2017 IEEE Third International Conference on Big Data Computing Service and Applications

Data-driven Water Quality Analysis and Prediction: A Survey

Gaganjot Kaur Kang
Department of Computer Engineering
San Jose State University

Jerry Zeyu Gao*
San Jose State University, USA
Taiyuan University of Technology, China
Mail: jerry.gao@sjsu.edu

Gang Xie
Taiyuan University of Technology, China
Taiyuan University of Science and Technology, China

Abstract. Water quality becomes one of the important quality factors for the quality life in smart cities. Recently, water quality has been degraded due to diverse forms of pollution caused by disposal of human wastes, industrial wastes, automobile wastes. The increasing pollution affects water quality and the quality of people's life. Hence, water quality evaluation, monitoring, and prediction become an important and hot research subject. In the past, many environmental researchers have dedicated their research efforts on this subject using conventional approaches. Recently, many researchers begin to use the big data analytics approach to studying, evaluating, and predicting water quality due to the advances of big data applications and the availability of environmental sensing networks and sensor data. This paper reviews the published research results relating to water quality evaluation and prediction. Moreover, the paper classifies and compares the applied big data analytics approaches and big data based prediction models for water quality assessment. Furthermore, the paper also discusses the future research needs and challenges.

Keywords– Water quality evaluation, big data analytics, data-driven water quality evaluation, and water quality prediction.

I. INTRODUCTION

Water is one of the most essential natural resources for the existence and survival of the entire life on this planet. We use water for drinking, cooking, personal hygiene, agricultural practices, and recreational purposes almost every day. In addition, plants and animals also depend on water for their basic survival. In short, all living organisms need large quantity and good quality of water to continue their life. According to the World Health Organization [1], an estimated 1.1 billion people lacked access to clean drinking water and 2.6 billion lacked access to basic sanitation in 2005. Hence, 2005-2015 was declared as the International Decade for Action: "Water for Life". The increasing population, its automobiles and industries are polluting all the water-bodies at an alarming rate. The effluents added by these pollution sources affect the pH value of water. This gives rise to many water-related problems such as water-borne diseases in organisms and deaths of aquatic animals like fish, crab, and so on. These pollutions eventually disrupt the food-chain and damage ecosystem in long run. Hence, water pollution is one of the most alarming concerns for us today. Addressing this concern, people have spent lots of research efforts in water quality evaluation and monitoring. In the past decades, many researchers have spent lots of time on studying and developing different models and methods in water quality analysis and evaluation.

Water quality evaluation has been conducted using conventional approaches in all these years. These approaches involve manual collection and assessment of raw data. M. A. Tirabassi [2] formulated a mathematical model using mathematical statistics to predict river water quality without using the chemical, biological, and physical relationships. His work focuses on a "black box" approach where a known input can be used to accurately and reliably predict the output. L. Hu, et al. [3], has talked about Grey Relational Analysis (GRA) which is based on the distance of a point to the interval. It is a simple method which has been used for the assessment of drinking water quality in Jiaozuo River, China. However, according to H. Liao, [24], these traditional methods are confusing in nature and have several shortcomings like:

a) the existence of complex mathematical calculations
b) equal treatment to the old data and new data
c) difficult prediction due to overlap of multiple variables

In recent times, many researchers have developed or used big data analytics models and machine learning based models to conduct water quality evaluation in order to achieve better accuracy in evaluation and perdition. This paper is written based on our recent literature survey and study on the existing publications which focused on water quality evaluation and prediction using big data approaches. The major objective is to presents a snapshot of the vast research work on data driven approaches and a classified comparative analysis on applied big data analytics models and machine learning based prediction models and methods. It provides a useful review on the current state-of-the-arts on applicable big data approaches and machine learning techniques for water quality evaluation and predication.

As a survey paper on big data quality evaluation, it begins with a general introduction to the concerns with water quality, causes and effects of water pollution. In Section II, it presents an understanding on water quality evaluation standards and their need. It further discusses the quality of water by dividing it into two categories and presents a comparison of different research papers for each of those categories. Section III reviews and compares the existing traditional models while section IV reviews and compares big data analytics models and research work on water quality evaluation. Section V covers and compares the machine learning models and research work for water quality evaluation. Section VI summarizes and compares the strong and weak points of the existing research work and publication results on water quality assessment. Section VII discusses the future research needs and directions in big data based water quality evaluation and prediction, and conclusion remarks.

II. WATER QUALITY EVALUTION

Water quality evaluation is an important way to monitor and control water pollution. The characteristics of a water supply affects its suitability for a specific use. Water quality evaluation shows how well the quality of water can meet the requirements of the user. It is defined in terms of certain physical, chemical and biological characteristics. The objective of water quality evaluation is to check the quality of water to know if a given sample of water is suitable enough for a given purpose. These characteristics are traditionally collected manually from different water resources (i.e. lakes, rivers, and oceans), and assessed manually. For example, out of two given drinking water samples of equally good quality, users might give preference to one sample over the other because of taste. So, water taste becomes a quality evaluation parameter to evaluate water quality and its

978-1-5090-6318-5/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataService.2017.40

acceptability. R. Rosly, et al. in [4] has pointed out that data mining techniques can be used to improve water quality predictive accuracy. There are various factors that may affect the quality of water which include thermal pollution, acidification, salinization, ion toxicity etc. Many water resources lack basic protection. Therefore, these resources are vulnerable to pollution from factory farms and industrial plants. Hence, water quality evaluation becomes necessary and important to ensure the quality of the water environment.

According to D. Yang, et al. [5], the conventional methods for water quality evaluation can be classified into two classes:

- Single factor based methods
- Comprehensive index based methods

In a single factor based method, the most impaired water quality parameter is considered to evaluate water quality. It cannot reflect the comprehensive quality of water based on all factors. In a comprehensive index method, each parameter is considered to have an equal contribution towards determining quality of water even though this is not always true in practical situations. The researchers in [5] have constructed a two-level index system, whose first-class indexes consist of physical indexes, organic matter, heavy metal, nutrients, oils, radioactive material, and new toxic pollutants as shown in Table 1. The actual indexes considered under this broad category are mentioned in the second class indexes.

