International Journal on
Advanced Science
Engineering
Information Technology

Vol.11 (2021) No. 6
ISSN: 2088-5334

Water Quality Prediction and Detection of the Vibrio Cholerae Bacteria

Camilo Enrique Rocha Calderón [a,*], Octavio José Salcedo Parra [a,b], Sebastián Camilo Vanegas Ayala [a]

[a] *Universidad Distrital Francisco José de Caldas, Faculty of Engineering, Intelligent Internet Research Group, Bogotá D.C, Colombia*
[b] *Universidad Nacional de Colombia, Department of Systems and Industrial Engineering, Faculty of Engineering, Bogotá D.C, Colombia*
*Corresponding author: *cerochac@correo.udistrital.edu.co*

***Abstract*— This document shows the results for two water quality-related trials based on the Physico-chemical characteristics given by the used dataset; both trials were carried out based on the same dataset from which the membership sets, and functions were defined the most relevant features. The first trial was a neural network method aimed to predict water quality through attributes as the pH, temperature, turbidity, salinity, among others; the second trial was a fuzzy logic system method for the detection of the Vibrio Cholerae in the water through the usual variables associated to its presence: temperature, salinity, phosphates, and nitrites' levels. The method for this research is divided into two phases. The first phase is developing suitable software using an iterative and incremental process model based on prototypes. The second phase or operative phase has an experimental characterization that allows for an adequation of the environment to establish the main features and properties that are relevant to the study object. The results showed effectiveness values of 99.99% (highest obtained value) for trial one and 70.23% for trial two; such values depict an accurate prediction on the quality of water and a valuable detection for Cholera related bacteria in water supplies. This research developed two highly interpretable and transparent systems to people through the graphic of the correspondences between the rules established and the membership functions in the input and output sets.**

***Keywords*—Fuzzy systems; neural networks; quality; Vibrio Cholerae; water.**

Manuscript received 28 Oct. 2020; revised 17 Mar. 2021; accepted 18 Jun. 2021. Date of publication 31 Dec. 2021.

I. INTRODUCTION

Water, and its consumption, are vital for all human activities [1]. Nevertheless, there are several issues related to its management (its growing scarcity, inappropriate use, lack of treatment for contaminated supplies, etc.) that profoundly impact humankind [2]. One of the most worrying problems related to water is the transmission of diseases as Cholera, which kills around 20.000 people every year oxygen [3], and is a direct outcome for the consumption of non-drinking water nursing the Vibrio Cholerae bacteria [4], this means water undergoing temperatures between 20 and 40 °C, high salinity, and high phosphates and nitrites' levels [5], [6]. Considering the need for an acknowledgment of water supply status the following physio-chemical variables are defined as the most relevant: pH; temperature; concentrations for phosphates, nitrites, nitrates, and ammonium; salinity; turbidity; and dissolved [7], [8].

This research was carried out following two trial proposals; the first one consists of a simulation and prediction trial on water quality according to the data extracted from a sample dataset [9]. The second trial was planned as a monitoring and detection model for Vibrio Cholerae bacteria on human's use water supplies [10], based on the main features that could foster its appearance [6].

As part of the applied computer science based on heuristic models, AI can offer the right tools to simulate, predict, and identify patterns from a single input dataset [11]. One of the most common techniques used to solve similar problems to our Water quality check is Neural Networking, which is well known for its high accuracy due to its multilevel sequence training process through different training, test, and assessment stages [12]. Besides, a Fuzzy Inference System (FIS) was used to approach the vibrio Cholerae bacteria detection because of the significance level on the output values this technique shows. The outcome is vital when relating the input/output datasets the established functions and rules [13].

To identify bacteria in water and evaluate its quality, different models have been presented using techniques as follows:

- Extreme Learning Machine (ELM) algorithm with Dolphine Swarm optimization [14].
- SVR and XGBoost algorithms [15].
- A 3D model water quality system composed of a data management system and environmental models [16].
- The central point triangular whitening weight functions (CTWF) method [17].
- Convolutional neural networks [18].
- LSTM algorithm with rules of cross correlation and a priori association [19].
- Techniques such as Adaptive Synthetic Sampling (ADASYN) and Principal Component Analysis (PCA) [20].
- A fuzzy logic system based on the Internet of Things (IoT) [21].
- IoT sensor system [22].
- A long-term and short-term memory neural network (LSTM NN) [23].

