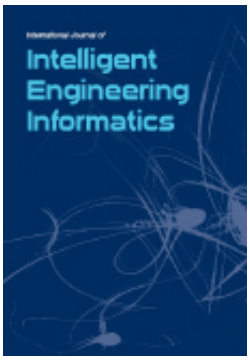


International Journal of Intelligent Engineering Informatics

ISSN online: 1758-8723 - ISSN print: 1758-8715
https://www.inderscience.com/ijiei

Explainable AI and sand cat optimisation algorithm for water quality classification

Gehad Ismail Sayed, Aboul Ella Hassanien

DOI: 10.1504/IJIEI.2024.10062813

Article History:

Received:	16 September 2023
Last revised:	13 January 2024
Accepted:	13 January 2024
Published online:	02 April 2024

Explainable AI and sand cat optimisation algorithm for water quality classification

Gehad Ismail Sayed*

School of Computer Science,
Canadian International College (CIC),
Cairo, Egypt
Email: gehad_sayed@cic-cairo.com
and
Scientific Research School of Egypt (SRSEG), Egypt
*Corresponding author

Aboul Ella Hassanien

College of Business Administration (CBA),
Kuwait University, Kuwait
and
Faculty of Computers and AI,
Cairo University,
Giza, Egypt
and
Scientific Research School of Egypt (SRSEG), Egypt
Email: aboitcairo@gmail.com

Abstract: Assessing river water quality is considered a critical task in enhancing water resource management plans. Therefore, an accurate prediction of the quality of the water has become highly needed to control water pollution. In this paper, a new water quality classification model is proposed based on explainable artificial intelligence (XAI) and an optimised artificial neural network (ANN). The sand cat optimisation algorithm (SCOA) is modified and applied for hyper-parameter optimisation of ANN. The proposed model is tested on a benchmark dataset of water quality taken from various places across India. The results are explained and interpreted using the XAI technique. The experimental results demonstrated that the modified SCOA can effectively find the optimal values of weights and bias coefficients for ANN. The proposed model can effectively classify the water quality. It obtained an overall accuracy of 98%, specificity of 99%, precision of 98%, sensitivity of 98%, and f-score of 98%.

Keywords: sustainable development goals; SDGs; machine learning algorithms; water quality index; water quality classification; swarm intelligence; sand cat optimisation algorithm; SCOA; metaheuristic optimisation algorithms; explainable artificial intelligence; XAI; artificial neural network; ANN; hyperparameters optimisation.

Reference to this paper should be made as follows: Sayed, G.I. and Hassanien, A.E. (2024) 'Explainable AI and sand cat optimisation algorithm for water quality classification', *Int. J. Intelligent Engineering Informatics*, Vol. 12, No. 1, pp.60–84.

Biographical notes: Gehad Ismail Sayed received her PhD in 2020 from the Department of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University. She received her MSc in 2016 and her BSc with honours in 2013 and, both from the Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University. Currently, she is an Assistant Professor at the School of Computer Science, Canadian International College (CIC). She has many published papers in respectful journals and conferences. Her research areas include swarm intelligence, medical image analysis, optimisation algorithms, machine learning and data mining.

Aboul Ella Hassanien is the Founder and Head of the Scientific Research Group in Egypt (SRGE) and a Professor of Information Technology at the Faculty of Computer and AI, Cairo University. He has more than 1,000 scientific research papers published in prestigious international journals and over 45 books covering such diverse topics as data mining, medical images, intelligent systems, social networks and smart environments. He is a collaborative researcher member of the Computational Intelligence Laboratory at the Department of Electrical and Computer Engineering, University of Manitoba. His other research areas include computational intelligence, medical image analysis, security, animal identification, space sciences, telemetry mining, and multimedia data mining.

1 Introduction

One of the basic needs of people is water. Water makes up 60% of the human body (Ben-Daoud et al., 2023). It is becoming a more precious resource due to industrialisation and population growth. On Earth, the surface has a relatively small water distribution. Management of water resources is therefore necessary. Conserving water resources, collecting water, planning how to use the net water resources, and distributing them to customers properly are all part of water management (Krishnan et al., 2022). To do the duties under a patchwork of controls also entails establishing policies and procedures. The traditional approaches and procedures were ineffective for carrying out the activities in an efficient manner. For the long-term sustainability of the water supply, water management strategies must be fully considered. The proportion of water – nearly 97% – is salty and unfit for human consumption (Koech and Langat, 2018). Water supplies are also impacted by pollution.

Water pollution is mostly caused by many industries, including intensive agriculture, wastewater, mining, industrial output, and untreated urban runoff (Berthet et al., 2021). To enhance water resource management strategies, evaluating the quality of river water is of utmost importance. To prevent water pollution and ensure sustainable water use, precise water quality prediction has become essential. Several sustainable development goals (SDGs) are perfectly aligned with this. First off, by aiming to ensure the

accessibility and sustainable management of clean water sources for all, it immediately supports goal 6: clean water and sanitation. Additionally, maintaining river water quality significantly supports goal 14: life below water since it directly affects marine life and coastal ecosystems, furthering the larger goal of protecting underwater biodiversity. Additionally, as robust river ecosystems are essential to terrestrial biodiversity, this supports goal 15: life on land.