Table 1. Comprehensive water evaluation index and method system [5]

First-class indexes	Scond-class indexes	Evaluation standards	Evaluation methods
physical indexes	Tm, salinity, suspended matter, pH	Marine Water Quality Standard(GB3097-1997)	Exponential evaluation methods ， Fuzzy comprehensive assessment method ， Artificial neural network ， Support vector machine ， Projection pursuit
organic matter	DO, COD_{Mn}, BOD_5	Marine Water Quality Standard(GB3097-1997)	
heavy metal	Hg, As, Cu, Pb, Zn, Cd, Cr	Marine Water Quality Standard(GB3097-1997)	
nutrients	NO^-_2-N, NO^-_3-N, NH^+_4-N, TN, activated phosphate, TP, reactive silicate, Chlorophyll a	Marine Water Quality Standard(GB3097-1997)	
oils		Marine Water Quality Standard(GB3097-1997)	
radioactive material	^{60}Co, ^{90}Sr, ^{106}Rn, ^{134}Cs, ^{137}Cs	Marine Water Quality Standard(GB3097-1997)	
new toxic pollutants	Polycyclic Aromatic Hydrocarbons, Perfluorooctane sulfonates, Perfluorooctanoic Acid, Organic Tin 、 Tetrachlorodibenzo-p-dioxin, hexabromobipheny, Pentabromodiphenyl ether, Gammexane, Chlordecone, Octabromodiphenyl ether, Pentachlorobenzene, Chlorcosane	Toxicology experiment or literature research	

A. Water Quality Model Processes

Water quality models can be applied to many different types of water systems, including streams, rivers, lakes, reservoirs, estuaries, coastal waters and oceans. These models describe the main water quality processes and typically require the hydrological and constituent inputs (the water flows or volumes and the pollutant loadings). In one of the studies by D. P. Loucks [6], the designated use of water is considered a qualitative description of the desired condition of a water body. A criterion is a measurable indicator surrogate for use attainment. The criterion may be positioned at any point in the causal chain of boxes shown in Figure 1.

To define a good model, they have to consider the following factors in [6]:

a) **Mass-balance principle:** The basic principle of water quality models is that of mass balance. A water system can be divided into different segments called ‘computational cells’. For each segment or cell, there must be a mass balance for each water quality constituent over time.

b) **Design stream-flows:** In streams and rivers, the water quality may vary significantly, depending on the water flow. It is therefore common practice to pick a low-flow condition for judging whether or not water quality standards are being met.

c) **Temperature:** Temperature affects almost all water quality processes taking place in water bodies. Thus, it is an important factor while creating a water process model.

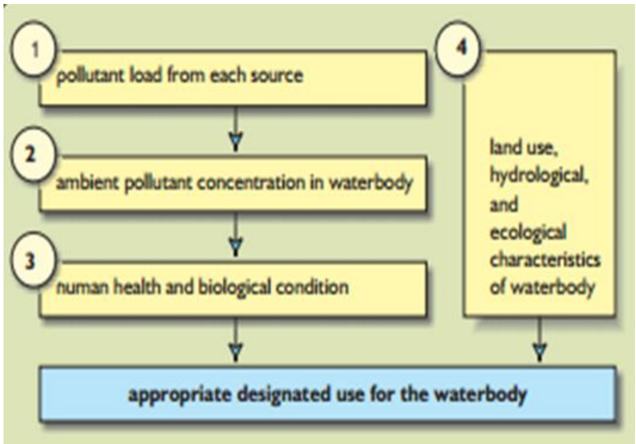
1
pollutant load from each source
2
ambient pollutant concentration in waterbody
3
human health and biological condition
4
land use, hydrological, and ecological characteristics of waterbody
appropriate designated use for the waterbody

Figure 1. Factors considered when determining designated use of water [6]

B. Water Quality Standards

The first step in water quality evaluation is to decide quality standards for a water body based on the desired uses of the water body (lake, stream, estuary etc.). In [6], D. P. Loucks defines “water designated use" as a term which refers to the most restrictive of the specific desired uses of a water body. Barriers to achieving the designated water use include the presence of pollutants and/or hydrological and geomorphic changes that affect the water quality. In addition, the standards set-up for upstream flow of water may also affect the uses of water downstream. To exemplify, small headwater streams may have aesthetic value but may not be able to support extensive recreational uses. However, their condition may affect the ability of a downstream area to achieve a particular designated use such as 'fishable' or 'swimmable'.

III. BIG-DATA WATER QUALITY ANALYSIS

These models use "Big Data" to optimally model water systems. The systems that are considered are highly dynamic, spatially expansive, and behaviorally heterogeneous. Nowadays, big data solutions have become efficient and receive more attention. According to a report from Aspen-Nicholas Water Forum [7], big data and the analysis creates opportunities to change the way the world looks at and acts on water. Sometimes there are mismatches between data needs and availability, such as discrepancies between the available and the desired levels of resolution. Key to making big data actionable is harnessing, standardizing, and integrating the enormous amount of data.

S. P. Sherchan et al. [8], has used integrated and intelligent sensors to operate in real-time, with the ability to recognize and diagnose daily minute water quality disturbances to monitor water quality. These sensors are integrated into contaminated water system to detect intentional or operational intrusion events, thus improving water security. Y. Liu et al. [9], quantify four parameters, namely, inorganic sediment particles, phytoplankton pigments, colored dissolved organic material, and Secchi disk depth of inland and near shore waters by means of remote sensing. It involved transformation of an image into water quality maps.

