Environ Monit Assess (2022) 194: 133
https://doi.org/10.1007/s10661-021-09698-4

Bacterial prediction using internet of things (IoT) and machine learning

Hamza Khurshid · Rafia Mumtaz · Noor Alvi · Ayesha Haque · Sadaf Mumtaz · Faisal Shafait · Sheraz Ahmed · Muhammad Imran Malik · Andreas Dengel

Received: 26 May 2021 / Accepted: 14 December 2021 / Published online: 28 January 2022

Abstract Water is a basic and primary resource which is required for sustenance of life on the Earth. The importance of water quality is increasing with the ascending water pollution owing to industrialization and depletion of fresh water sources. The countries having low control on reducing water pollution are likely to retain poor public health. Additionally, the methods being used in most developing countries are not effective and are based more on human intervention than on technological and automated solutions. Typically, most of the water samples and related data are monitored and tested in laboratories, which eventually consumes time and effort at the expense of producing fewer reliable results. In view of the above, there is an imperative need to devise a proper and systematic system to regularly monitor and manage the quality of water resources to arrest the related issues. Towards such ends, Internet of Things (IoT) is a great alternative to such traditional approaches which are complex and ineffective and it allows taking remote measurements in real-time with minimal human involvement. The proposed system consists of various water quality measuring nodes encompassing various sensors including dissolved oxygen, turbidity, pH level, water temperature, and total dissolved solids. These sensors nodes deployed at various sites of the

H. Khurshid · R. Mumtaz (✉) · N. Alvi · F. Shafait · M. I. Malik
School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
e-mail: rafia.mumtaz@seecs.edu.pk

H. Khurshid
e-mail: hkhurshid.bese16seecs@seecs.edu.pk

N. Alvi
e-mail: nalvi.bese16seecs@seecs.edu.pk

F. Shafait
e-mail: faisal.shafait@seecs.edu.pk

M. I. Malik
e-mail: malik.imran@seecs.edu.pk

A. Haque · S. Mumtaz
Dental College, HITEC-Institute of Medical Sciences, Taxila, Pakistan
e-mail: ayesha.haque@hitec-ims.edu.pk

S. Mumtaz
e-mail: sadaf.mumtaz@hitec-ims.edu.pk

S. Ahmed · A. Dengel
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Forschungsbereich Smarte Daten & Wissensdienste, Kaiserslautern, Germany
e-mail: sheraz.ahmed@dfki.de

A. Dengel
e-mail: Andreas.Dengel@dfki.de

A. Dengel
Fachbereich Informatik, Technische Universität Kaiserslautern, Kaiserslautern 67663, Germany

study area transmit data to the server for processing and analysis using GSM modules. The data collected over months is used for water quality classification using water quality indices and for bacterial prediction by employing machine learning algorithms. For data visualization, a Web portal is developed which consists of a dashboard of Web services to display the heat maps and other related info-graphics. The real-time water quality data is collected using IoT nodes and the historic data is acquired from the Rawal Lake Filtration Plant. Several machine learning algorithms including neural networks (NN), convolutional neural networks (CNN), ridge regression (RR), support vector machines (SVM), decision tree regression (DTR), Bayesian regression (BR), and an ensemble of all models are trained for fecal coliform bacterial prediction, where SVM and Bayesian regression models have shown the optimal performance with mean squared error (MSE) of 0.35575 and 0.39566 respectively. The proposed system provides an alternative and more convenient solution for bacterial prediction, which otherwise is done manually in labs and is an expensive and time-consuming approach. In addition to this, it offers several other advantages including remote monitoring, ease of scalability, real-time status of water quality, and a portable hardware.

Keywords Fecal coliform · Water quality index · Neural networks · Data analytic · Real-time data collection

Introduction

Issues related to water quality are rampant across the globe and contribute greatly to adversely affecting the health sector, particularly in developing countries. Though once a country with surplus water resources with an enormous resource of Indus water, Pakistan is now a country having a deficiency of water, with a deteriorated water quality supply. This is mainly due to the burgeoning population, ascending living standards of the civil society, and fast expansion of the number of industries, and agriculture development, which has led to water shortage and degrading its quality. In Pakistan, many cities are facing the deterioration of water quality causing serious health issues for the public. With the current water quality status in Pakistan, the accessibility of safe drinking water is available to only 20% of the population and 80% of the population is deprived of clean water (Daud et al., 2017). The primary renewable resources of drinking water are perennial rivers and groundwater. Due to climate changes in the Himalayas, there has been a decrease in river flows and a drying up of the Indus River, which is Pakistan's main source of fresh water. However, due to climate changes in the Himalayas, the flow of water in rivers has decreased, thus decreasing the availability of fresh water. Among these issues, sustaining the good quality of water has now become a recurring challenge.

The increased deforestation in the past years has led to soil and rock erosion which has caused the suspension of rock sediments in river water thus increasing its contamination level. Other reasons for water contamination include activities like agriculture development, rising urban sprawl, increased usage of fertilizers and pesticides, bacterial pathogens, and improper disposal of sewage and industrial wastes directly into the mainstreams, thus adversely affecting the public health (Shapiro, 2013; Nelliyat, 2016). Owing to the constant rise in population, there is tremendous stress on the current water reserves, which has eventually led to the depletion of groundwater. The excessive withdrawal of low levelled groundwater is contributing to water pollution as it thickens the saline layer and makes the water unusable for drinking and household consumption.

The quality of drinking water is mainly defined by the quality of water supply sources and the condition of pipelines laid for its flow (Baig et al., 2017). The quality of water pipelines declines due to weathering and rust, consequently increasing the development of microbial and the quantum of contaminants and toxins in water (Liu et al., 2016). In Pakistan, the cases of waterborne diseases are rampant. the population. According to the UN, around 30–40% of the deaths recorded in the country are associated with poor sanitation and low quality of water. It is also observed that the majority of these cases occurred due to the presence of fecal coliform bacteria in the water (Liu et al., 2015; Zeitoun et al., 2014). In addition to this, the use of different chemicals and fertilizers for crops and the contaminated water used for their irrigation can have serious implications on human health. Furthermore, several studies conducted related to the quality of water in Pakistan revealed the presence of microbes and metals in water, particularly lead and

arsenic, which have increased the water-related morbidity in the population (Daud et al., 2017).

According to WHO report 2017, around 361,000 children under the age of 5 die due to poor access to clean and good quality water (Naqvi, 2017). Unfortunately, Pakistan belongs to those countries, where water quality monitoring and control mechanisms are not functioning effectively due to a lack of technology-based solutions in place and an infrastructure to foster these solutions. Typically, water samples are tested manually in labs, which is a time-consuming and expensive method. In Pakistan, the water treatment plants are only present in a few cities, and they are not well maintained and fully functional which causes the microbial contamination to remain undetected (Hisam et al., 2014; World Bank, 2005).

Given the looming water crisis, there is an imperative need to devise a proper and systematic system to regularly monitor and manage the quality of water resources to arrest the related issues. Towards such end, we proposed an Internet of Things (IoT)–based system to monitor the quality of water in Pakistan, which provides near real-time assessment of the water to the community. Meanwhile, the water quality monitoring methods being used in most countries like Pakistan are not as efficient to log the water quality effectively. These methods are based more on human intervention than on technological and automated solutions. Most of the data is monitored in laboratories and takes time and effort and produces lower quantity of reliable results. Contrary to the manual-based method of water quality monitoring, IoT is a suitable alternative to such traditionally ineffective and complex approaches, and allows taking remote measurements in real-time with minimal human involvement. The parameters used in this research work for water quality assessment are pH level, turbidity, temperature, total dissolved solids, and dissolved oxygen. The selected parameter values are collected using the sensors deployed at different locations. Subsequently, the communication module transmits the data to the server for processing and analysis. In order to train the machine learning model and for trend analysis, the historic data of past years is obtained from the Rawal Lake Filtration Plant. For data visualization, heat maps and related charts are generated and displayed on the Web portal. Several machine learning algorithms including neural networks (NN), convolutional neural networks (CNN), ridge regression (RR), support vector machines (SVM), decision tree regression (DTR), Bayesian regression (BR), and an ensemble of all models are trained for fecal coliform bacterial prediction. A comparative analysis is done between these models, and results of these models on the test dataset are presented. A comparison is done using metrics such as R^2, mean average error, and mean squared error.

The major contributions of the paper are highlighted below:

- Development of an IoT-enabled portable water quality data collection node which can monitor 5 water quality parameters, while being deployed at river and lake banks for long durations.
- The node also has the ability to transmit real-time data to pre-deployed cloud servers.
- Integration of historical water quality and bacteria data with real-time water quality data for creating a dataset for bacterial prediction. The historical data contains all the real-time parameters in addition to the data for bacteria quantity.
- Development of bacterial prediction models that can predict quantity of bacteria in water using real-time water quality parameters detected directly by sensors, without the need of laboratory testing.

This prediction of the quantity of bacteria colonies in water directly through data from real-time sensors differentiates the proposed work from the existing work. Most of the existing technology-based solutions in the domain of remote water quality monitoring are either collecting water quality data using IoT devices or predicting future water quality data using past data, but the proposed system uses the past data along with data from IoT nodes to create a bacterial prediction system.

Related work

There are several studies conducted related to water quality assessments using a number of methods including (i) manual calculation and lab analysis; (ii) using different machine learning methodologies for water quality classification, prediction, and trend analysis; and (iii) IoT-based systems for water quality

monitoring and control in real-time. Each of these methods is described in the following sub-sections.

