Drinking water quality detection using genetic neural network

R. Isaac Sajan*

Department of Electronics and Communication Engineering,
Ponjesly College of Engineering,
Nagercoil, Tamilnadu, India
Email: isaacsajanr.001@gmail.com
*Corresponding author

V. Bibin Christopher

Department of Computing Technologies,
SRM Institute of Science and Technology,
Kattankulathur Campus, Chengalpattu Dt., Tamilnadu, India
Email: bibinchrist85@gmail.com

T.S. Akhila

Department of Electronics and Communication Engineering,
Mar Ephraem College of Engineering and Technology,
Elavuvilai, Marthandam, Tamilnadu, India
Email: akhilats31@gmail.com

M. Joselin Kavitha

Department of Electronics and Communication Engineering,
Marthandam College of Engineering and Technology,
Marthandam, Tamilnadu, India
Email: drjoselinkavitha@gmail.com

Abstract: Physical, chemical, and biological properties influence water quality. It assesses water treatment compliance versus standards. Most water quality standards assess ecosystem health, human safety, water pollution, and drinking water. Water quality affects supply. Microbial, chemical, and radioactive pollutants may damage drinking water. Drinking water pollution may affect babies, young children, pregnant women, the elderly, and those with impaired immune systems. Before consuming water, check its purity. Monitoring ensures water quality and identifies issues. Real-time ML algorithms may identify drinking water quality issues. Water quality may be checked continually and issues rectified immediately. This safeguards public health and drinking water. They may thereby improve water quality assessments. The MinMaxScaler class pre-processes data for our evolutionary neural network drinking water quality method. Also label encoding. The experiment yielded the best answer and 93% fitness function.

Keywords: genetic neural network; machine learning; neural networks; drinking water quality.

Reference to this paper should be made as follows: Sajan, R.I., Christopher, V.B., Akhila, T.S. and Kavitha, M.J. (xxxx) 'Drinking water quality detection using genetic neural network', *Int. J. Global Warming*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: R. Isaac Sajan received his BE and ME in Computer Science and Engineering and PhD in Information and Communication from the Anna University, Chennai, Tamilnadu, India, in 2006, 2008 and 2021 respectively. He is a Professor in the Department of Electronics and Communication Engineering and the Vice Principal at the Ponjesly College of Engineering, Tamilnadu, India. His research interests currently focus on wireless sensor networks, cloud computing, image processing and artificial intelligence. He is a life member of the Indian Society for Technical Education (ISTE) and International Association of Engineers: HK (IAEG). He has published papers in international, national conferences and indexed journals.

V. Bibin Christopher is an Assistant Professor in the Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, since March 2022. He received his BE in Electrical and Electronics Engineering and ME in Computer Science and Engineering from the Anna University, Tamilnadu, India, in 2006 and 2009. He received his PhD in Computer Science and Engineering from the Anna University, Chennai, India, in 2021. He has been in the teaching profession for the past 14 years. His areas of interest include cloud computing, wireless sensor networks, cryptography and network security. He has published papers in international, national conferences and indexed journals. He is a life member of the Indian Society for Technical Education (ISTE).

T.S. Akhila received her BE in Electronics and Communication Engineering and ME in Embedded System Technologies from the Anna University, India, in 2009 and 2011 respectively. She is an Assistant Professor in the Mar Ephraem College of Engineering and Technology, India. Her research interests are wireless sensor networks, wireless communication, VLSI design and embedded system.

M. Joselin Kavitha received her BE in Electronics and Communication Engineering and ME in Communication System from the Anna University, India, in 2008 and 2010 respectively. She is an Assistant Professor in the Marthandam College of Engineering and Technology, India and currently pursuing her PhD degree in the Anna University, Chennai. Her research interests are wireless sensor networks, wireless communication and VLSI design.

1 Introduction

Water quality (WQ) may be influenced by natural processes such as erosion, weathering, and biological interactions, as well as human activities like as industrial discharge, agriculture, urban runoff, and sewage disposal (Rao and Mamatha, 2004). Monitoring and managing WQ are important for maintaining public health, protecting ecosystems, and ensuring sustainable water resources for future generations.

According to its physical (Rahman et al., 2021), chemical, and biological (Bhateria and Jain, 2016) aspects, WQ may be evaluated to determine its state or characteristics. These characteristics influence whether water is suitable for different uses, including drinking, swimming, irrigation, or the sustenance of aquatic life.

Different uses of water have different quality requirements. Drinking water, for instance, must adhere to strict guidelines to guarantee that it is safe for human consumption, whereas water used for enjoyment, such as swimming, must be free of potentially dangerous microbes. Agricultural and industrial water uses also have specific quality requirements depending on the intended application.

