Journal of Informatics
Electrical and Electronics Engineering, 2023,
Vol. 04, Iss. 01, S. No. 007, pp. 1-9
ISSN (Online): 2582-7006

Assessment of Water Quality using Machine Learning and Fuzzy Techniques

Shashi Kant[1], Devendra Agarwal[2], Praveen Kumar Shukla[3]

[1,2,3]Artificial Intelligence Research Center, Department of CSE, School of Engineering, Babu Banarasi Das University, Lucknow, India
[1]shashikant3245@gmail.com, [2]devendragarwal@gmail.com, [3]drpraveenkumarshukla@gmail.com

How to cite this paper: S. Kant, D. Agarwal and P. K. Shukla, "Assessment of Water Quality using Machine Learning and Fuzzy Techniques," *Journal of Informatics Electrical and Electronics Engineering* (*JIEEE*), Vol. 04, Iss. 01, S No. 007, pp. 1–9, 2023.

https://doi.org/10.54060/jieee.v4i1.91

Received: 04/04/2023
Accepted: 023/04/2023
Published: 25/04/2023

Abstract

The water quality of river Ganga is an important concern due to its drinking, domestic uses, irrigation and also for aquatic life. But the extent of pollutants in river water has deteriorated the quality of river water. So, the assessment of river water becomes very important. But due to the involved subjectivity and uncertainty in the decision making parameter makes the task very complex. In this study, machine learning and fuzzy techniques are utilized to develop the river water quality assessment models. The quality of the water is grouped into three classes. Four machine learning algorithms namely decision tree, random forest tree, k-nearest neighbor and support vector machine are used and implemented on python and anaconda platform. Whereas, three fuzzy based models (fuzzy decision tree, wang-mendel and fast prototyping) are developed using Guaje open source software. All the seven models are analyzed in terms of accuracy, precision, recall and f1-score. The observed result shows that the fuzzy decision tree-based assessment model performs more accurately as compared with the machine learning based models.

Keywords

River water model, machine learning, fuzzy system, Ganga River

1. Introduction

River water is the one of the most vital resources for all kinds of life. It is playing important role for sustaining a good health in human life and also for the growth of nation's economy. However, it is in persistent danger of pollution by life itself. Industrial wastes, marine dumping, radioactive waste, atmospheric deposition, domestic discharge and many more are the

reasons which led to the deterioration of water quality to an extreme level. The consumption of poor quality water in daily life is a major factor for the increase in diseases like Cholera, Diarrhoea, Malaria, Typhoid, and Filariasis [1]. The poor water quality also causes a GDP loss of country every year [2].

Considering the above horrific consequences, it becomes very necessary to assess the quality of water. At present, the quality of river water is assessed through an expensive and time-consuming process since it includes sample collection, time to move the sample to labs and a considerable amount of time taken for experiments and statistical analysis. In this regard to overcome this inefficient approach, a model based on machine learning (ML) and fuzzy logic has been implemented and studied for the assessment of water quality in real time. The assessment of water quality is particularly uncertain due to the constant change in the values of decision parameters [4]. The types of water quality parameters are summarized in Table 1.

Table 1. Classification of water quality parameters [5].

S. No.	Chemical Parameters
1	pH
2	Acidity
3	Alkalinity
4	Chloride
5	Chlorine residual
6	Sulfate
7	Nitrogen
8	Fluoride
9	Iron and manganese
10	Copper and zinc
11	Hardness
12	Dissolved oxygen
13	Biochemical oxygen demand (BOD)
14	Chemical oxygen demand (COD)
15	Toxic inorganic substances
16	Toxic organic substances
17	Radioactive substances

S. No.	Physical Parameters
1	Turbidity
2	Temperature
3	Color
4	Taste and odor
5	Solids
6	Total Solids
7	Total dissolved solids
8	Total suspended solids
9	Electrical conductivity (EC)
Biological Parameters	
1	Bacteria
2	Algae
3	Viruses
4	Protozoa

Out of above identified parameters, four parameters are selected as decision parameter for the assessment of water quality. These are: Dissolved Oxygen (DO), Bio-Chemical Oxygen Demand (BOD), Total Coli, and pH.