Use of manual lab-based analysis

A study related to manual calculation and lab analysis of various water samples collected from across all the provinces of Pakistan is discussed (Daud et al., 2017). In this study, different samples were tested for various parameters and were compared against National Environmental Quality Standards (NEQS) and WHO standards. The majority of the samples indicated the presence of fecal coliforms, *Escherichia coli* (*E. coli*) and total coliforms. The presence of these contaminants was mainly due to sewage and industrial waste disposal. It was recommended that treatment plants be installed for better enforcement of NEQs. In another study (Alamgir et al., 2015), 46 samples of piped water were taken across various places around Orangi Town, Karachi, and tested for physiochemical and bacteriological analysis using standard methods for the examination of water and wastewater. For testing purposes, WHO and National Standards for Drinking Water Quality (NSDWQ) standards have been used as a benchmark for comparative analysis. Statistical analysis was carried out for each of these parameters, where the analysis revealed that except sulfates, all other physiochemical parameters were within safe limits. However, total coliform and total fecal coliform counts were at critically high levels which reflected the bad hygienic conditions. It was recommended to continuously monitor the water and revamp the sewage systems. Another research was conducted on river Ravi by sampling the data of the river water for 3 years. The dataset ranged from January 2005 to March 2007 and was sampled from 14 stations (Ejaz et al., 2011). There were 12 parameters namely biochemical oxygen demand (BOD), dissolved oxygen (DO), chemical oxygen demand (COD), phosphorus, sodium, suspended solids, total nitrogen, chloride, nitrate, oil and grease nitrite, and total coliforms were tested. The Standard Methods for the Examination of Water and Wastewater (1991 USA) has been used for testing the aforementioned parameters and NEQS (National Environmental Quality Standards of Pakistan) has been used to perform the comparative analysis. In this study, an expensive lab analysis has been used which is considered the major limitation of this work. In another study (Batabyal et al., 2015), another research study is discussed that was conducted in the Kanksa-Panagarh area situated in West Bengal. The samples from 98 tube wells were collected for the post-monsoon period and pre-monsoon period from November to December 2011, and from May to June 2012 respectively. The samples were tested for 13 parameters namely pH, total dissolved solids (TDS), HCO3, total hardness, SO4, Cl, F, NO3, Mg, Ca, Mn, Fe, and Zn against WHO (1993) and Indian (BIS, 1991) standards. The correlation analysis of these parameters was performed and, in addition to that, the water quality index (WQI) was calculated using the Indian method to manually calculate the WQI.

Use of machine learning approaches

There are several studies conducted which incorporated machine learning methodologies for water quality classification, prediction, and trend analysis. In a study (Ali & Qamar, 2013), research is described which is conducted on the Rawal watershed, situated in Islamabad. The data sampling was done with 663 water samples, which were collected from 13 different stations, and they were tested for appearance, temperature, pH, alkalinity, turbidity, hardness (CaCO3), conductance, calcium, TDS, chlorides, fecal coliforms, and nitratres against WHO standards. A correlation analysis was performed between the parameters to find the related parameters. Afterwards, the monthly and quarterly trends of water quality were analyzed by employing regression models. For classification of water quality, unsupervised learning was applied. The results obtained from classification showed higher concentrations of fecal coliform in specific months of March, June, July, and October. The model shows a clear limitation as none of the other parameters was outside the safe limits. Additionally, the model was a little biased and ensured accuracy mostly on turbidity and fecal coliforms.

In another study (Sakizadeh, 2016), research was conducted on a dataset that was acquired from the Ministry of Iran. The dataset had a record of 47 wells and spring for the time period (2006–2013). A total of 16 parameters of water quality were considered for this study and Horton Method (1965) was used for calculating WQI. They employed three methodologies on the data: artificial neural network (ANN) with early stopping, ANN with ensemble averaging, and ANN with Bayesian regularization. The correlation

coefficients were calculated between the predicted and observed values of WQI and were found to be 0.94 and 0.77. It was concluded that ANN with Bayesian regularization generalizes the dataset better than others. However, the developed model is prone to overfitting as it has fewer samples; therefore, the study has to focus further on efficient generalization.

Another study (Gazzaz et al., 2012) was conducted based on 255 samples taken from the Kinta River in Malaysia. These samples were obtained by Malaysia's Department of Environment. The dataset had a total of 9180 points of data that were taken from the sample measurements. There were 30 parameters acquired from those samples which were reduced to 23 through principal factor analysis (PFA). Subsequently, the WQI was calculated manually using the Malaysian WQI method and then ANN was used to train the model. The dataset was partitioned into 3 parts. In total, 80% of the data was reserved for training while 10% was reserved for validation and the remaining 10% was left for testing purposes. This was helpful for explaining 99.5% of predictions accurately. The only drawback to the proposed approach was to have a large dataset in order to achieve satisfactory accuracy.

In another study (Verma et al., 2013), the study conducted has acquired 73 datasets from the Jharia coalfield situated in Jharkhand, India. The researchers have used 58 of those datasets for training and the remaining 15 for testing. They have used three-layered feed-forward back propagation neural network and trained it for 1000 epochs. Their model took six inputs namely temperature, total solids, pH, total soluble solids, oil and grease, and dissolved oxygen, and produced two outputs BOD and COD. Their results have shown root mean squared error values for the BOD and COD to be 0.114 and 9.83% and corresponding correlation coefficients to be 0.976 and 0.981 respectively. It was concluded that an artificial neural network with Bayesian regularization generalizes the model best.

In another study (Rankovic et al., 2010), a research conducted on Gruza reservoir, Serbia, is discussed. There were 180 data samples acquired by monthly sampling for 3 years (2000–2003) through monitoring. For training purposes, 152 data samples were used and the remaining 28 were for testing. The input parameters were pH, temperature, total phosphate, chloride, nitrites, ammonia, manganese, iron, and electrical conductivity and the predicted parameter was DO. The feed-forward neural networks (FNN) models have been used to predict the DO. The FNN was trained using the Levenberg-Marquardt algorithm and 15 hidden neurons were established to give the optimal results. The results of FNN models have been compared with the data on the basis of mean squared error (MSE), correlation coeffecients (r) and mean absolute error. The main limitation of this study was the use of a small dataset which made it prone to overfitting. In another study (Dogo et al., 2019), a review of both traditional ML and DL (deep learning) approaches was carried out. A review was performed on the progress made in the detection of anomalies in water quality data using ML techniques. It was concluded that for the tasks related to feature learning, DL approaches are better than ML approaches in terms of accuracy and producing fewer false positives. It was also concluded that the application of ELM (extreme learning machine) was sparsely exploited in the domain of water quality anomaly detection. A hybrid DL-ELM was presented to be a potential solution that could be worked on to be used in the detection of anomalies in water quality data.

Use of IoT technology

There is a variety of studies conducted related to the use of the Internet of Things (IoT) technologies for water quality monitoring. In a study (Geetha et al., 2016), the authors discuss a generic real-time IoT system for monitoring water quality. The system comprised multiple sensors for reading parameter values. These readings were transmitted to a controller using wireless communication. The controller also used wireless communication to store these values at a central database. These readings were then shown in a custom application. There were 4 parameters namely conductivity, turbidity, water level, and pH acquired through respective sensors. For connectivity, they used TI CC3200 which is a single-chip micro-controller with an in-built WiFi module and ARM Cortex M4. It can be connected to a WiFi hotspot for the purpose of transmitting the data to the cloud or storage for analysis. The sensors, if not connected to the controller, can also be connected to the LoRa sensors.

In another study (Shafi et al., 2018), an IoT-enabled solution is proposed to monitor the quality of water

in real-time. A prototype system was developed having sensors for monitoring of turbidity and pH. The system communicated with the cloud server to transmit data in real-time. A mobile app was also developed for data analysis.

In another study (Vijai & Sivakumar, 2016), a general framework is introduced for real-time monitoring of water quality, forecasting of demand, and detection of anomalies. The parameters considered for this research include chlorine, turbidity, nitrates, ORP, pH, temperature, and conductivity. For connectivity, several options including 3G, Bluetooth, Zigbee, etc., were proposed. When connected together, these components form a centralized system. The system requires a steady supply of power to keep itself online. The two other components proposed were anomaly detection and demand forecasting. For anomaly detection, ANN and fuzzy systems were used. The system proposed is a quite general system and it was not tested using any dataset.

In another study (Vijayakumar & Ramya, 2015), the proposed system used Raspberry Pi to develop a real-time water quality monitoring solution. The system used numerous sensors attached to the micro-controller to read values of different parameters. The system was not cost-effective in terms of scalability and it only provided the feature of monitoring.

Another low-cost water quality monitoring system was proposed (Adamo et al., 2015). The system used a cost-effective seawater probe to measure temperature, conductivity, chlorophyll-a, and turbidity values of water. In another study (Agarwal et al., 2018), the proposed system used the technology of drone with water quality sensors for water quality monitoring. The values of water quality parameters recorded using the sensors were sent to the master drone, which synchronized these values with the server for analyzing the levels of water pollution.

In another study (Das & Jain, 2017), the ZigBee module was used in the proposed water quality monitoring system for transferring the water sensor data wirelessly to the micro-controller and then the GSM module sends it to the user's mobile phone. The system can generate real-time alerts to officials by employing proximity sensors in case someone tries to contaminate the water.

In another study (Saravanan et al., 2017), the authors used the LoRa module in their proposed water grid management system. Real-time data was collected by deploying sensors at various locations and SMS/email alerts were sent to authorities to notify them about any issues. The collected data was analyzed using a prediction algorithm and uploaded on a Web page for public viewing.

The study discussed (Saha et al., 2017) analyzed the polluted water using sensors like the UV sensors to detect the UV light, the temperature, and pH sensors to study the temperature of the water and its surroundings along with the pH level. The sensed data is sent to the cloud for further analysis.

In [32], the authors proposed smart, low-cost, and less complex solutions for monitoring the water quality based on IoT technology. Their proposed model collects the water samples, uploads over the Internet, and generates alerts for users in case of any deviations in the water quality parameters. Most of the studies discussed a similar pattern of analyzing water quality. However, if anomaly detection algorithms are introduced along with the abovementioned techniques, the results can be improved further.

In another study (Tadokoro et al., 2017), the authors proposed a water control system to monitor the water supply and the sewage facilities using Hitachi's IoT Gateway Platform. They used optical fiber sensing for gathering data from inside sewage pipes and the collected data was further analyzed using artificial intelligence algorithms applied on image and voice data.

In another study (Khatri et al., 2018), the authors proposed an IoT-based pH monitoring and control system. The proposed approach is similar to the approach of monitoring water quality where the sensed data is sent to the cloud and alerts are generated on any issues faced.

From the above discussion, it is observed that there exist several studies, prototypes, and commercial systems developed for water quality monitoring across the globe but Pakistan still relies on the old traditional systems to monitor the quality of water. These systems are only installed at a few locations across the country and do not provide ubiquitous coverage. This situation in turn compels the research community to keep on collecting the water samples manually from various locations for lab testing and subsequent analysis. In order to address these issues, we proposed an automated and scalable IoT-based water contamination monitoring system, capable of providing near

real-time status of water quality parameters. The collected data is sent to the server for analysis and prediction and the Web portal with a dashboard of Web services are developed to display the heat maps and related info-graphics.