Recognising the importance of safeguarding both human health and ecosystems, WQ monitoring, pollution prevention, and appropriate water treatment measures are vital. Implementing effective water management practices, promoting sustainable agricultural practices, treating wastewater before discharge, and reducing pollution sources are some of the strategies employed to protect WQ and mitigate potential risks to both human and ecosystem health.

Based on many WQ criteria, a water quality index (WQI) (Tyagi et al., 2013) is a numerical tool used to evaluate and describe the overall quality of water. It provides a standardised approach to evaluate WQ and is often used for comparing different water sources or monitoring changes in WQ over time.

The proportional relevance of each characteristic for the intended use of water (such as drinking, recreation, or irrigation) is taken into account when calculating the WQI. Parameters that have a bigger influence on the environment or on human health typically have higher weights.

The WQI serves as a valuable tool for assessing the effectiveness of water treatment processes, monitoring trends in WQ, identifying pollution sources, and supporting decision-making related to water resource management and protection.

In recent years, machine learning approaches have drawn a lot of interest for the assessment of WQ. Machine learning algorithms can be applied to WQ data to develop predictive models, classification models, and anomaly detection systems. Here are some ways machine learning is utilised in WQ. Overall, machine learning has the potential to enhance WQ monitoring, assessment, and management by providing accurate predictions, early warnings, and valuable insights for informed decision-making.

Multivariate analysis techniques are used to identify patterns, relationships, and correlations among multiple WQ parameters. These statistical techniques provide a more thorough investigation of the data on WQ and can aid in identifying the underlying causes of fluctuations in WQ.

Principal component analysis (PCA), cluster analysis, discriminant analysis, and factor analysis are common multivariate analytic techniques used in WQ evaluation. These methods analyse the interrelationships between different WQ parameters and identify key factors or components that explain the majority of the variability in the dataset. They can help in locating pollution sources, assessing the success of pollution management methods, and comprehending the general state of WQ.

Assessments of WQ are strengthened and made more illuminating by combining the WQI with multivariate analysis. While multivariate analytic techniques provide for a better knowledge of the complex interactions and factors causing WQ fluctuations, the WQI offers a condensed overview of overall WQ.

The creation of a real-time system that models and predicts WQ using cutting-edge artificial intelligence (AI) techniques is a viable replacement for expensive and time-consuming traditional laboratory and statistical assessments. This alternative approach aims to address the urgent need for faster and more cost-effective methods of assessing WQ, particularly in situations where contamination with disease-causing waste can have catastrophic consequences.

However, there are challenges associated with these AI-based modelling approaches. One notable challenge is the omission of factors that can significantly influence WQ. It is crucial to consider all relevant factors that contribute to WQ variations, such as environmental conditions, pollutant sources, hydrological factors, and seasonal variations. Neglecting these factors can lead to incomplete or inaccurate predictions, limiting the effectiveness of the AI models in assessing WQ.

To overcome this challenge, it is essential to enhance the AI models by incorporating comprehensive datasets that encompass a wide range of influential factors. Data from several sources, such as meteorological data, pollutant sources and emissions, hydrological data, land use data, and historical WQ data, must be gathered and integrated in order to do this. By incorporating these factors into the modelling process, AI algorithms can capture the complex relationships and interactions that affect WQ.

2 Literature survey

Five factors were used to calculate the WQI in the Uttar Pradesh, India districts of Muzaffarnagar and Shamli: pH, total dissolved solids (TDS), total hardness, chloride, and sulphate. The overall WQ class is 'good', according to the computed WQI findings, and the water is deemed suitable for household use (Krishan et al., 2016).

For irrigation reasons, the FFMLP model was utilised to forecast the quality of the water. The coefficient of multiple determinations (R2) and the root mean squared error (RMSE), two widely used metrics, were employed to assess the model's performance. It was noted that the training set produced superior outcomes. To confirm the model's dependability and generalisability, it was critical to assess its performance using separate and untainted data (Ubah et al., 2021).

Oladipo et al. (2021) have performed a study and found that fuzzy logic inferential (FLI) assessment is superior to the WQI because it considers both measured values and surface WQ standards. FLI is more effective in managing uncertainties and incorporating expert knowledge. It is also easily understandable for individuals without technical knowledge.

