sulphates and total suspended solids (EPA 2001; Bhandari & Nayal 2008; Verma & Singh 2012; Ali & Qamar 2013; Patel & Vaghani 2015).
- **Turbidity:** Turbidity of water is the measurement of non-filterable, divided solids in the water. This may also interfere with the treatment of water. It is mostly measured in Nephelometric Turbidity Units (NTUs). It is measured through a nephelometer or turbidimeter. It is significantly correlated with total hardness, electrical conductivity, sulphates, total dissolved solids and chemical oxygen demand (EPA 2001; Bhandari & Nayal 2008; Verma & Singh 2012; Patel & Vaghani 2015).
- **Temperature:** Temperature is one of the most important parameters, which has a considerable amount of effect on aquatic life. It also affects gas transfer rates and the amount of dissolved oxygen. It may alter the form of some of the elements or their concentration. It is mostly measured in Celsius. Its measurement is carried through a thermistor or thermometry in the field. It is highly correlated with electrical conductivity and loosely correlated with pH (EPA 2001; Verma & Singh 2012; Ali & Qamar 2013; Khatoon *et al.* 2013).
- **Chloride (Cl):** It is naturally present in water and while its excess is not harmful to humans normally, but the water's taste grows towards the saltier range if it increases to more than 250 mg/l and may be harmful to agricultural activities. It is mostly measured through titration and measured in mg/l. It is highly correlated with total hardness, electrical conductivity, total dissolved solids, biological oxygen demand and chemical oxygen demand (EPA 2001; Bhandari & Nayal 2008; Khatoon *et al.* 2013; Patel & Vaghani 2015).
- **Electrical conductivity (EC):** This indicates the water's potential to conduct electric current. It is not directly useful in terms of water quality. Nevertheless, it helps more in terms of water's ionic content, which in turn determines the hardness, alkalinity and some of the dissolved solids. The conductivity varies with the water source and is also correlated with pH, temperature, turbidity, chlorides, sulphates, dissolved oxygen, total dissolved solids and chemical oxygen demand. It is measured through the electrometric method (EPA 2001; Bhandari & Nayal 2008; Ali & Qamar 2013; Khatoon *et al.* 2013; Patel & Vaghani 2015).
- **Dissolved oxygen (DO):** This indicates oxygen's solubility in water. Water mostly absorbs oxygen from the atmosphere or produces it through photosynthesis. It is quite important for aquatic life. It is mostly measured through an electrometric meter or Winkler titration. It is highly correlated with electrical conductivity,

Table 1 | Water quality parameter correlation matrix (EPA 2001; Bhandari & Nayal 2008; Verma & Singh 2012; Ali & Qamar 2013; Khatoon *et al.* 2013; Patel & Vaghani 2015)

	PH	Turb	Temp	Cl	EC	DO	TH	TSS	TDS	BOD	COD	FC	TC
PH					✓		✓	✓					
Turb					✓		✓		✓		✓		
Temp					✓								
Cl					✓		✓		✓	✓	✓		
EC	✓	✓	✓	✓		✓			✓		✓		
DO					✓					✓			
TH	✓	✓		✓					✓	✓	✓		
TSS	✓								✓				
TDS		✓		✓	✓		✓	✓			✓		
BOD				✓		✓	✓						
COD		✓		✓	✓		✓		✓				
FC													✓
TC												✓	

biological oxygen demand and sulphates (EPA 2001; Patel & Vaghani 2015).

- **Total hardness (TH):** This as an important parameter to determine water's suitability for domestic and industrial use. It is mostly the amount of concentrations of calcium and magnesium present in the water. Their concentrations in rocks lead to significant hardness levels in water. It is significantly correlated with pH, turbidity, chloride, total dissolved solids, biological oxygen demand and chemical oxygen demand. It is measured through titration with EDTA and measured in mg/l $CaCO_3$ (EPA 2001; Bhandari & Nayal 2008; Ali & Qamar 2013; Patel & Vaghani 2015).
- **Total solids (TS):** This 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 through gravimetric (dried at stated temperature) method and measured in mg/l (EPA 2001; Verma & Singh 2012).
- **Total suspended solids (TSS):** This is the amount of remains of inorganic and organic solid material suspended in the water. The increase in TSS makes the water prone to the high absorption of light which increases the water temperature and in turn decreases the water's capability to hold oxygen. It highly affects aquatic life. It is measured through gravimetric (filtration, with drying at stated temperature) method and measured in mg/l. It is significantly correlated with pH and total dissolved solids (EPA 2001; Verma & Singh 2012; Patel & Vaghani 2015).
- **Total dissolved solids (TDS):** 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 highly correlated with turbidity, chlorides, electrical conductivity, total hardness, total suspended solids and chemical oxygen demand. It is measured through gravimetric (dried at stated temperature after filtration) method and measured in mg/l (EPA 2001; Bhandari & Nayal 2008; Ali & Qamar 2013; Khatoon *et al.* 2013; Patel & Vaghani 2015).
- **Biological oxygen demand (BOD):** This is the amount of oxygen consumed by biological activities in the water, particularly protozoa and bacteria. If the BOD level is quite high and surpasses DO, then other organisms die due to a shortage of oxygen. It is quite an important factor indicating water quality and is significantly correlated with chloride, dissolved oxygen and total hardness. It is measured through the incubation technique with oxygen determinations by oxygen meter or Winkler Method and measured in mg/l (EPA 2001; Bhandari & Nayal 2008; Verma & Singh 2012; Ali & Qamar 2013; Khatoon *et al.* 2013; Patel & Vaghani 2015).
- **Chemical oxygen demand (COD):** This is the amount of oxygen consumed during breaking down of organic

material and during oxidation of present inorganic material. Just like BOD, it is also an important factor representing the status of water quality and is highly correlated with turbidity, chlorides, electrical conductivity, total hardness and total dissolved solids. It is measured through microdigestion and colorimetry and reflux distillation with acid potassium dichromate followed by titrimetric and is measured in mg/l (EPA 2001; Verma & Singh 2012; Patel & Vaghani 2015).

- **Fecal coliform (FC):** Fecal coliforms are bacteria that are found in human and animal waste and mostly originate in the intestines of warm-blooded species. They indicate possible fecal contamination of water. They are measured through the membrane filtration method and most probable number: multiple tube method and measured in number of organisms/100 ml. It is loosely correlated with ammonia and significantly correlated with total coliform (EPA 2001; Cabral & Marques 2006).
- **Total coliform (TC):** Total coliforms consist of fecal coliforms and other types of similar non-fecal bacteria mostly found in soil. The total coliforms reflect the possible presence of pathogenic micro-organisms and are correlated with fecal coliforms. They are measured through the membrane filtration method and most probable number: multiple tube method and measured in number of organisms/100 ml (EPA 2001; Cabral & Marques 2006).

