Vol.15, Issue No 1, 2025

NATURE BASED PREDICTION MODEL OF BUG REPORTS BASED ON ENSEMBLE MACHINE LEARNING MODEL

[1]Mr. P ISAAC PAUL, [2]DARSI MANIDEEP, [3]BOGANI SURESH, [4]MUDUNURI AMAR NATH VARMA, [5]MARRI GANESH

[1](ASSISTANT PROFESSOR), CSE, RISE KRISHNA SAI GANDHI GROUP OF INSTITUTIONS ONGOLE

[2345]B.TECH, SCHOLAR, CSE, RISE KRISHNA SAI GANDHI GROUP OF INSTITUTIONS ONGOLE

ABSTRACT

The prediction of bug reports is a critical task in software development that helps in efficiently managing and resolving issues. This paper proposes a nature-based prediction model for bug reports using an ensemble machine learning approach. The model aims to predict the likelihood of bug occurrence and the potential severity of reported bugs by analyzing historical bug data. By incorporating natural language processing (NLP) techniques to process bug descriptions and using ensemble learning methods such as Random Forest, Gradient Boosting, and XGBoost, the system is able to enhance the accuracy of predictions by combining the strengths of multiple algorithms.

The nature-based model focuses on extracting features related to the environment, development context, and code characteristics that could influence bug occurrence, such as code complexity, developer experience, and recent changes in the codebase. These features, combined with NLP techniques, allow the model to analyze textual data from bug reports and accurately predict bug patterns, allowing developers to prioritize tasks and allocate resources more effectively.

By utilizing ensemble methods, the system benefits from improved performance and robustness, minimizing the limitations inherent in individual machine learning models. The model not only predicts the occurrence of bugs but also estimates their impact on the software development process, thereby aiding in proactive bug management and decision-making. This approach enhances the overall software quality and accelerates the development lifecycle by enabling more informed planning and resource management.

KEYWORDS: Bug Prediction, Ensemble Learning, Machine Learning, Natural Language Processing (NLP), Software Development, Code Complexity, Bug Severity, Random Forest, XGBoost, Gradient Boosting.

1.INTRODUCTION

Access to water is a critical component of human lives and is now considered a basic human right. Access to clean water is also one of the 17 Sustainable Development Goals (SDG) set up by the United Nations in 2015 to achieve a better future for all [1]. Specifically, the sixth goal, which is to ensure and sustain the availability of water and sanitation to all [2]. Potable water can also be linked to the third SDG goal _ good health and well-being, as contaminated water can be a transmission medium for diseases such as cholera, typhoid, and diarrhoea, which are jointly the highest cause of mortality (especially children) in developing nations of Africa and Asia [3]. Water is also important in agriculture and food production. Recent statistics shows that about 10% of the world population is malnourished, with developing countries being hit the hardest, with starvation resulting in about 45% of infant mortality [5]. Ensuring global food security is thus of utmost importance. Food security has been recognized as a critical requirement, hence its inclusion as one of the SDG (goal 2), with specific focus on ending hunger, by promoting sustainable agriculture and improving food distribution. Food production and agriculture in general rely heavily on water, both for irrigation and for animal consumption. It is thus pertinent to ensure the availability and sustainable management of water _t for agricultural use.

There are several sources of water for both drinking and irrigation use, including rivers, streams, rain, and groundwater (accessed through wells and boreholes). The nature and characteristics of a source of water are often critical factors that influence the constituents of water samples obtained therein. Beyond natural factors, chemical wastes from human activities such as mining, crude oil extraction, and industrial wastes, most often end up in streams, rivers, and other sources of water, changing the nature and properties of these waters. These waters then end up in homes or farms, where they are used for domestic purposes, drank, fed to livestock, or used to water crops. Consuming this type of water can have dire health consequences or result in death. It is therefore paramount that a proper process be put in place to ensure end to- end monitoring of the water right from the source to its last point of use. At each monitoring point, samples of water need to be collected to assess the quality or ``fitness for use" for human (and animal) consumption, irrigation and domestic (or industrial) uses.