In order to forecast WQ indicators impacted by urbanisation in the Johor River, Malaysia, Ahmed et al. (2019a) have suggested employing improved wavelet de-noising techniques using neuro-fuzzy inference systems (WDT-ANFIS). The WDT-ANFIS model fared better than competing models and showed promise in detecting temporal trends in WQ. The WDT-ANFIS model performed admirably for AN while also accurately forecasting pH and SS values.

Using four input parameters – temperature, turbidity, pH, and TDS – the research investigates the application of supervised machine learning algorithms to predict the WQI and water quality class (WQC). While multi-layer perceptron (MLP) had the best accuracy in identifying the WQC, gradient boosting and polynomial regression performed most well at predicting the WQI. This shows promising accuracy and points to

the possibility of using this technology in real-time WQ detection systems (Ahmed et al., 2019b).

Nasir et al. (2022) looks at how different AI algorithms are used to categorise WQ data and forecast the WQI. The CATBoost model had the best accuracy (94.51%), and using several meta-classifiers to stack ensemble models, accuracy was increased to 100%. The results imply that CATBoost is a trustworthy algorithm for classifying WQ and that AI can be a useful tool for enhancing WQ.

Haghiabi et al. (2018) compares AI methods for foretelling Tireh River water condition. The outcomes demonstrate that ANN and SVM perform well, with SVM being the most accurate. SVM performs better than other algorithms since it has the lowest DDR value.

Muhammad et al. (2015) concluded that the lazy model with the KStar algorithm achieved the highest accuracy of 86.67% in classifying the WQC of the Kinta River in Malaysia. This algorithm is considered the best choice for classifying the WQC of the Kinta River in Perak, Malaysia.

A 'feedback effect' was discovered in the study on the Gorgan Rod River in Iran, where changes in the analytical variables had an impact on the WQ. It was possible to pinpoint crucial monitoring stations and WQ factors using PCA (Khaledian et al., 2018).

This study classified nine monitoring stations according to WQ characteristics using multivariate statistical approaches (CA, PCA, and DA). The research divided the stations into two groups with comparable features of the WQ. It was discovered that one variable (NH3-N) could accurately and completely discriminate between the clusters. The analysis corroborated the measurement findings and supported the DOE's assessment of the WQ (Azhar et al., 2015).

Data on WQ from the Danube River were analysed in this study using multivariate analytic techniques, including PCA. The results indicated discrete clusters of sample locations based on their cross-sections but no stratification within the water columns. It was discovered that the location of a cross-section significantly affected WQ indicators. The study's findings generally imply that integrating WQ data with multivariate analysis, particularly PCA, might yield useful information on WQ changes in alluvial rivers (Horvat et al., 2021).

This study used PCA and CA to analyse WQ in DWSSs serving multiple municipalities. Key parameters affecting WQ were identified, and clusters based on similar characteristics were identified. PCA reduced dataset dimensionality and provided insights into influential parameters. CA helped create maps and group monitoring sites. These methods are useful for data interpretation and improving knowledge of DWSSs, especially in complex systems. They can also be applied to analyse data from SCADA systems and cyber-physical water systems (Maiolo and Pantusa, 2021).

By conducting PCA and factor analysis, researchers can identify latent factors or underlying dimensions that explain the correlations among observed variables. These techniques help in understanding the complex relationships between WQ parameters and the factors driving their variations. The identification of season-specific factors can provide insights into the specific processes or sources of pollution that are influential during different times of the year (Pejman et al., 2009).

PCA and correspondence analysis (CA), two multivariate analytic approaches, allowed for the significant information to be extracted from complicated datasets. To glean more data from the datasets, PCA and CA were performed. Three main factors, which together accounted for 60.0% of the variation, were found. Ionic composition was the primary component, followed by organic load and faecal contamination as the second and third components. The results of the PCA were confirmed by the CA, which produced three clusters of variables that matched the components. It was suggested that to correctly describe roof-collected water, at least one parameter from each group should be monitored (Vialle et al., 2011).

The creation and use of cutting-edge AI models, notably ANFIS, FFNN, and KNN, for forecasting and categorising water quality (WQI) models provide a more comprehensive and accurate understanding of water pollution processes, aiding decision-makers in formulating effective strategies and making timely decisions for water resource management and pollution control (Hmoud Al-Adhaileh and Waselallah Alsaade, 2021).