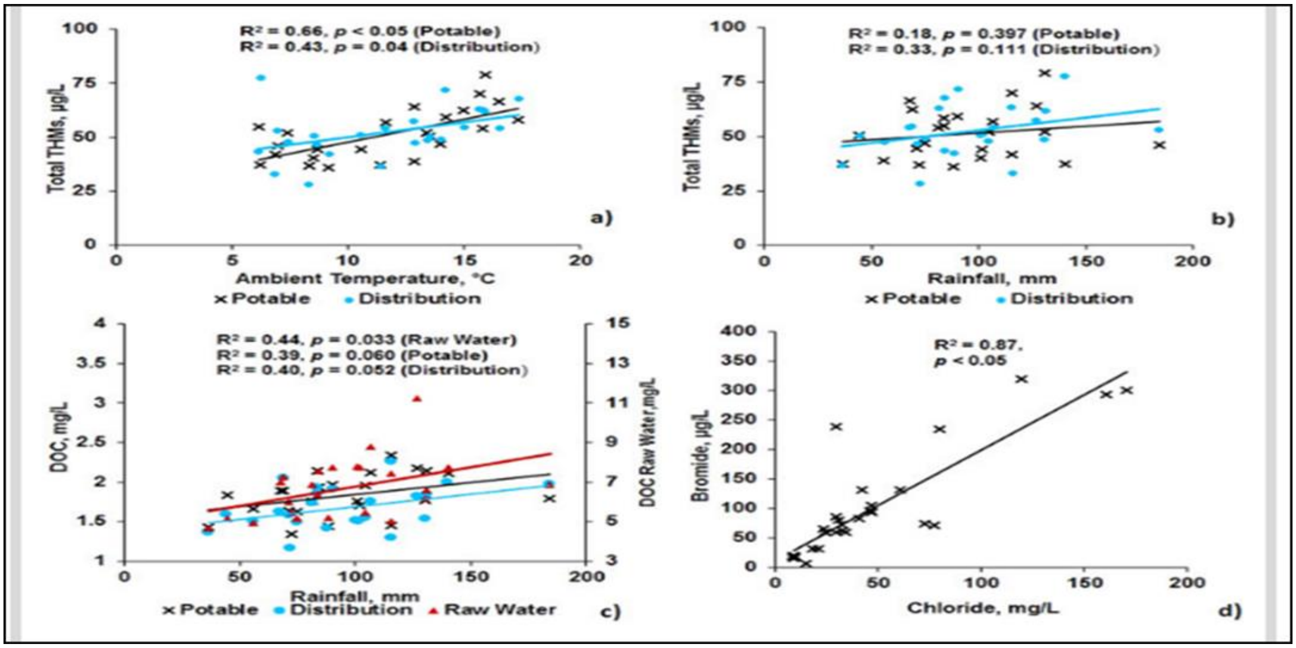

Figure 2. Correlation plots between monthly average concentrations (Jan. 2011- Jan. 2013) for (a) THMs and ambient temperature, (b) THMs and rainfall and (c) DOC and rainfall. (d) Correlations between bromide concentrations (Jan. 2011- Jan. 2013) [10]

M. Valdivia et al., [10] assessed ninety-three full-scale Scottish water treatment plants (WTPs) from Jan 2011 to Jan 2013 to identify factors that promote Trihalomethane (TMH) formation using Pearson's correlation analysis. Correlation analysis showed that ambient temperature was the primary THMs formation predictor in potable water ($r^2 = 0.66$, $p < 0.05$), while dissolved organic carbon ($r^2 = 0.55$, $p < 0.001$) and chloride (indicating marine influence; $r^2 = 0.41$, $p < 0.001$) also affected THMs formation. The results are graphically depicted in Figure 2. Y. Zhong et al. [11] presented a new big data processing algorithm based on fast fuzzy C-means (FCM) clustering for water quality analysis in Three Gorges Reservoir area with massive amounts of data collected by ultra-large-scale WSN. The water quality parameters considered were DO, COD, Mn, and NH3–N. In the FCM clustering algorithm, the researchers set an initial cluster center randomly. H. Jing-wei et al. [12] attempted to develop a better approach for spatial evaluation of drinking water quality. They used intelligent evaluation method integrating a geographical information system (GIS) and an ant colony clustering algorithm (ACCA) on drinking water samples from 29 wells in Zhenping County, China. 35 parameters on water quality were selected, such as chloride or sulphate concentration, total hardness etc.

M. Tarique [13] integrated sensors, transmitters, receivers, myRIO microcontroller, and IEEE 802.11 Wi-Fi technology to generate water quality data including pH, conductivity, and temperature. The real-time data was then sent wirelessly to a local control unit for analyzing, recording, and displaying. The system was also able to send alarm messages automatically to a remote management center when water quality falls below the required standard.

In one of the reserachs by T. C. Lobato et al. [14], the researchers constructed novel water quality index and quality indicator for reservoir water of Amazon region. Quality curves for each selected parameter were then created and the constructed QI and WQI were then applied to investigate the water quality at the reservoir. The hydrological cycle was shown by the indices to directly affect reservoir water quality, and the WQI was further useful in identifying anthropogenic impacts in the area. A. Newton et al. [15] collected historical data from the Ria Formosa lagoon and classified it according to the EEA 2001 guidelines. Water samples were significantly enriched in nitrogen (NH4+, NO2- and NO3-) with respect to the adjacent coastal waters indicating that inputs from sewage, agricultural runoff and benthic fluxes were not fully assimilated within the lagoon. Dissolved oxygen undersaturation (mean 75% during daylight hours) was associated with the area close to the sewage outlets of Faro. Table 2. summarises the results of all these big-data models and presents a snapshot of their accuracy and water parameters considered.

