Water Quality Research Journal

Water Quality Research Journal Vol 58 No 3, 230 doi: 10.2166/wqrj.2023.031

A Fuzzy Inference System for enhanced groundwater quality assessment and index determination

Isaac Sajan R.[a] and V. Bibin Christopher[b,*]

[a] Department of Electronics and Communication Engineering, Ponjesly College of Engineering, Alamparai, Tamil Nadu 629 003, India
[b] Department of Computing Technologies, School of Computing, College of Engineering and Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar Kattankulathur, Chennai, Tamil Nadu, India
*Corresponding author. E-mail: bibinchrist85@gmail.com

ABSTRACT

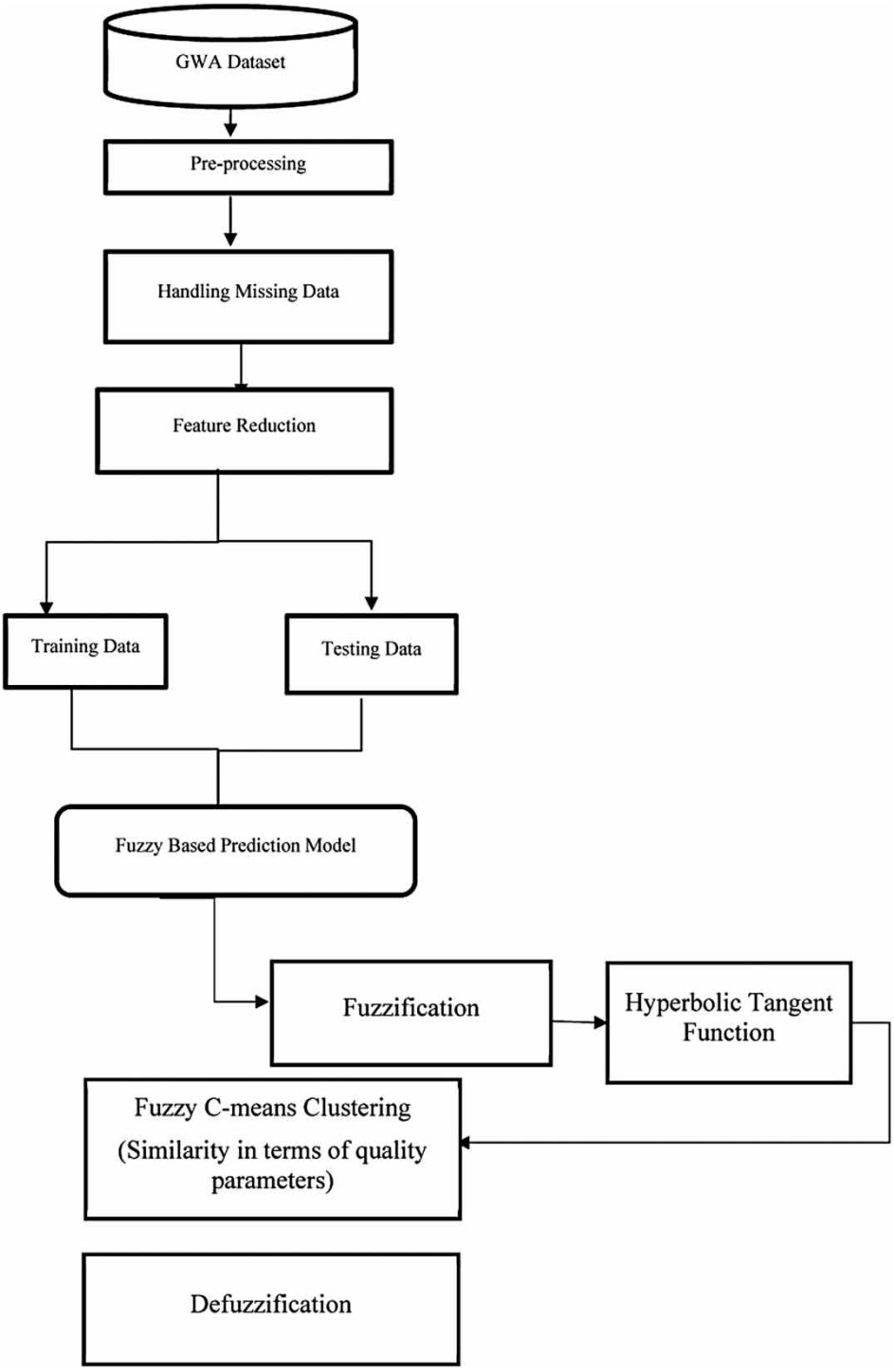
Groundwater is a vital resource for human consumption, particularly in rural areas with limited access to treated water. The conventional Water Quality Index models used for this purpose have limitations related to data volatility and judgment uncertainties. To overcome these limitations, our study introduces a novel approach that employs a Fuzzy Inference System to determine the Water Quality Index. The dataset used in our research includes multiple parameters such as pH, EC, TDS, Ca, Mg, Na, K, HCO_3, Cl, SO_4, TH, DWQI, and other physio-chemical and chemical parameters. Our approach utilizes linguistic variables, fuzzy rules, and the hyperbolic tangent set function to handle imprecise and uncertain water quality data. By employing Fuzzy C-Means clustering, we group similar water samples based on quality parameters and map membership values to linguistic terms representing water quality categories. Suitable defuzzification methods are then applied to convert fuzzy outputs into precise results. This proposed approach provides a comprehensive framework for accurate water quality assessment, enabling informed decision-making and more reliable and precise evaluations of groundwater quality.

Key words: defuzzification, fuzzification, Fuzzy C-Means clustering, Fuzzy Inference System, hyperbolic tangent set function, total dissolved solids

HIGHLIGHTS

- A unique usage of the Fuzzy Inference System (FIS) to determine the Water Quality Index (WQI).
- This research evaluated quality using pH, total hardness, total dissolved solids, calcium, and manganese.
- The recommended design method was compared against deterministic results to determine its feasibility.
- This paper aims to provide a fuzzy-based paradigm for evaluating groundwater safety for human consumption.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Water covers 70% of the earth's surface. The most vital component of our existence is water. It has to be used for various purposes like drinking, cooking, washing, plantation, irrigation, etc. Recently, water has also been used in many sports

events and entertainment, which gives some revenue (Jennings 2007). There are various sources of water on earth, namely the sea, rivers, lakes, and groundwater.

Among the various sources of water, groundwater is the most readily and easily available for all humans (Lall *et al.* 2020). Since the sea water and the river water are not accessible for people who do not live near the delta, groundwater is the only source that is available to everyone, and it is available almost everywhere at a cost. Deltas are important natural features that exhibit dynamic environments, continuously evolving through the interaction of sediment deposition, erosion, and the forces of water and tides. The term 'delta' originates from the Greek letter delta (Δ), chosen due to its resemblance to the triangular shape commonly associated with these landforms.

Deltas possess distinctive ecosystems and habitats, making them significant in the natural world. They support a variety of flora and fauna and serve as crucial breeding grounds and nurseries for aquatic species. Due to their ever-changing nature, deltas are dynamic landscapes that play a vital role in shaping the surrounding environment.

The water is highly polluted by various effects of mankind, like the increase in population, industrial revolution, fertilizer, etc. (Cabral Pinto *et al.* 2020). The quality of the water is highly affected, and it impacts the survival of many organisms, which also affects our health. Drinking polluted water can lead to serious health problems (Zeilhofer *et al.* 2007), as well as an increase in mortality (Kahlown *et al.* 2007). It has an impact on major countries such as Egypt and China.