To deal with the inefficiencies of traditional approach, machine learning and fuzzy techniques can be used to automate the process of assessment of water quality. The model also ensures the improvement in accuracy for the assessment of water quality. Arthur Samuel commented that machine learning algorithms helps system to learn from data to make decision, and also improves itself without being programmed. Fuzzy approaches, on the other hand, lead to the creation of fuzzy rule-based systems for assessing water quality in order to deal with the inherent ambiguity and subjectivity.

The main contributions of this study are as follows:

1. Several parameters are identified and analyzed for the assessment of water quality.
2. Different machine learning and fuzzy based models (classification) are developed and tested on the pre-identified dataset. Each model is verified quantitatively.

The rest of this paper is structured as follows: Section 2 depicts relevant work in the domain. To evaluate the model, we used machine learning methods and fuzzy based algorithms in Section 3. Section 4 implements the suggested model on a dataset to assess water quality, and the results are analyzed in terms of assessment metrics such as accuracy, precision, recall, and f1-score. In Section 5, we ended the paper research and discussed future work.

2. Related Work

A hybrid prediction model utilizing the random forest tree, M5P, reduced error pruning tree, and 12 hybrid data algorithm has been proposed [6]. Using 10 fold cross-validation procedures, the dataset is split into training and testing datasets in a 70:30 ratios. The hybrid approach increased the forecast accuracy of a number of independent models. A machine learning model was created to analyze the water quality of the Indiana River [7]. After preprocessing the dataset provided by the Central Pollution Control Board of India, a total of 8 characteristics were chosen. The model has a 96.1% accuracy rate.

For assessing the water quality of the Ganga, a fuzzy knowledge-based method has been created [8]. Four parameters are used as decision parameters for the prediction: dissolved oxygen (DO), biochemical oxygen demand (BOD), pH, and total coli-form. The Wang-Mendel approach for rule creation produces more precise results. In [9], a machine learning approach has been proposed as the solution to anomalies occurring on water quality time series data. Several models such as support vector machines, artificial neural network, logistic regression, etc. are applied to water quality data. The experimental study shows that SVM model performs better when applied to highly imbalanced dataset.

Two classification models, Random Forest and Random Tree algorithm, for water quality have been developed using WEKA data mining tool [10]. The simulation result showed that Random Forest performs better as compared to Random Tree algorithm. Support vector machine, a machine learning approach, is used to suggest a model for predicting the water quality of the Ganges [11]. The SVM classifier is created using a kernel called the Radial Basis Function (RBF). The prediction made by the model was 96.66% accurate. To assess the water quality, a comparison of classification methods has been carried out [12]. The ground water decision parameter was chosen to be the electrical conductivity. The outcome revealed that the SVM-implemented model was superior to the other classification methods.

3. Proposed Model

An assessment model based on machine learning and fuzzy techniques are proposed to assess the water quality. The different algorithms used for the assessment of water quality are given in table 2.

Table 2. Different algorithms used for the Assessment of Water Quality.

S. No.	Proposed Approach	Algorithms used
1.	Machine Learning Based Modelling	Decision tree [13] Random forest tree [14] k-nearest neighbor [15] support vector machine [16]
2.	Fuzzy Based Modelling (hfp partition)	Fuzzy decision tree [17] Wang-Mendel [18] Fast prototyping

The four decision parameters which are used as input to the model are given in table 3. The output of classification model is divided into three classes as given in table 4.

Table 3. Decision Parameters for the Assessment of Water Quality.