Table 2. A Comparison of Research Papers with Big Data Based Models

ID	PURPOSE AND AREA OF STUDY	BIG DATA MODEL	REAL-TIME MONITORING	WATER PARAMETERS	DATA-SOURCE	ACCURACY
[8]	Detection of Intentional Bacterial Spore Contamination of Potable Water	Real time water quality sensor model	Real-time sensing using BioSentry in-line sensor	turbidity, pH, temp., total organic carbon and conductivity	Deionized water as tap water using 1 µm pore size	detected *B. thuringiensis* spores with a detection limit of 102 spores/ml
[9]	Quantification of water parameters of inland and near shore waters by means of remote sensing	Remote sensing Model	Not real-time: Remotely sensed images	inorganic sediment particles, dissolved organic material, Secchi disk depth etc.	collected using Geographical Information System	Remotely sensed reflectance values have average *R*2 of 0.81 for Landsat data
[10]	detect and analysis of Trihalomethanes in drinking water in Scotland	Pearson's coefficient Analysis Model	Supports 24 hr. real time monitoring	Temp., dissolved organic carbon, chloride	93 water treatment plants assessed from 2011-'13	Temperature (r2 = 0.66, $p <$ 0.05)
[11]	To process big data for water monitoring of Three Gorges Reservoir area	Fast fuzzy C-mean clustering	Real-time wireless monitoring	DO, CO3, NH3-N	Data collected by ultra-large scale WSN	The water in which DO is higher than 11mg/L is best quality
[12]	Spatial quality evaluation of drinking water	GIS and Ant-colony algorithm	Supports 24 hr. real time monitoring	chloride or sulphate concentration, total hardness etc.	Data from 29 wells in Zhenping County, China	more than 95% area has the same water quality, which proves that the ACCA is feasible
[13]	To monitor water quality from a remote location with minimum supervision, initiating immediate corrective actions	Wireless System for water quality monitoring	Real-time data sent wirelessly	pH, conductivity, temperature	Water quality monitoring sensors	pH of water is almost constant at 8.8 till 11:45 a.m. The same rises to 10.00 at 12:15 p.m.
[14]	To construct a novel water quality index and quality indicator for reservoir water of Amazon region	Water Quality Indicator Model	Not real time	Physio-chemical parameters and metal concentration	11 water sample stations located upstream	determines seasonal differences in water quality with hydrological cycle
[15]	to evaluate the effect of water management of the environmental legislative Directives in Ria Formosa lagoon	EEA 2001 Guidelines	Not real-time	NH4+, NO2-, NO3-	Historical data for the lagoon	In the shallow west end of lagoon during summer, DO supersaturation reached 140% during day

IV. MACHINE-LEARNING PREDICTION MODELS

Machine learning (ML) is the branch of computer science which makes computers capable of performing a task without being explicitly programmed. There are many research papers that focus on classification of water quality evaluation using machine learning algorithms. Most of these articles use different scientific methods, approaches and ML models to predict water quality. S. Y. Muhammed et al. in [16] points out that machine learning algorithms are best suited for water quality prediction. Some of them are discussed below.

1. Artificial Neural Network (ANN) Model

Artificial neural Network model tries to simulate the structures and networks within human brain. The architecture of neural networks consists of nodes which generate a signal or remain silent as per a sigmoid activation function in most cases. A. Sarkar et al. in [17] points out that the ANNs are trained with a training set of input and known output data as shown in Figure 3. For training, the edge weights are manipulated in order to reduce the training error. One common training strategy is back propagation network with two hidden units.

One of the studies carried out by Y. Khan [18] aims to develop a water quality prediction model for four parameters, namely, dissolved oxygen, chlorophyll, specific conductance and turbidity. This research project uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS) uses Artificial Neural Networks (ANN) with Nonlinear Autoregressive (NAR) time series model for efficient water quality prediction. The motivation behind using ANN goes to the fact that ANN is most suited for classification of complex datasets such as those of environmental processes. The model has the ability to efficiently describe the non-linear relationship of the complex water quality datasets. Moreover, it has strong adaptability to depict the changes that might occur in the water environment of a particular area. The specified monitoring station is a channel situated in the State of New York. Their method includes two steps listed below.

- For this study, data was divided into training data (60%), testing data (20%) and validation data (20%). As seasonal variation affects the time series forecasting, it was made sure that data for training, testing and validation is from the same or nearly similar seasons. For better analysis and results, data was scaled to fall between the ranges of [0,1].
- A test was conducted in order to forecast the selected water quality factors based upon their past values. In these tests, the input and output are represented by the values of same parameters at different times.

The proposed model comprising of ANN-NAR proves to a reliable one with the prediction accuracy indicating much improved values, with the lowest MSE being 3.7x10-4 for turbidity and the best Regression value for Specific Conductance (0.99).

One of the studies carried out by X. Li and J. Song [19] uses artificial neural network and Markov chain approach to develop a new hybrid methodology for predicting the biochemical oxygen demand which is the main indicator of water quality. ANN produces the primary values and then the results are modified by three regression methods using the Markov transitional probability matrices respectively. They use a 27-year water quality data set of Tolo Harbor, Hong Kong which only has a total of 439 samples to test our method.

Their study process includes the following steps:

- They use 351 records of water quality data of TM3 station in Tolo Harbor to train the ANN model to predict primary values and generally the results in test data set is not well fit with the measured values.
- Then the Markov chain module provides the state transition probabilities and here they use a second order Markov chain state transition matrix.
- The final predictions of BOD are given by combing the primary determinations by ANN and the state transition probabilities of these predicted values by Markov chain approach through the specific regression methods.

The results are validated and a better prediction accuracy of the new ANN-Markov Chain Methodology is demonstrated through three criteria.

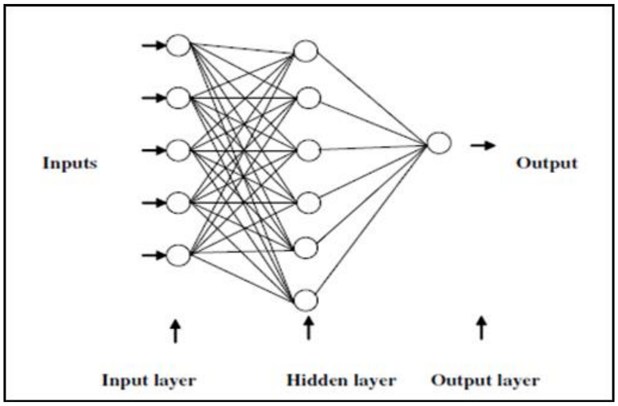

Figure 3. Artificial Neural Networks [17]

2. Radial basis Function Network (RBFN) Model

Radial basis function network model uses radial basis functions as activation functions as shown in Figure 4. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. As indicated by L. Ma et al.in [20], RBFN networks enjoy the best approximation property among all feed-forward networks. These have an input layer, a hidden layer of radial units and an output layer of linear units. The advantage of RBFN is found to be the best for overall performance on accuracy, robustness, problem types, sample size, efficiency, and simplicity. It uses a water utility in the southern of China. This network has 6909 nodes, no tank, and 7452 links, totaling 513 km of pipe. The authors in [20] use the following approach for water quality prediction:

- A new combinational algorithm (known as CORS-RBF-GA) is designed based on the framework of CORS method. It combines the CORS framework (known as RBFN network) to calibrate the parameters of water quality model.
- This new algorithm uses radial basis function metamodeling to decrease the time involved in water quality simulation. It can quickly realize the calibration of pipe wall reaction coefficients of chlorine model for large-scaled water distribution system. It is supplied from the four treatment plants for this study.