Water quality is of utmost importance for irrigation as it should not contain excessive salinity, harmful chemicals, or minerals that could adversely affect both the irrigation process and the surrounding ecosystem. Different industries also have specific water quality requirements depending on the minerals needed. Therefore, accurate prediction and maintenance of good water quality are essential.

To enhance water quality, several fundamental steps have been outlined in a study by Chen *et al.* (2017). These steps involve determining appropriate cropping patterns, selecting suitable irrigation systems, and implementing effective water purification methods, especially for industrial use. The quality of water plays a central role in water conservation efforts.

Given the widespread accessibility and affordability of groundwater, our objective is to assess its quality. It is important to note that groundwater may potentially contain toxic elements, such as potential toxic elements (PTEs), primarily due to the release of industrial waste, posing significant health hazards (Cabral Pinto *et al.* 2020).

Recognizing the significance of water purity, the Indian government has established the Indian Pollution Control to regularly monitor water quality through designated stations.

Water quality checking is very expensive and time-consuming, involving taking the water to the lab and running certain tests on machines that are costly and time-consuming. Water quality has reached an alarming level and it should be inexpensive so every living organism can access safe water. This motivates us to introduce the alternate solution, namely, a fuzzy-based prediction model which uses the Tamil Nadu Quality dataset, which is provided by the Tamil Nadu state government (Rama *et al.* 2021). We have implemented fuzzy logic because it can handle the complex linguistic data, which is deployed in the environmental monitoring system, and is simple (Ellina *et al.* 2020). The main contribution of this paper lies in its comprehensive approach to water quality assessment, incorporating the utilization of parameters such as the Water Quality Index (WQI) and the Trophic Level Index (TLI) (Liu *et al.* 2021). However, the distinctive aspect of this study is the development and implementation of a fuzzy inference model for predicting water quality. In addition to addressing missing data through pre-processing techniques and performing feature reduction, the proposed fuzzy inference model plays a crucial role in evaluating and predicting the quality of the water. By leveraging fuzzy logic and linguistic variables, this model effectively handles the inherent uncertainties and imprecisions associated with water quality assessment.

The article is structured as follows: (i) a review of the relevant literature and work conducted in the field; (ii) a discussion of the outcomes and findings obtained from applying the fuzzy inference model to water quality assessment; and (iii) a comprehensive conclusion summarizing the accomplishments of the research and outlining potential avenues for future work.

By integrating a fuzzy inference model into the water quality assessment process, this study offers an innovative and valuable contribution to the field, providing a robust framework for evaluating water quality and enhancing decision-making in water resource management.

2. RELATED WORK

Sahu *et al.* (2011) introduced the ANFIS (Adaptive Neuro-Fuzzy Inference System) in ground water near mines, which tends to be more contaminated. It uses PCA (Principle Component Analysis), which converts the correlated to the incorrelated data

and produces a fuzzy set of the quality of the water. It therefore shows better accuracy. But this process requires lots of training.

To predict the water quality in aquaculture, Liu *et al.* (2013) proposed SVM. But in SVM, choosing parameters and settings is an issue, so they introduced RGA-SVR (Real Value Genetic Algorithm Support Vector Regression), which is a genetic algorithm for choosing the parameters. This algorithm proves to be effective in nonlinear time series problems. But it needs lots of training and different types of mutations need to be set for different problems.

Tools like Fuzzy Logic (FL) and Fuzzy Inference System (FIS) are used to calculate the water quality in the reservoirs. It uses only eight parameters, so it is easy and cheaper. FIS has a total of 633 rules and seven verbal categories. Also, it has shown the best results in accuracy (Sedeño-Díaz & López-López 2016).

Khan & See (2016) has used Artificial Neural Network (ANN) with Nonlinear Autoregressive (NAR) time series and Scaled Conjugate gradient (SCG) as a training algorithm which uses four parameters: Chlorophyll, DO (Dissolved oxygen), turbidity, and specific conductance. This algorithm shows improved results in both performance and accuracy. Implementing it is a bit costlier.

A Fuzzy Wavelet Neural Network (FWNN) prediction model was proposed by Huang *et al.* (2018), introduced to check the water quality in rivers. It is based on both genetic and the gradient descent algorithm. This algorithm helps to handle the fluctuations and the non-seasonal time data with better accuracy, performance, and robustness.

Two prediction methods, namely the Improved Grey Relational Analysis (IGRA) algorithm and a Long-Short Term Memory (LSTM) neural network, were introduced by Zhou *et al.* (2018). He used IGRA for the feature selection and LSTM which helps to identify the water quality. But the main disadvantage he found is that it consumes more historical data and training time.

Ahmed *et al.* (2019) compared various artificial intelligence prediction techniques, such as ANN (Artificial Neural Network), GMDH (Group Method of Data Handling), and SVM (Support Vector Machine). According to the DDR indices, the SVM's data dispersion is less than the other two. Overall, GMDH and the SVM are more reliable compared to the ANN.

Ahmed *et al.* (2019) have recommended the technique WDT-ANFIS (Wavelet DeNoising Technique using ANFIS), which mainly depends on historical data to calculate the WQI. Two scenarios were introduced, which calculate the performance and accuracy and show better value when compared to the machine learning models.

The Bootstrap Wavelet Neural Network (BWNN) was developed to predict the ammonia nitrogen and DO in China monthly. Its performance was compared with that of ANN, WNN (Wavelet Neural Network), and bootstrapped ANN. The BWNN shows better results when there is a fluctuation in seasonal time series. It can handle missing data and produce a better result when the other can only produce a good result when all of the data are present on a regular basis.

The related work highlights the diverse range of prediction models utilized in water quality assessment, showcasing their strengths and limitations in various environmental contexts. These advancements contribute to the understanding and management of water resources, fostering informed decision-making and promoting sustainable water quality practices.

3. METHODOLOGY

Investigations dealing with the effects of human and climatic change may benefit tremendously from groundwater level time series. These time series are of significant relevance for a wide range of groundwater studies. Before moving on to any further applications, such as trend analysis, it is necessary to perform quality assurance checks on the groundwater level measurements. The quality management of data is often confined to the removal of outliers or the eradication of whole time series from a dataset, despite the fact that such procedures significantly diminish the geographical coverage of datasets that were once enormous. Studies often have a tendency to offer data that has already undergone quality control, but they frequently fail to illustrate how the data were chosen, evaluated, and altered as shown in Figure 1.

The collection of obtained dataset includes districts that may be found in every region of the state of Tamil Nadu. The most significant source of water in these areas, groundwater provides the majority of the water required for household and agricultural purposes. The Water Resource Department collects information on the quality of the groundwater both before and after the monsoon season on a regular basis. The department then analyses the nature of the information collected. In the course of this study project, the time period covered by the dataset ranged from 2010 to 2018. The dataset is comprised of 34 parameters, each of which may be classified into one of two subgroups: numeric or non-numeric. The parameters used