In another study (Sithole et al., 2019), an IoT-based water quality monitoring system was developed to collect experimental data. Five common sources of water contamination were used which were washing powder, vinegar, chlorine, salt, and soil. Sensors for turbidity, pH, conductivity, and flow rate of water were used to record water quality data parameters. Furthermore, the water CR (consumption risk) was calculated using deviation defined by WHO (World Health Organization) and DWA (Department of Water Affairs).

In another study (Ahmed, 2020), the authors proposed an IoT system for sensing water quality data, which was a low-cost and an efficient solution as compared to other solutions. The authors also suggest that the use of ML techniques such as artificial neural networks, regression, correlation analysis, clustering, and state vector machines can help predict water quality index, and can ultimately help in learning trends of water quality and detecting anomalies.

In another study (Faruq et al., 2017), the authors developed a water quality monitoring device using a micro-controller and sensors for measuring the pH, turbidity and temperature of the water. The micro-controller was used as a CPU (Central Processing Unit) while an LCD (Liquid Crystal Display) was used to show the readings from these sensors.

Methodology

Selection of site points for water quality monitoring nodes

The field visits of the selected study area, Rawal Dam, is conducted and the existing arrangements and methods for measuring water quality parameters such as pH level, turbidity, temperature, fecal coliform count, TDS, and dissolved oxygen were observed. Additionally, the locations/points at the dam site, where we intended to install our hardware, are explored and identified. During these field visits, some of the samples are collected for off-site analysis.

We identified 4 points for data collection along Rawal Lake. All of these points are far from tourist spots, have ample space for setting up the monitoring kits, and are connected closely to lake's water intake or release. These characteristics made these points optimal for data collection. There are three points, which are along the input streams of Rawal Dam, while one point is along the output stream coming from dam spillway. The images of points along streams are shown in Appendix. The representation of collection points on a map can be seen in Fig. 1.

During the development of this work package, the temporal changes in water quality from Rawal Lake were analyzed using satellite remote sensing. Additionally, for temporal analysis of the water quality parameters, the historic data was acquired from PCRWR and Rawal Lake Filtration Plant. The statistical analysis was performed to analyze the variation of the water quality parameters over time. The variations in the concentration of these parameters in terms of Water Quality Index (WQI) was compared to the acceptable limits under environmental standards defined by WHO.

Development of water quality monitoring wireless sensor network

This work involved setting up of wireless sensor network for data collection of the various water quality attributes for detecting the contamination of water at the selected sites. The IoT sensors are installed on the selected points on the Rawal Dam, pertaining to inlet and outlet streams. Each sensor node is be equipped with several water quality sensors as mentioned earlier along with the GSM communication module. These monitoring nodes are connected to a localized gateway by communicating through Arduino as a micro-controller. The Arduino is programmed to send the parameter readings to a central cloud server. Machine learning and data analytic are applied to the recorded data, and heat maps are generated accordingly. The system architecture showing the organization of various components and their interconnections are shown in Fig. 2.

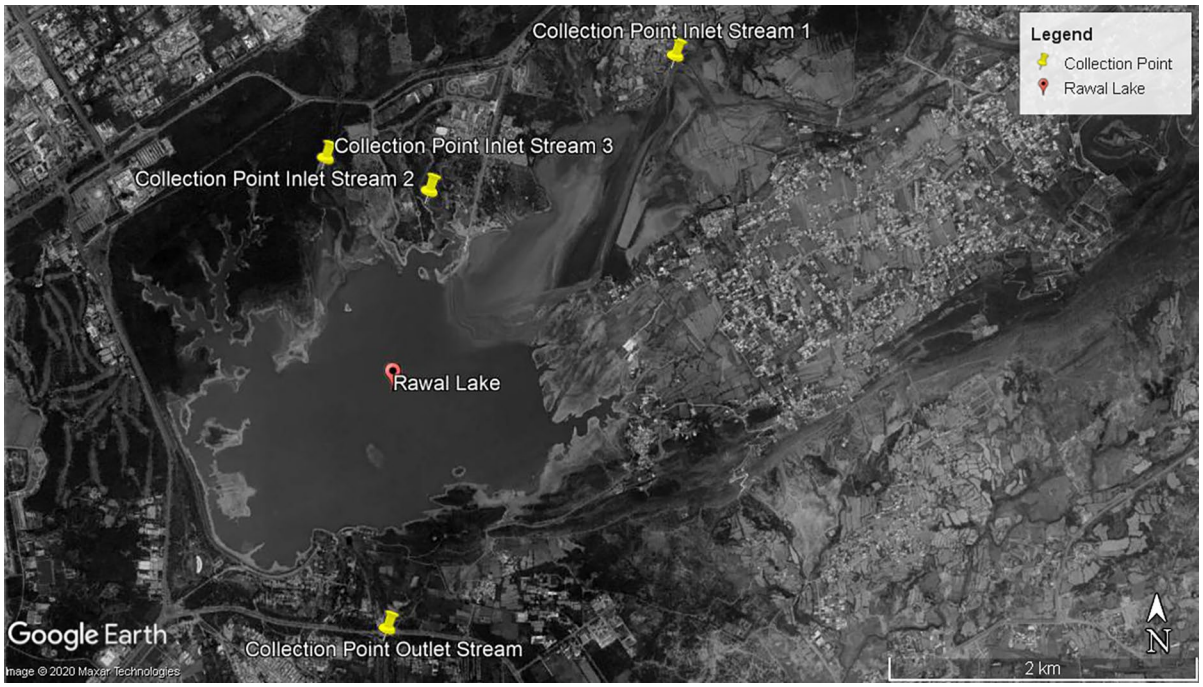


Fig. 1 Map showing data collection points of study area

WQI-based classification of water quality monitoring

The data gathered from IoT sensors at the centralized server is integrated with the historic data collected from alternate sources such as the Rawal Lake Filtration Plant. This data is used to classify water quality data based on water quality indices.

The historical data contains surface water quality parameters and bacterial population estimates. This data is split and an ANN along with other regression models are trained to calculate quantity of bacteria. TODO: check if required later Subsequent to classification, heat maps are generated using the data. For this purpose, a color-coding scheme is used to display the quality of water in terms of WQI at different points of the study area, wherein red represents the data points beyond safe limits and green represents data points within safe ranges. The data is mapped to classes of water quality as specified in the literature (Dascalescu et al., 2017), and in Table 1. The water quality classification map with color codes is also shown in Fig. 6.

Later on, a predictive model is developed based on machine learning regression analysis to forecast the water quality based on fecal coliform bacteria.

Data visualization and analysis

A Web portal is developed to display the data transmitted by IoT nodes to the server. The Web portal consists of a dashboard of services featuring several visual elements. The portal uses all the data transmitted by IoT nodes, as well as historic data obtained from other sources, to visualize and analyze the water quality in the form of info-graphics. These include annotated charts for comparative analysis, data sheets for visualization of historic and real-time water quality data, and color-coded heat maps (based on WQI) to visualize the Rawal lake streams. The WQI is used to classify different data points into six distinctive classes as shown in Table 1. The map is capable of displaying data sources and parameters. The radius of the data points shows the range of data collection performed near the inlet and outlet streams of Rawal Lake. The snippets from the Web portal can be seen in Figs. 3, 4, 5, and 6.

Bacterial prediction

In order to analyze the quality of water, the data has been collected from two sources including historic

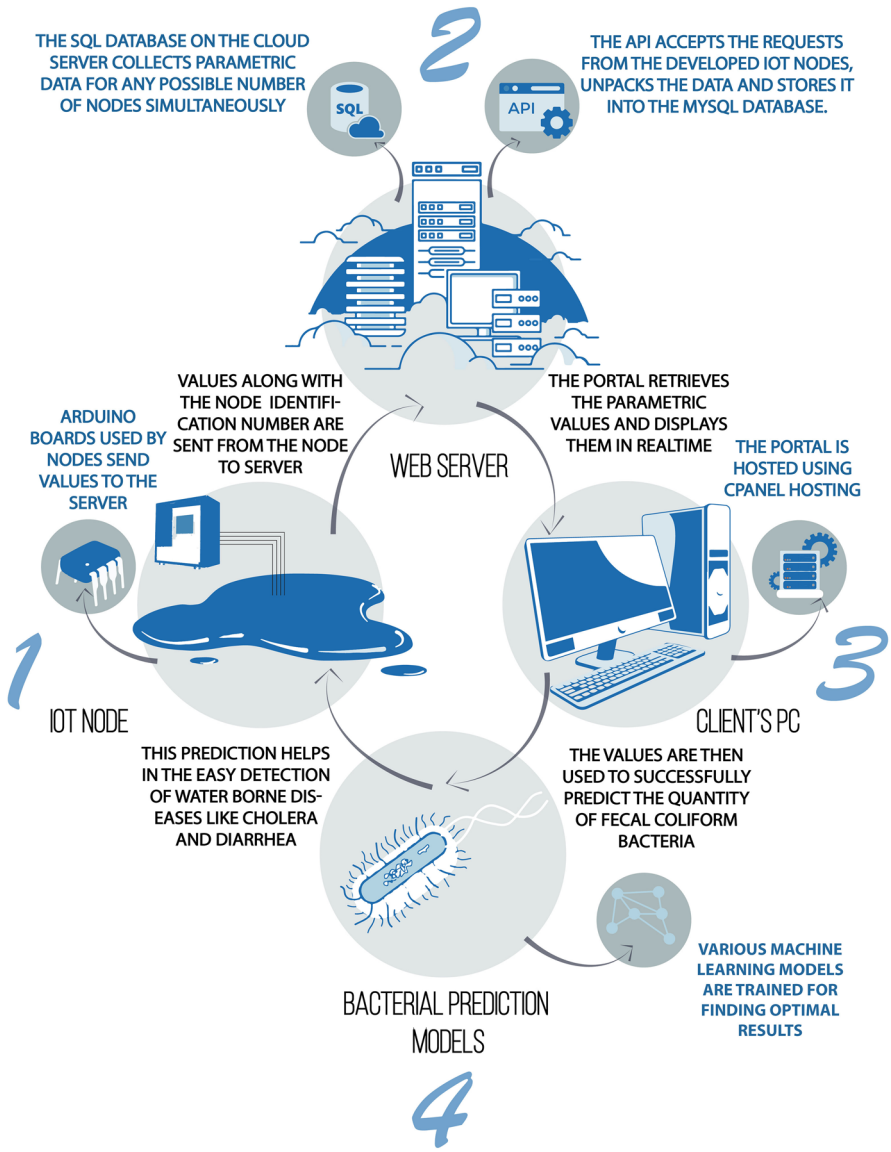
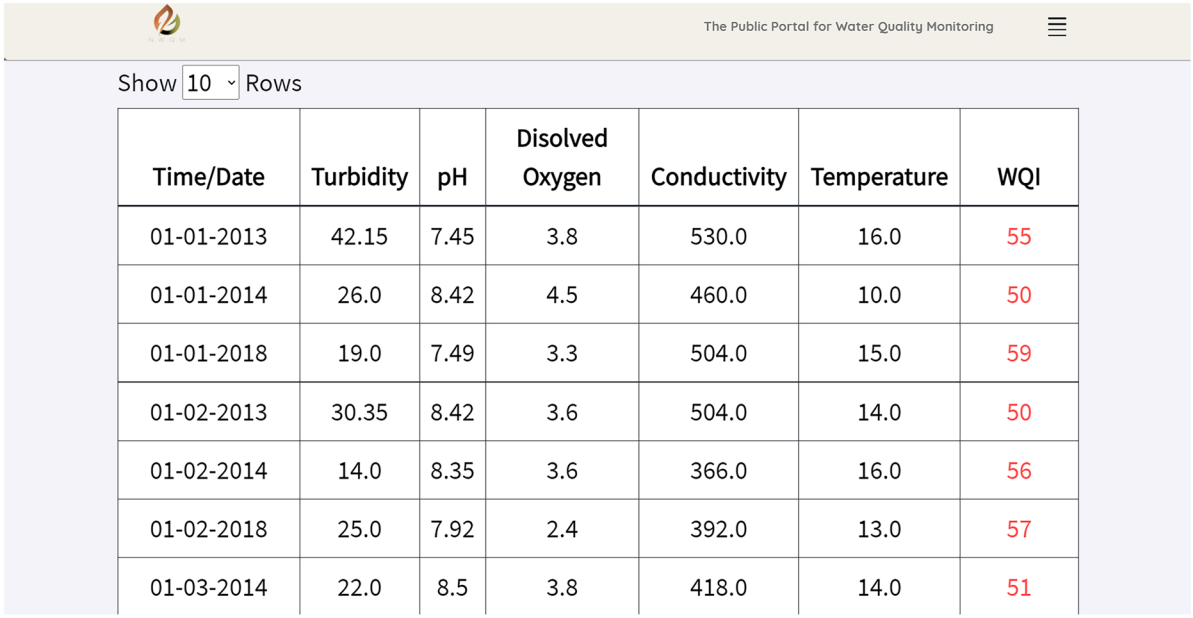