Water quality detection systems

According to experts, the world has become more cautious after 9/11 about resources, particularly water, that could be intentionally polluted to stir up chaos amongst the masses. It eventually brought up the need to have real-time monitoring and contamination detection systems in place. Consequently, several systems were developed, where 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 the Environmental Protection Agency (EPA), National Homeland Security Research Center. It is currently deployed at Greater Cincinnati Water Works (GCWW). It provides several open source components, most 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 that allows the user to update its default algorithms. It is also Representational State Transfer (REST) web service friendly, allows for XML input and output. It supersedes other systems in certain major aspects including the algorithms' transparency, the capability to directly integrate operational data into its event detection component and to have centralized processing on a single computing system as well as supporting multiple sensors. Another such system is ***OptiEDS*** by Elad Salomons, which helps to detect anomalous water quality conditions in real time. It is also capable of water quality monitoring and water contamination detection in real time.

Bluebox is another system that can identify the behavior of water quality parameters that cause abnormal behavior. It produces a reliable output even if some of the parameters are missing. It initially performs normalization, calculates the points' distance amongst the parameters in each data point and plots the frequency curves of the distances to visualize. However, it is quite expensive, costing up to $92,500 and does not have the capability to directly integrate operational data into event detection. Another system, ***Event monitor***, was created by the Hach Company, which has a heuristic ability to learn events, automatically tune itself and define what constitutes an abnormality in the system. However, it is also quite expensive. **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 a dashboard to reflect real-time water quality parameters (Canary 2010; EPA 2013).

The comparison of a few water quality and contamination detection systems based on the important parameters is given in Table 2.

LITERATURE SURVEY

The literature highlights the problem of water contamination and, due to its gravity, a lot of research work has been done to find a suitable solution encompassing state of the art technology. A survey of various local and international research papers related to this problem area is summarized in Table 2. It has been observed that water

Table 2 | Comparison of water contamination detection systems currently available (Canary 2010)

	Water contamination detection systems		
Comparison parameters	**HACH, Guardian Blue**	**S::can con::stat**	**Canary**
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
Cost: event detection software (10 Stations)	$92,500	$60,000	$0.00
Cost: required computing hardware (10 Stations)	$0.00 (included)	$0.00 (included)	$3,500
Total cost	$92,500	$60,000	$3,500

quality measurement, analysis, prediction, and anomaly detection has been addressed by researchers through the techniques of manual calculation and laboratory analysis, machine learning methodologies employed to learn the trends of water quality parameters and Internet of Things-based solutions. In this regard, the following sections discuss in detail the specialized work that has been carried out in the respective areas.

Research concerning statistical inference

The research work concerning manual calculation and laboratory analysis of various water samples is highlighted in this section. Daud *et al.* (2017), in a research study, collected various water samples across all the provinces of Pakistan. Various samples that were tested for different parameters were compared against the National Environmental Quality Standards (NEQS) and World Health Organization (WHO) standards. Most of the samples had a presence of total coliform, fecal coliform, and *Escherichia coli* primarily due to the mixing of sewerage water and secondly due to industrial waste. The authors recommended installation and maintenance of treatment plants for the contaminated water and taking measures to ensure enforcement of NEQS. Alamgir *et al.* (2015) have collected 46 piped water samples across different places of Orangi town, Karachi, Pakistan and tested them for bacteriological and physio-chemical analyses using standard methods for the examination of water and wastewater. The standards against which the samples are compared included the WHO and National Standard for Drinking Water Quality (NSDWQ). The authors have calculated mean, median, minimum, maximum, standard deviation, quartile range and standard error for each of the parameters and have found physio-chemical parameters to be well in limits except sulphates. Bacteriological parameters such as total fecal coliform and total coliform counts for these samples were critically high, reflecting poor hygienic and sanitation conditions. The authors recommended continuous monitoring of water quality and revamping of sewerage systems.

Ejaz *et al.* (2010), in their study, have conducted research on the dataset of the river Ravi by sampling its data for 3 years, from January 2005 to March 2007, from 14 sampling stations. The samples have been tested for 12 different parameters including biological oxygen demand (BOD), dissolved oxygen, chemical oxygen demand (COD), suspended solids, phosphorus, chloride, sodium, total nitrogen, nitrate, nitrite, oil and grease, and total coliforms. The standard methods for the examination of water and wastewater (Andrew *et al.* 1995) have been utilized for testing the aforementioned parameters. For the comparison of their parameter reading, the National Environmental Quality Standards of Pakistan (NEQS) has been used. For the said purpose, expensive laboratory analysis has been carried out, which is a limitation of their work as well. In this research, it has been recommended to install more treatment plants and ensure enforcement of NEQS to improve the quality of water.

Batabyal & Chakraborty (2015) conducted their research in the Kanksa-Panagarh area of West Bengal. They collected samples from 98 tube wells from November to December 2011 for the post-monsoon period and from May to June 2012 for the pre-monsoon period.

They tested the samples for 13 parameters including pH, total dissolved solids (TDS), total hardness, HCO_3, Cl, SO_4, NO_3, F, Ca, Mg, Fe, Mn, and Zn against WHO (1993) and Indian (BIS 1991) standards. In the analysis, the authors have performed correlation analysis amongst the parameters to investigate the correlation. In addition to this, they have also calculated the WQI using the attained parameters, demonstrating in detail the Indian method to manually calculate the WQI.