3 Methodology

3.1 Dataset contains

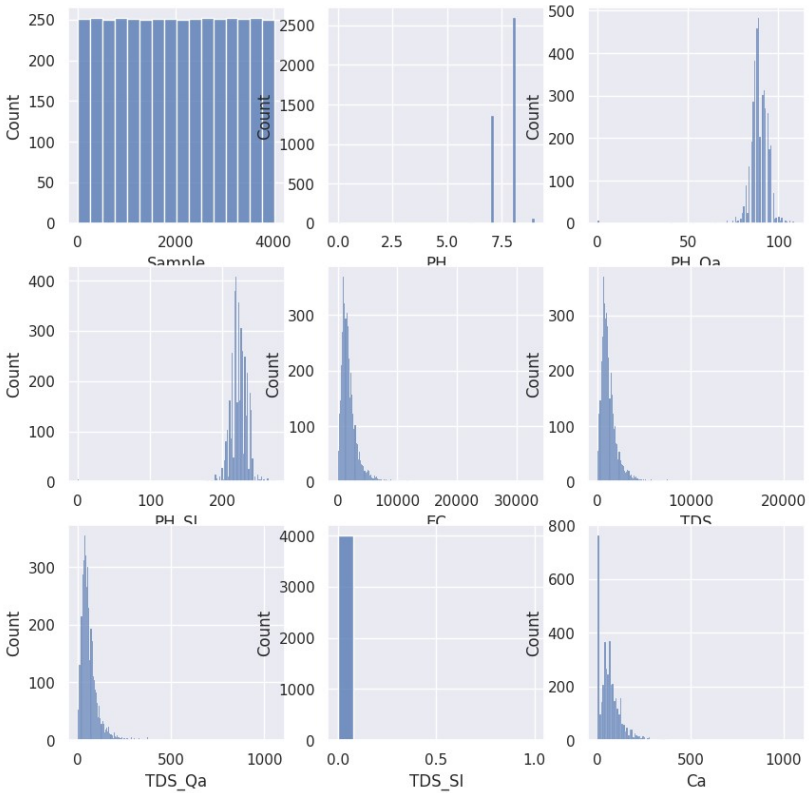
Based on the dataset (Balamurugan et al., 2020) we obtain contain 34 columns representing different features and a target variable ('label'). Here's a breakdown of the columns: columns 0 to 32 represent different features related to WQ, such as pH, electrical conductivity (EC), TDS, calcium (Ca), magnesium (Mg), sodium (Na), potassium (K), bicarbonate (HCO_3), chloride (Cl), sulphate (So4), and total hardness (TH). These features have both numerical (int64 and float64) and categorical (int64) data types. Columns 33 represents the water quality index (DWQI), is a calculated index indicating the overall WQ based on the various features. The histogram of the overall variable distribution in Figure 1.

Column 34 ('label') represents the target variable, appears to be a binary classification label with integer values. Figure 2 shows the box plot distribution of the variables to determine the dependent and independent variables involved in the parameters conducted for WQ. Label encoder is used to categorise the WQ on the basis.

3.2 Data pre-processing

To get started, load the dataset that has all of the information you need regarding the water's quality. Decouple the goal variable, which stands for the WQ level, from the input characteristics (Ramírez-Gallego et al., 2017), which include pH, temperature, and chemical concentrations such as EC, TDS, calcium (Ca), magnesium (Mg), sodium (Na), potassium (K), bicarbonate (HCO3), chloride (Cl), sulphate (So4), and total hardness (TH). In order to assess the effectiveness of the classifier, the dataset should be segmented into training and testing sets.

Figure 1 Histogram (see online version for colours)