Their study testifies that some margin inputs, for example [-0.5, -6.0, -6.0, -3.5, -4.2, -6.0], will bring more total error because short of initial fitting points cannot fit the RBF.

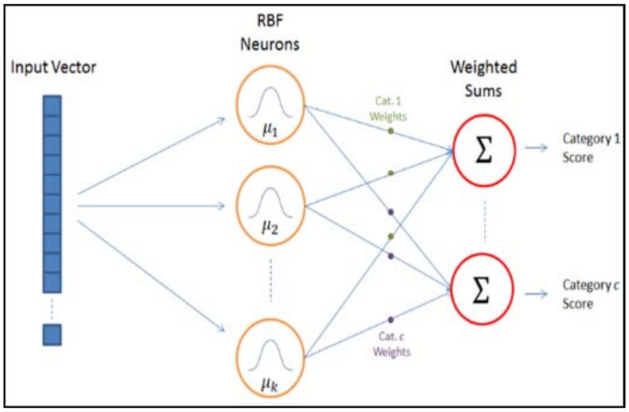

Figure 4. Radial Basis Function Network Model [21]

3. Deep Belief Network Model

A Deep Belief Network (DBN) is an unsupervised learning technique with multiple layers of hidden units. The layers are connected but the units are not connected. DBN can learn to probabilistically reconstruct its inputs when trained in unsupervised way. DBN are constructed by using restricted Boltzmann machine (RBM) arranged in a stack as shown in Figure 5. A. Solanki et al., 2015 in [22] used a deep belief network model to analyze and predict value of chemical features of water, in particular, dissolved oxygen and ph value. Their study shows that deep learning techniques provide accurate results as compared to the techniques based on supervised learning. Their research was carried out using the WEKA tool based on the secondary data collected from a third party for Chaskaman River located near Nasik, Maharashtra, India. The comparison of results show that robustness can be achieved by DBN and it can also successfully handle the variability in the data. Their used process is listed below.

- After collecting the data, clustering technique was applied on it based on three seasons: winter, summer, and monsoon.
- Then data cleaning was applied, in which missing values were replaced by using the mean of available values.
- After applying data mining, results of classical methods were gathered using Weka tool and the results of deep learning algorithms were collected by using the code developed in java.

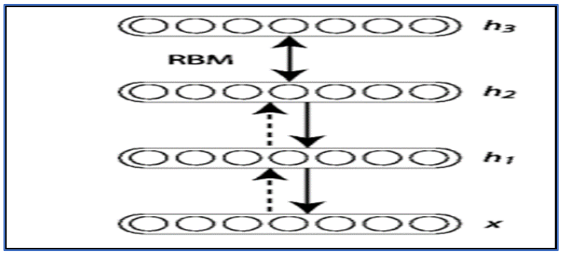

Figure 5. Deep Belief Network [22]

Their results show that the predicted value of dissolved oxygen is 3.92 as opposed to the actual value of 6.9 and the predicted value of pH is

9.9 as opposed to actual value of 8.3. The standard error is -1.6 and 2.98 respectively.

4. Decision Tree Model

Decision tree model is one of the most effective approaches in knowledge discovery and data mining. Unlike the previous model, it is a supervised learning technique which uses a predictive model to map observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves) as shown in Figure 6. Jaloree et al., 2014 [23] uses this model to predict quality of Narmada river, M.P and data is taken from 1990 to 2010. It takes into account the five chemical factors of water, namely, ph, and dissolved oxygen, BOD, NH3_N and NO3_N. The following procedure is used for water quality prediction:

- In the experiment, whole dataset was trained as a training set for developing a model. J48 decision tree classifiers were used in this study. J48 is an open source Java implementation.
- C4.5 is used which is a program that creates a decision tree based on a set of labeled input data.

Then rules are extracted from WEKA generated tree. For instance, if pH value goes to less than or equal to 8.1 unit and the amount of NO3_N increases, then the quality of surface water decreases at one class level. That means amount of NO3_N increasing increase pollution in surface water. The study has found out that the correct classification of instances is 95.4545% and incorrect classification is 4.5455%.

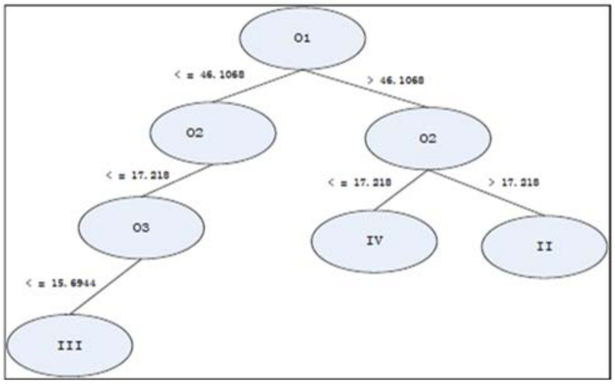

Figure 6. Decision tree model [24]

5. Improved Decision Tree Model

This model combines artificial neural networks and decision tree algorithm as shown in Figure 7. The main advantage of this approach lies in the clustering of data when processing the data with stress on inter-dependence between parameters and an aim at reducing rough disadvantages at the same time. Therefore, more accurate predictions can be made by using this method. In a study carried out by H. Liao and W. Sun [24] for water quality of Chao Lake, China, the improved decision tree model was employed to handle multiple variables of data. Weekly or biweekly samples of DO, BOD5, PH, temperature, Ammonia-Nitrogen and other selected water quality variables were collected at 12 stations. The procedure followed for prediction was listed below:

- Software was used to calculate and give the result of prediction. They are directly predicting the score of water quality by using the three key training parameters—O1, O2 and O3.
- Meanwhile, there is a relationship with the score and water quality level. So, they can get the water quality level directly.
- In summary, the prediction correct rate of IDTL model is 85% as the experiment shows by using See5 soft package which is better than decision tree model having the prediction correct rate at 70%.