Research concerning machine learning

This section explains the application of machine learning techniques in water quality prediction and trend analysis. Najafzadeh & Ghaemi (2019) in their study have employed several such machine learning techniques to predict two important water quality parameters; namely, five-day biological oxygen demand (BOD) and chemical oxygen demand (COD). They have used least square-support vector machine (LS-SVM) and multivariate adaptive regression spline (MARS) techniques and found them significantly more accurate than other conventional machine learning methods such as artificial neural networks. LS-SVM with polynomial kernel yielded the most accurate estimations in predicting BOD_5 with 5.463 root mean square error (RMSE) while LS-SVM with RBF kernel predicted COD better with RMSE of 4.461. They have used nine water quality parameters as input which are quite expensive to employ in an IoT system. Najafzadeh *et al.* (2018) have conducted their study on the dataset of Karoun river, Iran. They have used nine independent water quality parameters to estimate three significant parameters namely, dissolved oxygen (DO), biological oxygen demand (BOD) and chemical oxygen demand (COD). They have employed three machine learning techniques, Model Tree (MT), Evolutionary Polynomial Regression (EPR) and Gene Expression Programming (GEP). The performance of GEP was slightly better in predicting BOD with RMSE of 5.388 while ERP's performance proved to be better in predicting COD and DO with RMSE of 4.997 and 4.728 respectively. The research concurs that an in-depth understanding of the input and output parameters is necessary when dealing with empirically derived equations for estimation, while the alternative of artificial intelligence (AI) techniques is much more convenient to employ. The only hurdle is the availability of readings of nine input parameters for prediction, which might turn out to be expensive if it's to be used in an IoT system. Shafi *et al.* (2018) have used Internet of Things (IoT) for real-time monitoring of water quality and applied different machine learning methodologies to predict water quality. The authors have devised a low-cost real-time water quality monitoring kit using the ATMega328 microcontroller, pH sensor, turbidity sensor, and temperature sensors. The data analytics part of the research work is carried out on a dataset collected from 11 different water sources in Pakistan. To predict water quality, machine learning algorithms, including Support Vector Machine (SVM), k Nearest Neighbor (kNN), Artificial Neural Network (ANN) and deep neural networks have been used. In their findings, deep neural networks yield the highest accuracy of 93% while the second-best prediction algorithm is SVM with an accuracy of 91%. In this research work, the parameters of accuracy, precision, and recall have been used for performance evaluation purposes.

Sakizadeh (2016) has conducted his research on the dataset of 47 wells and springs acquired from a ministry from Iran in the time duration of 2006–2013. His study considers 16 water quality parameters. He has used the method proposed by Horton (1965) to calculate the WQI. Three methodologies have been employed in this research work: ANN with early stopping, ANN with ensemble averaging, ANN with Bayesian Regularization. The correlation coefficients between the predicted and observed values of WQI were calculated to be 0.94 and 0.77 and concluded that ANN with Bayesian Regularization generalizes the dataset better than the rest of the techniques. However, this model is prone to over-fitting due to a lesser number of samples so the study must 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 ANN. BOD and COD are predicted using easily attainable parameters like pH, temperature (T), total suspended (TS) and total suspended solids (TSS). This study has been conducted on the data acquired from the Ekbatan wastewater treatment plant, Iran. In order to validate the model, two

prominent evaluation criteria were used including MSE 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 the prediction of COD was RMSE = 49.4 mg/L, r = 0.81. It was established that both the 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 various websites have been gathered in a data repository for monitoring and analysis. The machine learning algorithms, including pixel-based adaptive segmenter and bag of words approach, are used on this data to aid a user to make informed decisions. The authors have conducted their study on Dublin Bay and have used YSI 6600EDS for continuous monitoring of turbidity, optical dissolved oxygen, temperature, conductivity, and depth. In this research, the authors have modified and used pixel-based adaptive segmenter from the image processing domain to detect anomalous events from a continuous data stream. Once anomalous events are detected, the features of anomalous events are then extracted and clustered to perform decision making.

Ali & Qamar (2013), in their research, have mapped the water quality detection problem to the machine learning domain. They have conducted their research on Rawal watershed, situated in Islamabad, Pakistan. They collected 663 water samples from 13 different stations and tested them for features of appearance, temperature, turbidity, pH, alkalinity, hardness as $CaCO_3$, conductance, calcium, total dissolved solids, chlorides, nitrates and fecal coliforms against WHO standards. The data was initially preprocessed; the missing values were filled out by the attribute mean and the outliers replaced by attribute median, followed by correlation analysis to draw out the correlation amongst the parameters. In this paper, regression models are employed to check seasonal water quality trends (monthly and quarterly) and since there was no WQI in the data, authors have employed unsupervised learning: Average Linkage (within groups) method of Hierarchical Clustering using Euclidean distance to categorize water quality. In results, the higher values of fecal coliforms were found in the months of March, June, July, and October. However, their model had a clear limitation since no other parameters except fecal coliforms and turbidity were out of the standard limits in the data set. Hence, the accuracy was mainly dominated by turbidity and fecal coliforms.

Gazzaz *et al.* (2012) have conducted their research on 255 samples from the Kinta river, Malaysia, obtained by their Department of Environment. This dataset comprises 9,180 data points derived from measurements on those samples. Thirty parameters from these samples have been acquired and reduced to 23 through Principal Factor Analysis (PFA). Initially, using the WQI method, the authors have calculated the WQI manually and then using ANN with a setting of 23-34-1 to train their model. The dataset was partitioned into three parts; that is, 80% for training, 10% for validation and 10% for testing. The aforementioned setting explained 99.5% of the predictions and variations of accuracy are dependent on the size and variation of the dataset the data accurately. However, the accuracy is dependent on the size and variation of the dataset.

Verma & Singh (2012), in their study, have acquired 73 samples from Jharia coalfield situated in Jharkhand, India. They used 58 of those samples for training and the rest for testing. A three-layer feed-forward backpropagation neural network was used and was trained for 1,000 epochs. Their model takes in six input parameters, including temperature, pH, TS, TSS, DO, oil and grease and produces two outputs: BOD and COD. The results reflect the RMSE values for the BOD and COD to be 0.114 and 9.83 respectively, and corresponding coefficients of correlation to be 0.976 and 0.981. They concluded that ANN with Bayesian Regularization generalizes in the best possible way. One of the major limitations of this work is that the prediction of WQI is not carried out, but only estimation of BOD and COD is done, which might add to the error if it is used to predict WQI.

Mahapatra *et al.* (2011) have proposed to use a fuzzy system to predict the WQI. Generally, conventional fuzzy systems are efficient; however, complexity affects efficiency. For this reason, authors have proposed a cascaded fuzzy system that works better with complex problems. The proposed fuzzy system takes multiple inputs and gives out multiple outputs by using multiple fuzzy sub-systems. The

authors have validated their system on data collected from the Central Pollution Control Board (CPCB), India. They have used the data of six Indian rivers and estimated WQI using three water quality calculation methods i.e. Indian, Malaysian, and USA. Six parameters have been used for their case study, including pH, biological oxygen demand, dissolved oxygen, fecal coliform, electric conductivity, ammoniacal nitrogen, and temperature. Three fuzzy subsystems have been used, each for a different water quality criterion. Evidently, predictions of the system are quite close to the actual WQIs of each criterion making the proposed system more fit, to the problem at hand, than conventional fuzzy systems.