S.No.	Parameter	Description
1.	Dissolved Oxygen (DO)	This is how much dissolved oxygen is present in the water. The quality of water improves as DO content rises.
2.	Bio-chemical Oxygen Demand (BOD)	This is the amount of oxygen needed in a certain amount of water to break down organic materials. Higher BOD is indicated by the presence of more organic compounds in water.
3.	pH	The pH of water is a test to determine whether it is basic or acidic. Water has a pH that ranges from 0 to 14. Since pH 7 is considered neutral, a pH of less than 7 denotes an acidic environment, and a pH of more than 7 denotes a base solution. For home use, drinking water should have a pH of 6.5 to 8.5.
4.	Total Coliform	This is the collection of all pathogens that are related but not vulnerable. Their existence in water, however, implies that disease-causing microorganisms may be present. As a result, the Total Count of such pathogens is used to detect the presence of potentially dangerous pathogens in water.

The working of the proposed machine learning and fuzzy based model is shown in figure.1 and figure. 2.

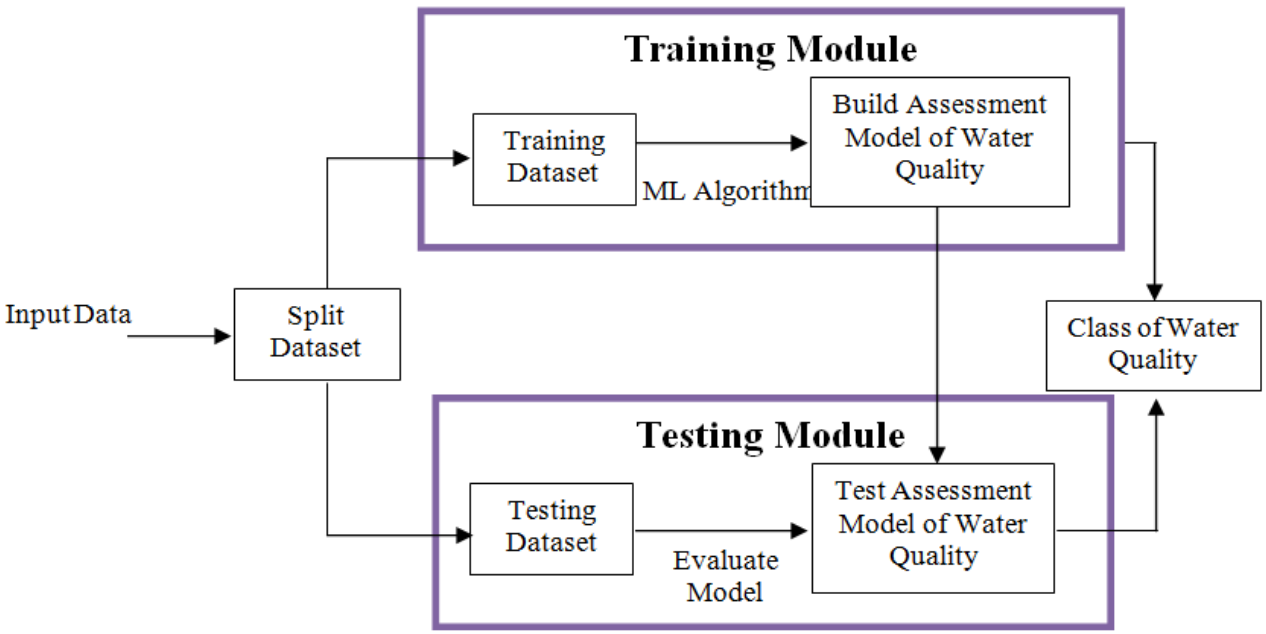

Figure 1. Framework of Proposed Machine Learning Model.

Table 4. Output (Class) of assessment Model.

S. No.	Class	Class Description
1.	Class 1	After sterilization, it is safe to drink.
2.	Class 2	After sterilization, suitable for drinking with conventional treatment.
3.	Class 3	Drinkable only after extensive treatment, including sterilization.