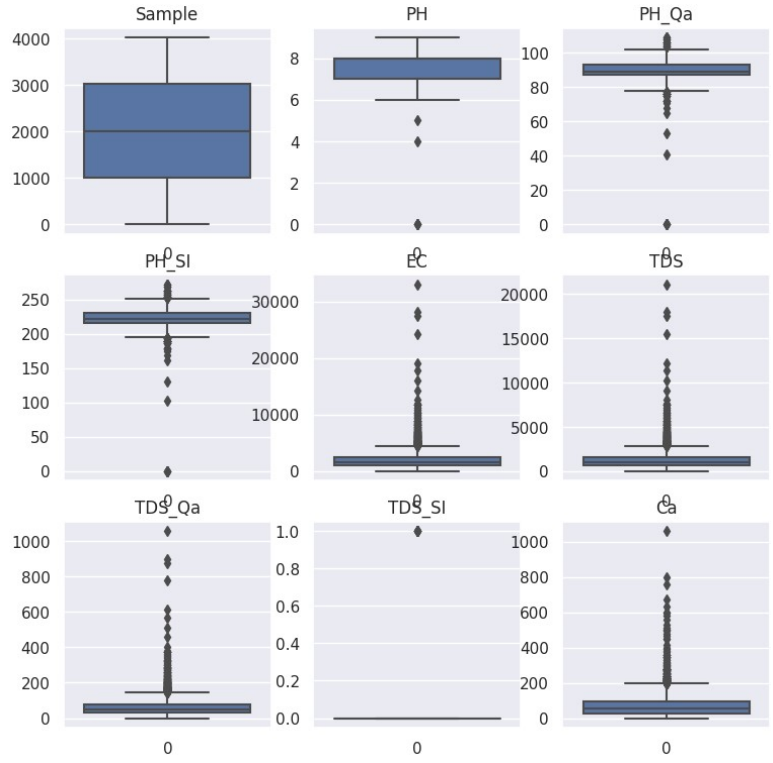
3.3 MLP classifier

It is necessary to configure the MLP classifier, which is an example of an artificial neural network. Provide details on the structure of the MLP, such as the number of input neurons (which should be determined depending on the total number of characteristics) and output neurons (which should reflect the overall WQ). Define the architecture of the hidden layers, including the number of layers that are present and the number of neurons that are included inside each layer. Determine how the neurons in each layer will process and send information by selecting the activation function for that layer. The architecture of the hidden layers in the MLP classifier used for WQ prediction consists of an input layer, two hidden layers, and an output layer (Ieracitano et al., 2020). The input layer takes the input features from the dataset. The first hidden layer has 32 neurons, and the second hidden layer has 16 neurons. Each neuron applies the rectified linear unit (ReLU) activation function, which sets negative values to zero and keeps positive values unchanged.

$$f(x) = \max(0, x),$$

The output layer has a single neuron and uses a linear activation function. This architecture allows the model to make predictions for binary classification of WQ levels (Thoe et al., 2014).

Figure 2 Box plot to find the distribution of features (see online version for colours)



3.4 Algorithm

1 Start with an empty array called *layer1_outputs* to hold the outputs of the neurons in the first hidden layer.

2 For each neuron in the first hidden layer:

- Initialise a variable called *weighted_sum* to zero, which will be used to calculate the weighted sum of inputs for the current neuron.

3 Iterate over each input feature and its corresponding weight:

- Multiply the input feature by its weight and add the result to the *weighted_sum*.

4 Add the bias term to the *weighted_sum*.

5 Apply the ReLU activation function to the *weighted_sum*:

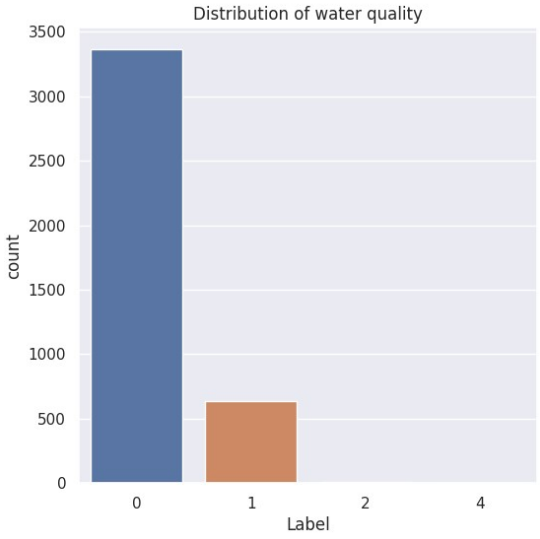
- If the *weighted_sum* is negative, set the output to zero.

- If the *weighted_sum* is positive or zero, keep the output as it is.

6 Store the output of the neuron in the *layer1_outputs* array.

7 After processing all the neurons in the first hidden layer, the *layer1_outputs* array will contain the outputs of each neuron.

Figure 3 Distribution of WQ (see online version for colours)



4 Genetic algorithm

Set the appropriate values for the Genetic Algorithm's parameters so that it can simulate natural selection and get optimal results in optimisation. Adjust the population size, which is what affects the number of possible solutions (MLP classifiers) that are generated by the algorithm with each iteration. Once the parameters have been configured, the genetic algorithm can be run to find a solution to the optimisation problem. The algorithm will continue to iterate until it reaches the termination criteria, which could be a fixed number of generations, a certain fitness threshold, or a time limit. Indicate the mutation rate, which determines the likelihood of the evolution process resulting in the introduction of arbitrary shifts or modifications to the solutions. Find out how many generations there are, which tells you how many iterations the GA will go through to enhance the solutions (Zhang et al., 2019a).

- Set the population_size to determine the number of individuals (networks) in each generation of the genetic algorithm.
- Define the mutation rate to control the probability of mutation occurring during the evolution process.

- Specify the num_generations to set the number of generations the genetic algorithm will evolve through.