Bucak & Karlik (2011) have emphasized the importance of real-time detection of water contamination and their research work is mostly focused on intentional contamination of water. The Cerebellar Model Articulation Controller Artificial Neural Network (CMACANN) has been utilized for contamination detection due to its fast learning capabilities. Five parameters have been monitored including pH, conductivity, chlorine residual, turbidity, and total organic carbon (TOC). To validate their model, they have intentionally introduced certain contaminants in the water such as sodium cyanide, sodium arsenate, sodium fluoroacetate, parathion, cryptosporidium parvum oocysts, and a surrogate of *Bacillus anthracis* spores. Their model then detects the effects of contaminants and classifies it as an anomaly. The proposed model works far better than conventional Multi-Layer Perceptron with Backpropagation (MLP with BP). Whereas the MLP achieves an accuracy of 98% after 1,000 iterations, the proposed model achieves 100% accuracy with far less iterations.

Yan *et al.* (2010) have used an Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict water quality status instead of the conventional Artificial Neural Networks and found ANFIS to be more efficient than the other. In this work, the dataset of major river basins of China was obtained from CNEMC, consisting of 845 observation samples. Three parameters have been selected for the classification model, including dissolved oxygen (DO), chemical oxygen demand (COD) and ammoniacal-nitrogen (NH_3-N). The employed model 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 make it easier to interpret and map; however, generation of those rules and their outcomes require expert knowledge which makes fuzzy logic unsuitable to our problem. However, ANN comes with certain adaptability, which enables ANFIS to combine the power of both algorithms. ANN allows ANFIS to learn and construct rules of fuzzy logic, which turns out to be more efficient than either of the models and classified 89.59% of the data correctly.

Rankovic *et al.* (2010) conducted their study on Gruza reservoir, Serbia. They acquired 180 data samples by a monthly sampling for 3 years (2000–2003) through monitoring. They used 152 of those data samples for training and 28 for testing. The input parameters considered in this work include pH, temperature, chloride, total phosphate, nitrites, nitrates, ammonia, iron, manganese, and electrical conductivity and the predicted parameter is dissolved oxygen (DO). The Feed-forward Neural Network (FNN) model has been used to predict the dissolved oxygen. The Levenberg–Marquardt algorithm is used to train the FNN and the researchers have established that 15 hidden neurons give optimal results. The results of FNN models have been compared to the measured data based on 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. Moreover, 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 including total dissolved solids (TDS), electrical conductivity, and turbidity. This study has been conducted on two monitoring stations, Johor River and Sayong River, situated in Malaysia. A different methodology has been employed for each parameter as well as for each monitoring station. For TDS, backpropagation with two hidden layers and Bayesian regularization, but with distinct transfer functions for each monitoring station, has been used. TDS using EC has been predicted, since both are highly correlated as is evident in their results. The same methodology has been employed for EC and they predicted it using TDS given their correlation. For turbidity, FNN using backpropagation has been

employed with a single hidden layer and backpropagation. A distinct function for each monitoring station was used. Turbidity has been predicted 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 ANNs to predict biochemical oxygen demand (BOD) and chemical oxygen demand (COD) using other water quality parameters, including, TOC, total suspended solids (TSS), total dissolved solids (TDS), phenol concentration, ammoniacal nitrogen (AMN) and Kjeldahl's nitrogen (KJN). The regression analysis has been employed to find the correlation of TOC with BOD and COD. After regression analysis, 12 different models of ANN have been run using different combinations of the aforementioned water quality parameters to predict BOD and COD. The Average Relative Error (ARE) is used to find the accuracy of the model. The model has seven hidden neurons in the hidden layer and a training count of 5,000 with TOC, phenol, TSS and AMN predicting the BOD most effectively, having an ARE of 11.66%. Similarly, the model having eight hidden neurons and a training count of 1,500 with TOC, phenol and TDS as input predicted COD in a better manner, having an ARE of just 6.97%. The model with six hidden neurons and a training count of 5,000 with TOC, phenol, TSS and TDS as input performed effectively for both BOD and COD with ARE for BOD as 8.20% and ARE for COD as 11.08%. The empirical relations formed amongst various parameters are quite reliable and bring versatility in the domain.

Research concerning internet of things (IoT)

The IoT domain proposes smart and low-cost solutions to the problem of water contaminant detection and water quality analysis. Geetha & Gouthami (2017) have proposed a generic IoT system for real-time water quality monitoring. It comprises sensors that take parameter readings, then those parameter readings are transmitted to a controller through wireless communication devices attached to the sensors. Later the controllers, through wireless communication technology, store those sensor readings to data storage, which are reflected in a customized application. This is followed by the implementation of an instance of the generalized IoT system. They used four parameter sensors for this, namely conductivity, turbidity, water level, and pH. For connectivity, they used TI CC3200, which is a single-chip microcontroller with built-in 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 the cloud or storage, and then an application could use 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. For the software end, the MySql database along with SOAP web services and applications developed in C# and Android are part of this solution. The C# application allows sending a request for sensor readings through Arduino and a multiplexer. Once requested, a sensor takes readings and sends them back to the computer through a ZigBee transmitter and the readings are then received by a ZigBee receiver attached to the computer. The received readings are then saved in the local database and are sent to the cloud through a web service and are eventually visualized in the Android application. AI is not used in the system, but it does set the base for AI to be used in the future for effective real-time decision making.

Raju & Varma (2017) have proposed a real-time monitoring system for aqua farmers which allows the farmers to be apprised of the anomalous events if the water body is contaminated. They have used Raspberry Pi 3 with built-in Wi-Fi module, a solar panel and a sensor node comprising various sensors including those of dissolved oxygen, ammonia, pH, temperature, salt, nitrate, and carbonates. The system continuously monitors and stores the sensor data and generates an alert for the farmer if any of the data deviates from the allowed range. There is a mobile application for the farmer through which he can monitor the sensor data in real-time and access historical data. Vijai & Sivakumar (2016) have proposed an IoT framework for real-time water quality monitoring, demand forecasting, and anomaly detection. In the proposed system, the parameters of turbidity, chlorine, oxidation-reduction potential (ORP), nitrates,

pH, conductivity, and temperature have been considered and respective sensors have been used to take the various measurements. The connectivity is achieved using either of several options i.e. 3G, Bluetooth, ZigBee, etc. All these components, when connected, make a centralized system requiring a steady power supply to keep the system online. The authors have also proposed two other components of Demand forecasting and Anomaly detection. For anomaly detection, the technique of ANN and fuzzy system has been utilized. However, the proposed system has not been tested with any data set.