Degradation of the level of water quality can badly affect the supplies of safe fresh water for irrigation and human consumption, thus affecting aquatic ecosystems. Oftentimes, developing countries such as India pass through many intervals of the expansion of the economy. This may result in a negative impact on the environment (Bui et al., 2020). Additionally, the speedy increase in population and wealth can lead to growing pressures on the fecundity of soils naturally due to increasing the demand for food production. Thus it will result in a high need for artificial fertilisers, which move to rivers and thus to oceans and lakes. This can cause irreparable damage to the environment and consequently to human health. As water species can permit a certain limit of pollution exceeding this percentage can threaten the existence of these creatures (Aldhyani et al., 2020). Due to the less hygienic qualities and lack of public awareness, the quality of the drinking water is badly affected. According to the report of the United Nations, every year, almost one and a half million people die as a result of contaminated water-driven diseases. The deaths resulting from contaminated water are higher than from terrorist attacks, accidents, and crimes (Pruss-Ustun, 2008).

Traditional water management techniques fall short of the need to efficiently use water from varied sources. The current water use practices are not very cost-effective (Zhao et al., 2020), and there is also a reluctance to use the most recent information and communication technology (Malviya and Jaspal, 2021). Monitoring and predicting water quality is a critical task to protect the environment, human health, and sustainable water management (Kamali et al., 2021).

Machine learning algorithms have demonstrated remarkable efficacy across diverse fields. They have been applied for forecasting strand settlement in brittle sand and geocell (Jeyanthi et al., 2023), for intrusion detection in mobile ad hoc networks (Singh and Vigila, 2023), for air quality monitoring and prediction (Priya and Khanaa, 2022), for detection of coronavirus disease (Bourouis, 2022), and for detecting seizures from electroencephalography (Patel et al., 2022). Meanwhile, machine learning algorithms have introduced a great solution for the treatment of wastewater (Gaya et al., 2017). With a specified objective, these algorithms can exponentially increase the learning process. Standard techniques would not scale exponentially to cover unidentified patterns in the new datasets.

Artificial neural networks or shortly artificial neural network (ANN) is one of the well-known machine learning algorithms. It plays a vital role in prediction and classification fields. It has been applied to find an optimal solution for many complex and nonlinear equations (Elshaboury et al., 2021). The general architecture of ANN consists of the input layer, the hidden layer, and the output layer. Each layer may have a different number of neurons. Additionally, each one of these neurons is assigned a bias and weight coefficient. Through the training process of ANN, the values of biasses and weights are updated several times to get the most accurate result. The backpropagation learning algorithm with the gradient method is the commonly used method to tune the weights and the bias coefficients. However, the main problem with this method is that it

may fail in the local optima. Thus, several studies have used metaheuristic optimisation algorithms to find the global optima.

Metaheuristic optimisation algorithms have garnered widespread utilisation for tackling a diverse array of optimisation problems such as for designing harmonic estimators (Mehta et al., 2022), and finding the optimal placement of wind turbines in a predefined wind farm (Kumar and Sharma, 2023). Metaheuristic algorithms, such as genetic algorithms and particle swarm optimisation, provide varied and effective methods for exploring the huge and frequently complex hyperparameter space. These algorithms excel in navigating high-dimensional and non-convex search spaces, allowing for the identification of optimal hyperparameter configurations. Sand cat optimisation algorithm (SCOA) is one of the metaheuristic algorithms recently proposed in 2023 (Seyyedabbasi and Kiani, 2023). The inspiration for the original SCOA came from the behaviour of sand cats to survive in nature. The authors of the original SCOA show the capability of the algorithm to explore the search to find the optimal solution using a population-based search mechanism. SCOA has demonstrated its effectiveness in locating the best answers to a variety of optimisation problems. Its advantage has also been demonstrated through literature comparisons with various metaheuristic algorithms. Additionally, it has shown its effectiveness in different optimisation problems such as in finding the optimal solution for the engineering design problem (Wu et al., 2022) and in evaluating the minimum safety factor of earth slopes under seismic and static loading circumstances (Iraji et al., 2022).

In this paper, an optimised version of ANNs based on the SCOA is introduced. SCOA is used as an alternative approach for optimising the biasses and the weight parameters of ANN. Additionally, explainable artificial intelligence (XAI) is employed to interpret the proposed water quality classification model results. Overall the proposed water quality classification model consists of three main phases; the data pre-processing phase, the optimised ANN based on the SCOA-hyperparameters optimisation algorithm phase, and the classification and result interpretation phase. The proposed model is tested on the Indian rivers benchmark dataset. To the best of our knowledge, this is the first time to develop an optimised version of ANN based on SCOA. This research aims to introduce a reliable model for precisely categorising water quality. The main contributions of this paper are summarised as follows:

- a new model for water quality classification is proposed
- a new version of SCOA is proposed to optimise the weights and biasses parameters of ANN
- SHAP XAI technique is utilised to explain and interpret the results of the proposed model.

The organisation of the rest of the paper is structured as follows. An overview of the previous studies is presented in Section 2. A brief description of the basic SCOA is introduced in Section 3. Section 4 describes the adopted dataset. In Section 5, the proposed water quality classification model using the ANN and SCOA is described in detail. Sections 6 and 7 provide the simulation results and discussions, respectively. Finally, conclusions and future work are proposed in Section 8.