As a consequence of the literature review, it can be said that this work refers to the method in previous studies [17], [22], which are based on the analysis, simulation, and prediction of data related to water quality through computer-aided techniques. The dataset used in this research is obtained from three measuring points across the water supply, each one presenting variations on the physio-chemical features in the water samples. The election for this dataset is confirmed through the comparison of conditions and variables [15], [16], [23]. Regarding the techniques used, the Neural Networks and Fuzzy Logic Systems are the most preferred to predict and assess water quality [16], [18], [19], [21]. Also, a comparison between the outcome system of this research and the ones in the literature review can be established. As a differentiating aspect, the detection for Vibrio Cholerae depended on the climate variables, and the rate of the population infected [20]. However, it does not consider the Physico-chemical characteristics of the water. Therefore, this research exploits such features for its detection, as discussed in Yan *et al.* [14] and Waleed *et al.* [21].

II. Material and Method

The methodology for this research is divided into two phases. The first one focuses on developing suitable software and is based on empirical, experimental, and incremental-iterative prototyping [24] [25]. The second (operative) phase has an experimental characterization, that allows for an adequation of the environment, to establish the main features and properties that are relevant to the main trial [26]. (Fig. 1 shows a flow chart of the stages of the research).

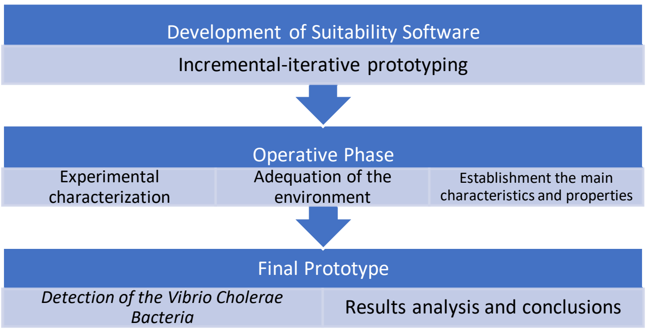

Development of Suitability Software
Incremental-iterative prototyping
Operative Phase
Experimental characterization
Adequation of the environment
Establishment the main characteristics and properties
Final Prototype
Detection of the Vibrio Cholerae Bacteria
Results analysis and conclusions

Fig. 1 Stages of the research.

The methodologic process to follow consists of four phases. Phase 1 is Neural Network and Fuzzy inference systems design. Phase 2 is Dataset load and pre-processing (extracting the most relevant attributes, deleting irregular entries, and normalization of values in a 0 to 1 range). Phase 3 is the algorithm is set depending on the main focus: prediction (neural network is configured, trained, and tested) or classification (Fuzzy system is configured and tested). Phase 4 is the final procedure are validated through different rubrics for prediction and classification. (Fig. 2 shows an exploited flow chart of the methodology process).