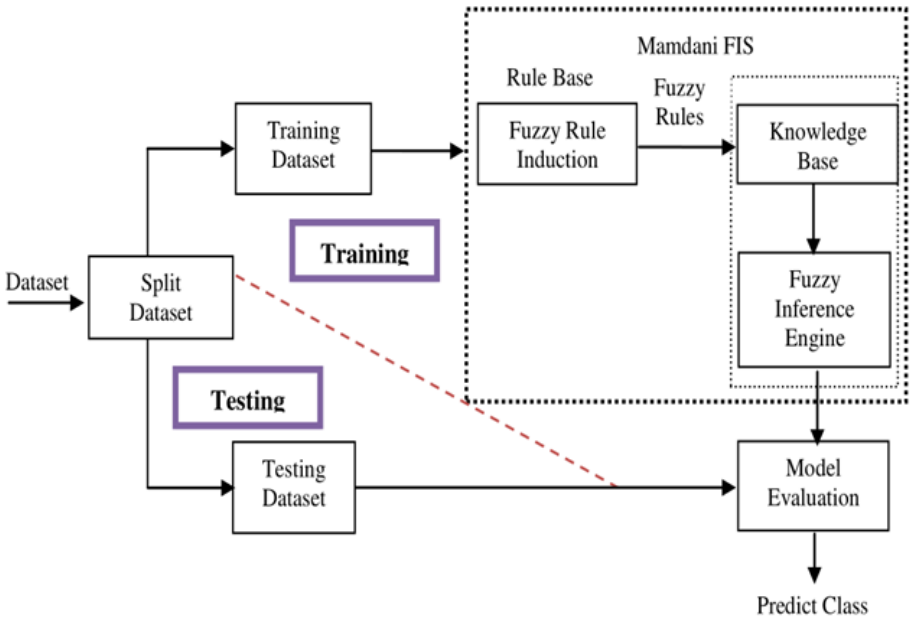

Figure 2. Framework of Proposed Fuzzy Model.

4. Implementation and Result Analysis

Using data acquired from the Namami Gange Yojna section of the government of India's official website, the Ganga's water quality is evaluated. 32 distinct Bihar and Uttar Pradesh locales are used to gather the samples. The suggested machine learning model is implemented using Python 3.6.3 and Anaconda 5.0.0, while the fuzzy-based assessment model is developed using Guaje open source software. The correctness of the suggested model is used to analyze the outcome.