These configuration parameters determine the size of the population, the rate of mutation and the number of generations the genetic algorithm will iterate over. They play a crucial role in shaping the search space and the exploration-exploitation trade-off in finding the optimal solution. In our methodology we have used a genetic algorithm with a population size of 100, a mutation rate of 0.01, and a number of generations of 1,000 would result in an algorithm with a large enough population to ensure a good diversity of solutions, a low enough mutation rate to prevent the algorithm from becoming unstable, and a sufficient number of generations to allow the algorithm to converge on a good solution to determine the assessment of WQ.

4.1 Fitness function

Develop a fitness function that can measure the effectiveness of an MLP classifier when used to the problem of predicting WQ. The MLP classifier should next be trained using the training data, and its accuracy, F1 score, recall, and precision should be measured using the validation set. Combine these performance indicators into a single fitness score in order to quantify the classifier's overall level of effectiveness. The fitness score indicates how well the classifier can forecast the levels of WQ based on the features that have been provided (Taha et al., 2017).

4.2 Initial population generation

- Generate the initial population of MLP classifiers with random configurations.
- Each MLP classifier represents a potential solution for the WQ prediction task.
- The population size is determined by the previously set parameter (Deniz and Kiziloz, 2019).

4.3 Evolution process

Commence the evolution process, which involves multiple generations. Perform an evaluation of the fitness of each MLP classifier that is a part of the population at each generation using the fitness function. After that, choose the parents of the next generation depending on how fit they are individually. Classifiers with a better overall performance level have a greater likelihood of being chosen. Then, construct offspring classifiers by using crossover and mutation techniques. In order to come up with brand new answers, the crossover technique combines the qualities of two different parent classifiers. After that, the mutation algorithm makes a few arbitrary tweaks to the offspring classifiers in order to investigate a variety of locations inside the solution space. By adding the newly formed offspring classifiers to the current population and replacing the existing population (Holladay and Robbins, 2007).

4.4 Best classifier selection

- After the evolution process, select the best MLP classifier from the final population based on its fitness score.
- The best classifier is the one that achieved the highest performance on the validation set during the evolution process.
- It represents the solution that is expected to provide the most accurate predictions for WQ levels (Zhang et al., 2019b).

4.5 Algorithm for the proposed methodology

1 Load the dataset and split it into input features (X) and the target variable (y).

2 Split the dataset into training and testing sets.

3 Scale the input features using a StandardScaler and label encoder.

4 Define a fitness function to assess the performance of an MLP classifier.

5 Set the GA configuration parameters: population size, mutation rate, and number of generations.

6 Create an initial population of random MLP classifiers.

7 Perform evolution by iterating through generations:
   - Evaluate the fitness of each classifier in the population.
   - Select parents based on their fitness scores.
   - Create offspring through crossover and mutation operations.
   - Replace the existing population with the offspring.

8 Identify the best classifier from the final population.

9 Evaluate the performance of the best classifier on the test set using suitable metrics such as accuracy, F1 score, recall, and precision.

10 Output the performance metrics.

5 Experimental results

We compare the best MLP classifier's results on a test set of data that it has never seen before. The performance of the classifier in making predictions about WQ is then evaluated using a variety of criteria, including accuracy, F1 score, recall, and precision. The classifier's capacity to generalise and provide accurate predictions on real-world data may be gauged using these measures. Show the results of the top MLP classifier's test-set performance metrics. The GA-MLP method for assessing WQ is evaluated using these indicators. The comparative measure give in Tables 1 and 2.The evaluation metrics for proposed methodology is shown in Figure 4 and the confusion matrix is shown in Figure 5.

Table 1 Comparison of proposed GNN with other methods

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1 score*
Genetic neural network	0.93	0.99	0.97	0.98
Random forest	0.9987	0.6875	0.75	0.7142857
Decision tree	0.9975	0.999259	0.99606299	0.99764495
ANN	0.9317	0.7172	1.0000	0.82051282

Table 2 Pros and cons of existing methodologies

Algorithm	*Advantages*	*Disadvantages*
Genetic neural network	• High accuracy • Can handle complex data • Can learn nonlinear relationships	• Time-consuming to train
Random forest	• High accuracy • Less time-consuming to train than genetic neural network • Can handle complex data	• May not be as accurate as genetic neural network
Decision tree	• Very accurate • Easy to interpret • Fast to train	• Can be sensitive to noise in the data • May not be able to handle complex data
ANN	• Can learn complex relationships • Can handle noisy data • May be overfitting the data	• Can be time-consuming to train

Figure 4 Evaluation metrics (see online version for colours)