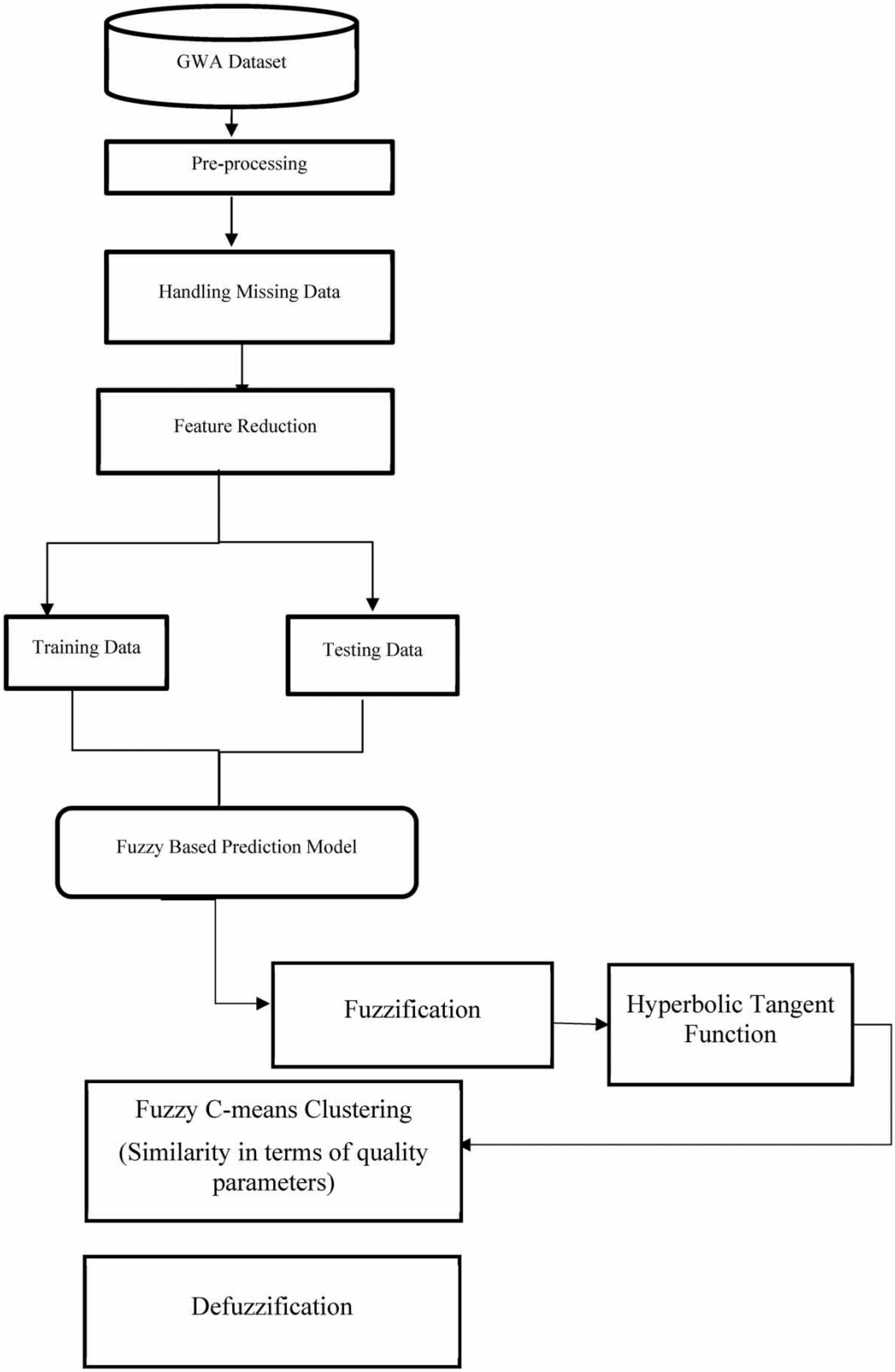


Figure 1 | Overall architecture of groundwater quality.

in the dataset are pH – hydrogen ion concentration, EC – electrical conductivity, TDS – total dissolved solids, Ca – calcium, Mg – magnesium, Na – sodium, K – potassium, HCO_3 – bicarbonate, Cl – chloride, SO_4 – sulfate, TH – total hardness, DWQI – Drinking Water Quality Index, physicochemical parameters: pH, EC, TDS. Chemical parameters: major ions – Ca, Mg, Na, K, HCO_3, Cl, SO_4; cations – Ca, Mg, Na, K; anions – HCO_3, Cl, SO_4, and TH is calculated by the addition of calcium and magnesium concentration in groundwater. $TH = Ca + Mg$.

The first step in the data mining process is called pre-processing, and it is used to prepare the data for the actual mining technique. The pre-processing is the foundation for a few strategies that enable us to offer the authentic information and increase the exactness of the information. These approaches are underpinned by the pre-processing.

Pre-processing is a crucial stage in any research project that is carried out to ensure that predictions are produced with a high degree of accuracy. The following procedures are used to complete the information pre-processing and produce the information in its final, well-formed state for use in further research endeavors. Figure 2 shows a representation of the flow of information that is being pre-processed.

In this part of the study, the process of cleaning the data substitutes any information that is absent from the dataset as well as any information that is particularly noisy. After that, the data integration process will combine the cleansed information with the dataset. The information is then merged into the suitable structure for the mining method from that point forward. We have used a data reduction approach in order to cut down on the number of ground water quality datasets (feature selection). The size of the data collection may be reduced by attribute selection by excluding redundant or superfluous information, and there is an additional advantage to extracting with the smallest possible number of characteristics (Han & Kamber 2006). Obtaining a smaller sized assortment of datasets has really been our primary objective in making use of the data mining application.

There are four stages that make up an attributes selection technique, and they are referred to as (1) subset creation, (2) subset evolution, (3) stop criteria, and (4) result validation (Dash & Liu 1997) subset generation is a searching strategy shown in Figure 2, and we have used the Best First Search Method by way of the DM tool in our investigation. The attribute selection technique, involving subset generation and evolution, plays a crucial role in preparing the dataset for input into the fuzzy inference model. Through this technique, the most relevant and significant characteristics are selected, streamlining, and optimizing the dataset. This optimization enhances the performance and efficiency of the subsequent fuzzy inference model. By identifying the essential features that significantly contribute to water quality prediction or evaluation, the attribute selection technique helps in determining which variables should be used as inputs within the FIS. This utilization of selected features enhances the accuracy and effectiveness of the fuzzy model in assessing groundwater safety and determining water quality. In summary, although the passage does not explicitly mention the direct connection between the attribute selection technique and the fuzzy model, it can be inferred that attribute selection plays a vital role in optimizing the dataset for improved performance of the fuzzy inference model in water quality assessment and prediction. Each newly generated subset underwent evaluation and comparison with the previous best one using a predetermined evolution criterion. If the new subset ends up being significantly superior to the older one, it replaces the older one as the finest subset. The procedure of developing new subsets and evolving existing ones is repeated until a predetermined quitting condition is satisfied. We were able to acquire 34 characteristics out of a total of 44 by using the chosen strategy.

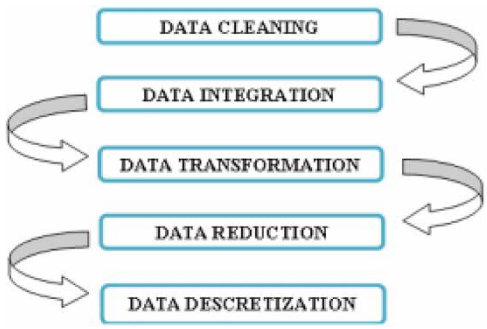


Figure 2 | Data flow diagram of data preprocesssing.

3.1. Fuzzy sets

Fuzzy sets were originally presented to the public in the middle of the 1960s, and since then, they have found use in a wide variety of domains, including control and decision-making. In order to apply fuzzy rules and fuzzy sets, one must first provide the rule arguments and the rule replies. Fuzzy sets were described in their index in terms of a membership function that converts a domain of interest to the range [0, 1]. This was done in accordance with the typical fuzzy rules that are used. Curves were used in the mapping process for each set's membership function. They demonstrated the extent to which a particular value was a member of the related set by stating:

$$\mu A: X \rightarrow [0, 1]$$

Alternately, fuzzy rules may be developed automatically, with the parameters inside the prospective rules being optimized to achieve the best match with available data.