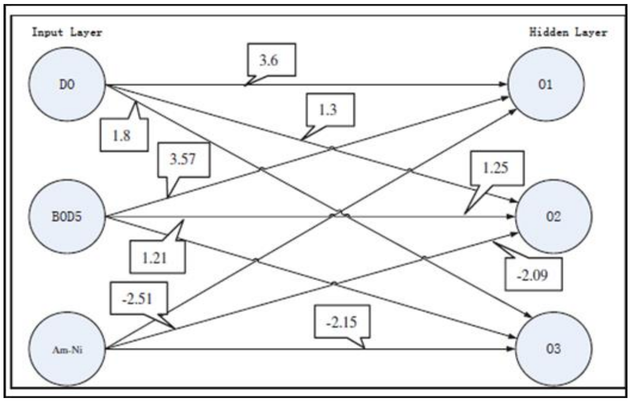

Figure 7. Improved Decision Tree Model [24]

6. Least squares Support Vector Machine Model

Least squares support vector machine model presents a supervised learning technique in which one finds the solution by solving a set of linear equations as shown in Figure 8. The research carried out by L. Yan-jun [25] uses this model. In this paper, the water samples were collected from 100 different places in different time and locations, including urban domestic sewage and surface water, specifically including river, lake, seawater, waste water from car washer, et al. The effective samples are 79 in urban domestic sewage, and 21 in surface water. Fluorescence data of water was obtained using portable analyzer developed by Zhejiang University. Its analysis and prediction process is as below”

- Before the training processing, a data pre-processing is performed.
- The characteristics of fluorescence data are high-dimensional, spectral overlap, nonlinear, and so on. To solve those problems, they run a blank assay to eliminate the interference of spectral data, such as ambient noise, and temperature drift.
- After the clustering process, LSSVM algorithm is used to establish the predicting regression model between spectral data and TOC index by TOC-VCSH of training samples for each cluster.
- After that, the final results can be given for a new sample and its vector angle cosine as a criterion is used to judge the appropriate cluster and regression model in order to get analysis value.

According to the prediction effectiveness, three performance criterions such as RMSE, MAE, MAPE are improved 40.7%, 2.1%, 8.9% and 41.9%, 17.7%, 16.5%, respectively, by method based on clustering (k-means and AP, respectively) than single directing LSSVM regression method.

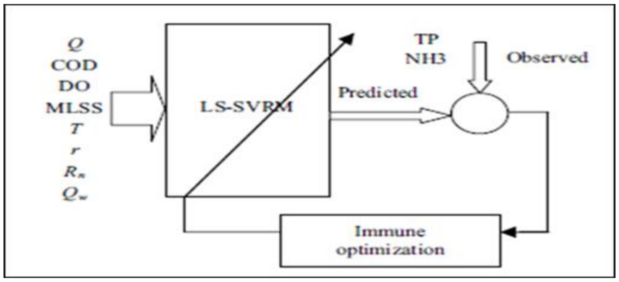

Figure 8. Least Square Support Vector Machine [25]

We present a comparison table in Table 3 which dives a tabular comparison of the research papers studied in this section. It talks about the purpose of study, model proposed, parameters considered and data-source referred. Finally, we present a Table 4 which gives a comparison of various pros and cons of all these models in terms of accuracy, complexity, robustness, data-set size etc. These results are based on the study of all the research papers referred so far and the propositions made by these studies.

Table 3. A comparison table of research papers with Machine learning based models to predict water quality

ID	PURPOSE AND AREA OF STUDY	ML MODEL	REGION	PARAMETERS	DATA-SOURCE
[18]	To develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis	Artificial neural network (ANN)	New York	Chlorophyll, Dissolved Oxygen, Specific conductance, turbidity	USGS online resource: National Water Information System
[19]	To develop a new hybrid methodology for predicting the biochemical oxygen demand which is the main indicator of water quality	Artificial neural network- Markov Chain	Tolo Harbor, Hong Kong	Chlorophyll, Dissolved Oxygen, Salinity etc.	27-year water quality data set
[20]	To use a new method combing both macro and detailed model to optimize the water quality parameters. It is optimized for the purpose of decreasing the times of time consuming water quality simulation	Radial Basis Network Function	South China, water utility	residual chlorine, turbidity, pH and temperature	network has 6909 nodes, no tank, and 7452 links. It is supplied from the four treatment plants
[22]	To provide fairly accurate predictions for variable data to evaluate water quality	Deep belief network	Chaskaman River, India	Dissolved oxygen, pH, turbidity	secondary data collected from a third part
[23]	To present a Classification data model using decision tree for the purpose of analyzing water quality data	Decision tree model	Narmada river, India	(NH3_N, NO3_N), pH, Temp _C, BOD, COD	Data collected from 1990 to 2010.
[24]	To present an improved decision tree learning method making water quality prediction easier and forecast more accurate	Improved Decision tree model	Chao Lake, China	O, O2, O3	Environment Protection Bureau of Anhui Province and Evaluation standard reference, Hong Kong
[25]	To address water quality predicting based on spectrometry	Least square support vector machine model	100 water samples from different places	Total Organic Carbon (TOC) criterion	Fluorescence data obtained by Zhejiang Univ.