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 or not. The pH sensor, along with the pH meter and turbidity sensor, has been used to sense these readings from the water body. These readings are sent to an analog to digital converter (ADC) which in turn sends the digital readings to 16F887A PIC microcontroller to be shown on an LCD. Their work is extendable to employ GSM technology for communication. Cloete *et al.* (2016), in their research, have designed a sensor node which consists of temperature, conductivity, pH, ORP and flow sensors. Since the commercially available sensors are expensive, they have implemented the sensor designs themselves, thus making the system cost-effective. The signals generated from the sensors then go through conditioning in order to be able to interface with the microcontroller. Apart from the sensor node, their proposed system makes use of a ZigBee module to receive and transmit the measurements 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 measurements goes out of its allowable limits.

Wong & Kerkez (2016) have emphasized the importance of using real-time data along with historical data for water quality detection. The authors have also discussed the flexibility that comes with using web services along with the IoT platforms. Since some of the water quality constituents are difficult to measure or their sensors are too expensive for a cost-effective solution, this solution utilizes adaptive sampling of water along with easily available sensors. Adaptive sampling, instead of sampling after predefined intervals, adapts and samples only when an event of interest occurs; for example, flood, and minimizes the number of samples to be taken. The NeoMote wireless sensing platform has been used which consists of an ARM-Cortex M3 microprocessor for computing and the Xively IoT platform, which is used, as an interface for the services. The sensor node was connected to an automated sampler (ISCO 3700) which had a 24-bottle capacity. To emphasize the flexibility of web services, the authors have used three web services, each of which is developed using a different programming language. One of them was used to receive commands for sampling and transmitting data; the second service used the adaptive sampling algorithm and sent the sampling commands to the first service. The third web service was used to interface with the IoT platform in order to access the historical data and communicate with the sensors. The 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 presented a prototype for measuring the water level in real-time, and consequently generate alarms for authorities and on social networks, in case of alarming events like floods etc. The ultrasonic sensors have been utilized along with a wireless gateway, ATmega328P controller and a cloud server. After frequent intervals, ultrasonic sensors determine the distance between the water level and the sensor by sending a sound wave and estimating the water level by its reflection. Once determined, the water level reading is transmitted to the cloud server through a wireless gateway, where it is stored in 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. In addition to that, the data regarding water level stored on a cloud server is visualized through a web application to learn the trends and perform decision making. Vijayakumar & Ramya (2015) have proposed an online water quality monitoring system, which employs five sensors including sensors of temperature, pH, turbidity, conductivity, and dissolved oxygen. The Raspberry Pi B+ and IoT module USR-WIFI232-X-V4.4 have been used to transmit the data collected from sensors to the cloud through a gateway. The proposed system provides water quality monitoring and is suggested to be installed in different locations in a pond to collect real-time water quality data.

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

Rasin & Abdullah (2009) have proposed a cost-effective online water quality monitoring system using wireless sensor networks (WSNs). Their system consists of two modules: a wireless node and a base monitoring station. The wireless node consists of a sensor unit and a microprocessor and is powered by a 9 V battery. They use the ZMN2405HP ZigBee module, which consists of a CC2430 transceiver IC. The inexpensive pH, temperature and turbidity sensors have been used and readings from these sensors go through signal conditioning to determine their validity. Once conditioned, the wireless sensor node sends the readings to the base monitoring station through the transceiver. The other ZigBee module, consisting of the transceiver at the base monitoring station, receives the readings and sends to the computer using the RS 232 protocol. The received data are then visualized on a custom GUI developed in C++. Wang *et al.* (2009) have proposed a low cost, low power, and long-distance supervisory system based on the WSN for aquaculture. Their proposed system consists of two modules, a coordinator and a sensor node. The coordinator is composed of a ZigBee based wireless communication module, which uses a CC2430 chip with an RF transceiver and an ADC, and a GPRS module to transmit the data to the monitoring computer, which stores the data and helps in visualization. The sensor node contains the sensors, which read the water quality parameters and apply signal conditioning on the readings to prepare them to be digitized. After signal conditioning, the readings are sent to the coordinator, where these are digitized and processed further. In addition to this, the system is modeled to consume low amounts of battery, as it goes into sleep mode when there is no request for data to be read.

Several methods employing statistical inference, machine learning and IoT were reviewed in the preceding sections. Most of the research concerning statistical inference used laboratory analysis such as electrometry for pH readings, titration for hardness and membrane filtration method for coliforms etc. followed by WHO range examination and WQI calculation through different empirical methods such as DoE-WQI and IWQI (EPA 2001; Gazzaz *et al.* 2012; Batabyal & Chakraborty 2015). Research concerning machine learning explored different supervised learning techniques such as ANN, SVM, LS-SVM, MARS, EPR, GEP, KNN, MLR, ANFIS, FNN, CMACANN and fuzzy systems to estimate water quality and other water quality parameters (Yan *et al.* 2010; Bucak & Karlik 2011; Mahapatra *et al.* 2011; Gazzaz *et al.* 2012; Abyaneh 2014; Najafzadeh *et al.* 2018; Shafi *et al.* 2018; Najafzadeh & Ghaemi 2019). Research concerning IoT explored several low cost IoT systems designed for water quality monitoring. Most of the employed hardware included parameter sensors, the TI CC3200 microcontroller, Arduino module, Raspberry Pi 3, 16F887A PIC, ARM-Cortex M3 and ATmega328P, while most systems established communication through LoRa, Zigbee, a built-in Wifi module and GSM (Perumal *et al.* 2015; Birje *et al.* 2016; Wong & Kerkez 2016; Encinas *et al.* 2017; Geetha & Gouthami 2017; Raju & Varma 2017). The reviewed methodologies built a foundation for a real-time water quality system using IoT and machine learning methodologies. It reflected that most substantial studies employing machine learning techniques to estimate water quality through WQI used at least six to nine parameters as input, which is expensive in terms of IoT hardware given the cost and availability of the sensors. Hence, we found a gap in machine learning techniques for estimating water quality with a lesser number of parameters in order to build low cost water quality systems.