3.2. Fuzzy set functions

Triangle set functions:

$$a_j(x) = \begin{cases} 1 - \dfrac{m_j - x}{l_j}, & \text{if } m_j - l_j \le x \le m_j \\ 1 - \dfrac{x - m_j}{r_j}, & \text{if } m_j < x \le m_j + r_j \\ 0, & \text{else} \end{cases} \tag{1}$$

Trapezoid set function

$$a_j(x) = \begin{cases} 1 - \dfrac{ml_j - x}{l_j}, & \text{if } ml_j - l_j < x < ml_j \\ 1, & \text{if } ml_j \le x \le mr_j \\ 1 - \dfrac{x - mr_j}{r_j}, & \text{if } mr_j < x \le mr_j + r_j \\ 0, & \text{else} \end{cases} \tag{2}$$

Gaussian set function

$$a_j(x) = exp\left\{ -\left(\frac{x - m_j}{d_j} \right)^2 \right\} \tag{3}$$

Cauchy set function

$$a_j(x) = \left[1 + \left(\frac{x - m_j}{d_j} \right)^2 \right]^{-1} \tag{4}$$

Sinx set function

$$a_j(x) = \sin\left(\frac{x - m_j}{d_j} \right) / \left(\frac{x - m_j}{d_j} \right) \tag{5}$$

Laplace set function

$$a_j(x) = \exp\left\{ -\left| \frac{x - m_j}{d_j} \right| \right\} \tag{6}$$

Logistic set function

$$a_j(x) = \frac{1}{D_j}(S_j(x - m_j + l_j) - S_j(x - m_j - l_j)) \tag{7}$$

Hyperbolic tangent set function

$$a_j(x) = \frac{1}{D_j}\left(\tanh\left(\frac{x - m_j + l_j}{d_j}\right) - \tanh\left(\frac{x - m_j - l_j}{d_j}\right)\right) \tag{8}$$

Equations (1)–(8) represent the fuzzy membership sets. When there is a great deal of unpredictability around a situation, a fuzzy system is used. The hyperbolic tangent (tanh) set function and Fuzzy C-Means (FCM) clustering can be effectively used together in fuzzy logic applications. The hyperbolic tangent set function is commonly used to define membership functions in fuzzy logic. It maps a range of input values to an output between −1 and 1, creating an S-shaped curve. This set function is suitable for representing degrees of membership or truth values in fuzzy sets.

On the other hand, FCM clustering is a popular algorithm used to partition data into clusters based on similarity measures. It assigns membership values to each data point, indicating the degree of belongingness to different clusters. The algorithm iteratively updates the cluster centers and membership values until convergence. By combining the hyperbolic tangent set function and FCM clustering, we can effectively handle membership assignment in fuzzy logic. The FCM algorithm can determine the cluster centers and membership values based on similarity measures, while the hyperbolic tangent set function can map these membership values to linguistic terms or fuzzy sets. In water quality assessment, the FCM algorithm can be employed to cluster water samples based on their similarity in terms of quality parameters. The algorithm assigns membership values to each sample indicating their degree of belongingness to different clusters representing distinct water quality categories. These membership values can then be mapped using the hyperbolic tangent set function to linguistic terms such as 'Excellent water,' 'Good water,' 'Poor water,' enabling the interpretation of water quality based on the assigned membership values. By leveraging the capabilities of both the hyperbolic tangent set function and FCM clustering, we can effectively handle membership assignment and representation in fuzzy logic, enhancing the accuracy and interpretability of fuzzy logic models in various applications, including water quality assessment. Table 1 and Figure 3 represent the Fuzzy sets with points and range and water quality indices

3.3. Algorithm

The algorithm for combining the hyperbolic tangent set function and FCM clustering in water quality assessment:

1. Input: Obtain the dataset containing water quality parameters.
2. Initialize the FCM algorithm:
   - Determine the desired number of clusters representing different water quality categories.
   - Set the fuzziness parameter to control membership assignment.
3. Apply FCM clustering:
   - Calculate the similarity or dissimilarity measures between data points using suitable distance metrics.
   - Initialize cluster centers randomly or based on prior knowledge.
   - Update membership values based on similarity measures and the fuzziness parameter.

Table 1 | Fuzzy sets with points and range

Points	Range	Water type/Fuzzy sets
0	<50	Excellent water
50	50–100	Good water
100	100–200	Poor water
200	200–300	Very poor water
300	>300	Water unsuitable for drinking

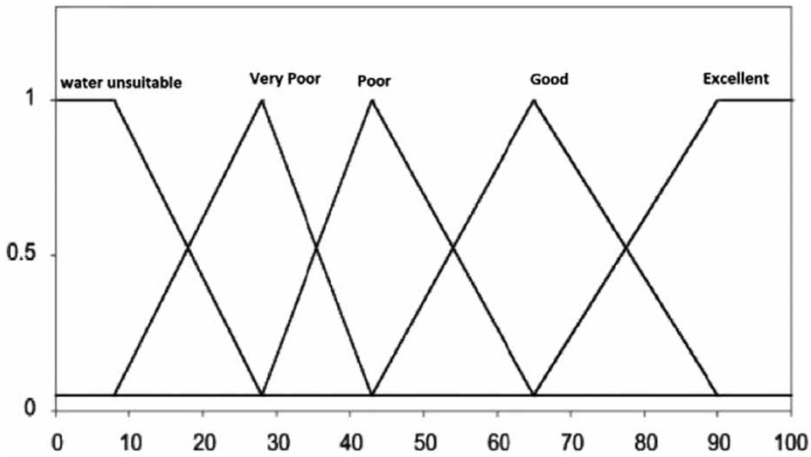


Figure 3 | Water quality indices.

- Update cluster centers using the current membership values.
- Repeat the previous two steps until convergence criteria are met.

4. Map membership using the hyperbolic tangent set function:
   - Utilize membership values obtained from FCM clustering.
   - Apply the hyperbolic tangent set function to map membership values to linguistic terms representing water quality categories.
   - Assign data points to the appropriate linguistic term based on the mapped membership values.
5. Output: Obtain water quality assessment results:
   - Analyze the distribution of data points among linguistic terms to evaluate overall water quality.
   - Interpret the results for decision-making or recommendations based on the water quality assessment.

3.4. Fuzzification

When converting precise numerical values of water quality parameters into fuzzy values, linguistic variables and membership functions are employed to account for the inherent uncertainty and imprecision in the data (Wu 2019). This process, known as fuzzification, involves the following steps:

1. Linguistic Variable Definition: Linguistic variables represent qualitative terms like 'Excellent,' 'Good,' 'Fair,' or 'Poor' that describe water quality. These terms offer a more intuitive and human-readable representation of the data, tailored to the specific context of the water quality assessment.
2. Membership Function Design: Membership functions determine the degree of membership or the extent to which a numerical value belongs to a particular linguistic variable. These functions are shaped based on the characteristics of the water quality parameter being evaluated, using various curve Gaussians, or sigmoid curves. The following Figure 4 shows the Gaussian membership function is used to calculate fuzzy membership values and is returned by the $y =$ gaussmf(x, params) function: $f(x; \sigma, c) = e - (x - c)$ 2 2 σ 2. Utilize the params variable to provide the standard deviation, as well as the mean, c, for the Gaussian function. The values of membership are determined for each input value in the variable x.

A sigmoidal membership function may be realised with the help of the Sigmoidal MF block shown in Figure 5. When the sign of an is positive, the shape of the curve moves to the right from left as it goes up. When the sign of an is negative, on the other hand, the curve will fall off to the right as it moves to the left. The point of the curve's inflection is determined by the value of the parameter c.