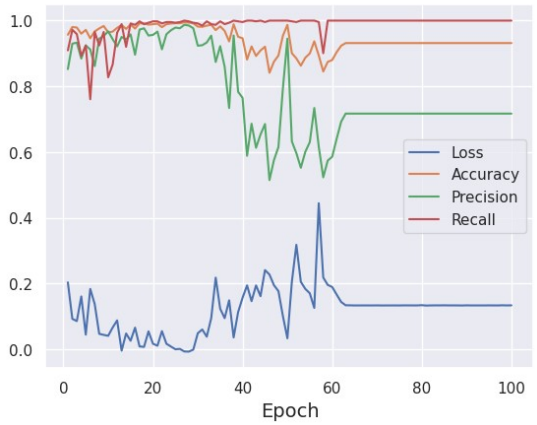
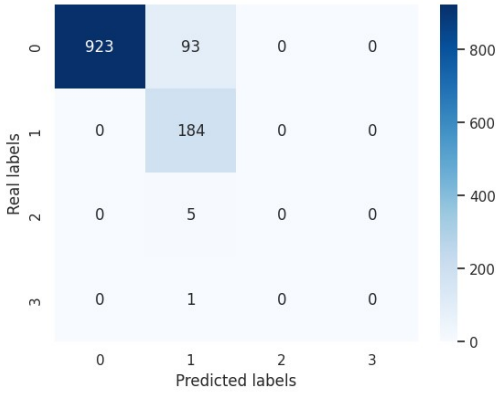


Figure 5 Confusion matrix (see online version for colours)



6 Conclusions

Accessing of drinking WQ is necessary for the survival of human beings in the world. WQ assessment encompasses a wide range of activities focused on evaluating and monitoring the condition and characteristics of water sources. It involves the analysis of physical, chemical, biological, and radiological factors to determine if water is suitable for specific purposes and to identify potential contaminants or risks. By delivering more precise and up-to-date data, ML-based techniques have the potential to improve WQ monitoring and management. But to achieve confidence and understanding in their applications, ML model interpretability and explain ability must be considered, as must the quality and representativeness of the training data. Here we are utilising a genetic algorithm-based artificial neural network for detecting WQ. Based on the given performance metrics for WQ classification using an MLP classifier, we can draw the following conclusions: impressive accuracy the MLP classifier achieves 93% accuracy, indicating a high level of correct WQ classification overall. Exceptional precision exhibits an outstanding precision score of 0.99, indicating its ability to accurately identify positive WQ samples with minimal false positives. With a recall score of 0.97, the MLP classifier effectively captures a significant proportion of actual positive WQ samples while minimising false negatives. The F1 score of 0.98, which combines precision and recall, reflects a reliable balance between the two measures and signifies the classifier's strong performance. In summary, the MLP classifier demonstrates robust performance in predicting WQ. Its high accuracy, precision, recall, and F1 score indicate its effectiveness in distinguishing between different WQ categories. Consequently, the MLP classifier shows promise as a valuable tool for WQ assessment and monitoring.

References

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M. et al. (2019a) 'Machine learning methods for better water quality prediction', *Journal of Hydrology*, Vol. 578, No. 1, p.124084.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R. et al. (2019b) 'Efficient water quality prediction using supervised machine learning', *Water*, Vol. 11, No. 11, p.2210.

Azhar, S.C., Aris, A.Z., Yusoff, M.K., Ramli, M.F. and Juahir, H. (2015) 'Classification of river water quality using multivariate analysis', *Procedia Environmental Sciences*, Vol. 30, pp.79–84.

Balamurugan, P., Kumar, P.S. and Shankar, K. (2020) 'Dataset on the suitability of groundwater for drinking and irrigation purposes in the Sarabanga River region, Tamil Nadu, India', *Data in Brief*, Vol. 29, No. 1, p.105255.

Bhateria, R. and Jain, D. (2016) 'Water quality assessment of lake water: a review', *Sustainable Water Resources Management*, Vol. 2, No. 1, pp.161–173.

Deniz, A. and Kiziloz, H.E. (2019) 'On initial population generation in feature subset selection', *Expert Systems with Applications*, Vol. 137, No. 1, pp.11–21.

Haghiabi, A.H., Nasrolahi, A.H. and Parsaie, A. (2018) 'Water quality prediction using machine learning methods', *Water Quality Research Journal*, Vol. 53, No. 1, pp.3–13.

Hmoud Al-Adhaileh, M. and Waselallah Alsaade, F. (2021) 'Modelling and prediction of water quality by using artificial intelligence', *Sustainability*, Vol. 13, No. 8, p.4259.