2 Literature review

A vital challenge for efficient water distribution in smart cities is water quality measurement (Kaddoura, 2022). One of the most important things to do is find the impurities in the water supply. Several papers have been proposed to classify the level of water quality. Following that, an overview of recently proposed models for water quality classification is investigated. Furthermore, current studies on the use of metaheuristic algorithms for tackling the hyperparameter optimisation problem are examined.

Aldhyani et al. (2020) used K-nearest neighbour (KNN), support vector machine (SVM), and naive Bayes to classify the level of water quality. The authors applied their approach to the Indian water quality dataset. The simulation results showed that SVM is the best machine-learning algorithm to classify water quality. It obtained an overall 97% classification accuracy. In Muhammad et al. (2015), the authors used five machine learning algorithms to classify the water quality of the Kinta River, Perak Malaysia. The simulation results revealed that the k-start algorithm obtained better results than J48, bagging, conjunctive rule, and naive Bayes. It obtained 86.67% classification accuracy. Another water quality classification model is proposed in Dilmi and Ladjal (2021). The authors used recurrent neural networks (RNNs) with Long Short Term Memory (LSTM) for Real-time water quality monitoring. The experimental results revealed that their proposed model obtained 99.72% classification accuracy. Sillberg et al. (2021) proposed another model based on SVM and integrating the attribute-realisation algorithm. They tested the performance of the model on Chao Phraya River's water quality dataset. The simulation results demonstrated that their proposed model is very promising and classifies the water quality of Chao Phraya River with 95% classification accuracy.

Although machine learning algorithms have shown effectiveness in a variety of applications, several of these algorithms require hyperparameter tweaking to improve performance. This is because the optimal tuning for these parameters can have a substantial impact on their overall performance. Metaheuristic optimisation algorithms can be employed to efficiently find a solution for the hyperparameter optimisation problem. Ahmadzadeh et al. (2017) used the particle swarm optimisation (PSO) algorithm with ANN to optimise the weights and biasses coefficient. The simulation results revealed that the proposed optimised ANN based on PSO can obtain better results compared with the traditional ANN and other multilayer regression models. In Elshaboury et al. (2021), the authors proposed an optimised version of ANN based on the teaching-learning-based optimisation algorithm (TLBO). They applied their optimised ANN to simulate the water network pipe condition. The performance of the TLBO algorithm is compared with PSO, sine cosine algorithm (SCA), and genetic algorithms (GA). The results showed that the proposed approach can effectively be used to plan the required maintenance and allocate the available budget for the water municipality. In Sayed and Hassanein (2023), the authors employed the war strategy optimisation algorithm to boost the performance of the ANN algorithm. The proposed optimised ANN is used for the classification of air pollutant species. Vadood et al. (2011) used the GA to tune the hyperparameters of ANN including the number of neurons, the number of hidden layers, activation functions, the learning rate, and the number of maximum fail epochs. The simulation results showed that using GA can significantly boost the performance of ANN. Another hybrid approach is proposed in Sayed et al. (2018). In this approach, a modified version of the optimal foraging algorithm is utilised to find the optimal values for the radial bias kernel function (RBF)

of the support vector machine. The results showed that the proposed approach can significantly boost the performance of the support vector machine algorithm.

It should be mentioned that the existing research papers mostly focus on the use of traditional machine-learning algorithms or deep-learning architectures for water quality classification, frequently ignoring the critical factor of adjusting their hyperparameter values. As a result, a significant research gap exists in the development of alternative approaches to improve water quality classification models. The goal of this paper is to fill that gap by introducing a hybrid approach based on employing a modified version of sand cat optimisation and the ANN algorithm. This paper also takes into account the interpretation of the results. As will be addressed more in the following sections, this hybrid approach has the potential to significantly increase the accuracy and efficiency of water quality classification, presenting a promising option for future research in this field. At the time of writing, there was no such hybridisation, including the use of XAI-based techniques.

3 Sand cat optimisation algorithm

In this section, the inspiration from the mathematical model of the original SCOA is presented. The inspiration analysis of SCOA is discussed in Section 3.1 and the mathematical model in Section 3.2.

3.1 Inspiration analysis

The original SCOA is one of the recent swarm intelligence algorithms proposed in 2023 by Seyyedabbasi and Kiani (2023). The inspiration for the original SCOA came from the behaviour of sand cats to survive in nature. The sand cat is one of the mammal families. These kinds of cats have a great ability to dig for prey and can remarkably recognise low frequencies less than 2 kHz. Foraging the prey and attacking the prey are the two main actions of sand cats. In the foraging behaviour, there are two stages. These stages are searching and attacking the prey. Next, the mathematical description of the original SCOA is presented.

3.2 Mathematical model

In the beginning, the SCOA starts with the random initialisation of sand cats' positions. Through searching for the prey (optimisation process), each sand cat optimisation is evaluated using a fitness function. When the sand cat swarm finds the best location of a sand cat close to the prey, next, all the sand cats tend to move toward that cat. Finally, the algorithm terminates and the best position with its corresponding best fitness value is reported.