The standard deviation of the residuals is equivalent (prediction errors) to the root-mean-square-error, or RMSE. The RMSE measures how scattered the residuals are, whereas the residuals measure how distant the data points are from the

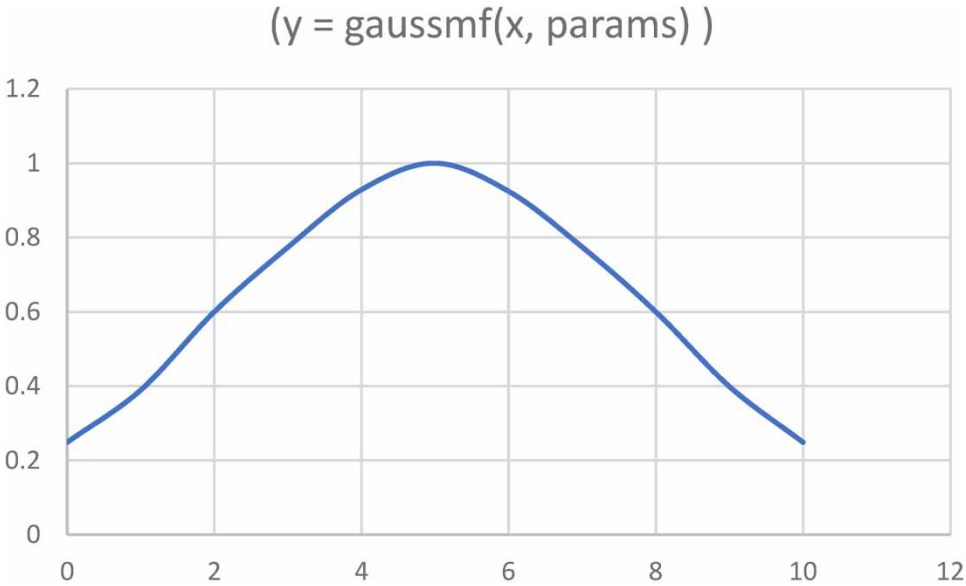


Figure 4 | Gaussian curve membership function.

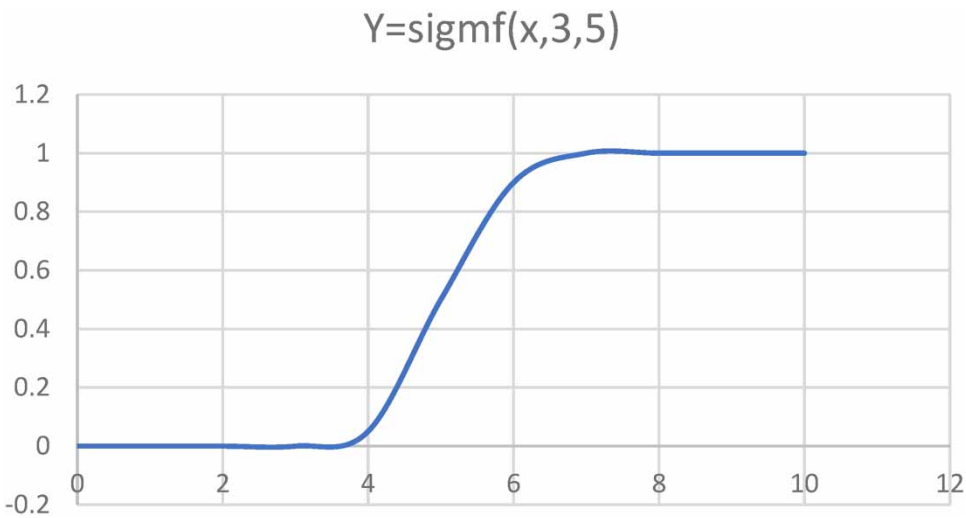


Figure 5 | Sigmoid-shaped MF.

regression line. To put it another way, it shows how closely the data are clustered around the line of best fit. A lower value of the RMSE suggests that a model is more capable of 'fitting' a given dataset. Figure 6 shows the comparison measures.

From the above figure we know that Gaussian membership functions possess low root-mean-square error, hence we are implementing fuzzy set with Gaussian membership function. The Gaussian membership function is used in the computation of fuzzy membership values by this function. This membership function may also be computed by utilizing a fismf object as the data source. There is a distinction to be made between a Gaussian probability distribution and a Gaussian membership function. A Gaussian membership function will never have a value higher than one as its maximum.

$$f(x) = e^{\frac{-(x-c)^2}{2\sigma^2}}$$

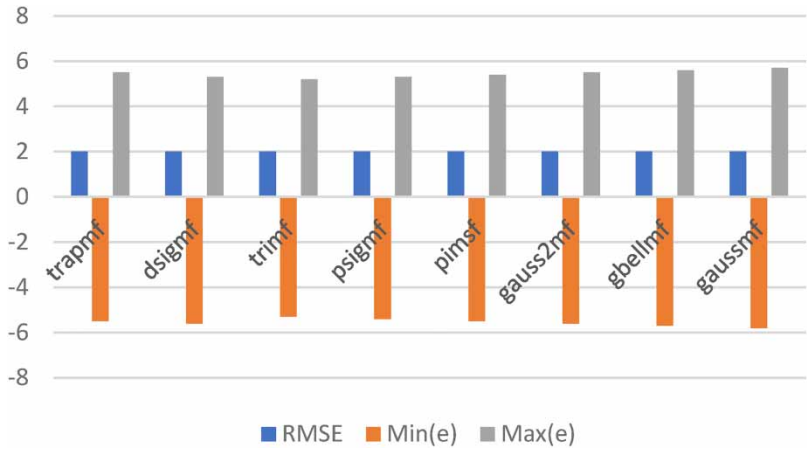


Figure 6 | Comparison measures of membership functions.

Utilize the params variable to provide the standard deviation, as well as the mean, c, for the Gaussian function.

3. Assignment of membership degrees: The hyperbolic tangent (tanh) set function is used to assign membership degrees to the linguistic variables based on the measured values of the water quality parameters. This function maps the measured values to membership degrees on a scale ranging from −1 to 1, where −1 represents no membership and 1 represents full membership. The shape of the hyperbolic tangent curve determines the gradual transition of the membership degrees.
4. Interpretation of membership degrees: The obtained membership degrees indicate the degree of association between the measured values and each linguistic variable. Higher membership degrees indicate a stronger association, while lower degrees indicate a weaker association. These membership degrees effectively capture the uncertainty and imprecision inherent in water quality data, recognizing that values can have multiple interpretations within different linguistic variables.

Through fuzzification, where crisp numerical values are transformed into fuzzy values, water quality assessments can accommodate the inherent uncertainty and imprecision in the data. This approach allows for a more flexible and comprehensive analysis of water quality, as fuzzy values consider multiple interpretations and accurately represent the nuanced nature of water quality parameters.