Several models have been developed to assess water quality, all of which consider various parameters, including chemical (such as hydrogen potential (pH), calcium, oxygen, sulphate levels etc.), microbial (such as E. coli, rotaviruses, Entamoeba etc.), and physical (temperature and clarity). These models produce a unit metric, known as the Water Quality Index (WQI), as output. Globally, different guidelines have been adapted for calculating WQI. For instance, in parts of Europe, the British Columbia Water Quality Index (BCWQI) and the Scottish Research Development Department (SRDD) are used, while in North America, the Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI) and National Sanitation Foundation Water Quality Index (NSFWQI) are predominant. In Asia, specifically India, the Bureau of Indian Standards (BIS) is prominent, while in Africa, notable standards include the South African National Standard for drinking water (SANS 241-1) and the Kenya Bureau of Standards (KEBS). A number of these models have been reviewed in [6]. It is important to note that many of these national standards are mostly local adaptations of the standards defined by the World Health Organization (WHO) [7]. This work is based on the South African and WHO standards.

Indeed, measuring water parameters for diverse water samples can be a laborious and daunting task, as it often involves adhering to a stringent set of rules in collecting the water samples, maintaining set conditions during transportation to the test laboratories, following standard methodologies in analysing the samples, and generally ensuring quality control. Some of these processes (and corresponding guidelines) are given in [8], [9]. The output of these processes indicates if the water sample is potable or non-potable. In this work, we propose a Cyber-physical network architecture for real-time monitoring of water parameters across a city and an alternative model based on machine learning to determine potability of water samples. Like [10] _[13] [14], our work also only focuses on the physical and chemical parameters of water, while ignoring the biological. This is because our model is meant to be sensor based (in the context of the Internet of Things), and to our knowledge, there are no physical sensors for measuring biological parameters, such as the presence of E. coli in water. We do not trivialize the importance of microbial water parameters, and our proposed model can indeed be adapted to consider these parameters by simply incorporating suitable physical sensors (if available) or virtual / soft sensors, such as the one proposed in [15] into our model.

Figure 1 gives a high-level depiction of our proposed architecture which is built upon 4 layers. The constituent components of this architecture are described as follows.

:1) Sensing Layer: As depicted in the figure, the sensing layer interacts directly with the water samples in a river, stream, dam etc. to measure water parameters. It is built into a vertical pole tagged ``sensor probe'' and consists of numerous sensors bundled together. These sensors might include pH, conductivity, turbidity, temperature, residual chlorine etc., similar to those offered by Labellum [16]. All telemetry data measured by these sensors are sent to the Fog Nodes (FNs), wired or wirelessly, via the sending unit. In scenarios were installing sensors in water source(s) is extremely difficult or when the required sensors are not readily available, water parameter readings can be collected from the associated water treatment plants.

2) Edge Layer: This layer consists of low-end processing devices (edge modules), such as single board computers (e.g., Raspberry Pi or Nvidia Jetson), or microcontrollers (e.g. Arduino, ESP32). These devices act as i.) data pre-processing units, responsible for the

collection, aggregation, filtration, and shaping of data received from the sensing layer; ii) network gateway to ``ferry'' telemetry data to the FNs, through 3G/4G/5G cellular or other low powered long-range network solutions.

3) Fog / Cloud Layer:

_ Fog Nodes (FNs): these are small sized distributed cloud computing nodes that bring computing and storage closer to the data source, thus reducing latency resulting from transmission delay to/from the remote Cloud [17]. The FN is responsible for classification of water samples using machine learning models such as the ones proposed in this work. Due to the limited computing power at the Fog (compared to the Cloud), only the most influential parameters need to be considered when classifying water samples. This can be beneficial as less sensors would be required (since not all parameters are being measured) and by extension.

lower computing resources would be needed for the classification process. Furthermore, resource management, scheduling etc. can also be carried out on FNs. When long term storage and/or advanced computations are required, which are beyond the Fog's capacity, data are forwarded to the Cloud data centre.