Consider each sand cat position denoted as y_i at i^{th} iteration. Using the ability of the sand cat to detect low frequencies noise emission, the sensitivity range S_r is declared for the whole swarm. The value of S_r linearly decreases from two to zero. The mathematical definition of S_r is defined in equation (1). The authors of the original

SCOA defined another sensitivity range parameter R for each cat that will be used in the updating positions of cats. It is defined in the following equation.

$$S_r = S_h - \left(\frac{2 \times S_h \times i}{Max_{iter}}\right) \tag{1}$$

$$R = S_r \times r \tag{2}$$

where S_h is a constant parameter setted to 2 and Max_{iter} is the maximum number of iterations. To control the balancing between the exploration and exploitation phases, T is introduced as in equation (3).

$$T = 2 \times S_r \times r - S_r \tag{3}$$

where r is a random number generated between zero and one. Each sand cat updates its position based on the best candidate position Y_b. The mathematical equation of the new sand cat position through searching for the prey phase (exploration) is defined in equation (4).

$$Y_{i+1} = R \times (Y_b - r \times Y_i) \tag{4}$$

The mathematical definition of the updating position of the sand cat in attacking the prey phase (exploitation) is defined in equation (5).

$$Y_{i+1} = Y_b - R \times r \times (Y_b - Y_i) \times \cos(\Theta) \tag{5}$$

As the sensitivity range is defined as a circle, Θ is randomly chosen between 0 and 360 to guarantee that sand cats move in all directions in the search space.

Figure 1 Samples of the adopted dataset (see online version for colours)

	DO	PH	Conductivity	BOD	NI	Fec_col	Tot_col
2	6.7	7.5	203	0	0.1	11	27
3	5.7	7.2	189	2	0.2	4953	8391
4	6.3	6.9	179	1.7	0.1	3243	5330
5	5.8	6.9	64	3.8	0.5	5382	8443
6	5.8	7.3	83	1.9	0.4	3428	5500
7	5.5	7.4	81	1.5	0.1	2853	4049
8	6.1	6.7	308	1.4	0.3	3355	5672
9	6.4	6.7	414	1	0.2	6073	9423
10	6.4	7.6	305	2.2	0.1	3478	4990
11	6.3	7.6	77	2.3	0.1	2606	4301
12	7.1	7.1	176	1.2	0.1	4573	7817
13	6.7	6.4	93	1.4	0.1	2147	3433
14	7.4	6.8	121	1.7	0.4	11633	18125
15	6.9	7	620	1.1	0.1	3500	6300
16	6	7.5	72	1.6	0.2	4995	9517
17	7.3	7	247	1.5	0.2	1095	2453
18	7.3	7	188	1	0.1	1286	3048
19	7	6.9	224	1.2	0.3	3896	6742

4 Dataset description

In this paper, a dataset was collected from predefined historical locations in India. The Indian government obtained the dataset to assess the quality of the supplied drinking water during the period from 2005 to 2014 (Anbarivan, 2018). It has seven significant parameters biological oxygen demand (BOD), faecal coliform, dissolved oxygen (DO), total coliform (TC), pH Level (a measure of acidity), conductivity, and nitrate. The total number of records in this dataset is 7,245 records. Figure 1 shows a sample of the used dataset.

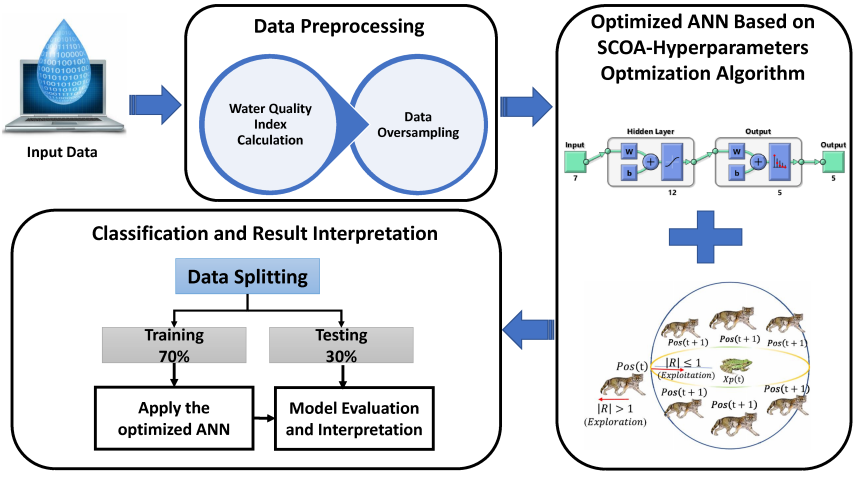
5 The proposed water quality classification model