Fig. 2 System architecture showing the organization of the components of the proposed system

data and real-time data. As discussed earlier, the historic data is collected from the Rawal Lake Filtration Plant which comprises 12 parameters. However, the data obtained from IoT nodes consists of 5 parameters. In this way, the historic data allowed us to perform a more detailed analysis by combining it with

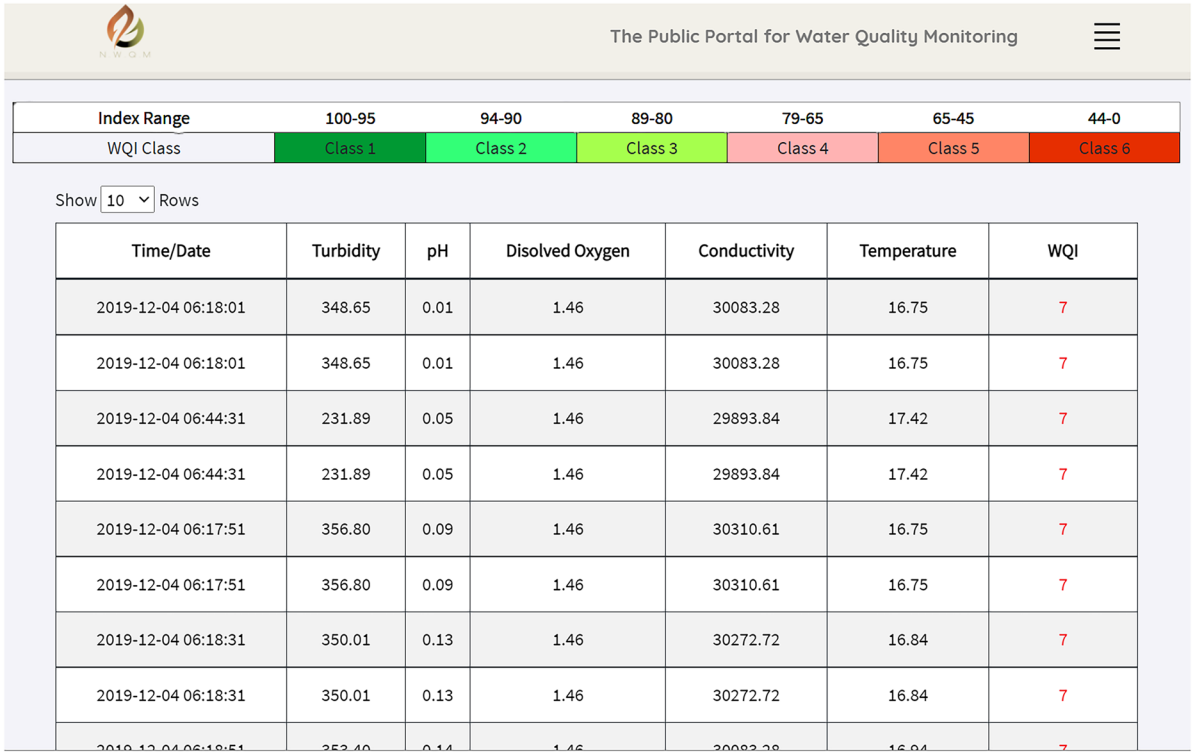
Table 1 Water quality classes

Class	WQI	Description
Class 1	100-95	Excellent water quality
Class 2	94-90	Very good water quality
Class 3	89-80	Good water quality
Class 4	79-65	Medium water quality
Class 5	64-45	Polluted water
Class 6	44-0	Very polluted water



Time/Date	Turbidity	pH	Disolved Oxygen	Conductivity	Temperature	WQI
01-01-2013	42.15	7.45	3.8	530.0	16.0	55
01-01-2014	26.0	8.42	4.5	460.0	10.0	50
01-01-2018	19.0	7.49	3.3	504.0	15.0	59
01-02-2013	30.35	8.42	3.6	504.0	14.0	50
01-02-2014	14.0	8.35	3.6	366.0	16.0	56
01-02-2018	25.0	7.92	2.4	392.0	13.0	57
01-03-2014	22.0	8.5	3.8	418.0	14.0	51

Fig. 3 Data sheets showing historic data collected from the Rawal Lake Filtration Plant



Time/Date	Turbidity	pH	Disolved Oxygen	Conductivity	Temperature	WQI
2019-12-04 06:18:01	348.65	0.01	1.46	30083.28	16.75	7
2019-12-04 06:18:01	348.65	0.01	1.46	30083.28	16.75	7
2019-12-04 06:44:31	231.89	0.05	1.46	29893.84	17.42	7
2019-12-04 06:44:31	231.89	0.05	1.46	29893.84	17.42	7
2019-12-04 06:17:51	356.80	0.09	1.46	30310.61	16.75	7
2019-12-04 06:17:51	356.80	0.09	1.46	30310.61	16.75	7
2019-12-04 06:18:31	350.01	0.13	1.46	30272.72	16.84	7
2019-12-04 06:18:31	350.01	0.13	1.46	30272.72	16.84	7

Fig. 4 Data sheets showing real-time data collected using our IoT nodes

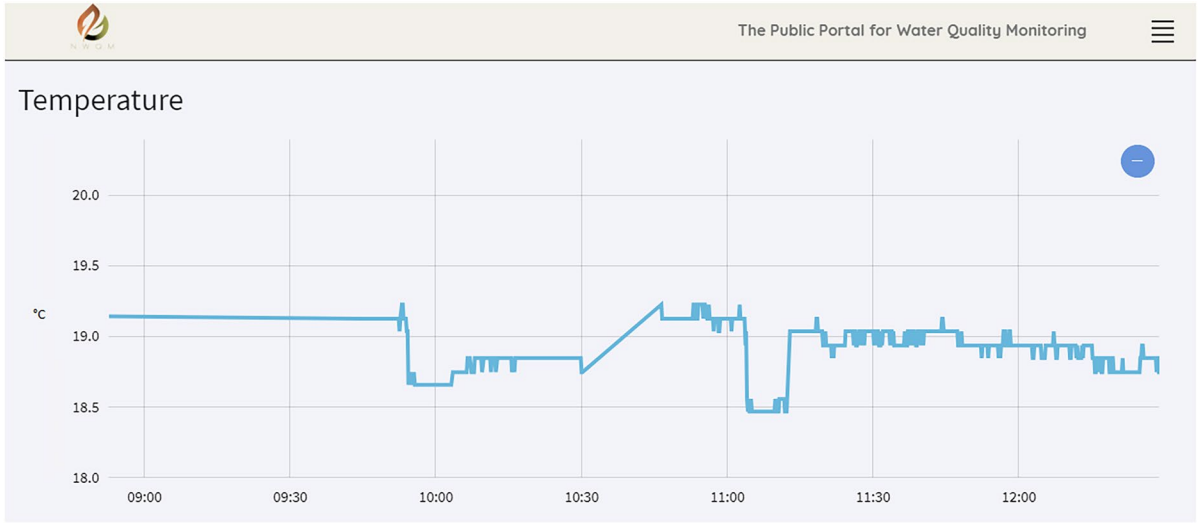


Fig. 5 A sample data chart on Web portal showing temperature values

real-time data. In the historic data, one of the most important water quality parameters, fecal coliform, is also included. Fecal coliform is a common bacteria found in the feces of humans and animals. The presence of this bacteria in water directly relates to risks of water-borne diseases such as diarrhea, cholera, and typhoid.

For the purpose of predicting bacteria, common parameters of the historic data and real-time data are used to create regression models. The 5 overlapping parameters are provided as input to the models and the number of fecal coliform bacteria is projected as the output. The dataset formed using these parameters is normalized or scaled to bring the data to a common range. While feature scaling or normalization simplifies the computation, at the same time it makes the data lose its true representation and meaning owing to data normalization in the range of 0 to 1. To address this, the results generated by the model are later de-normalized before analyzing the model outputs. The resulting dataset has 7 variables that are time stamp, temperature, turbidity, pH, conductivity, dissolved oxygen, and fecal coliform, while it has 1114 records of data as can be seen in Table 2.