Table 3 provides a detailed description and analysis in the form of a comparison of various research papers encompassing the domains of IoT, machine learning and manual laboratory analysis. The comparison of the research papers has been carried out based on the parameters of the methodology employed by researchers for the problem, limitations of the research paper, the dataset used, employed water

Table 3 | The comparison of various research papers related to water quality detection

Sr. no.	Paper	Methodology	Limitations	Dataset	Parameters	Results	Hardware
1	Najafzadeh & Ghaemi (2019)	Estimating BOD_5 and COD using LS-SVM	Uses 9 WQPs which is impractical for an IoT system	200 datasets from Karoun River	Ca^{2+}, Na^{+}, Mg^{2+}, NO^{-2}, NO^{-3}, PO_4^{3-}, EC, pH and turbidity	$RMSE_{BOD5} = 5.463$ and $RMSE_{COD} = 4.461$	N/A
2	Najafzadeh *et al.* (2018)	Estimating BOD, COD and DO using MT, EPR and GET	While it is certainly a better alternative than empirical calculations it still uses 9 parameters, rendering its use for IoT system inconvenient	Dataset of Karoun River, Iran	Ca^{2+}, Na^{+}, Mg^{2+}, NO^{-2}, NO^{-3}, PO_4^{3-}, EC, pH and turbidity	$RMSE_{BOD} = 5.388$, $RMSE_{COD} = 4.997$ and $RMSE_{DO} = 4.728$	N/A
3	Shafi *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. The standard WQI has not been used	Dataset of 667 samples collected from PCRWR	pH, turbidity, temperature	Accuracy: DNN 93% SVM 91% NN 86% kNN 76%	ATMega328, LCD and parameter sensors
4	Geetha & Guthami (2017)	Monitoring using sensors and cloud infrastructure	Only monitoring, no prediction	N/A	Conductivity, turbidity, water level, pH	N/A	TI CC3200 controller and parameter sensors
5	Daud *et al.* (2017)	General review of water quality across all provinces of Pakistan	Only manual laboratory analysis	Manual samples gathered across Pakistan	Total coliform, fecal coliform, *E. coli*	Excessive total coliform due to sewerage	N/A
6	Encinas *et al.* (2017)	Water quality monitoring using sensors and SOAP web services	Only monitoring, no prediction	N/A	Temperature, pH, dissolved oxygen	N/A	Parameter sensors, Arduino module and ZigBee transceivers
7	Raju & Varma (2017)	Real-time monitoring system and mobile application for aqua farmers to be apprised of contamination	Only provides monitoring, does not process data for trends	N/A	DO, ammonia, pH, temperature, salt, nitrate and carbonates	N/A	Raspberry Pi3 with built-in Wi-Fi module, a solar panel and a sensor node

(*continued*)

Table 3 | continued

Sr. no.	Paper	Methodology	Limitations	Dataset	Parameters	Results	Hardware
8	Wong & Kerkez (2016)	Adaptive sampling of water using adaptive sampling algorithm, Xively IoT platform and web services to monitor water quality	This work does not monitor water quality in real-time but through sampling, it just provides monitoring. No predictive analysis	N/A	N/A	N/A	NeoMote wireless sensing platform: ARM-Cortex M3 microprocessor and Xively IoT platform. Automated sampler (ISCO 3700 with 24 bottle capacity)
9	Vijai & Sivakumar (2016)	Artificial neural network (ANN) and fuzzy systems	Proposes a generic IoT system without any dataset and results	N/A	Turbidity, chlorine, ORP, nitrates, pH, conductivity temperature	N/A	Sensors, connectivity: 3G, Bluetooth, & ZigBee
10	Sakizadeh (2016)	ANN with early stopping, ANN with ensemble averaging and ANN with Bayesian Regularization	Prone to overfitting with fewer 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
11	Alamgir *et al.* (2015)	Bacteriological and physio-chemical analyses	Only manual laboratory analysis	Forty-six samples of piped water in Orangi town, Karachi, Pakistan 2014	pH, TSS, TDS, turbidity, TCC, TFC, TFS	Well within limits except for sulphates and total fecal coliform	N/A
12	Batabyal & Chakraborty (2015)	Calculates 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 water quality was attributed to high contents of TDS, NO3, and Cl	N/A
13	Vijayakumar & Ramya (2015)	Monitoring employing IoT through sensors and cloud	Only provides monitoring	N/A	Temperature, pH, turbidity, conductivity, DO	N/A	Sensors, Raspberry Pi B +, IoT module USR-WIFI232-X-V4.4
14	Abyaneh (2014)	Multivariate linear regression (MLR), artificial neural networks (ANN), RMSE	Only predicts BOD, which does not completely reflect water quality	Data acquired from Ekbatan wastewater treatment plant, Iran	pH, temperature, total suspended, total suspended solids	Both models predicted BOD better than COD and pH had the most effect on the prediction	N/A

(*continued*)

Table 3 | continued

Sr. no.	Paper	Methodology	Limitations	Dataset	Parameters	Results	Hardware
15	Zhang *et al.* (2014)	Continuous monitoring, pixel-based adaptive segmenter, and bag of words	Does not predict water quality, just clusters possible anomalous events	Dublin Bay	Turbidity, optical dissolved oxygen, temperature, conductivity, depth	N/A	YSI 6600EDS
16	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, alkalinity, hardness as $CaCO_3$ conductance, calcium, TDS, chlorides, nitrates, fecal coliform	High fecal coliforms were found in the months of March, June, July, and October	N/A
17	Verma & Singh (2012)	ANN with Bayesian regularization: 1,000 epochs	Does not calculate WQI but predicts BOD and COD	73 samples (58 for training and 15 for testing)	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% and correlation is 0.976 and 0.981	N/A
18	Gazzaz *et al.* (2012)	Artificial neural network, 23-34-1	Requires larger dataset	9,180 data points, 255 samples	30 parameters reduced to 23 through PFA	Predictions explain almost 99.5% of the variations	N/A
19	Rankovic *et al.* (2010)	FNN. Levenberg–Marquardt algorithm is used to train the FNN. 15 hidden neurons	WQI not calculated but predicts DO, which might result in error ahead if WQI is calculated using it	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

quality parameters, results, and hardware used in the proposed solution.

WATER QUALITY MONITORING AND DETECTION SYSTEM (WQMDS)

This section highlights in detail the water quality monitoring and detection system that we have proposed, which not only monitors the water quality in real time but also predicts the trends of water quality and recognizes anomalous events. The high-level architecture of the proposed system comprising various modules including the sensing module, the coordinator module, the data processing and analysis module and the storage and core analytics module is shown in Figure 1.

The detailed description of each of the modules involved in the proposed system is given next.