Overall the proposed water quality classification model consists of three main phases; the data pre-processing phase, the optimised ANN based on the SCOA-hyperparameters optimisation algorithm phase, and the classification and result interpretation phase. Figure 2 shows the architecture of the proposed water quality classification model with a visual representation of the content per each phase. In this figure, the visual representations with the layer dimensions of each layer in the proposed ANN architecture are provided. The structure of the proposed ANN architecture and its specific configurations can be examined as follows: The proposed ANN architecture consists of two layers, each with a single input and output. Bias, input, layer, and output connect vectors delineate the network's connections. Input delays, layer delays, and feedback delays are important aspects, with a total of 161 weight factors adding to the model's complexity. These factors are tuned using the proposed SCOA-based hyperparameters optimisation algorithm. The proposed ANN's performance is evaluated using the mean squared error function. The Bayesian regularisation algorithm is used in the training process. The structure of the ANN is encapsulated in its weight and bias values, with layer weights (LW), input weights (IW), and biasses (b) all contributing to the overall model configuration. First, the original data is passed through the data pre-processing phase. In this phase, the water quality index is calculated. Then, based on the obtained value, the class level is determined as excellent, good, poor, very poor, and unsuitable for drinking. After that, the original dataset, consisting of a total of 7,245 samples, is divided into a 70% training set comprising approximately 5,072 samples and a 30% testing set constituting around 2,173 samples. Then, an oversampling method is applied to the training set by generating artificial samples for the minority classes. Then, the weights and bias coefficients of ANN are tuned using the proposed SCOA-based hyperparameters optimisation algorithm. Finally, the optimised ANN based on the modified version is employed to evaluate the performance of the overall water quality classification model using the testing set. Next, the detailed description of each phase is discussed.

5.1 Data pre-processing phase

Data pre-processing is considered a very important task in data analysis. Additionally, it has a significant impact on the performance of an algorithm. In this phase, the calculation of the water quality index is presented. Then the classes of WQI values are determined. Next, an oversampling method is applied.

Figure 2 The proposed water quality classification model architecture (see online version for colours)



5.1.1 *Water quality index calculation*

This paper uses seven parameters of the adopted dataset in the water quality index (WQI) calculation. Equation (6) defines the WQI formula.

$$WQI = \frac{\sum_{j=1}^{S} w_j \times q_j}{\sum_{j=1}^{S} w_j} \tag{6}$$

where S is the total number of parameters, which in this paper is 7. w_j is the weight of each parameter. It is mathematically defined in equation (7).

$$w_j = \frac{K}{N_j} \tag{7}$$

where N_j is the desired standard value of each j^{th} parameter. It is determined 8.5 for pH, 10 mg/l for dissolved oxygen, 45 mg/l for nitrate, 1,000 µS/cm for conductivity, 5 mg/l for biological oxygen demand, 100 cfu/100 ml for faecal coliform, and 1,000 cfu/100 ml for total coliform. K is a constant variable calculated using the following equation.

$$K = \frac{1}{\sum_{j=1}^{S} N_j} \tag{8}$$

$$q_j = 100 \times \left(\frac{V_j - V_{Ideal}}{N_j - V_{Ideal}} \right) \tag{9}$$

where V_j is the measured value of the j^{th} parameter, V_{Ideal} is the optimal value of the j^{th} parameter in pure water. It equals 0 for all the parameters in the dataset except pH equals 7.0 and DO equals 14.6 mg/l.

In this paper, the water quality index has been classified into five types. Table 1 shows the assigned class for each range of the water quality index value. As can be observed the paper considers five different classes of water quality index, namely excellent, good, poor, very poor, and unsuitable for drinking. Additionally, this table shows the classes' annotations that will be used later in the paper.

Table 1 Classification of water quality index

WQI range	*Class name*	*Class abbreviation*
$[0, 25]$	Excellent	Class 1
$[26, 50]$	Good	Class 2
$[51, 75]$	Poor	Class 3
$[76, 100]$	Very poor	Class 4
>100	Unsuitable for drinking	Class 5

5.1.2 Data oversampling

The dataset is the most significant component affecting the performance of a machine learning algorithm. Thus, the imbalanced dataset can cause a major problem in the training process. An imbalanced dataset means that a class or more classes have fewer samples (minority class) compared with the samples in the other classes (majority class). The variation of the class distribution can make a classifier skewed to the majority classes in the learning phase and ignore partially fully minority classes (Sayed et al., 2021). Data sampling methods are the most commonly used methods to address the data imbalance problem by adjusting the amount of samples in each class. According to which class to modify, they can be classified into oversampling and under-sampling methods. In this paper, the self-adaptive synthetic over-sampling method (Gu et al., 2020) is utilised to generate synthetic samples of the minority class for the training set. The SASYNO algorithm follows three phases; Identifying pairwise neighbouring samples, Creating explorations by Gaussian disturbance, and Creating interpolations for synthetic data generation.

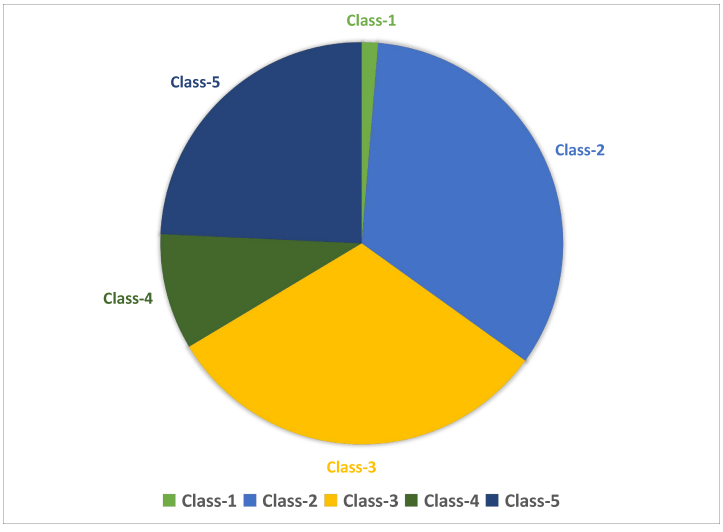
The class distribution percentage of the training set for the first class is 1.31%, for the second class is 33.6%, for the third class 31.49%, for the fourth class is 9.34%, and for the fifth class is 24.26%. Figure 3 shows a comparison between before and after applying the self-adaptive synthetic over-sampling method in the proposed model in terms of the class distribution. As can be observed, the distribution percentage for the first class and the fourth class is very low compared with the percentage of the other classes. Additionally, it can be observed that after applying the self-adaptive synthetic over-sampling method, the distribution percentage of each class became equal.