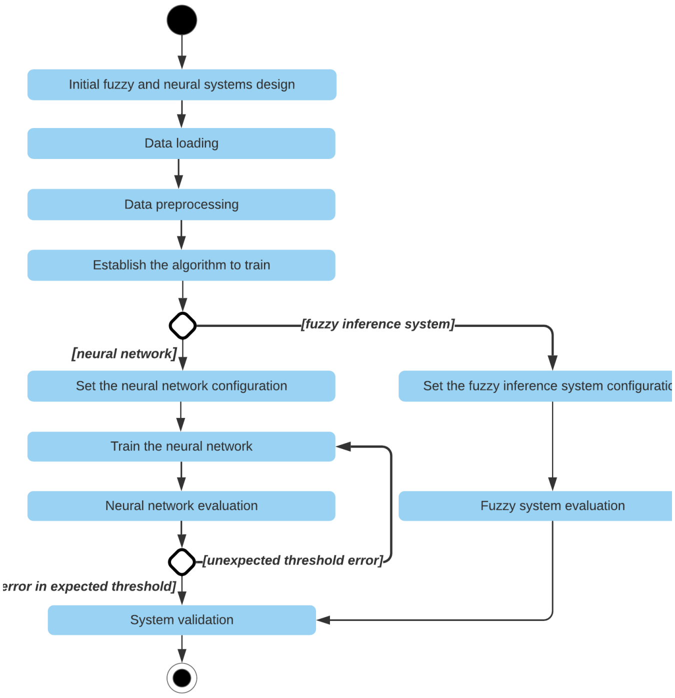

Initial fuzzy and neural systems design
Data loading
Data preprocessing
Establish the algorithm to train
[fuzzy inference system]
[neural network]
Set the neural network configuration
Set the fuzzy inference system configuratic
Train the neural network
Neural network evaluation
Fuzzy system evaluation
[unexpected threshold error]
error in expected threshold]
System validation

Fig. 2 Research method.

A. *Design and Application*

Network and Fuzzy Inference systems are designed considering both of the approaches in this research: water quality assessment and detection of the Vibrio Cholerae bacteria respectively. The dataset used contains the Physico-chemical features of the Elwha River [8]. TABLE shows the selected parameters relevant to this project. After finishing the pre-processing of the dataset the initial structures for the neural network and fuzzy inference systems are defined.

TABLE I
Dataset Parameters

Parameter	Measure
Phosphates concentration	mg/l
Nitrates concentration	mg/l
Ammoniums concentration	mg/l
Salinity	g/l
Turbidity	ncu
Temperature	°C
Dissolved Oxygen	mg/l
Dissolved Oxygen Percentage	%
ICA	Weighing
Vibrio Cholerae presence	(0 o 1)

B. Design of the Neural Network to Assess the Water Quality

Seven different configurations, varying in the number of layers and neurons per layer (TABLE). It was created to compare their performances and select the best performing one. Every system's neurons operate under the function TANSIG. The threshold ranges were defined in terms of the minimum and maximum input parameter value (nine parameters) and the output corresponds to the Dataset parameters. The designed systems follow a feedforward scheme and use the training method Backpropagation during 500 cycles.

TABLE II
NEURAL NETWORKS STRUCTURE.

NN	Layers	Neurons per layer
1	6	5
2	5	5
3	4	5
4	4	8
5	4	9
6	3	10
7	3	5

C. Design of the fuzzy Inference System to Detect the Presence of Vibrio Cholerae Bacteria

The designed fuzzy system is based on a Mamdani system to establish the input set for detecting the Vibrio cholera bacteria. The variables considered are temperature, phosphates and nitrates levels, and salinity. The output set is defined in terms of the presence or absence of the bacteria as shown in Fig. 3.

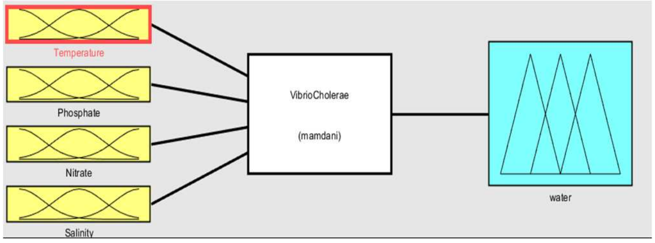


Fig. 3 Fuzzy system

Membership functions are defined for the input and output sets as follows:

1) Temperature (Fig. 4): the linguistic values low, medium, and high are adapted in three triangular membership functions in a 0 to 37 range.

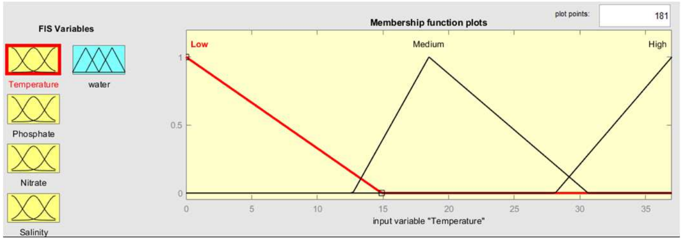


Fig. 4 Fuzzy sets for temperature classification.