Sensing module

The sensing module contains several sensors to measure four of the most important water quality parameters that are used to detect water quality. This module is responsible

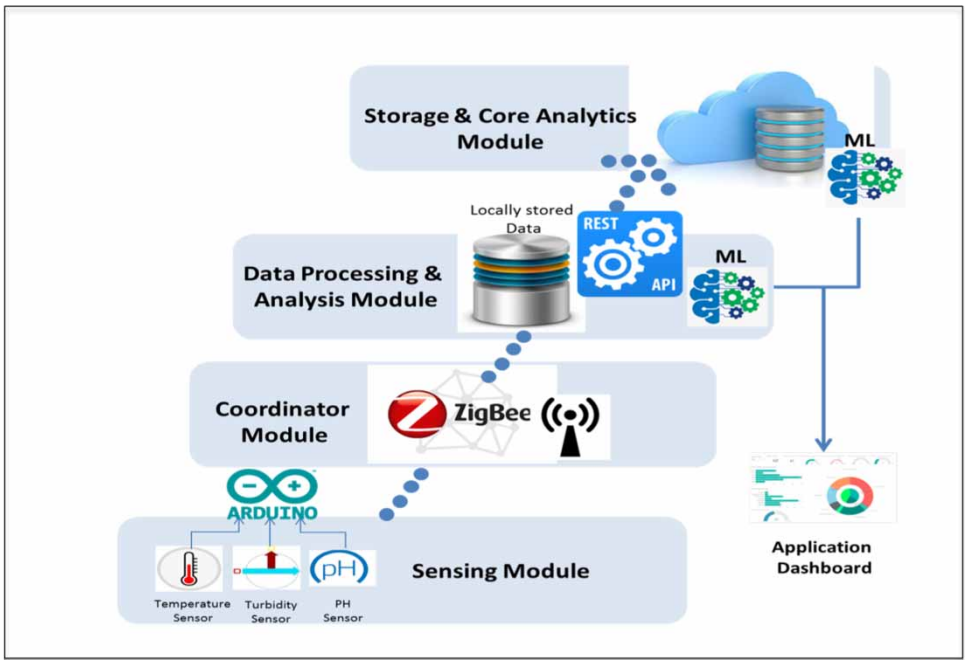

Figure 1 | The architecture of the proposed water quality detection and monitoring system (WQDMS).

for measuring the parameter readings and transmitting these readings to the coordinator module. The following four sensors have been utilized: the pH, turbidity, temperature, and total suspended solids sensors, which sense the respective parameters from the water body.

Coordinator module

The coordinator module is responsible for coordinating between the sensing module and the data processing and analysis module. The coordinator module uses the Arduino microcontroller, which signals the sensors to take the real-time readings and in return receives all the parameter measurements from the various sensors that are connected to it. Once sensor measurements are received, the coordinator module transmits them through a ZigBee transceiver to the on-site computer.

Data processing and analysis module

The data processing and analysis module, which is connected to the coordinator module through a transceiver, comprises various web services, including the data pre-processing service, storage service, and the data analysis service. Once the real-time measurements are received from the coordinator module, they are stored locally in a MySql database using the storage service. In addition to this, the data pre-processing service processes the data that is received in real time, including the filtering of useful data.

Since a large amount of data is available at the local server, analysis can be carried out to find out the hidden trends and some useful information could also be explored. This process is also carried out at this layer.

Storage and core analytics module

The storage and core analytics module fulfills two major responsibilities; that is, firstly to ensure the long-term storage of water parameter readings and secondly to predict water quality trends using machine learning techniques and detect anomalous events.

Once the data have passed the pre-processing stage, it is transferred to the cloud using the REST web service.

Various machine learning algorithms are applied to predict water quality. The detection of anomalous events, including out of range parameter readings or any other malfunctioning, is also detected through this module. The machine learning module implements two machine learning models deployed on the cloud along with the data; one of them is trained using ANN, which detects the anomalies and informs the dashboard about them. The other is trained using an unsupervised learning technique; that is, K-Means clustering, which classifies water quality into three clusters: Class 0, Class 1 and Class 2. The normal data points would be assigned to cluster 1 and would represent water fit for drinking. Similarly, another cluster would represent water fit for uses other than drinking and the last cluster would represent water unfit for any use. For any new arriving point, if it passes the anomaly detection, the system would query the deployed K-Means model to be sure what cluster it should belong to. This helps to get a rough measure of its deviation from normal, if there is any. The aforementioned machine learning models can be queried through the web service written in Java.

Application dashboard

The proposed system has an application dashboard, which is used for visualization of the water quality data on desktop, web and mobile platforms. Once the real-time data is received and stored, this dashboard, which has been developed in Java, visualizes that information in the form of graphs and heat maps. The safety ranges will also be displayed on the interface as well as alerts notifying the admin regarding any anomalous events. The web dashboard has two instances, one of which is deployed on the local desktop computer on the site, which is connected to the local backup database and is accessible only on site. The other instance is deployed on the internet and uses the REST web service to access the cloud for the data, and uses that for visualizations on the dashboard. The Android application also accesses the data from the cloud through the same web service and plots visualizations. Both applications generate an on-screen alert if anything seems to be gravely out of limit and suggest an informed measure in case of an anomaly.

CONCLUSION AND FUTURE WORK

Water is one of the most essential resources for survival and its quality is determined by the WQI, which is measured through various water quality parameters depending upon the type of standard used. Conventionally, to measure water quality parameters, expensive and time-consuming laboratory analysis is performed, which makes timely contaminant recognition and its ramifications difficult. Alternatively, an IoT-based system can be employed to monitor water quality in real time, which is an efficient and low-cost solution to the problem. Several such systems, such as CANARY, are deployed at various places using IoT effectively and they have proved to be an effective alternative to expensive manual laboratory analysis. While IoT systems are employed for real-time water quality monitoring, machine learning techniques such as artificial neural networks (ANN), support vector machines (SVM), regression, correlation analysis, hierarchical clustering etc. aid in learning the trends of the water quality parameters, predicting WQI, and detecting anomalous events like intentional contamination to enable real-time contamination detection and action. The proposed system makes use of IoT for real-time monitoring and uses machine learning algorithms to learn various trends in the data to incorporate its learning in the system and aid in decision making. Additionally, the review found a gap in the machine learning methodologies to estimate water quality with a lesser number of parameters, which can be easily employed in a low cost IoT system using minimal number of parameter sensors.

REFERENCES

Abyaneh, H. Z. 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.

Alamgir, A., Khan, M. A., Hany, O. E., Shaukat, S., Mehmood, K., Ahmed, A., Ali, S., Riaz, K., Abidi, H., Ahmed, S. & Ghori, M. 2015 Public health quality of drinking water supply in Orangi Town, Karachi, Pakistan. *Bulletin of Environment, Pharmacology, and Life Sciences* **4** (11), 88–94.

Ali, M. & Qamar, A. M. 2013 Data analysis, quality indexing and prediction of water quality for the management of Rawal Watershed in Pakistan. In: *Eighth International Conference on Digital Information Management (ICDIM 2013)*.