5.2 Optimised ANN based on SCOA-hyperparameters optimisation algorithm phase

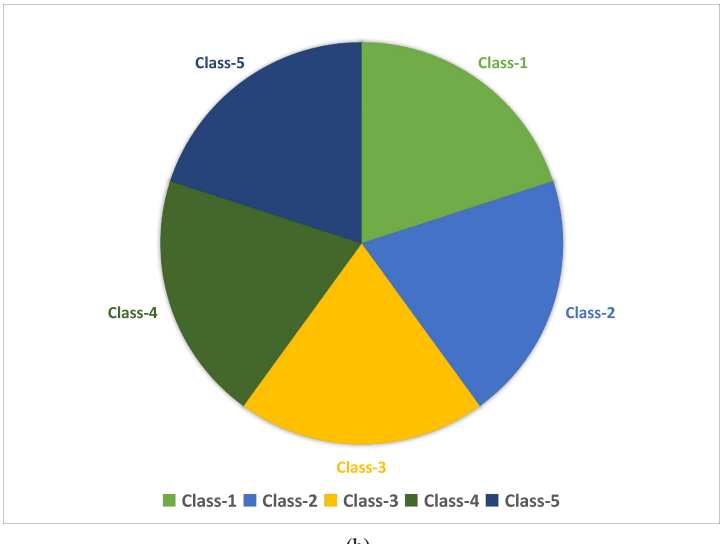
ANN is one of the machine learning algorithms that simulates the anatomy of the human brain. The feed-forward neural network is one of the ANN architectures. The input data is used to feed the network, which is received by the input layer and processed by the hidden layers. The output layer is used to predict the outcome. However, the traditional ANN proves its efficiency in many applications; it needs to be trained well with different

control parameters to obtain the desired accuracy. Moreover, the initial values of the weights and biasses have a significant effect on the overall performance of ANN.

Figure 3 The classes distribution (a) before applying the oversampling method and (b) after applying the oversampling method (see online version for colours)



(a)



(b)

In this paper, a modified version of the SCOA is proposed to find the optimal values of weights and bias parameters for each hidden neuron. Each sand cat position is a column vector that represents all network neurons' biasses and weights. The size of each sand cat position is calculated using the following formula:

$$dim = I_s \times n + n + n + 1 \tag{10}$$

where I_s is the input data size after the pre-processing phase and n is the number of hidden layers. The first part in equation (10) represents the number of input weights, the second part represents the number of input biasses, the third part represents the number of output weights, and the fourth part represents the number of output biasses. The rest of the parameter settings for this algorithm are listed in Table 2.

Table 2 Parameters setting of the modified SCOA

Parameter	*Value*
The maximum sensitivity range	2
Population size	30
Maximum number of iterations	20
Searching boundary	[–10, 10]

Through the optimisation process, the fitness value for each sand cat position is calculated. In this paper, mean square error (MSE) is used to evaluate how good the sand cat position is. It is mathematically defined as follows:

$$MSE(y, x) = \frac{\sum_j (y_j - x_j)^2}{N} \tag{11}$$

where y_j refers to the actual value and x_j refers to the predicted value. N is the total number of observations.

The sand cat position with the minimum fitness value is known as Y_b which has the optimal values of weights and biasses. With every iteration, the position of sand cats is updated using equations (4) and (5). The optimisation process is repeated over and over until a termination criterion is met. In this paper, the algorithm terminates when the maximum number of iterations is satisfied.