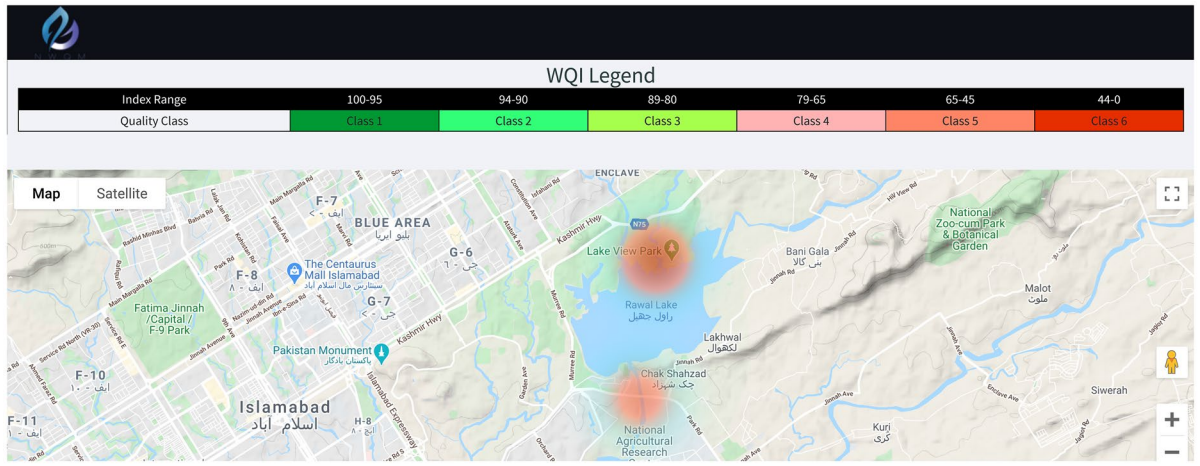


Fig. 6 Map showing water quality geographically

Table 2 Overview of water quality dataset

	Time Stamp	Temperature	Turbidity	pH	Conductivity	Dissolved Oxygen	Fecal Coliform
0	16-04-2012	23.0	18.00	7.19	736.0	3.5	170.0
1	01-01-2013	16.0	42.15	7.45	530.0	3.8	53.0
2	02-01-2013	16.0	46.70	7.99	538.0	2.5	63.0
3	03-01-2013	16.0	47.15	8.05	496.0	2.3	55.0
4	04-01-2013	10.0	22.00	8.18	398.0	4.6	57.0
...	...	...	...	...	...	...	...
1109	30-12-2018	22.0	14.00	7.45	460.0	4.9	59.0
1110	31-12-2018	21.0	16.00	7.79	590.0	5.6	72.0
1111	09-05-2019	21.0	11.00	7.24	496.0	3.3	140.0
1112	07-07-2019	24.0	22.00	7.25	520.0	6.6	122.0
1113	17-09-2019	22.0	13.00	7.18	370.0	3.2	133.0

[1114 rows x 7 columns]

For predictive analytic, multiple models are trained using different regression techniques including neural networks (NN), convolutional neural networks (CNN), support vector machines (SVM), ridge regression (RR), decision trees regression (DTR), and Bayesian regression (BR). The primary purpose of training multiple models is to have a comparative analysis for determining the most optimal model for our dataset. These techniques are most commonly used and are state-of-the-art for regression problems, which is the reason for their selection. The details of these models along with their configuration details are mentioned below:

Neural networks (NN) The NN are well-known for learning relations between dependent data and is useful for providing a model for predicting output values on unseen data. The NN model used for this research work takes 5 parameters as input namely turbidity, temperature, conductivity, dissolved oxygen, and pH as described before, while the quantity of fecal coliform is predicted as the output of the network. The NN architecture for bacterial prediction is shown in Fig. 7. The network contains 3 layers of neurons with 2 hidden layers. A smaller number of layers are chosen to ensure that the network is not very deep and complex to avoid over-fitting on training data. The two hidden layers of the network have 200 and 100 neurons respectively. The number of neurons is high to encourage maximum feature learning by a shallow neural network. Adam's optimizer is used with a learning rate of 0.003, while the model is trained for 5000 epochs.

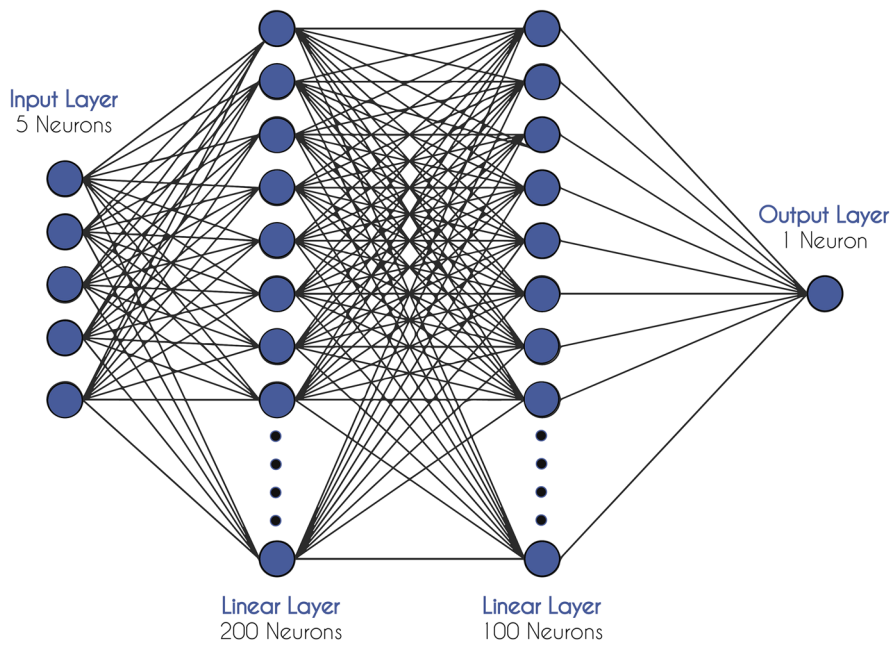


Fig. 7 Neural network architecture for bacteria prediction

Convolutional neural networks (CNN) The CNNs are used as state-of-the-art algorithms in the domain of computer vision. However, the performance of CNNs in finding patterns in data and identifying important intermediate features to help in the prediction of better results makes them a suitable candidate for trials with regression problems. As our data is one-dimensional data, so one-dimensional convolutions are used in the CNN. Additionally, the designed architecture contained a very small number of layers as our data is prone to over-fitting especially when applying CNNs. For training the model, Adam's optimizer was used with a learning rate of 0.003, while the model was trained for 1500 epochs. The architecture of the CNN used for this purpose is shown in Fig. 8.

Ridge regression (RR) A model for ridge regression was also trained. This technique is used when there is a chance of input parameters having correlations with each other (multi-collinearity). This technique was used due to the fact that some of our input parameters are interdependent such as pH and conductivity, dissolved oxygen, and temperature. The training of the model was done while keeping the alpha value of the model equal to 0.001.

Decision trees regression (DTR) The decision trees are normally used for classification problems; however, it can also be used for regression problems. They are used where the input data has relation to the output which can be determined by a set of rules. The model was trained keeping the max depth of the decision tree to infinite.

Bayesian regression (BR) The BR is an approach to linear regression having the advantage of getting a whole range of inferential solutions rather than point estimates. This technique was applied to our data to get the results of linear regression and compare it with others.The model was trained for 300 iterations while keeping alpha-1, alpha2, gamma-1, and gamma-2 all set to 10^{-6}.

Ensemble of all approaches Ensemble is used to reduce the generalization error of prediction. It was applied to all models described above to get the benefit of all of them and reduce the errors.

Results and discussion

As discussed in Section 3.5, multiple regression models are trained to identify the optimal model for bacterial prediction. The NN model after training has shown a high mean squared error (MSE) of 0.77845, a high mean average error (MAE) of 45.85944, and R^2 of 0.50493 on the test set. The model plotted using the test set is shown in Fig. 9. It can be seen that the regression line closely follows the trend of the actual data points despite having a large MSE value. Considering this, R^2 is misleading in this case where the predictions computed on the unseen data seem promising.

The CNN model has a high MSE of 0.63301, a high MAE of 41.54191, and an R^2 of -1.50159. The model plotted using the test data is shown in Fig. 10. It can be observed that the CNN model is showing regression results on test data with