_ Cloud Data Centre: The Cloud is a remote high performance computing infrastructure, which provides computing on demand [18]. In our system, the Cloud serves as a data warehouse as well as a platform for performing advanced data analytics, dash boarding, and hosting for relevant services and software.

4) Application Layer: serves as an interface between users (water management authorities, end users / customers, other stakeholders) and software / services running in the Cloud. Relevant software for water parameter monitoring is hosted at this layer and made available to users through mobile and web platforms.

The water monitoring network proposed in this work is to be deployed in the City of Cape Town in Western Cape, South Africa, with the intention of monitoring water parameters in water storage dams and/or water treatment plants across the city. Data gathered by the monitoring network are then passed through Machine Learning (ML) models to determine their suitability for consumption or irrigation purposes. The specific contributions of this work can be summarized as Follows V

1) Build a network for real-time collection and monitoring of water quality across water storage dams in the city of Cape Town. This network takes into consideration the unique geographical features of Cape Town, such as mountains and elevations that might obstruct radio frequency propagation.

2) Curate ample sized datasets on drinking and irrigation water that can be used to train (and test) machine learning models to automatically determine the ``fitness for use'' of a sample of water for drinking and/or irrigation purposes.

3) Build models that determine the most critical parameters that influence the accuracy of machine learning models in analysing water for drinking or irrigation.

2.EXISTINGSYSTEM

In [12], a network for measuring and monitoring water parameters in a metal producing city in Brazil was developed. Twelve water monitoring stations were setup to measure several physico-chemical water parameters, including pH, dissolved solids, Zinc, Lead etc. Finally, obtained results were analysed using principal component analysis. In a similar manner, [13] developed a system to monitor water quality in Limpopo River Basin in Mozambique and set up 23 monitoring stations to measure physico-chemical and microbiological parameters, and ultimately assess the quality of water in the river basin. To address the challenges of optimal placement of gauges and sampling frequencies, which are often faced when developing water monitoring systems, the authors in [14] developed an economically viable model that combined genetic algorithm with 1-D water quality simulation. Though the work was only simulated by using genetic algorithm, the authors were able to solve the NP hard problem of optimally placing monitoring stations.

Monitoring water parameters often entails periodically sampling a body of water to capture relevant metrics. These metrics might include physico-chemical and microbiological measurements, such as potential of hydrogen (pH), temperature, sodium levels etc. In a water monitoring network, measured parameters need to be transferred to a base station where relevant decision(s) would be taken. Due to the sparse nature of transmitted data, light weight communication protocols capable of transmitting relatively small data over long distance are required for water monitoring networks. From literature, Low Power Wide Area Network (LPWAN) technologies have been favoured for such applications. An extensive discussion on LPWAN technologies was done in [19]. The work compared a few sub-GHz solutions including Sig- Fox, LoRa, Ingenu and Telensa, with respect to their range,

transmission rate, and channel count. Ingenu was reported to have the longest range in city settings at 15 km, followed by Sigfox at 10 km (in cities) and 50 km (in rural areas); then LoRa at 5 km (in cities), and 15 km in rural settings.

Regarding the assessment of communication technologies, there has been a long-drawn debate over the efficacy of software simulations versus real-world testing. Though this debate still rages, several researchers have shown that simulation results are often at par with real-world tests. For instance, using LoRa, the authors in [20] compared simulation results with real world test for intervehicle communication. They used NS3 as a simulation platform and an Arduino UNO C Dragino LoRa module for the real-world tests, while Propagation loss, coverage Packet Inter-reception (PIR), Packet Delivery Ratio (PDR) and Received Signal Strength Indicator (RSSI) level were used as benchmark metrics. They concluded that the results of the simulator were consistent with those of the real-world tests. In a similar work,Hassan [21] also compared the efficacy of simulation results (from Radio Mobile simulator) with real-world tests (using micro controllers C LoRa modules) when using LoRa as a bridge for Wi-Fi. Unlike [20], [21] did not give a side-by side comparison of simulated vs. real-world results for each metric considered but concluded that the simulator performed well. [22] set up seven pairs of XBee modules and compared communication performance using both the 800/900MHz and 2.4GHz frequencies. They concluded that simulation results from the Radio Mobile simulator corroborated with those of real-world tests.