2) Phosphate levels (Fig. 5): two triangular membership functions are set to classify them in a 0 to 1 range (low and high).

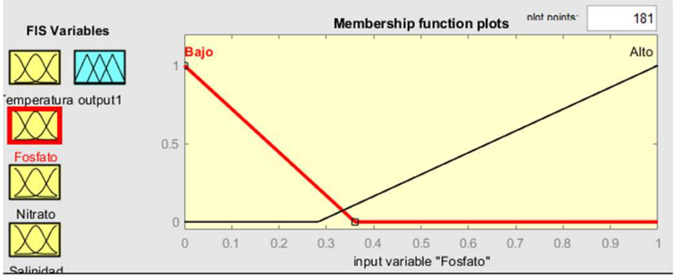


Fig. 5 Fuzzy sets for Phosphate levels.

3) Nitrate levels (Fig. 6): two triangular membership functions are set to classify them into high (0-8 range) and low(8-20 range).

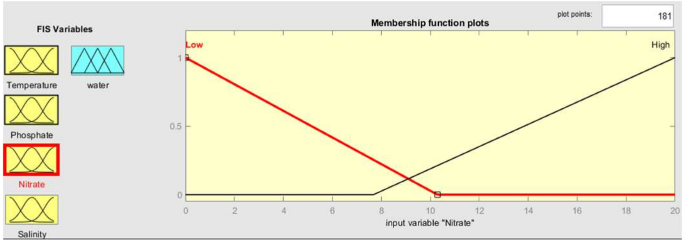


Fig. 6 Fuzzy sets for Nitrate levels.

*4) Salinity (Fig. 7)***:** two triangular membership functions are set to classify them into high (10-20 range) and low (0-10 range).

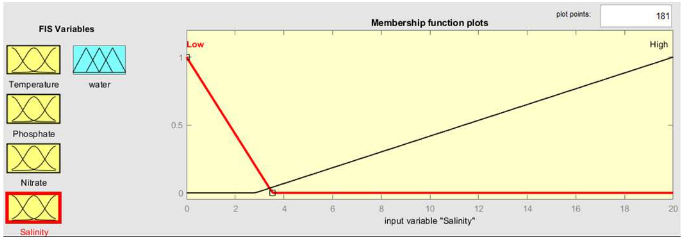


Fig. 7 Fuzzy sets for Salinity levels.

The output set (Fig. 8) is defined in a 0 to 1 range through triangular membership functions and assesses the water as sane or contaminated.

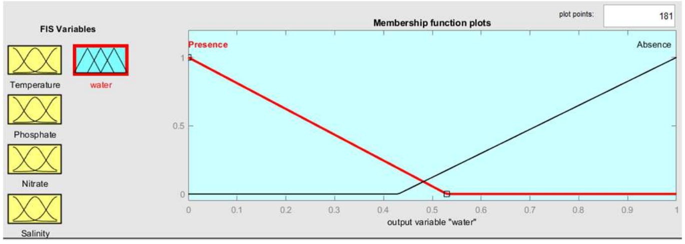
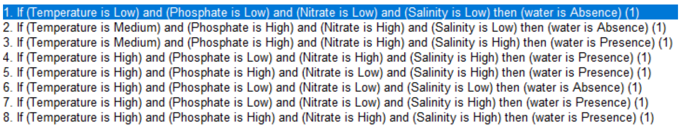


Fig. 8 Output set.

The fuzzy inference system associates the input sets and the established system rule set (Fig. 9) to show the water conditions and behaviors related to the presence and reproduction of the Vibrio Cholerae bacteria.



1. If (Temperature is Low) and (Phosphate is Low) and (Nitrate is Low) and (Salinity is Low) then (water is Absence) (1)
2. If (Temperature is Medium) and (Phosphate is High) and (Nitrate is High) and (Salinity is Low) then (water is Absence) (1)
3. If (Temperature is Medium) and (Phosphate is High) and (Nitrate is High) and (Salinity is High) then (water is Presence) (1)
4. If (Temperature is High) and (Phosphate is Low) and (Nitrate is High) and (Salinity is High) then (water is Presence) (1)
5. If (Temperature is High) and (Phosphate is High) and (Nitrate is Low) and (Salinity is High) then (water is Presence) (1)
6. If (Temperature is High) and (Phosphate is Low) and (Nitrate is Low) and (Salinity is Low) then (water is Absence) (1)
7. If (Temperature is High) and (Phosphate is Low) and (Nitrate is Low) and (Salinity is High) then (water is Presence) (1)
8. If (Temperature is High) and (Phosphate is High) and (Nitrate is High) and (Salinity is High) then (water is Presence) (1)

Fig. 9 Fuzzy system rules set.