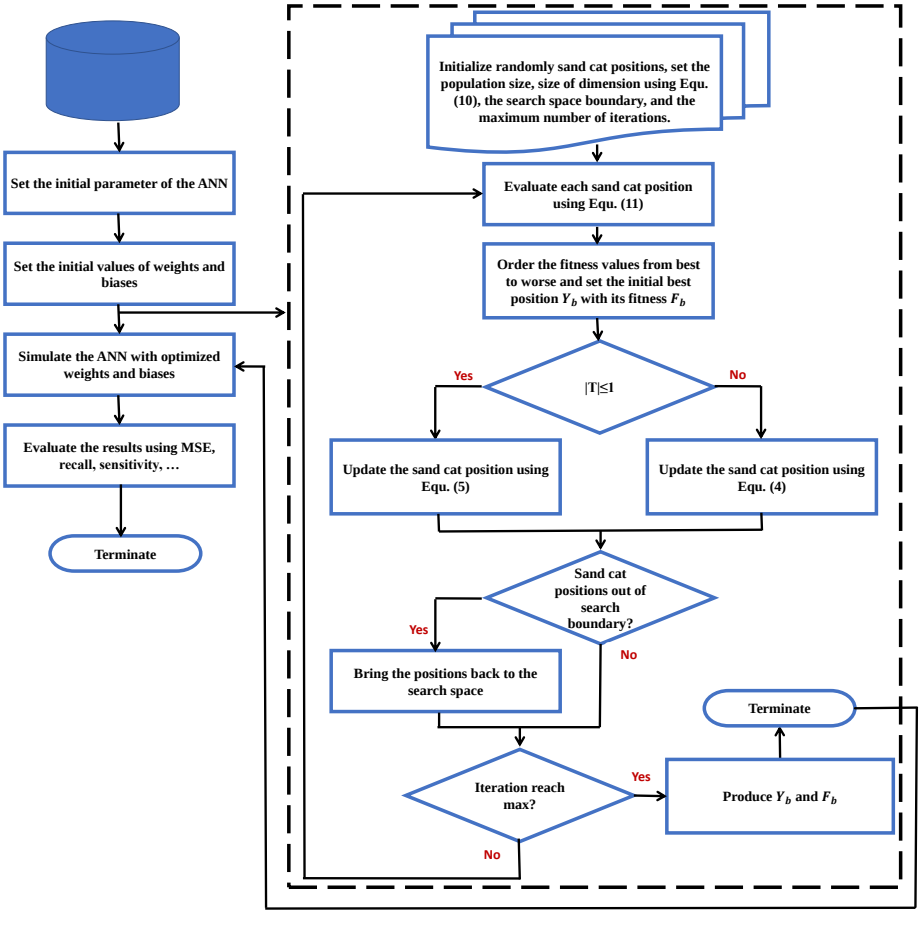
5.3 Classification and result interpretation phase

In this section, the proposed SCOA-hyperparameters optimisation algorithm is used to tune ANN. First, the dataset is divided into training and testing sets. Around 30% and 70% of the data are used for testing and training, respectively. The training set is used in the previous phase. The output from the previous phase is the best values of weights and bias parameters for each hidden neuron in ANN. The rest of the parameters setting of the proposed optimised version of ANN based on SCOA is shown in Table 3. The testing set is used to evaluate the overall proposed water quality classification model. Several metrics are used for the evaluation. These metrics are accuracy, sensitivity, specificity, f-score, convergence curve, and receiver operating characteristic curve. Figure 4 shows the overall architecture of the proposed optimised ANN based on the SCOA-hyperparameters optimisation algorithm.

Table 3 Parameters setting of the optimised ANN algorithm

Parameter	*Value*
No. of hidden units	10
Network training function	Bayesian regularisation
Neural transfer function	Hyperbolic tangent sigmoid transfer function
Performance function	Mean squared error function
No. of iterations	1,000
Neural network type	Feedforward network (pattern recognition network)

Figure 4 The proposed optimised ANN based on SCOA-based hyperparameters optimisation algorithm (see online version for colours)



Finally, the result interpretation is done using explainable AI (XAI). XAI has been recognised as a crucial component in the domains of model evolution and interpretation. Models of artificial intelligence (AI) have proved challenging to

comprehend, particularly those with high degrees of complexity like deep neural networks. In this situation, XAI serves as a lighthouse, cracking through the mystery and exposing model behaviour. By producing explanations for model predictions, XAI approaches are used to better comprehend the thinking underlying AI decisions. Through visuals, feature importance attribution, and interactive interfaces, XAI offers a thorough knowledge of how input data is turned into output predictions (Letzgus et al., 2022). In this paper, the proposed water quality classification model is explained by using Shapley additive explanations (SHAP), one of the XAI techniques. By quantifying the importance of each feature in the prediction process, SHAP values offer a precise and clear knowledge of how input features impact model results.

6 Simulation results

In this section, the conducted results from the proposed water quality classification model are reported and analysed. The experiments in this section are divided into three main experiments. The main objective of the first experiment is to evaluate the proposed model's performance before and after applying oversampling methods. Additionally, it aims to compare different oversampling methods on a real dataset. The objective of the second experiment is to evaluate the efficiency of the proposed SCOA-based hyperparameter optimisation algorithm. Finally, the last experiment aims to evaluate the overall proposed water quality classification model. Also, it aims to compare the performance of the proposed model with other proposed models in the literature. In all the experiments, the significance and the reasoning behind why these algorithms are adopted in the proposed model are presented and discussed. The best-obtained results are highlighted in bold format. All the conducted experiments are implemented on MATLAB 2020 with Core i7 and 16 GB RAM.

6.1 Data pre-processing results

In this section, the conducted experiment aims to highlight the importance of tackling the imbalanced data problem using oversampling methods. Moreover, the performance of the proposed water quality classification model is evaluated by applying an oversampling method in the data pre-processing phase and without using it. It should be noted that in this experiment, the rest of the phases including SCOA hyperparameter optimisation and classification phases are considered. Table 4 compares the performance of the proposed water quality classification model before and after applying the SASYNO oversampling method in terms of sensitivity, accuracy, specificity, f-score, and precision. This table shows how employing oversampling techniques such as SASYNO can considerably decrease the problem of class imbalance within a dataset. Machine learning models tend to be biased toward the majority class in cases where one class considerably outnumbers the others, resulting in inferior performance for minority classes. Oversampling is the practice of generating artificial samples of a minority class. During model training, this augmentation ensures a more equal representation of classes, preventing the algorithm from favoring the majority class. As can be seen from Table 4, after applying oversampling, the model's ability to learn from minority class patterns is improved by providing it with a more equitable distribution of instances, resulting in enhanced overall performance and increased accuracy across all classes. As can be

observed, the difference between before using SASYNO and after using it is almost 10%. This is due to the imbalance in the class distribution, where the first class and the fourth class are the minority classes, despite the rest of the classes.