3. PROPOSED SYSTEM

The water monitoring network proposed in this work is to be deployed in the City of Cape Town in Western Cape, South Africa, with the intention of monitoring water parameters in water storage dams and/or water treatment plants across the city. Data gathered by the monitoring network are then passed through Machine Learning (ML) models to determine their suitability for consumption or irrigation purposes.

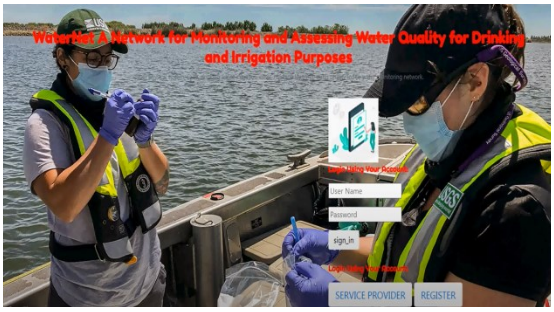
1) Build a network for real-time collection and monitoring of water quality across water storage dams in the city of Cape Town. This network takes into consideration the unique geographical features of Cape Town, such as mountains and elevations that might obstruct radio frequency propagation.

2) Curate ample sized datasets on drinking and irrigation water that can be used to train (and test) machine learning models to automatically determine the `fitness for use" of a sample of water for drinking and/or irrigation purposes.

3) Build models that determine the most critical parameters that influence the accuracy of machine learning models in analysing water for drinking or irrigation.

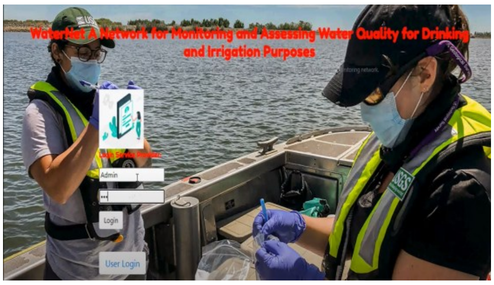
4.SCREEN SHOTS

LOGIN PAGE



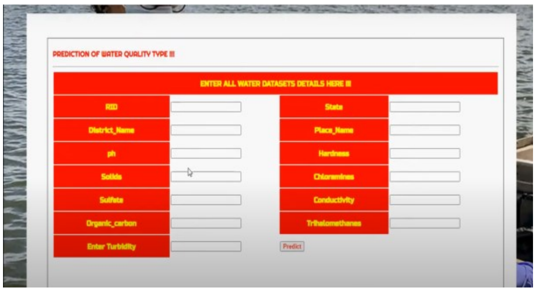
WaterNet A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes

SERVICE PROVIDER LOGIN



WaterNet A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes

PREDICTION



PREDICTION OF WATER QUALITY TYPE !!!

ENTER ALL WATER DATASETS DETAILS HERE !!!

RID		State	
District_Name		Place_Name	
ph		Hardness	
Solids		Chloramines	
Sulfate		Conductivity	
Organic_carbon		Trihalomethanes	
Enter Turbidity		Predict	