Holladay, K.L. and Robbins, K.A. (2007) 'Evolution of signal processing algorithms using vector based genetic programming', in *2007 15th International Conference on Digital Signal Processing*, IEEE, July, pp.503–506.

Horvat, Z., Horvat, M., Pastor, K., Bursić, V. and Puvača, N. (2021) 'Multivariate analysis of water quality measurements on the Danube River', *Water*, Vol. 13, No. 24, p.3634.

Ieracitano, C., Mammone, N., Hussain, A. and Morabito, F.C. (2020) 'A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia', *Neural Networks*, Vol. 123, No. 1, pp.176–190.

Khaledian, Y., Ebrahimi, S., Natesan, U., Basatnia, N., Nejad, B.B. et al. (2018) 'Assessment of water quality using multivariate statistical analysis in the Gharaso River, Northern Iran', *Urban Ecology, Water Quality and Climate Change*, pp.227–253.

Krishan, G., Singh, S., Kumar, C.P., Garg, P., Suman, G., Ghosh, N.C. and Chaudhary, A. (2016) 'Assessment of groundwater quality for drinking purpose by using water quality index (WQI) in Muzaffarnagar and Shamli Districts, Uttar Pradesh, India', *Hydrology: Current Research*, Vol. 7, No. 227, p.2.

Maiolo, M. and Pantusa, D. (2021) 'Multivariate analysis of water quality data for drinking water supply systems', *Water*, Vol. 13, No. 13, p.1766.

Muhammad, S.Y., Makhtar, M., Rozaimee, A., Aziz, A.A. and Jamal, A.A. (2015) 'Classification model for water quality using machine learning techniques', *International Journal of Software Engineering and Its Applications*, Vol. 9, No. 6, pp.45–52.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M. et al. (2022) 'Water quality classification using machine learning algorithms', *Journal of Water Process Engineering*, Vol. 48, No. 1, p.102920.

Oladipo, J.O., Akinwumiju, A.S., Aboyeji, O.S. and Adelodun, A.A. (2021) 'Comparison between fuzzy logic and water quality index methods: a case of water quality assessment in Ikare community, Southwestern Nigeria', *Environmental Challenges*, Vol. 3, No. 1, p.100038.

Pejman, A.H., Bidhendi, G.N., Karbassi, A.R., Mehrdadi, N. and Bidhendi, M.E. (2009) 'Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques', *International Journal of Environmental Science and Technology*, Vol. 6, No. 1, pp.467–476.

Rahman, A., Jahanara, I. and Jolly, Y.N. (2021) 'Assessment of physicochemical properties of water and their seasonal variation in an urban river in Bangladesh', *Water Science and Engineering*, Vol. 14, No. 2, pp.139–148.

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M. and Herrera, F. (2017) 'A survey on data preprocessing for data stream mining: current status and future directions', *Neurocomputing*, Vol. 239, No. 1, pp.39–57.

Rao, S.M. and Mamatha, P. (2004) 'Water quality in sustainable water management', *Current Science*, Vol. 5, No. 1, pp.942–947.

Taha, A., Alsaqour, R., Uddin, M., Abdelhaq, M. and Saba, T. (2017) 'Energy efficient multipath routing protocol for mobile ad-hoc network using the fitness function', *IEEE Access*, Vol. 5, No. 1, pp.10369–10381.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L. and Boehm, A.B. (2014) 'Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions', *Water Research*, Vol. 67, No. 1, pp.105–117.

Tyagi, S., Sharma, B., Singh, P. and Dobhal, R. (2013) 'Water quality assessment in terms of water quality index', *American Journal of water resources*, Vol. 1, No. 3, pp.34–38.

Ubah, J.I., Orakwe, L.C., Ogbu, K.N., Awu, J., Ahaneku, I.E. and Chukwuma, E.C. (2021) 'Forecasting water quality parameters using artificial neural network for irrigation purposes', *Scientific Reports*, Vol. 11, No. 1, p.24438.

Vialle, C., Sablayrolles, C., Lovera, M., Jacob, S., Huau, M.C. et al. (2011) 'Monitoring of water quality from roof runoff: interpretation using multivariate analysis', *Water Research*, Vol. 45, No. 12, pp.3765-3775.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J. and Atkinson, P.M. (2019a) 'Joint deep learning for land cover and land use classification', *Remote Sensing of Environment*, Vol. 221, No. 1, pp.173–187.

Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L.B. and Pan, B. (2019b) 'Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network', *Water Research*, Vol. 164, No. 1, p.114888.