3.5. FCM clustering

The FCM clustering algorithm is utilized in water quality assessment to group water samples based on their similarity in terms of quality parameters. This algorithm takes into account the multidimensional nature of the data and assigns membership values to each data point, indicating their degree of belongingness to different clusters representing distinct water quality categories. To determine similarity, the FCM algorithm calculates measures of similarity or dissimilarity between pairs of water samples using appropriate distance metrics like Euclidean distance. This calculation helps assess the proximity or similarity between samples based on their quality parameter values. Initially, membership values are randomly assigned to each water sample, representing their initial degree of belongingness to each cluster. These membership values are typically assigned as random values ranging between 0 and 1. The algorithm then iteratively updates the membership values for each data point based on the similarity measures and a fuzziness parameter (usually denoted as 'm'). This parameter controls the level of fuzziness or overlap between clusters, with higher values of 'm' resulting in fuzzier membership assignments. After updating the membership values, the cluster centers are recalculated using the weighted average of the water samples, where the membership values act as weights. This process determines the center of each cluster, representing the centroid of the water samples within that cluster. The iterative process continues with repeated updates to the membership values and cluster centers until convergence is achieved. Convergence is determined based on a predefined stopping criterion, such as a maximum number of iterations or a small change in the cluster centers or membership values. The FCM algorithm in water quality assessment employs similarity calculations, random initialization of membership values, iterative updates of membership values based

on similarity measures and a fuzziness parameter, and recalculation of cluster centers. This process iterates until convergence is reached, allowing for the grouping of water samples into distinct clusters representing different water quality categories.

3.6. Defuzzification

The process of converting fuzzy outputs obtained from membership mapping into precise and actionable results is known as defuzzification. This involves summarizing the fuzzy outputs and obtaining a single crisp value that represents the overall

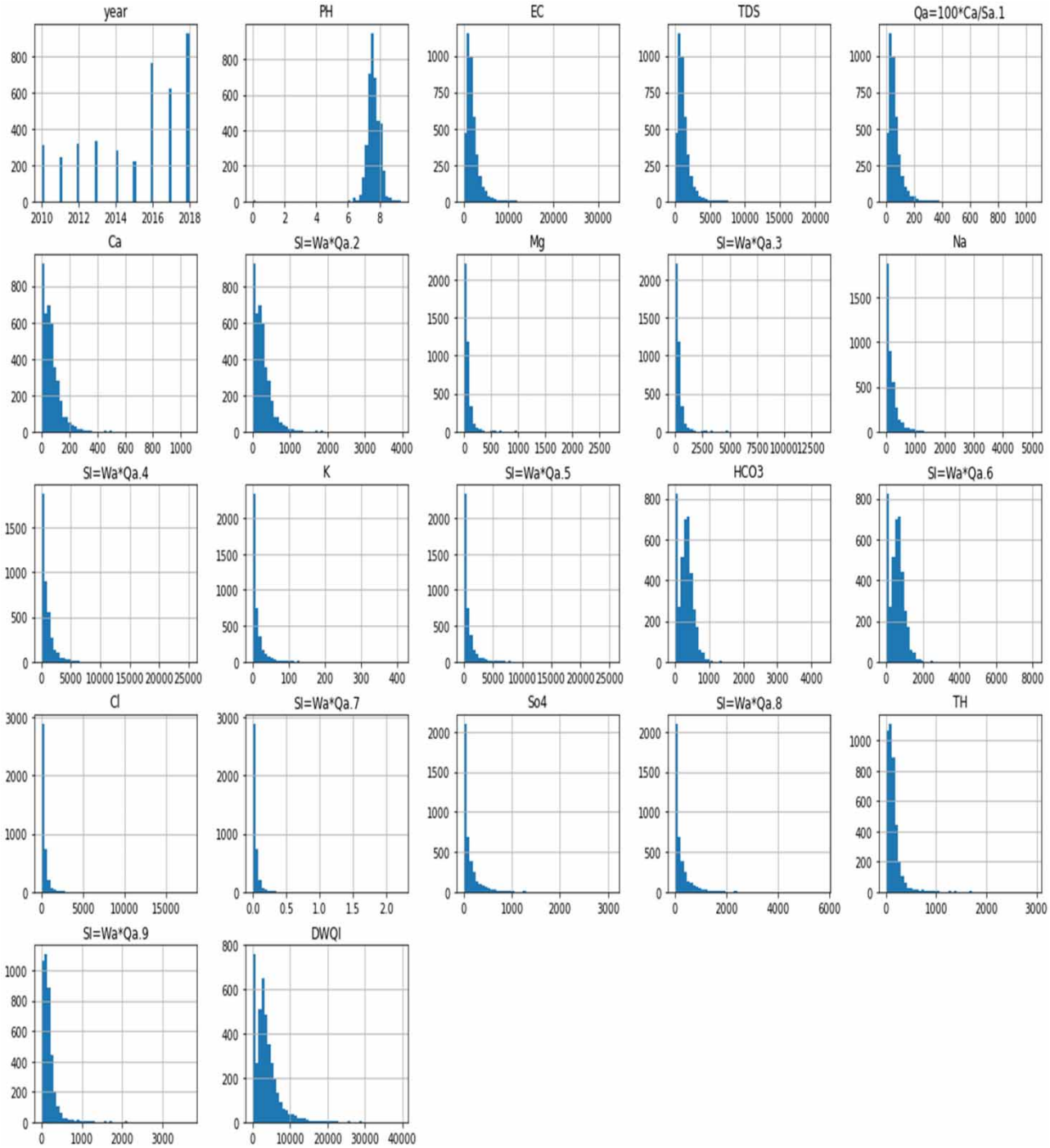


Figure 7 | Histogram of ground water quality.

water quality assessment outcome. Different defuzzification methods, such as centroid methods, height methods, or area methods, can be employed. Centroid methods determine the crisp value by calculating the center of gravity or centroid of the fuzzy set distribution. This is achieved by finding the weighted average of the positions of linguistic terms based on their membership values. Height methods choose the highest membership value among the fuzzy set and assign the corresponding crisp value. This method assumes that the highest membership value signifies the dominant water quality category. Area methods consider the area under the curve of the fuzzy set to estimate the crisp value. The area represents the extent of membership across linguistic terms. Techniques like center of area or mean of maximum can be used to calculate the crisp value based on the area. The selection of a specific defuzzification method depends on the application's

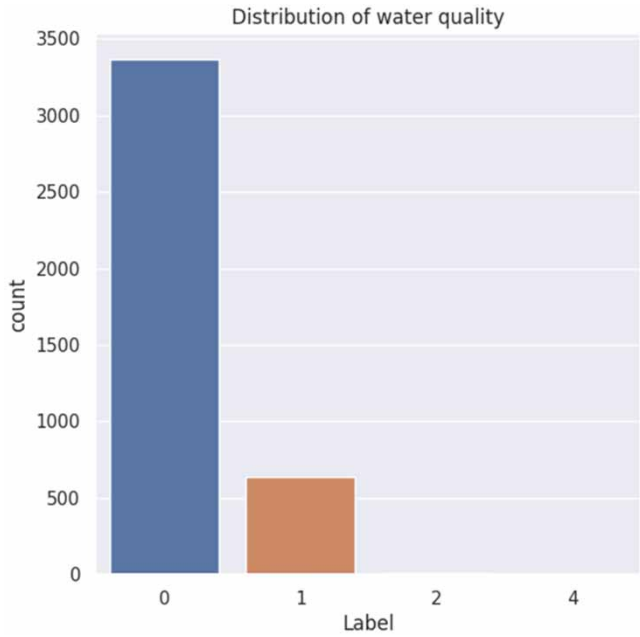


Figure 8 | Type of water quality vs. number of observations in Tamil Nadu.