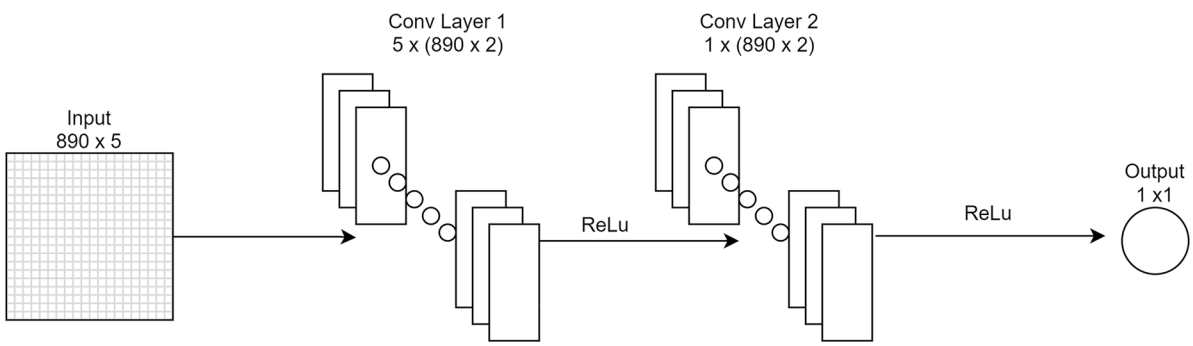


Fig. 8 CNN architecture for bacterial prediction

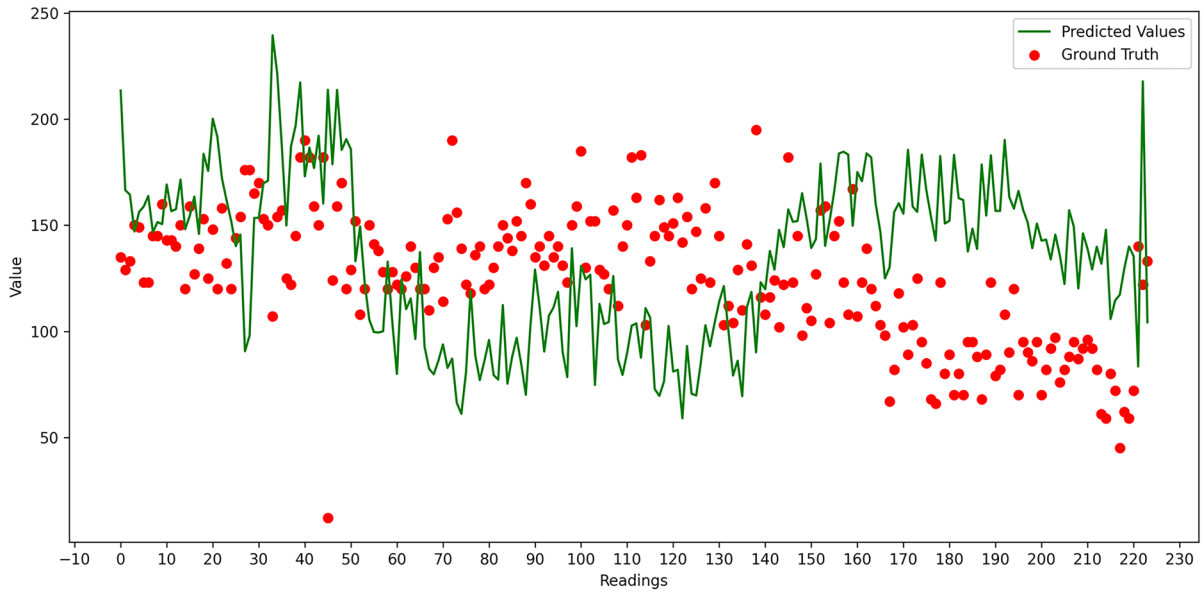


Fig. 9 NN regression results for fecal coliform prediction

a relatively poor R^2 score than NN model. The regression line moves away from a major chunk of ground truth values and weakly approximates the others.

The SVM model has a low MSE of 0.35575, a low MAE of 30.76134, and an R^2 of −0.40588. The model plotted using the test data is shown in Fig. 11. It can be observed that the output of the SVM model is showing the prediction of data points with a low error. The R^2 score is negative in this case which may be attributed to the predictions computed on the unseen data or indicates a poorly fitted model. However, R^2 is

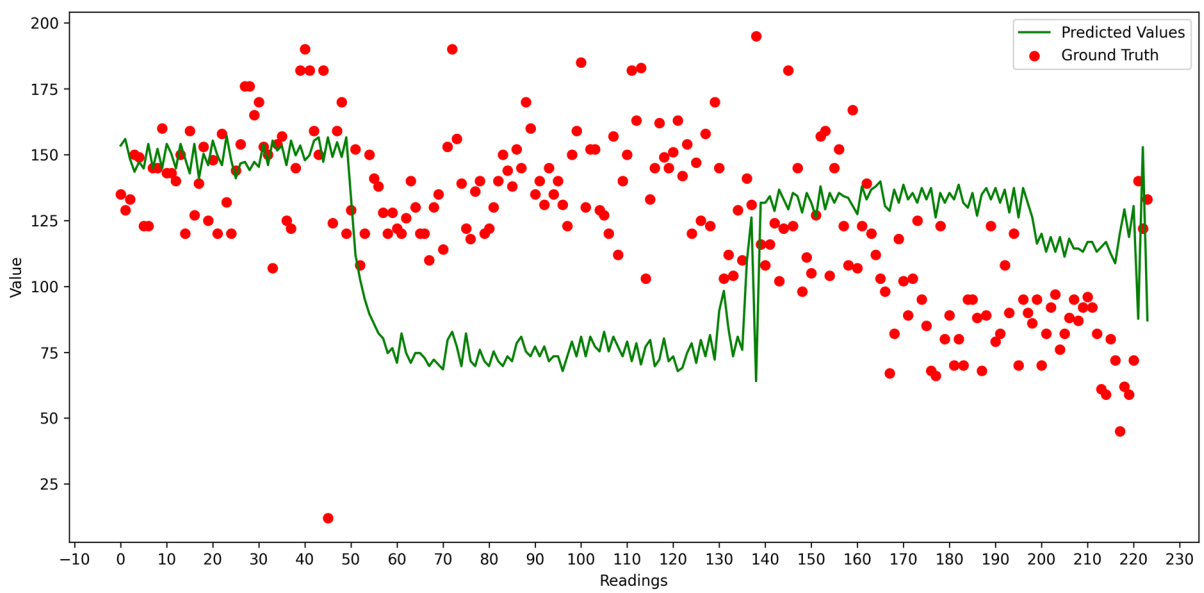


Fig. 10 CNN regression results for fecal coliform prediction

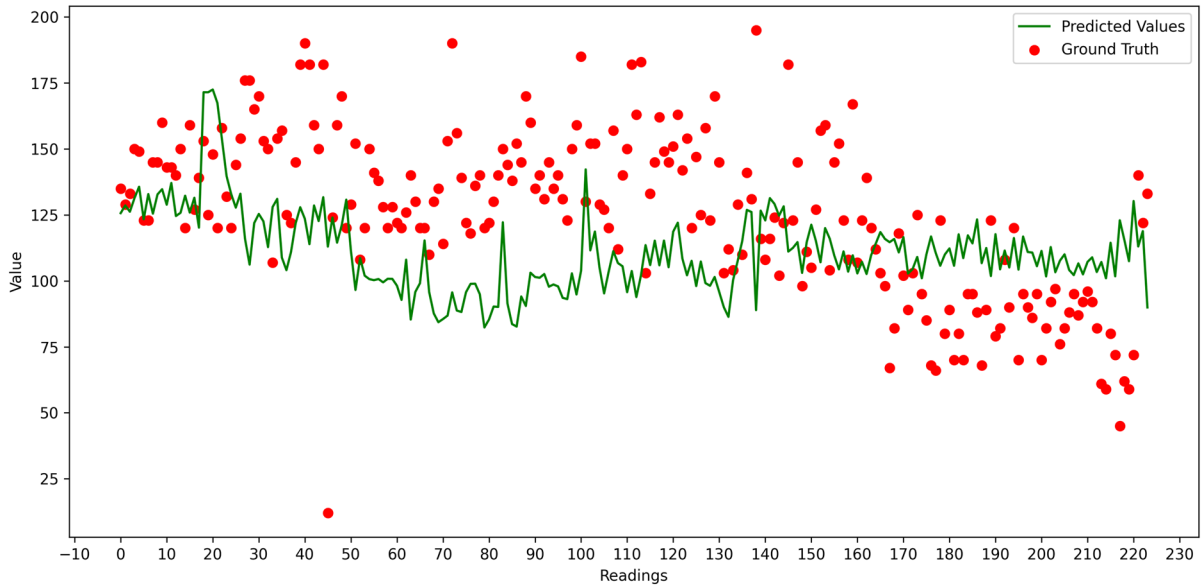


Fig. 11 SVM regression results for fecal coliform prediction

misleading here in measuring the performance of the model, where the regression line is passing close to the ground truth and projects the bacteria values fairly accurate except for the outliers. Hence, the regression line of the SVM model indicates the good performance of the model and looks generic and closer to actual data points.

The RR trained model has an MSE of 0.40282, an MAE of 32.63781, and an R^2 of −0.59190. The model plotted using the test set is shown in Fig. 12. It can be seen that the regression line of RR fits the ground truth data points quite well despite having a negative R^2. The plotted regression line indicates that RR model performs well at predicting the output.

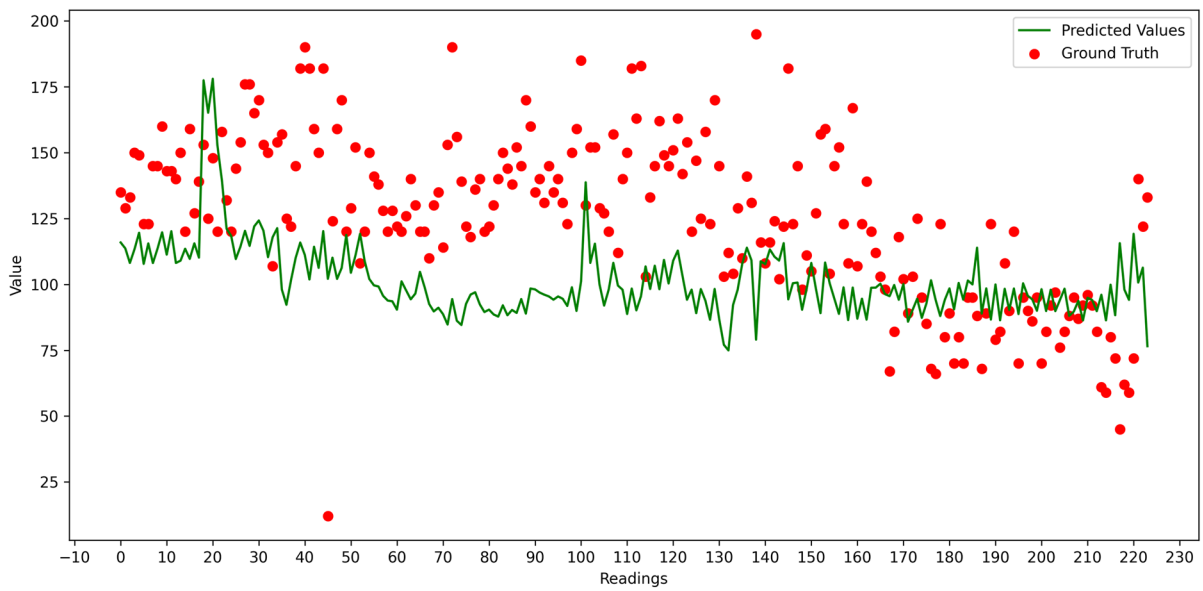


Fig. 12 RR results for fecal coliform prediction

The resulting model of DTR returned an MSE of 1.06162, an MAE of 53.36011, and R^2 of −3.19540 on the test data. The model plotted using test set can be seen in Fig. 13. The regression line shows that the model does not fit well to the ground truth values. DTR model trains well when there are a set of concrete rules which stem from the input parameters and propagate to the output. The results indicate that no concrete set of steps can be learned from our dataset to have a suitable DTR model. The DTR model performed the worst compared to all the other models in terms of both MSE and R^2.

The model resulting from BR has an MSE of 0.39566, an MAE of 31.25342, and an R^2 of −0.56363 on the test data. The model plotted using the test set is shown in Fig. 14. Despite having a negative R^2 value, the regression line fits well to the ground truth values. However, it has missed several data points, exhibiting the non-flexible nature of this model. The MSE of BR is the lowest after SVM. This reveals that the BR model can be ranked second for showing promising results with low MSE.