Table 4. Comparison of features of various ML based models

MODEL /ALGORITHM	ARTIFICIAL NEURAL NETWORK	RADIAL BASIS FUNCTION	DEEP BELIEF NETWORK	DECISION TREE	IMPROVED DECISION TREE	SUPPORT VECTOR MACHINE	IMPROVED SUPPORT VECTOR MACHINE
Big-data based	Y	Y	Y	N	Y	N	Y
Water quality factors	Physical, Bio, Chemical	Physical Chemical	Biological Chemical	Physical, Bio, Chemical	Biological Chemical	Chemical	Chemical
Structured data-sets	Y	Y	Y	N	Y	Y	Y
Training data	60%	82%	76%	55%	70%	60%	78%
Testing data	20%	54%	54%	30%	30%	40%	45%
Real-Time Prediction	Y	N	Y	N	Y	Y	Y
Simplicity	Y	Y	N	Y	N	Y	Y
Accuracy	Err: 3.7x10-4	R^2 =0.6020	Err: -1.6 (O2)	70%	85%	RMSE: 40.7%	Err.0.089
Upstream flow	Y	N	Y	N	Y	Y	Y
Downstream flow	Y	N	Y	Y	N	Y	Y
Regression based	Y	Y	Y	Y	N	N	N
Supervised Learning	Y	N	N	Y	Y	Y	Y
Sensors used	N	Y	N	Y	N	Y	Y
Robustness	Y	Y	Y	N	N	Y	Y
Flexibility	Y	N	Y	N	Y	N	N
Data clustering	Possible	Impossible	Possible	Possible	Impossible	Possible	Possible

V. ISSUES, CHALLENGES, and NEEDS

Since last year, we have begun to conduct water quality evaluation for San Francisco Bay using selected big data analysis approaches and machine learning models. Based on literature and experience, we highlight some research issues, challenges, and future needs in water quality evaluation and prediction.

- ***Issues #1: Data quality and validation issue*** - As pointed out in [34], there are lots of sensor data quality issues which affect the accuracy of water quality evaluation and assessments due to device faults, battery issues, and sensor network communication problems. This brings the first need below.
- ***Need #1: Research demand in big data quality assurance*** - There is a strong need in big data quality assurance research in data quality modeling, automatic real-time validation methods, and tools to increase the accuracy of water quality evaluation and prediction.
- ***Issue #2: Real-time water quality monitor and supervision for water resources*** - As the advance of smart sensing and IoT, more and more environmental sensors (including water sensors and networks) have installed and deployed for many water resources, such as, lacks, rivers, creeks, ocean bays and coasts. However, there is a lack of integrated real-time big data based water evaluation and monitor environments for smart cities to support dynamic water quality evaluation, monitor, and supervision management. Water in a city could be considered as a multi-level water system, covering surface water and underground water. The water quality on both levels usually affect each other. This brings the second demand below.
- ***Need #2: Research and development of real-time water quality monitor and evaluation systems supporting water quality evaluation and analysis on multiple levels.*** This demand is caused by the lack of the existing research work addressing the water quality impacts on different water levels (i.e. surface water, and underground water) due to water pollution from a special water source. This suggest the demand on an integrated real-time water quality monitor and evaluation system based on sensor networks and IoT infrastructures at the different levels.
- ***Issue #3: Big data modeling issues for dynamic water quality monitor and analysis at the different levels for smart cities -*** Most published research work applied big data analytics approaches and used one specific machine learning technique for water resource at specific level in a limited location (or a region) during an interested time period. The water system for a future smart city must support real-time water quality monitor, evaluation, and prediction for diverse water sources at different levels to support environmental management and supervision needs. This implies that we need study and develop integrated and dynamic water quality models using hybrid machine learning models to address the following factors: a) the nature of dynamic water flow, b) both single-input time series and multiple input time series, c) dynamic quality impacts on different water levels.

VI. CONCLUSIONS

With the advance of IoT infrastructures, big data technologies, and machine learning techniques, real-time water quality monitor and evaluation is desirable for future smart cities. This paper reports our recent literature study, reviews and compares current research work on water quality evaluation based on big data analytics, machine learning models and techniques. Finally, it highlights some observations on future research issues, challenges, and needs.

ACKNOWLEDGEMENT

This research work is supported by Futurewei for our excellence research Center on Smart Technology, Computing, and Complex Systems at San Jose State University (Reference link: http://smartcenter.sjsu.edu/index.html). We would like to sincerely thank Futurewei for the continuous support for future research in smart cities and complex systems.

REFERENCES

[1] World Health Organization, "Meeting the MDG drinking water and sanitation target: the urban and rural challenge of the decade", Geneva, 2006.

[2] M. A. Tirabassi, "A statistically based mathematical water quality model for a non-estuarine river system1." *JAWRA Journal of the American Water Resources Association, Vol.* 7, pp. 1221-1237, December 1971.

[3] L. Hu, C. Zhang, C. Hu, and G. Jiang, "Use of grey system for assessment of drinking water quality: a case study of Jiaozuo city, China", *Advances in Grey Systems Research*, Springer Berlin Heidelberg, pp. 469-478, 2010.

[4] R. Rosly, M. Makhtar, M. K. Awang, M. N. A. Rahman, and M. M. Deris, "The Study on the Accuracy of Classifiers for Water Quality Application", *International Journal of u- and e- Service, Science and Technology*, Vol. 8, No. 3, pp.145-154, 2015.

[5] D. Yang, L. Zheng, W. Song, S. Chen, and Y. Zhang, "Evaluation indexes and methods for water quality in ocean dumping areas", *Procedia Environmental Sciences: Proc. of the 7th International Conference on Waste Management and Technology*, Vol. 16, pp.112-117, December 2012.

[6] D. P. Loucks and E. V. Beek, "Water Quality Modelling And Prediction," *Water Resources Systems Planning And Management: An Introduction To Methods, Models And Applications*, Paris: UNESCO, pp. 381-425, 2005.

[7] Aspen-Nicholas Water Forum, "Data Intelligence for 21st Century Water Management: A Report from the 2015 Aspen-Nicholas Water Forum", 2015.