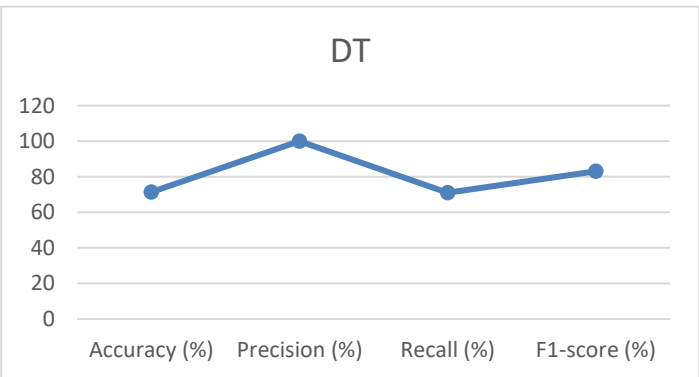

Figure 3 (a). Comparing Results of decision tree

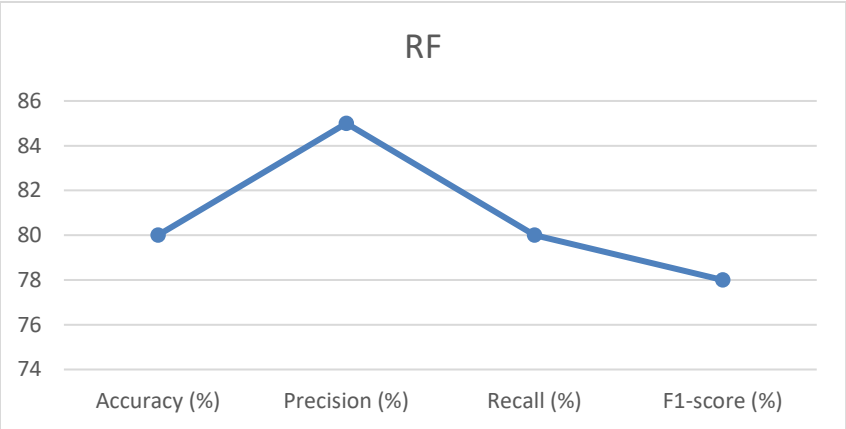

Figure 3 (b). Comparing Results of random forest

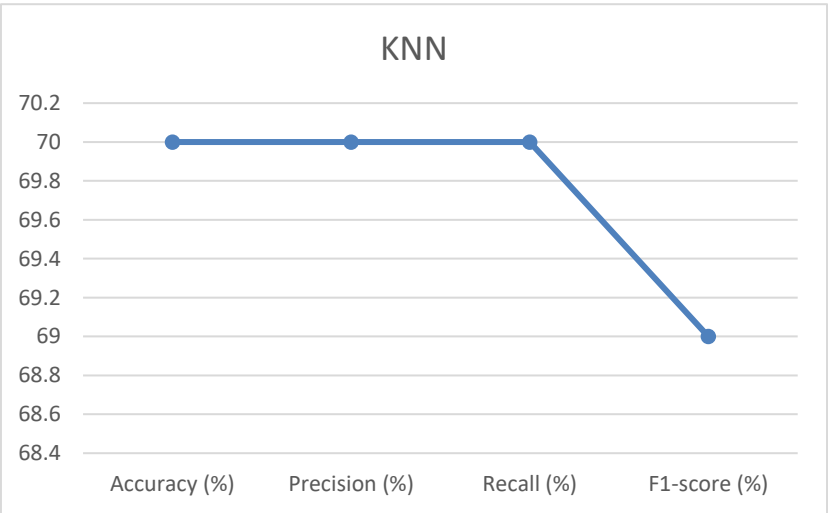

Figure 3(c). Comparing Results of K nearest neighbor

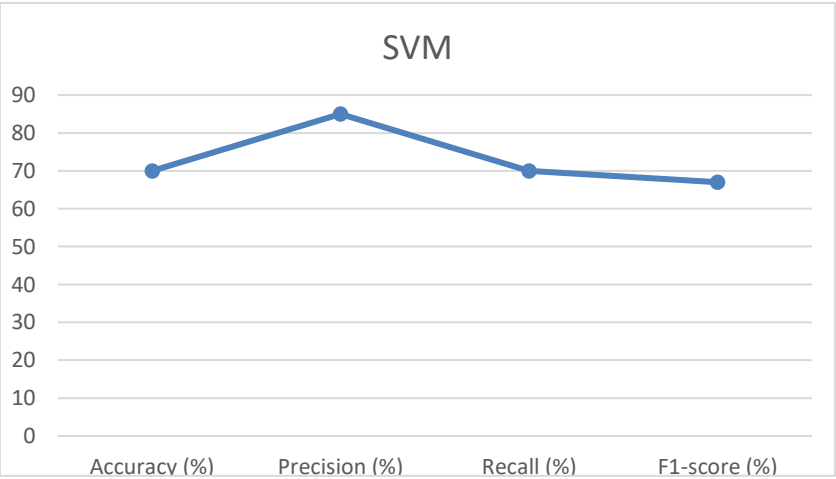

Figure 3(d). Comparing Results of support vector machine

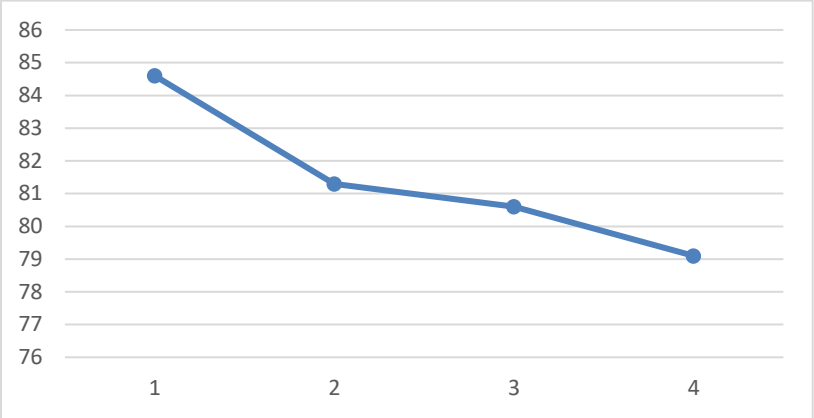

Figure 4(a). Comparing Results of Fuzzy Decision

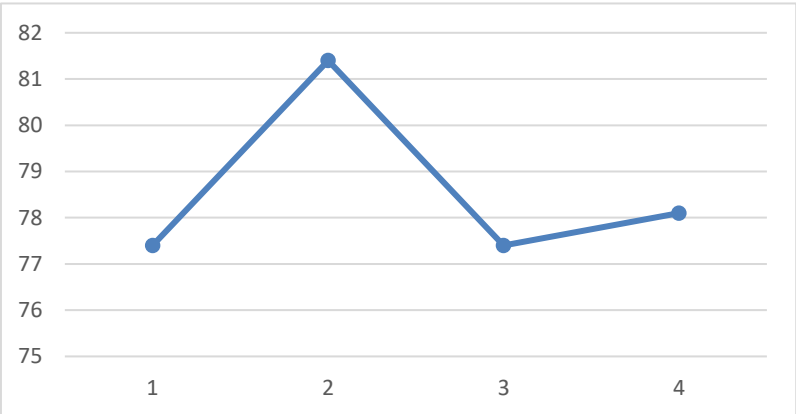

Figure 4(b). Comparing Results of Wang-Mendel

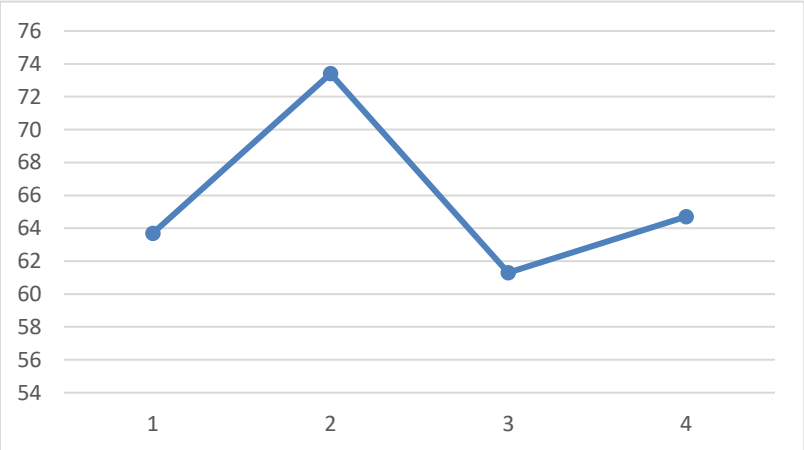

Figure 4(c). Comparing Results of Fast Prototyping

Following findings are observed after analyzing the result of both the model:

1. The river water quality assessment model is created using machine learning methods based on decision trees, random forest tree, k-nearest neighbors, and support vector machines. The water quality has been more properly evaluated using the random forest approach.
2. To develop a fuzzy rule-based model to evaluate water quality, Wang-Mendel, Fast Prototyping Rule Induction, and Fuzzy Decision Tree algorithms were employed. Fuzzy decision tree has given more accurate result as compared with other competent fuzzy algorithms.

5. Conclusion and Future Scope

The decision making parameters for assessing river water quality are imprecise and subjective due to frequent changes in their values. The result presented in this paper strongly approves a competent model based on fuzzy techniques because it can handle the subjectivity of the problem. The random forest approach of machine learning based model has performed quite well but fuzzy decision tree-based water quality assessment model has been found more precise and suitable with higher accuracy. The author is interested in developing a more accurate and interpretable assessment model by optimizing the different decision-making parameters of water quality in the coming future.