The model resulting from the ensemble of all approaches has an MSE of 0.43834, an MAE of 35.22278, and an R^2 of −0.73229. The model plotted in Fig. 15 has shown a good fit on data points of the test set despite having negative R^2. The MSE of the model is better than DTR, CNN and NN.

A comparison of the evaluation metrics of all the models trained is listed in Table 3.

It is pertinent to mention here that the conventional range of R^2 metric is between 0 and 1. However, the above regression models showed values beyond this range for test-set evaluations. R^2 is a common metric used to evaluate regression models. The metric helps us to compare our model with a constant baseline to assess how good the model is. Conventionally, R^2 has a range of 0 to 1, but it can go negative if a model is tested using out of sample data. This explains that the negative R^2 values in our test-set evaluations are due to the fact that the regression line is drawn on the unseen data and R^2 is calculated on the data points on which the model is not actually trained (Chiu, 2019). R^2 is the measure of the quality of the fitted line and a negative R^2 also means that the selected model poorly fits the data and is not suitable for the context of the application.

It is also a fact that R^2 has limitations and it is not always useful for all kinds of data-sets. It is strongly tied up with the quality and size of the data and cannot be used to determine whether the model predictions are biased. R^2 does not indicate if a regression model provides an adequate or good fit to your data. It is possible that a good model can have a low R^2 value (as discussed above for the case of SVM, BR, RR, and ensemble model) whereas a biased

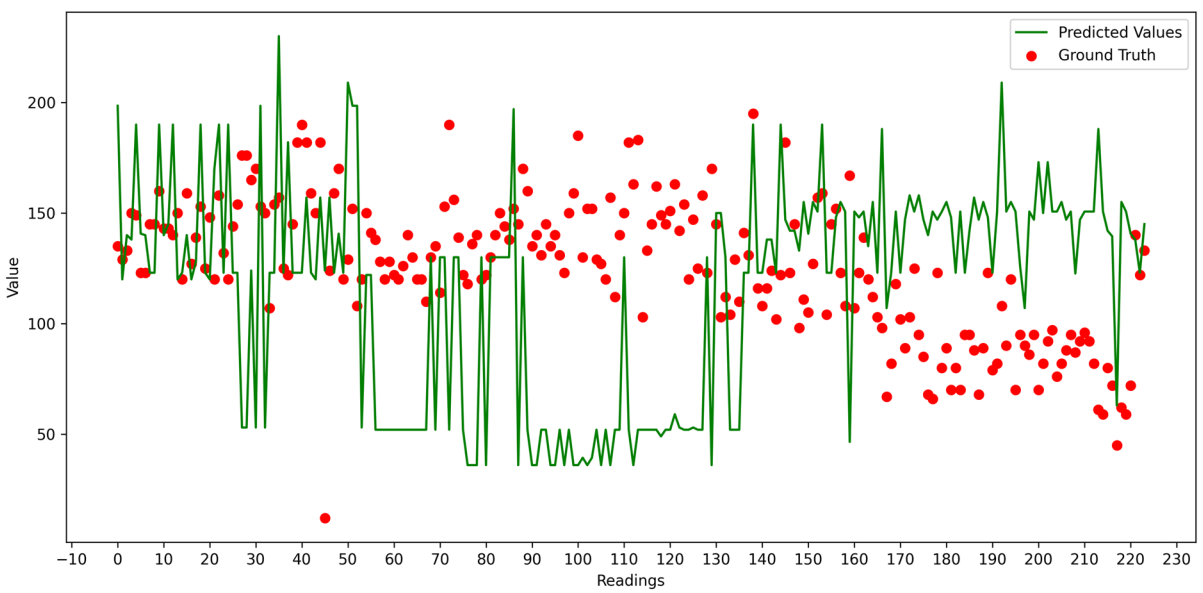


Fig. 13 DTR results for fecal coliform prediction

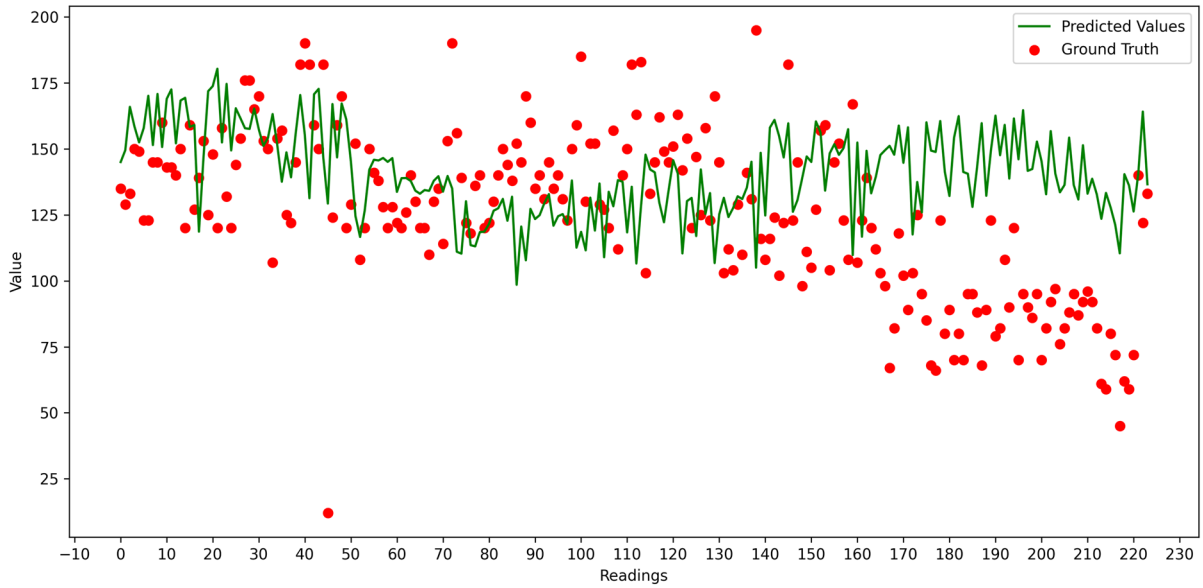


Fig. 14 BR results for fecal coliform prediction

model can have a high R^2 value (Spanos, 2019; Frost, 2018; Clay, 2015). R^2 does not explain and suggest anything about the future predictive performance, whereas MSE does and is considered as a good measure of prediction. Also, it must be noted that MAE results are directly proportional to the results shown by MSE. Considering these facts, the models' performance has been evaluated based only on MSE and it can be concluded from Table 3 that SVM and Bayesian regression models have shown the optimal performance with MSE of 0.35575 and 0.39566 respectively.

On the other hand, the algorithms such as neural networks and convolutional neural networks have

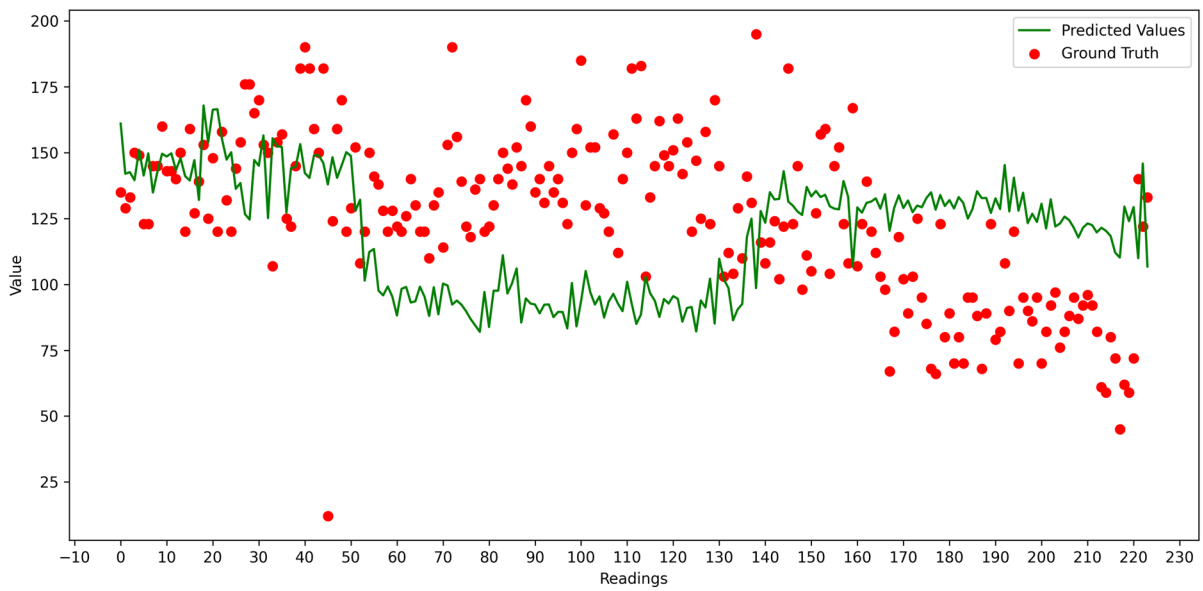


Fig. 15 Ensemble results for fecal coliform prediction

Table 3 Performance evaluation of models

Model	MSE	MAE	R^2 (Test)	R^2 (Train)
3-layer neural network	0.77845	45.85944	0.50493	0.43053
Convolutional neural network	0.63301	41.54191	−1.50159	−0.05989
Support vector machines	0.35575	30.76134	−0.40588	0.25756
Ridge regression	0.40282	32.63781	−0.59190	0.00747
Decision tree regression	1.06162	53.36011	−3.19540	0.99994
Bayesian regression	0.39566	31.25342	−0.56363	0.46823
Ensemble of all approaches	0.43834	35.22278	−0.73229	0.56926

shown high MSE, where decision tree regression has shown the worst results compared to other models.

A possible explanation of SVM and Bayesian regression performing better than all the other algorithms is that both SVM and Bayesian regression are less prone to overfitting on small datasets relative to other algorithms. For this reason, they performed better than the other algorithms owing to a small dataset. However, both ANN and CNN are the algorithms that are more suitable for a dataset of large sizes; therefore, they are more prone to over-fitting on the small datasets. This could be the reason that they did not perform well on our dataset. Also, the decision tree regression might be performing worse on our dataset because it is more suitable for discrete data, and for classification problems; therefore, it is not able to provide a good fit to our dataset.