III. Results and Discussion

A. Water Quality Prediction

All of the neural networks used were trained and tested using 70% and 30% of the input dataset. The outcomes were validated through different rubrics used in regression and prediction models such as the Mean Square Error (MSE) and Maximum and Minimum error values. TABLE shows all the error values obtained as well as the effectiveness percentage represented by the RMSE for each network; the effectiveness range goes from 98.68% for the less accurate trial to 99.99% for the top one. According to the results, the network with the best performance (MSE 1.16X10^{-10}, and 99.99% effectiveness) is network number 2, configured using five hidden layers and five neurons per layer, as shown in Fig. 10. When compared, the data collected using the best performing neural network for the obtained results and the expected ones on the prediction of the status of the Elwha River are significantly similar as shown in Fig. 11.

TABLE III
Results for The Water Quality Prediction

NN	Minimum Error	Maximum Error	MSE	RMSE	Effectiveness (%)
1	1.05E-07	8.92E-03	1.68E-06	1.30E-03	99.87036
2	4.73E-08	3.34E-05	1.16E-10	1.079E-05	99.99892
3	3.01E-08	1.75E-03	2.82E-08	1.68E-04	99.98321
4	2.05E-12	1.67E-02	8.64E-06	2.94E-03	99.70609
5	7.67E-14	4.11E-02	9.71E-05	9.86E-03	99.01439
6	2.28E-10	8.05E-02	1.72E-04	1.31E-02	98.68912
7	5.95E-08	6.66E-02	4.03E-05	6.35E-03	99.36515

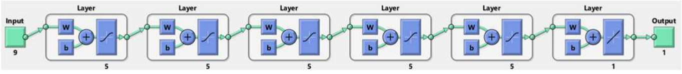


Fig. 10 Neural Network configuration #2

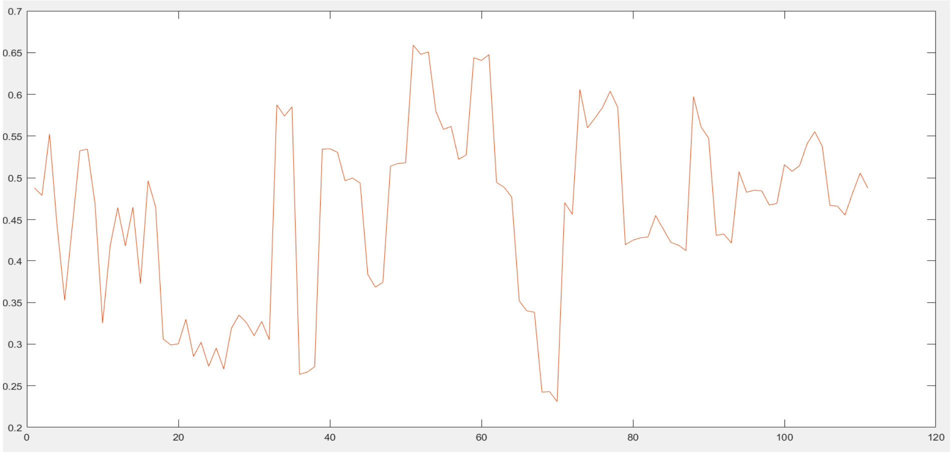


Fig. 11 Expected data compared to the results of the best performing neural network.