[8] S. P. Sherchan, P. G. Charles, and L. P. Ian, "Evaluation of real-time water quality sensors for the detection of intentional bacterial spore contamination of potable water." *Journal of Biosensors & Bioelectronics* 2013, 2013.

[9] Y. Liu, M. Islam, and J. Gao, "Quantification of shallow water quality parameters by means of remote sensing", *Progress in Physical Geography*, Vol. 27, No. 1, pp. 24-43, March 2003.

[10] M. Valdivia, D.W. Graham, and D. Werner, "Climatic, Geographic and Operational Determinants of Trihalomethanes (THMs) in Drinking Water Systems" *Scientific reports*, Vol. 6, 2016.

[11] Y. Zhong, L. Zhang, S. Xing, F. Li, and B. Wan, "The big data processing algorithm for water environment monitoring of the three Gorges reservoir area" *Abstract and Applied Analysis*, Vol. 2014, Hindawi Publishing Corporation, 2014.

[12] Hou Jing-Wei, MI Wen-Bao, and Long-Tang Li, "Spatial quality evaluation for drinking water based on GIS and ant colony clustering algorithm", *Springer-Verlag Berlin Heidelberg*, March 25, 2015.

[13] M Tarique, H. Khaleeq, and A. A. ElNour, "A Reliable Wireless System for Water Quality Monitoring", Vo. 8, No. 3, 2016.

[14] T. C. Lobato, R. A. Hauser-Davis, T. F. Oliveira, A. M. Silveira, H. A. N. Silva, M. R. M. Tavares, and A. C. F. Saraiva, "Construction

of a novel water quality index and quality indicator for reservoir water quality evaluation: A case study in the Amazon region", *Journal of Hydrology*, *522*, pp. 674-683, 2015.

[15] A. Newton, and M. M. Stephen, "Lagoon-sea exchanges, nutrient dynamics and water quality management of the Ria Formosa (Portugal)" *Estuarine, Coastal and Shelf Science*, pp. 405-414, 2005.

[16] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal, "Classification Model for Water Quality using Machine Learning Techniques" *International Journal of Software Engineering and Its Applications*, pp. 45-52, 2015.

[17] A. Sarkar and P. Pandey, "River water quality modelling using artificial neural network technique" *Aquatic Procedia*, Vol. 4, pp. 1070-1077, 2015.

[18] Y. Khan and C. S. See, "Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model," *IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, 2016.

[19] X. Li and J. Song, "A New ANN-Markov Chain Methodology for Water Quality Prediction*," International Joint Conference on Neural Networks*, pp. 12-17 July, 2015.

[20] L. Ma, K. Xin, and S. Liu, "Using Radial Basis Function Neural Networks to Calibrate Water Quality Model," *World Academy of Science, Engineering and Technology International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering*, Vol. 2, No. 2, 2008.

[21] C. McCormick, "Radial Basis Function Network (RBFN) Tutorial" [Online] Available: http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/, [Accessed: 10- Oct- 2016], 2013.

[22] A. Solanki, H. Aggarwal, & K. Khare, "Predictive Analysis of Water Quality Parameters using Deep Learning", *International Journal of Computer Applications*, vol. 125, no. 9, pp. 0975-8887, Access from Google Scholar, Sept. 2015.

[23] Jaloree, Shailesh, A. Rajput, and Sanjeev Gour. "Decision tree approach to build a model for water quality." *Binary Journal of Data Mining & Networking* 4.1 (2014): 25-28.

[24] H. Liao and W. Sun. "Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method." *Procedia Environmental Sciences* 2 (2010): 970-979.

[25] L. Yan-jun and M. Qian. "AP-LSSVM modeling for water quality prediction." *Control Conference (CCC), 2012 31st Chinese*. IEEE, 2012.

[26] X. Wang, G. Wang, and X. Zhang, "Prediction of Chlorophyll-a content using hybrid model of least squares support vector regression and radial basis function neural networks," *Sixth International Conference on Information Science and Technology*, pp. 366-371, May 6-8, 2016.

[27] W. Lin, L. Ran, T. Youcai, and L. Kefeng, "Research on Prediction of Water Quality of Water Reservoir with Combined Multiple Neural Networks Model," *International Conference on Electric Technology and Civil Engineering (ICETCE)*, pp. 4376-4379, 2011.

[28] S. Song, X. Zheng, and F. Li, "Surface Water Quality Forecasting Based on ANN and GIS for the Chanzhi Reservoir, China," *IEEE 2nd International Conference on information Science and engineering*, pp. 4094-4097, 2010.

[29] G. A. C. Cordobaa, L. Tuhovcak, and M. Taus, "Using artificial neural network models to assess water quality in water distribution networks," *12th International Conference on Computing and Control for the Water Industry*, pp. 399-408, 2014.

[30] L. Ying, Z. Jiti, W. Xiangrui, and Z. Xiaohui, "Water quality evaluation of nearshore area using artificial neural network model", *3rd International Conference on Bioinformatics and Biomedical Engineering*, pp. 11-13, June 2009.

[31] N. Chang, & B. Vannah, (2015) "Comparative Data Fusion between Genetic Programming and Neural Network Models for Remote Sensing Images of Water Quality Monitoring", *IEEE International Conference on Systems, Man, and Cybernetics, Manchester*, pp. 1046-1051, 2013.

[32] M. R. Estuar et al., "Towards Building a Predictive Model for Remote River Quality Monitoring for Mining Sites", *TENCON 2015 2015 IEEE Region 10 Conference, Macao*, and pp. 22159-3442, Nov. 2015.

[33] M. Osborne et al., "A Machine Learning Approach to Pattern Detection and Prediction for Environmental Monitoring and Water Sustainability", *Workshop on Machine Learning for Global Challenges (ICML2011*), Bellevue, WA, 2011.

[34] Jerry Gao, Chunli Xie, and Chuanqi Tao, "Big Data Validation and Quality Assurance - Issues, Challenges, and Needs", IEEE Symposium on Service-Oriented System and Engineering, IEEE Computer Society, Oxford, UK, April, 2016.