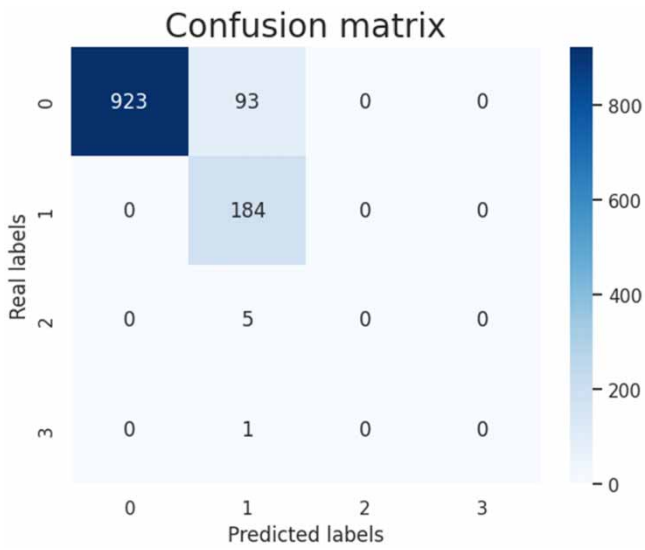


Figure 9 | Confusion matrix.

requirements and the desired interpretation of the water quality assessment result. By employing an appropriate defuzzification method, the fuzzy outputs are transformed into a single, precise value that offers clear and actionable information for decision-making, classification, or further analysis in water quality assessment.

4. PERFORMANCE EVALUATION

Several features and indicators that have been offered by a wide range of organizations and agencies are used to determine whether or not water is appropriate for drinking. The histogram graphical representation of groundwater quality is described

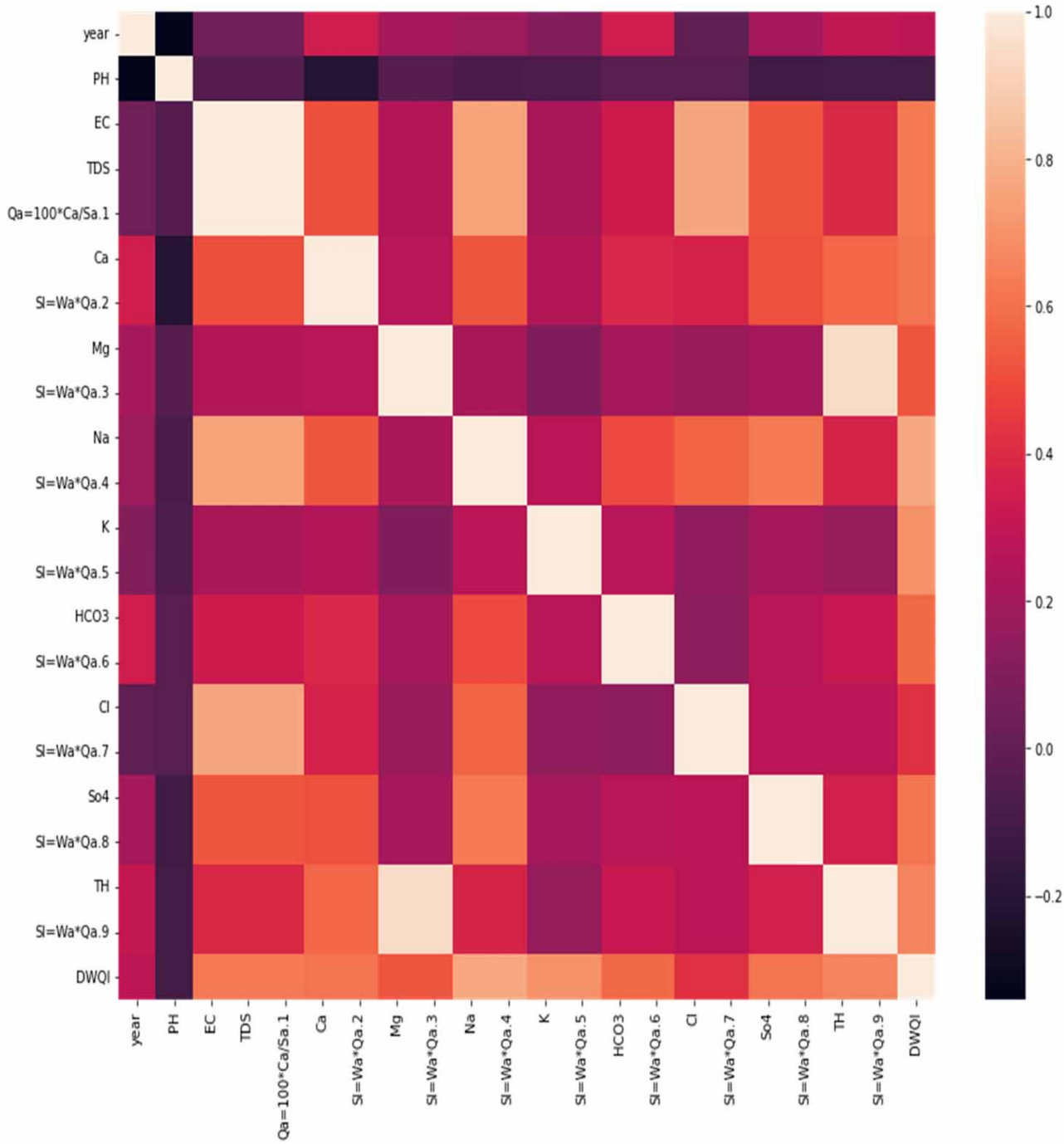


Figure 10 | Heatmap co-relation map of ground water quality indices.

in Figure 7. Type of water quality vs. number of observations in Tamil Nadu is visualized in Figure 8. The confusion matrix is shown in Figure 9. The heatmap co-relation map of ground water quality indices is evaluated in Figure 10. The training and testing accuracy of the FIS is presented in the figure. TDS is the accumulation of ion concentrations in the water.

$$\text{TDS} = \sum(\text{Cations} + \text{anions})$$

The WQI was determined by conducting tests on 14 different physicochemical and bacteriological drinking water quality parameters. These parameters included turbidity, pH, EC, TDS, total alkalinity, TH, calcium (Ca), chloride (Cl), sulfate (SO_4), magnesium (Mg), sodium (Na), potassium (K), nitrate (NO_3), and total coliform. It was discovered that turbidity, EC, total alkalinity, and TH had a greater impact on the quality of drinking water. The detection of groundwater quality is extremely beneficial in a variety of current issues. The proper quantification of the amount of residential sewage that enters the groundwater from the various bodies of water is one way to help prevent groundwater pollution. Increases in the quantity of groundwater that is accessible for abstraction may be determined with the use of the GIS tool. This can be accomplished by enhancing the natural replenishment capacity and increasing the percolation of surface waters into aquifers. Continuous monitoring of the groundwater table level, in conjunction with studies on the water's quality, can help reduce the likelihood of future degradation. In prior study, it was noted that cross-effects between explanatory factors, such as the cross-correlation between land covers and the cross-correlation between land cover and climate in impacting stream water quality, have not been taken into consideration in water quality studies. Studies on water quality have this significant disparity. Conventional statistical models do not have the benefit that machine learning models have, since machine learning models may leverage input variables to improve model prediction accuracy shown in Figure 11. For instance, it is quite likely that physicochemical factors, in addition to environmental factors and groundwater pollution, had an impact on the water quality of groundwater; as a consequence, the forecasting precision may be increased.