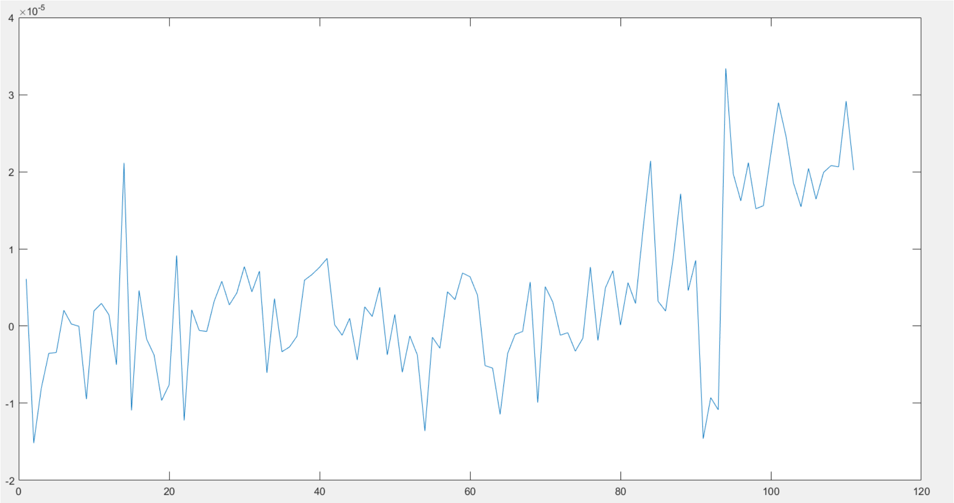


Fig. 12 Best performing Network error.

Fig. 12 shows the error graph for the water quality prediction made by the best performing network; the minimum difference between the expected and obtained values explains the 10-5 order value for the error.

B. Detection of the Vibrio Cholerae Bacteria

The fuzzy inference system used in the second trial was tested with the totality of the input dataset, and its performance was validated using specialized classification rubrics such as Accuracy, Precision, recall, and f1-score. The resulting outcome for the fuzzy inference system is shown in

Fig. 13 through a Classification Confusion Matrix that depicts the hit ratio for each of the possible options: presence (0) and absence (1) of the vibrio cholera bacterium.

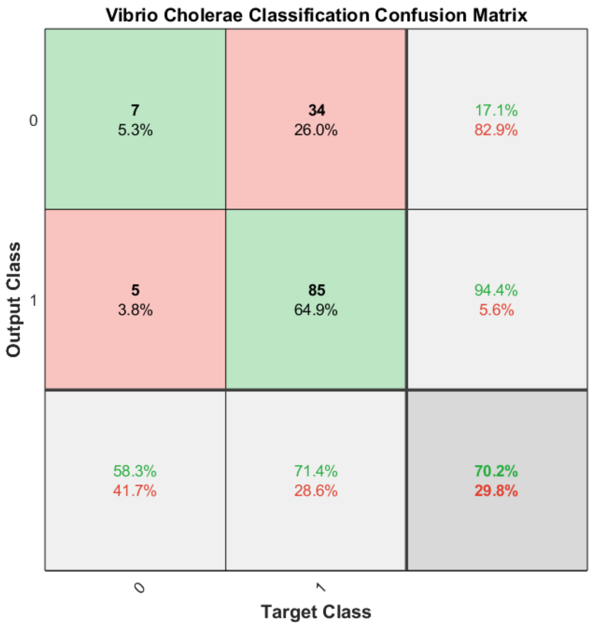


Fig. 13 Fuzzy system confusion matrix

All the calculated values to assess the system's performance, using every defined rubric, are shown in TABLE .

TABLE IV
FUZZY SYSTEM RESULTS.

Accuracy	Precision	Recall	F1-score
70.23%	71.43%	94.44%	81.34%

C. *Discussion*

The fuzzy inference system used to detect the presence of the vibrio cholera bacteria in a water supply showed high interpretability as the rules that deal with the output set (temperature, salinity, sodium, and phosphorus levels) are directly related to the circumstances under which these specific bacteria can proliferate and survive. Nevertheless, as this AI technique is not highly accurate, the system's performance got a 70.23% score under the Accuracy rubric. Also, it can be inferred that greater accuracy levels can be achieved by using an input dataset completely focused on detecting the Vibrio Cholerae in water supplies.