Andrew, D. E., Lenore, S. C. & Arnold, E. G. 1995 *Standard Methods for the Examination of Water and Wastewater*, 19th edn. American Public Health Association, American Water Works Association and Water Environment Federation, Washington, DC, USA.

Batabyal, A. K. & Chakraborty, S. 2015 Hydrogeochemistry and water quality index in the assessment of groundwater quality for drinking uses. *Water Environment Research* **87** (7), 607–617.

Bhandari, N. S. & Nayal, K. 2008 Correlation study on physico-chemical parameters and quality assessment of Kosi river water, Uttarakhand. *Journal of Chemistry* **5** (2), 342–346.

Birje, S. V., Bedkyale, T., Alwe, C. & Adiwarekar, V. 2016 Water pollution detection system using pH and turbidity sensors. *International Journal of Advanced Research in Computer and Communication Engineering* **5** (4), 530–533.

Bucak, I. O. & Karlik, B. 2011 Detection of drinking water quality using CMAC based artificial neural networks. *Ekoloji Dergisi* **20** (78), 75–81.

Bureau of Indian Standards 1991 *Indian Standard Drinking Water Specification. 1st rev*. Bureau of Indian Standards, New Dehli, India.

Cabral, J. P. & Marques, C. 2006 Faecal coliform bacteria in Febros river (northwest Portugal): temporal variation, correlation with water parameters, and species identification. *Environmental Monitoring and Assessment* **118** (1–3), 21–36.

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

Cao, F., Jiang, F., Liu, Z. & Yang, Z. 2014 Application of ISFET microsensors with mobile network to build IoT for water environment monitoring. In: *International Conference on Intelligent Environments*, Shanghai, China.

Cloete, N. A., Malekian, R. & Nair, L. 2016 Design of smart sensors for real time water quality monitoring. *IEEE Access* **4**, 3975–3990.

Daud, M. K., Nafees, M., Ali, S., Rizwan, M., Bajwa, R. A., Shakoor, M. B., Arshad, M. U., Chatha, S. A. S., Deeba, F., Murad, W., Malook, I. & Zhu, S. J. 2017 Drinking water quality status and contamination in Pakistan. *BioMed Research International* **2017**, 1–18.

Ejaz, N., Hashmi, H. N. & Ghuman, A. R. 2010 Water quality assessment of effluent receiving streams in Pakistan: a case of River Ravi. *Mehran University Research Journal of Engineering & Technology* **30** (3), 383–396.

Encinas, C., Ruiz, E., Cortez, J. & Espinoza, A. 2017 Design and implementation of a distributed IoT system for the monitoring of water quality in aquaculture. *Wireless Telecommunications Symposium (WTS)*, pp. 1–7.

Environmental Protection Agency 2001 *Parameters of Water Quality, Interpretation and Standards*. Available from: 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*. Available from: https://www.epa.gov/sites/production/files/2015-07/documents/water_quality_event_detection_system_challenge_methodology_and_findings.pdf (accessed 19 November 2018).

Gazzaz, N. M., Yusoff, M. K., Zaharin Aris, A., Juahir, H. & Firuz, M. 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.

Geetha, S. & Gouthami, S. 2017 Internet of things enabled real time water quality monitoring system. *Smart Water* **2** (1), 1–19.

Horton, R. K. 1965 An index number system for rating water quality. *Journal of Water Pollution Control Federation* **37** (3), 300–306.

Khatoon, N., Khan, A. H., Rehman, M. & Pathak, V. 2013 Correlation study for the assessment of water quality and its parameters of Ganga River, Kanpur, Uttar Pradesh, India. *IOSR Journal of Applied Chemistry* **5** (3), 80–90.

Mahapatra, S. S., Nanda, S. K. & Panigrahy, B. K. 2011 A cascaded fuzzy inference system for Indian river water quality prediction. *Advances in Engineering Software* **42** (10), 787–796.

Najafzadeh, M. & Ghaemi, A. 2019 Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental Monitoring and Assessment* **191** (6), 380.

Najafzadeh, M., Ghaemi, A. & Emamgholizadeh, S. 2019 Prediction of water quality parameters using evolutionary computing-based formulations. *International Journal of Environmental Science and Technology* **16** (10), 6377–6396.

Najah, A., Elshafie, A., Karim, O. A. & Jaffar, O. 2009 Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of Scientific Research* **28** (3), 422–435.

Patel, J. Y. & Vaghani, M. V. 2015 Correlation study for assessment of water quality and its parameters of par river Valsad, Gujarat, India. *IJIERE* **2**, 150–156.

Perumal, T., Sulaiman, M. N. & Leong, C. Y. 2015 Internet of things (IoT) enabled water monitoring system. In: *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*.

Raju, K. R. S. R. & Varma, G. H. K. 2017 Knowledge based real time monitoring system for aquaculture using IoT. In: *IEEE 7th International Advance Computing Conference (IACC)*.

Rankovic, V., Radulovic, J., Radojevic, I., Ostojic, A. & Comi, L. 2010 Neural network modeling of dissolved oxygen in the Gruza reservoir, Serbia. *Ecological Modelling* **221** (8), 1239–1244.

Rasin, Z. & Abdullah, M. R. 2009 Water quality monitoring system using Zigbee based wireless sensor network. *International Journal of Engineering & Technology IJET* **9** (10), 24–28.

Rene, E. R. & Saidutta, M. B. 2008 Prediction of water quality indices by regression analysis and artificial neural networks. *International Journal of Environmental Research* **2** (2), 183–188.

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

Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M. & Khurshid, H. 2018 Surface water pollution detection using internet of things. In: *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*.

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

Vijai, P. & Sivakumar, P. B. 2016 Design of IoT systems and analytics in the context of smart city initiatives in India. *Procedia Computer Science* **92** (2016), 583–588.

Vijayakumar, N. & Ramya, R. 2015 The real time monitoring of water quality in IoT environment. In: *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*.

Wang, Z., Wang, Q. & Hao, X. 2009 The design of the remote water quality monitoring system based on WSN. In: *2009 5th International Conference on Wireless Communications, Networking and Mobile Computing*.

Wong, B. P. & Kerkez, B. 2016 Real time environmental sensor data: an application to water quality using web services. *Environmental Modelling & Software* **84**, 505–517.

World Health Organization 1993 *Guideline for Drinking Water Quality*, 2nd edn, Vol. 1. World Health Organization, Geneva, Switzerland.

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

Zhang, D., Sullivan, T., Briciu-Burghina, C., Murphy, K., McGuinness, K., O'Connor, N. E., Smeaton, A. & Regan, F. 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.

First received 1 May 2019; accepted in revised form 23 September 2019. Available online 11 October 2019