5. CONCLUSION

In conclusion, this study presents a novel approach utilizing an FIS for evaluating groundwater safety and determining the WQI. By incorporating linguistic variables, fuzzy rules, and the hyperbolic tangent set function, we address the uncertainties and imprecision inherent in water quality data. The application of FCM clustering allows for the grouping of water samples based on similarity in quality parameters, enabling the identification of patterns and categorization of samples into distinct water quality categories. The results demonstrate the effectiveness of the proposed approach compared to deterministic

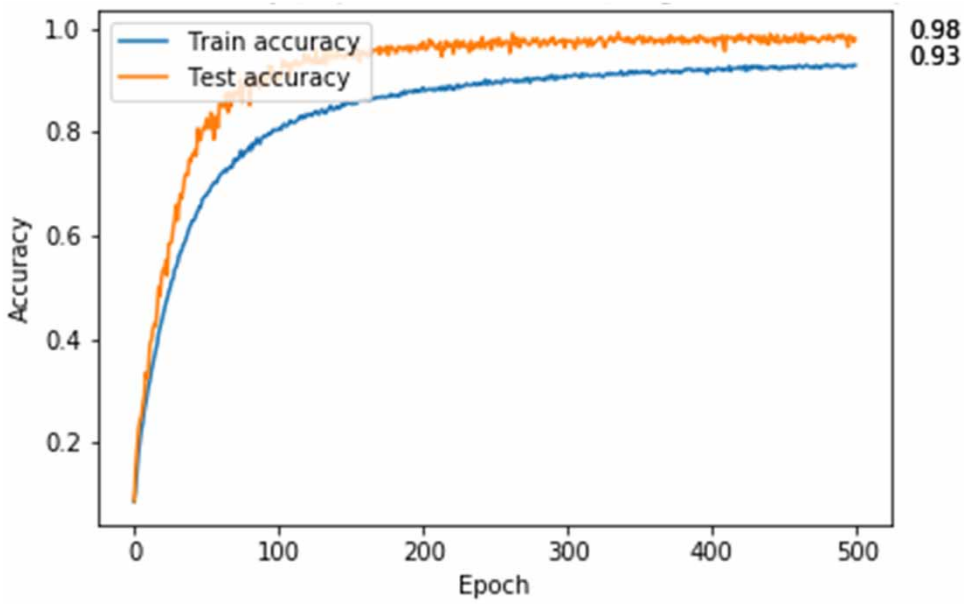


Figure 11 | Testing and training accuracy of fuzzy.

models, with the FIS exhibiting a lower mistake rate in assessing the safety of groundwater samples for human consumption. The utilization of defuzzification methods further converts the fuzzy outputs into crisp and actionable results, providing a clear representation of the overall water quality assessment. This research contributes to the understanding and management of water resources, particularly in areas with limited access to treated water. By providing a fuzzy-based paradigm for evaluating groundwater safety, it enhances the assessment and monitoring of water quality, promoting economic development, and safeguarding human health. Future research can focus on expanding the application of the FIS to other water quality assessment parameters and exploring advanced defuzzification methods. Overall, the proposed approach offers a robust framework for accurate water quality evaluation, contributing to effective decision-making and the sustainable management of water resources. The developed fuzzy-based water quality assessment model achieves an impressive test accuracy of 98% and a train accuracy of 93%. By incorporating fuzzy logic techniques, linguistic variables, and fuzzy rules, the model accurately evaluates water quality parameters while handling uncertainties and imprecision. Its high accuracy underscores its reliability and effectiveness in predicting water quality, making it a valuable tool for decision-making and resource management.

ACKNOWLEDGEMENTS

We would like to show our gratitude to our institution for sharing their pearls of wisdom with us during the course of this research work. We are also immensely grateful to the well-wishers for their comments on an early version of the manuscript, although any errors are own and should not tarnish the reputations of these esteemed individuals.

FUNDING STATEMENT

The authors received no specific funding for this study.

AUTHORS CONTRIBUTIONS

I.S.R. conceptualized the study and prepared the methodology and the original draft. V.B.C. implemented and supervised the manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are available from https://www.kaggle.com/datasets/adityakadiwal/water-potability.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M. & Elshafie, A. 2019 Machine learning methods for better water quality prediction. *Journal of Hydrology* **578**, 124084.

Cabral Pinto, M., Ordens, C. M., Condesso de Melo, M. T., Inácio, M., Almeida, A., Pinto, E. & Ferreira da Silva, E. A. 2020 An interdisciplinary approach to evaluate human health risks due to long-term exposure to contaminated groundwater near a chemical complex. *Exposure and Health* **12** (2), 199–214.

Chen, X., Chen, Y., Shimizu, T., Niu, J., Nakagami, K. I., Qian, X., Jia, B., Nakajima, J., Han, J. & Li, J. 2017 Water resources management in the urban agglomeration of the Lake Biwa region, Japan: an ecosystem services-based sustainability assessment. *Science of the Total Environment* **586**, 174–187.

Dash, M. & Liu, H. 1997 Feature selection for classification. *Intelligent Data Analysis* **1** (1–4), 131–156.

Ellina, G., Papaschinopoulos, G. & Papadopoulos, B. K. 2020 Research of fuzzy implications via fuzzy linear regression in data analysis for a fuzzy model. *Journal of Computational Methods in Sciences and Engineering* **20** (3), 879–888.

Han, J. & Kamber, M. 2006 *Data Mining: Concepts and Techniques*, 2nd edn. University of Illinois at Urbana Champaign, Morgan Kaufmann.

Huang, M., Tian, D., Liu, H., Zhang, C. & Yi, X. 2018 A hybrid fuzzy wavelet neural network model with self-adapted fuzzy-means clustering and genetic algorithm for water quality prediction in rivers. *Complexity* **1**, 1–11.

Jennings, G. 2007 *Water-based Tourism, Sport, Leisure, and Recreation Experiences*. Routledge, London, pp. 1–20.

Kahlown, M. A., Tahir, M. A. & Rasheed, H. 2007 National Water Quality Monitoring Programme. In: *Fifth Monitoring Report (2005–2006). Pakistan Council of Research in Water Resources Islamabad*. Islamabad, Pakistan.

Khan, Y. & See, C. S. 2016 Predicting and analyzing water quality using Machine Learning: a comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1–6.

Lall, U., Josset, L. & Russo, T. A. 2020 A snapshot of the world's groundwater challenges. *Annual Review of Environment and Resources* **45** (1), 171–194.

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L. & Wei, Y. 2013 A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Mathematical and Computer Modelling* **58** (3–4), 458–465.

Liu, W., Ziliang, G., Da'an, W., Zhang, M. & Zhang, Y. 2021 Spatial-temporal variation of water environment quality and Pollution source Analysis in Hengshui Lake. *Huan Jing Ke Xue = Huanjing Kexue* **42** (3), 1361–1371.

Rama, A., Rajakumari, S. & Selvamani, P. 2021 Performance evaluation of machine learning algorithms in forecasting water quality indices: study in Tamilnadu water bodies. *European Journal of Molecular & Clinical Medicine* **7** (5), 1892–1900.

Sahu, M., Mahapatra, S. S., Sahu, H. B. & Patel, R. K. 2011 Prediction of water quality index using neuro fuzzy inference system. *Water Quality, Exposure and Health* **3** (3), 175–191.

Sedeño-Díaz, J. E. & López-López, E. 2016 Fuzzy logic as a tool for the assessment of water quality for reservoirs: a regional perspective (Lerma River Basin, Mexico). *Lake Sci Clim Change* **5**, 155–174.

Wu, H. C. 2019 Fuzzification of real-valued functions based on the form of decomposition theorem: applications to the differentiation and integrals of fuzzy-number-valued functions. *Soft Computing* **23** (16), 6755–6775.

Zeilhofer, P., Zeilhofer, L. V. A. C., Hardoim, E. L., Lima, Z. M. D. & Oliveira, C. S. 2007 GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil. *Cadernos de Saude Publica* **23**, 875–884.

Zhou, J., Wang, Y., Xiao, F., Wang, Y. & Sun, L. 2018 Water quality prediction method based on IGRA and LSTM. *Water* **10** (9), 1148.

First received 28 November 2022; accepted in revised form 26 July 2023. Available online 7 August 2023