IV. CONCLUSIONS

The neural network designs' results show high effectiveness in assessing the quality of a water supply, taking into account all the Physico-chemical parameters established in the input dataset. According to the measurements and the MSE, the systems operate on an effective range of 98.68% to 99.99%. Furthermore, the best performing configuration (5 layers with five neurons per layer) shows a 1.079x10^{-5} MSE value, implying a high capacity to assess the water supply as drinkable. The fuzzy system designed to detect the presence of the vibrio cholera bacteria showed 71.43% and 70.23% values for precision and accuracy, respectively. This implies that the system can recognize the presence of the bacteria in the water regarding the Physico-chemical characteristics that would foster its proliferation and survival (temperature, phosphates and nitrates' levels, and salinity). As the system graphically shows the correspondences between the rules established and the membership functions in the input and output sets, it can be said that the system has great interpretability and is transparent to the people approaching its design and setting. For future research direction, in detecting the bacteria in the water supply, we recommend an increase in the accuracy of the fuzzy system through a reevaluation of the membership functions in the input and output sets. This can be achieved using optimization techniques such as genetic and gradient algorithms, which tend to converge towards a local or global minimum based on the difference in the expected and obtained values.

REFERENCES

[1] J. Juntunen, P. Meriläinen, and A. Simola, “Public health and economic risk assessment of waterborne contaminants and pathogens in Finland,” *Sci. Total Environ.*, vol. 599–600, pp. 873–882, Dec. 2017, doi: 10.1016/j.scitotenv.2017.05.007.

[2] L. E. Armstrong and E. C. Johnson, “Water intake, water balance, and the elusive daily water requirement,” *Nutrients*, vol. 10, no. 12, Dec. 2018, doi: 10.3390/nu10121928.

[3] C. Troeger *et al.*, “Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a systematic analysis for the Global Burden of Disease Study 2016,” *Lancet Infect. Dis.*, vol. 18, no. 11, pp. 1211–1228, Nov. 2018, doi: 10.1016/S1473-3099(18)30362-1.

[4] M. Ali, A. R. Nelson, A. L. Lopez, and D. A. Sack, “Updated Global Burden of Cholera in Endemic Countries,” *PLoS Negl. Trop. Dis.*, vol. 9, no. 6, p. e0003832, Jun. 2015, doi: 10.1371/journal.pntd.0003832.

[5] A. Richterman, M. F. Franke, G. Constant, G. Jerome, R. Ternier, and L. C. Ivers, “Food insecurity and self-reported cholera in Haitian households: An analysis of the 2012 Demographic and Health Survey,” *PLoS Negl. Trop. Dis.*, vol. 13, no. 1, p. e0007134, Jan. 2019, doi: 10.1371/journal.pntd.0007134.

[6] S. Lonappan, R. Golecha, and G. Balakrish Nair, “Contrasts, contradictions and control of cholera,” *Vaccine*, vol. 38, pp. A4–A6, Feb. 2020, doi: 10.1016/j.vaccine.2019.08.022.

[7] M. M. Foley, A. Ritchie, P. B. Shafroth, J. J. Duda, M. M. Beirne, and R. Paradis, “Water quality in the Elwha River estuary, Washington, from 2006 to 2014,” *U.S. Geol. Surv. data release*, vol. 8, no. 2, pp. 552–577, Aug. 2017, doi: 10.1002/ecm.1268.

[8] U.S. Government Works, “Water quality in the Elwha River,” Feb. 2018.

[9] L. Li, S. Rong, R. Wang, and S. Yu, “Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review,” *Chem. Eng. J.*, vol. 405, p. 126673, Feb. 2021, doi: 10.1016/j.cej.2020.126673.

[10] J. G. Nayak, L. G. Patil, and V. K. Patki, “Development of water quality index for Godavari River (India) based on fuzzy inference system,” *Groundw. Sustain. Dev.*, vol. 10, p. 100350, Apr. 2020, doi: 10.1016/j.gsd.2020.100350.

[11] Y. Zhao, T. Li, X. Zhang, and C. Zhang, “Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future,” *Renew. Sustain. Energy Rev.*, vol. 109, pp. 85–101, Jul. 2019, doi: 10.1016/j.rser.2019.04.021.