5.CONCLUSION

This work focused on two major concept, firstly, the proposal of a real-time water monitoring network for gathering data on water parameters from water bodies. Secondly, the application of machine learning (ML) models as means of assessing water quality. The developed water monitoring network is based on Lo Ra, a low power long range protocol for data transmission, and was developed using the City of Cape Town as case study. Results of the simulation done in Radio Mobile, revealed a partial mesh network topology as the most adequate network to cover the city. Data gathered from this monitoring network would ideally be aggregated on a Cloud server, where ML models can then be applied to assess the water's fitness of use for drinking or irrigation purposes. Due to the absence of relevant datasets, two suitable datasets were built in this work and used to training and testing three ML models considered, which are Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM). Results of the test showed that LR performed best for drinking water, as it gave the highest classification accuracy and lowest false positive and negative values, while SVM was better suited for irrigation water. Finally, a model for identifying the most influential water parameter(s) w.r.t classification accuracies of the ML models was then explored using recursive feature elimination (RFE). Obtained results showed that pH, and total hardness were the least influential parameters in drinking water, while SSP was the least for irrigation water.

Though the authors acknowledge the possible application of deep learning models, these were not used in this work. In future works, deep learning models such as the various variants of neural networks could be considered as expansion to this work. Furthermore, water quality indices were manually calculated and used to assess the ``fitness for use'' of water, future works could explore the application of unsupervised ML models as alternatives to manually calculated water quality indices. In the same vein, rather than using RFE, other approaches such as multi criteria decision making could also be considered to identify influential parameters. Finally, incorporating usage prediction models and microbial monitoring into the water network as well as tracking sources of water contaminates could also be avenues to further this work.

6.REFERENCES

[1] B. X. Lee, F. Kjaerulf, S. Turner, L. Cohen, P. D. Donnelly, R. Muggah, R. Davis, A. Realini, B. Kieselbach, L. S. MacGregor, I.Waller, R. Gordon, M. Moloney-Kitts, G. Lee, and J. Gilligan, ``Transforming our world: Implementing the 2030 agenda through sustainable development goal indicators,'' J. Public Health Policy, vol. 37, no. S1, pp. 13_31, Sep. 2016.

[2] Integrated Approaches for Sustainable Development Goals Planning: The Case of Goal 6 on Water and Sanitation, U. ESCAP, Bangkok, Thailand,2017.

[3] WHO. Water. Protection of the Human Environment. Accessed: Jan. 24, 2022. [Online]. Available: www.afro.who.int/health-topics/water

[4] L. Ho, A. Alonso, M. A. E. Forio, M. Vanclooster, and P. L. M. Goethals, ``Water research in support of the sustainable development goal 6: A case study in Belgium,'' J. Cleaner Prod., vol. 277, Dec. 2020, Art. no. 124082.

[5] Global Nutrition Report 2016: From Promise to Impact: Ending Malnutrition by 2030, International Food Policy Research Institute, Washington, DC, USA, 2016, doi: 10.2499/9780896295841.

[6] N. Akhtar, M. I. S. Ishak, M. I. Ahmad, K. Umar, M. S. Md Yusuff, M. T. Anees, A. Qadir, and Y. K. A. Almanasir, ``Modi_cation of the water quality index (WQI) process for simple calculation using the multicriteria decision-making (MCDM) method: A review,'' Water, vol. 13, no. 7, p. 905, Mar. 2021.

[7] World Health Organization. (1993). Guidelines for Drinking-Water Quality. World Health Organization. Accessed: Jan. 12, 2022. [Online]. Available: http://apps.who.int/iris/bitstream/handle/10665/44584/9789241548151-eng.pdf

[8] Standard Methods for the Examination of Water and Wastewater, Federation WE, APH Association, American Public Health Association (APHA), Washington, DC, USA, 2005.

[9] L. S. Clesceri, A. E. Greenberg, and A. D. Eaton, ``Standard methods for the examination of water and wastewater,'' Amer. Public Health Assoc. (APHA), Washington, DC, USA. Tech. Rep.21, 2005.

[10] M. F. Howladar, M. A. Al Numanbakth, and M. O. Faruque, ``An application of water quality index (WQI) and multivariate statistics to

evaluate the water quality around Maddhapara granite mining industrial area, Dinajpur, Bangladesh," Environ. Syst. Res., vol. 6, no. 1, pp. 1_8, Jan. 2018