References

[1]. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, “Efficient Water Quality Prediction Using Supervised Machine Learning,” Water, vol. 11, no. 11, p. 2210, Nov. 2019.

[2]. S. Desbureaux, R. Damania, A. Rodella, J. Russ, and E. Zaveri, “The Impact of Water Quality on GDP Growth,” World Bank, Washington, DC, Jan. 2019.

[3]. A. H. Haghiabi, A. Nasrolahi, and A. Parsaie, “Water quality prediction using machine learning methods,” Water Quality Research Journal, vol. 53, no. 1, pp. 3–13, Jan. 2018.

[4]. S. Babaei et al., “Water quality index development using fuzzy logic: A case study of the Karoon River of Iran,” African Journal of Biotechnology, vol. 10, no. 50, pp. 10125–10133, Sep. 2011.

[5]. N. H. Omer, “Water Quality Parameters,” Water Quality - Science, Assessments and Policy, Oct. 2019.

[6]. D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, “Improving prediction of water quality indices using novel hybrid machine-learning algorithms,” Science of The Total Environment, vol. 721, p. 137612, Mar. 2020.

[7]. M. Ahmad, “Machine Learning Approach for Predicting the Quality of Water,” 2020.

[8]. P. K. Shukla, “Development of Fuzzy Knowledge-Based System for Water Quality Assessment in River Ganga,” Advances in Intelligent Systems and Computing, pp. 17–26, 2020.

[9]. F. Muharemi, D. Logofătu, and F. Leon, “Machine learning approaches for anomaly detection of water quality on a real-world data set,” Journal of Information and Telecommunication, vol. 3, no. 3, pp. 294–307, Feb. 2019.

[10]. S. Nafi, A. Mustapha, S. A. Mostafa, S. H. Khaleefah, and M. N. Razali, “Experimenting Two Machine Learning Methods in Classifying River Water Quality,” Communications in Computer and Information Science, pp. 213–222, 2020.

[11]. A. K. Bisht, R. Singh, R. Bhutiani, and A. Bhatt, “Application of Predictive Intelligence in Water Quality Forecasting of the River Ganga Using Support Vector Machines,” Predictive Intelligence Using Big Data and the Internet of Things, 2019.

[12]. R. Prakash, V. P. Tharun, and S. Renuga Devi, “A Comparative Study of Various Classification Techniques to Determine Water Quality,” IEEE Xplore, 01-Apr-2018.

[13]. S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology”, IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660–674, 1991.

[14]. M. Pal, “Random forest classifier for remote sensing classification”, International Journal of Remote Sensing, vol. 26, no. 1, pp. 217–222, Jan. 2005.

[15]. K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, “A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques”, CSI Transactions on ICT, vol. 4, no. 2, pp. 313–319, Dec. 2016.

[16]. R. Besrour, Z. Lachiri, and N. Ellouze, “ECG beat classifier using support vector machine”, 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, pp. 1–5, 2008.

[17]. H. Ichihashi, T. Shirai, K. Nagasaka, and T. Miyoshi, “Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning”, Fuzzy Sets and Systems, vol. 81, no. 1, pp. 157–167, Jul. 1996.

[18]. L. Wang and J. Mendel, “Generating fuzzy rules by learning from examples”, IEEE Transactions on Systems, Man, and Cybernetics, vol. 22, no. 6, pp. 1414–1427, Nov. 1992.

[19]. P. K. Shukla and S. P. Tripathi, “A new approach for tuning interval type-2 fuzzy knowledge bases using genetic algorithms”,” Journal of Uncertainty Analysis and Applications, vol. 2, no. 1, pp. 1–15, 2014.

[20]. P. K. Shukla and S. P. Tripathi, “A Review on the Interpretability-Accuracy Trade-Off in Evolutionary Multi-Objective Fuzzy Systems (EMOFS)”,” Information, vol. 3, pp. 256–277, 2012.