[12] Y. Deng, H. Xiao, J. Xu, and H. Wang, “Prediction model of PSO-BP neural network on coliform amount in special food,” *Saudi J. Biol. Sci.*, vol. 26, no. 6, pp. 1154–1160, Sep. 2019, doi: 10.1016/j.sjbs.2019.06.016.

[13] I. Škrjanc, J. Iglesias, A. Sanchis, D. Leite, E. Lughofer, and F. Gomide, “Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A Survey,” *Inf. Sci. (Ny).*, vol. 490, pp. 344–368, Jul. 2019, doi: 10.1016/j.ins.2019.03.060.

[14] H. Yan, Y. Liu, X. Han, and Y. Shi, “An evaluation model of water quality based on DSA-ELM method,” in *ICOCN 2017 - 16th International Conference on Optical Communications and Networks*, Nov. 2017, vol. 2017-January, pp. 1–3, doi: 10.1109/ICOCN.2017.8121280.

[15] K. Joslyn and J. Lipor, “A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection,” in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, Jan. 2019, pp. 2511–2514, doi: 10.1109/BigData.2018.8622628.

[16] P. Huang *et al.*, “An integrated modelling system for water quality forecasting in an urban eutrophic estuary: The swan-canning estuary virtual observatory,” *J. Mar. Syst.*, vol. 199, p. 103218, Nov. 2019, doi: 10.1016/j.jmarsys.2019.103218.

[17] A. Delgado, A. Aguirre, E. Palomino, and G. Salazar, “Applying triangular whitenization weight functions to assess water quality of main affluents of Rimac river,” in *Proceedings of the 2017 Electronic Congress, E-CON UNI 2017*, Jun. 2017, vol. 2018-January, pp. 1–4, doi: 10.1109/ECON.2017.8247308.

[18] N. S. K. Gunda, S. H. Gautam, and S. K. Mitra, “Artificial Intelligence Based Mobile Application for Water Quality Monitoring,” *J. Electrochem. Soc.*, vol. 166, no. 9, pp. B3031–B3035, Mar. 2019, doi: 10.1149/2.0081909jes.

[19] P. Wang *et al.*, “Exploring the application of artificial intelligence technology for identification of water pollution characteristics and tracing the source of water quality pollutants,” *Sci. Total Environ.*, vol. 693, p. 133440, Nov. 2019, doi: 10.1016/j.scitotenv.2019.07.246.

[20] J. Leo, E. Luhanga, and K. Michael, “Machine Learning Model for Imbalanced Cholera Dataset in Tanzania,” *Sci. World J.*, vol. 2019, 2019, doi: 10.1155/2019/9397578.

[21] A. S. Khalid Waleed, P. D. Kusuma, and C. Setianingsih, “Monitoring and classification system of river water pollution conditions with fuzzy logic,” in *Proceedings - 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2019*, Jul. 2019, pp. 112–117, doi: 10.1109/ICIAICT.2019.8784857.

[22] U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar, and H. Khurshid, “Surface Water Pollution Detection using Internet of Things,” in *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT and IoT, HONET-ICT 2018*, Nov. 2018, pp. 92–96, doi: 10.1109/HONET.2018.8551341.

[23] Y. Wang, J. Zhou, K. Chen, Y. Wang, and L. Liu, “Water quality prediction method based on LSTM neural network,” in *Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2017*, Jul. 2017, vol. 2018-January, pp. 1–5, doi: 10.1109/ISKE.2017.8258814.

[24] R. S. Pressman, *Ingenieria del Software. Un Enfoque Practico*. 2010.

[25] N. Gilbert and A. Rusli, “Single object detection to support requirements modeling using faster R-CNN,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 830–838, Apr. 2020, doi: 10.12928/TELKOMNIKA.V18I2.14838.

[26] H. Espitia, J. Soriano, I. Machón, and H. López, “Design Methodology for the Implementation of Fuzzy Inference Systems Based on Boolean Relations,” *Electronics*, vol. 8, no. 11, p. 1243, Oct. 2019, doi: 10.3390/electronics8111243.