Conclusion

The deteriorating water quality is increasingly becoming a serious concern for the world today. While freshwater sources are rapidly depleting, the pollution is ascending and making its way into almost all water bodies. Despite being a global issue, the solutions available to monitor and control water pollution are limited and no proper mechanism to monitor the water quality and mitigate the sources of contamination is available at a local scale. To address this issue, a system for monitoring and reporting the water quality in real-time is proposed and developed in this research work. The developed system is used to collect data for water quality parameters at Rawal Dam in real-time and the historical data of water quality is also obtained from the organizations such as the Rawal Lake Filtration Plant. This data is further used to calculate the water quality index. The historical data is used to train a number of models to predict the number of a specific type of bacteria, called fecal coliforms. Six different models are trained for predicting the quantity of this bacteria which is substantial to estimate the risks to public health. An ensemble of these models is also formed. Among these models, SVM and BR models have shown the optimal performance with MSE of 0.35575 and 0.39566 respectively. The most useful application of the bacterial prediction model is the projection of healthcare risks in advance. The bacteria are not monitored in real-time and generally require lab analysis to compute its quantity. However, the developed system provides an inexpensive way of predicting the quality of fecal coliform bacteria from the real-time data transmitted by the IoT nodes. This can help the system generate alerts and early warnings to trigger remedial actions for arresting the contamination and inhibiting its further propagation. Additionally, this would help to determine the risks of onseting specific diseases caused by this type of bacteria. The proposed system provides an alternative and more convenient solution for bacterial prediction, which otherwise is done manually in labs and is an expensive and time-consuming approach. The proposed solution offers several advantages including remote monitoring, ease of scalability, real-time status of water quality and subsequent bacterial prediction, portable hardware, etc.

Limitations and future work

The primary limitation of this work is that the data obtained by IoT nodes is point data, which was collected by deploying these nodes at the banks of Rawal Lake. That means that the data of water at the center of the lake and at water distribution points were not recorded. Primarily, this was due to these points being difficult to access. Another limitation is that the dataset used for training the bacterial prediction model is not large enough to train complex neural networks and

any other algorithms that require a large dataset. In the future, data recording techniques such as water imagery, water vehicles, and GIS (Geographic Information System) will be explored to be able to record more data and ultimately enhance the performance of the prediction models. Additionally, the role of other data parameters related to water quality and bacteria will be investigated, which may enhance the performance of the proposed work. Furthermore, more data will be collected from the Rawal Lake Filtration Plant to enhance the performance of the predictive models trained as part of this work.

Acknowledgements This research work is conducted in NUST-SEECS, IoT Lab, Islamabad, Pakistan. We are indebted to the staff members of Rawal Lake Filtration Plant for extending their generous support in providing historic data.

Author contributions 1. Hamza Khurshid: conceptualization, methodology, validation, investigation, writing–original draft, writing–review and editing, visualization. 2. Rafia Mumtaz: conceptualization, methodology, validation, investigation, writing–original draft, writing–review and editing, visualization, supervision. 3. Noor Alvi: conceptualization, methodology, investigation, writing–review and editing. 4. Ayesha Haque: software, validation, data curation, writing–original draft, writing–review and editing, visualization. 5. Sadaf Mumtaz: methodology, software, validation, investigation, writing–original draft. 6. Faisal Shafait: investigation, writing–review and editing, visualization. 7. Sheraz Ahmed: data curation, writing–review and editing, supervision. 8. Muhammad Imran Malik: investigation, methodology, validation, writing–review and editing. 9. Andreas Dengel: conceptualization, validation, visualization, writing–review and editing

Funding Information This research work is funded by DAAD, Germany

Availability of data and materials The data can be made available by requesting the authors through email.

Declarations

Ethics approval The project does not involve any human or animal as the subject to be studied. Therefore, ethical approval is not applicable.

Consent to participate The project does not require any involvement of living beings, therefore the consent to participate is not applicable.

Consent to publish All the authors of the paper have given the consent to publish this work

Conflict of interest The authors declare that they have no conflict of interest.

References

Adamo, F., Attivissimo, F., Carducci, C. G. C., & Lanzolla, A. M. L. (2015). A smart sensor network for sea water quality monitoring. *IEEE Sensors Journal, 15*(5), 2514–2522.

Agarwal, A., Shukla, V., Singh, R., Gehlot, A., & Garg, V. (2018). Design and Development of Air and Water Pollution Quality Monitoring Using IoT and Quadcopter. *In Intelligent Communication, Control and Devices* (pp. 485-492). Springer, Singapore.

Ahmed, U., Mumtaz, R., Anwar, H., Mumtaz, S., & Qamar, A. M. (2020). Water quality monitoring: from conventional to emerging technologies. *Water Supply, 20*(1), 28–45.

Alamgir, A., Khan, M. A., Hany, O. E., Shaukat, S., Mehmood, K., Ahmed, A., & 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)* (pp. 108-113). IEEE.

Baig, S. A., Lou, Z., Baig, M. A., Qasim, M., Shams, D. F., Mahmood, Q., & Xu, X. (2017). Assessment of tap water quality and corrosion scales from the selected distribution systems in northern Pakistan. *Environmental Monitoring and Assessment, 189*(4), 194.

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.

Chiu, W. (2019). *When is R squared negative?* https://www.quora.com/When-is-R-squared-negative/answer/William-Chiu

Clay, F. (2015). *Is R-squared Useless?* University of Virginia Library. https://data.library.virginia.edu/is-r-squared-useless/

Das, B., & Jain, P. C. (2017). Real-time water quality monitoring system using Internet of Things. *In 2017 International conference on computer, communications and electronics (Comptelix)* (pp. 78-82). IEEE.

Dascalescu, I. G., Morosanu, I., Ungureanu, F., Musteret, C. P., Minea, M., & Teodosiu, C. (2017). Development of a versatile water quality index for water supply applications. *Environmental Engineering and Management Journal, 16*(3), 525–534.

Daud, M. K., Nafees, M., Ali, S., Rizwan, M., Bajwa, R. A., Shakoor, M. B., & Malook, I. (2017). Drinking Water Quality Status and Contamination in Pakistan. *BioMed Research International*.

Dogo, E. M., Nwulu, N. I., Twala, B., & Aigbavboa, C. (2019). A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water Journal, 16*(3), 235–248.

Ejaz, N. A. E. E. M., Hashmi, H. N., & Ghumman, A. R. (2011). Water quality assessment of effluent receiving streams in Pakistan: A case study of Ravi River. *Mehran University Research Journal of Engineering & Technology, 30*(3), 383–396.

Faruq, M. O., Emu, I. H., Haque, M. N., Dey, M., Das, N. K., & Dey, M. (2017). Design and implementation of cost effective water quality evaluation system. *In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 860-863). IEEE.

Frost, J. (2018). How To Interpret R-squared in Regression Analysis. *Statistics By Jim*. https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (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. J. S. W. (2016). Internet of things enabled real time water quality monitoring system. *Smart Water, 2*(1), 1–19.

Hisam, A., Rahman, M. U., Kadir, E., Tariq, N. A., & Masood, S. (2014). Microbiological contamination in water filtration plants in Islamabad. *Journal of the College of Physicians and Surgeons-Pakistan, 24,* 345–350.

Horton, R. K. (1965) An index-number system for rating water quality, *J. Water Pollut. Control Fed.,* 37, pp. 300–306.

Khatri, N., Sharma, A., Khatri, K. K., & Sharma, G. D. (2018). An IoT-based innovative real-time pH monitoring and control of municipal wastewater for agriculture and gardening. *In Proceedings of First International Conference on Smart System, Innovations and Computing* (pp. 353-362). Springer, Singapore.

Liu, J., Chen, H., Yao, L., Wei, Z., Lou, L., Shan, Y. & Zhou, X. (2016). The spatial distribution of pollutants in pipe-scale of large-diameter pipelines in a drinking water distribution system. *Journal of Hazardous Materials, 317*, 27–35.

Liu, L., Oza, S., Hogan, D., Perin, J., Rudan, I., Lawn, J. E., & Black, R. E. (2015). Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet, 385*(9966), 430–440.

Naqvi, J. (2017). *Exposure to pollution kills millions of children*. The Washington Post: WHO reports find https://www.washingtonpost.com/news/to-your-health/wp/2017/03/05/exposure-to-pollution-behind-millions-of-childrens-deaths-who-reports-find/.

Nelliyat, P. (2016). *Water pollution: extent, impact, and abatement. In Indian Water Policy at the Crossroads: Resources, Technology and Reforms* (pp. 131-151). Springer, Cham.

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

Saha, H. N., Auddy, S., Chatterjee, A., Pal, S., Pandey, S., Singh, R., & Maity, A. (2017). Pollution control using internet of things (IoT). *In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 65-68). IEEE.

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

Saravanan, M., Das, A., & Iyer, V. (2017). Smart water grid management using LPWAN IoT technology. *In 2017 Global Internet of Things Summit (GIoTS)* (pp. 1-6). IEEE.

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)* (pp. 92-96). IEEE.

Shapiro, J. (2013). *Chinese research perspectives on the environment*. D. Yang (Ed.). Brill.

Sithole, M. P. P., Nwulu, N. I., & Dogo, E. M. (2019). Dataset for a wireless sensor network based drinking-water quality monitoring and notification system. *Data in Brief, 27,*.

Spanos A. (2019). *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*. Cambridge University Press. p. 635

Tadokoro, H., Jp, P. E., Nakamura, N., Nishimura, T., Uemura, K., Kikuchi, N., & Hatayama, M. (2017). Monitoring and Control Systems for the IoT in the Water Supply and Sewerage Utilities. *Hitachi Review, 66,* 704–711.

Verma, A. K., & Singh, T. N. (2013). 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,* 583–588.

Vijayakumar, N., & Ramya, R. (2015). *The real time monitoring of water quality in IoTenvironment. In Circuit, Power and Computing Technologies (ICCPCT)*, 2015 International Conference on (pp. 1-4). IEEE.

World Bank. (2005). *Pakistan: Country Water Resources Assistance Strategy, Water Economy: Running Dry*. Washington DC. World Bank. https://openknowledge.worldbank.org/handle/10986/8343 License: CC BY 3.0 IGO.

WHO (1993). *Guidelines for drinking-water quality*: second edition. Geneva: World Health Organization. ISBN: 924154460O.

Zeitoun, M. M., & Mehana, E. E. (2014). Impact of water pollution with heavy metals on fish health: overview and updates. *Global Veterinaria, 12*(2), 219–231.