Table 4 Comparison of the performance of the proposed water quality classification model before and after applying SASYNO

	Before	*After*
Precision (%)	86.63	95.85
Sensitivity (%)	86.41	100
Specificity (%)	96.61	98.71
Accuracy (%)	86.43	94.89
F-score (%)	86.43	97.85

Table 5 compares the results of two different oversampling methods namely SASYNO and synthetic minority oversampling technique (SMOTE) (Barua et al., 2011). SMOTE uses a k-nearest neighbour algorithm to search for the neighbours in the minority class. Then, it synthesises the selected sample and its neighbours to generate new samples. As can be observed, the employed SASYNO oversampling method is more suitable to the characteristics of the adopted dataset. Thus, it significantly reflected the classification performance. As can be seen, the employed SASYNO oversampling method obtained better results than using SMOTE in terms of precision, sensitivity, specificity, f-score, and accuracy.

Table 5 The proposed water quality classification model using SASYNO vs. using SMOTE

	SMOTE	*SASYNO*
Precision (%)	91.44	95.85
Sensitivity (%)	87.24	100
Specificity (%)	98.02	98.71
Accuracy (%)	91.23	94.89
F-score (%)	89.29	97.85

6.2 Optimised ANN based on SCOA-hyperparameters optimisation algorithm results

In this experiment, the performance of the proposed SCOA-based hyperparameter optimisation is tested. Several metrics are used for evaluation purposes. These metrics are precision, sensitivity, specificity, accuracy, f-score, elapsed time, and convergence curve. Table 6 compares the performance of the overall proposed water quality classification model before using the modified version of SCOA and after using it. As can be seen, tuning the weights and bias parameters of the traditional ANN can significantly boost its performance and the overall performance of the water quality classification model. This is due to, the inappropriate setting of the initial values of biasses and weight coefficients can make the CPU take a long time through the learning process. This can lead to the backpropagation getting stuck and producing incorrect results.

Table 6 The performance of the proposed water quality classification model before and after employing the optimised ANN based on the SCOA algorithm

	Before	*After*
Precision (%)	95.96	98.11
Sensitivity (%)	95.96	98.11
Specificity (%)	99.00	99.53
Accuracy (%)	95.99	98.11
F-score (%)	95.95	98.11

Figure 5 The elapsed time in seconds before and after employing the optimised ANN based on the SCOA algorithm (see online version for colours)

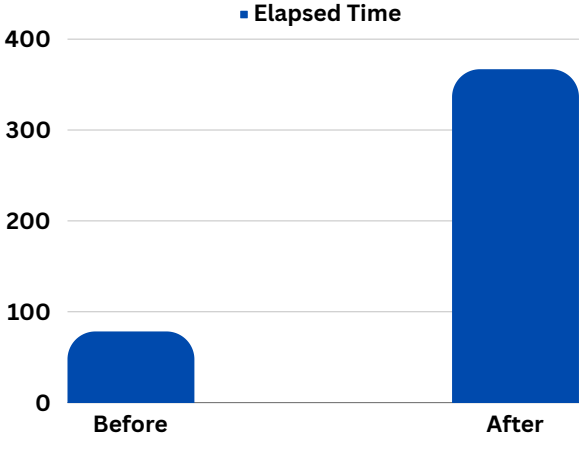
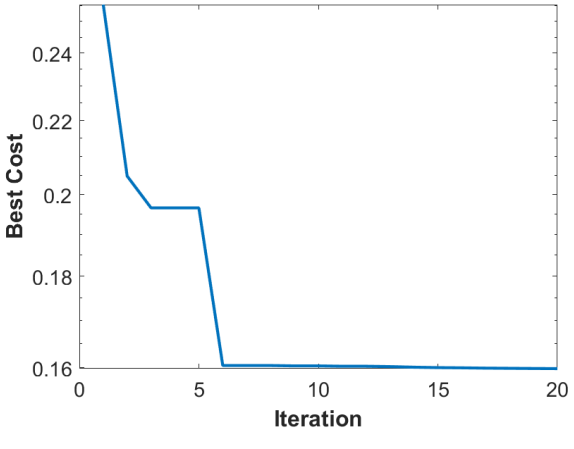


Figure 6 Conversion curve of the proposed SCOA-based hyperparameters optimisation algorithm (see online version for colours)



To comprehensively assess the performance of the proposed water quality classification model, comparisons can be done before and after employing the optimised ANN based

on the SCOA-based hyperparameters optimisation algorithm, taking into account elapsed time. It is demonstrated from Figure 5 that applying the SCOA-based hyperparameter optimisation algorithm increases the elapsed time. However, when evaluating the accuracy, precision, specificity, sensitivity, and f-score, the SCOA-based hyperparameter optimisation algorithm is preferred, especially in water quality classification applications where accuracy metrics are more important than elapsed time considerations.

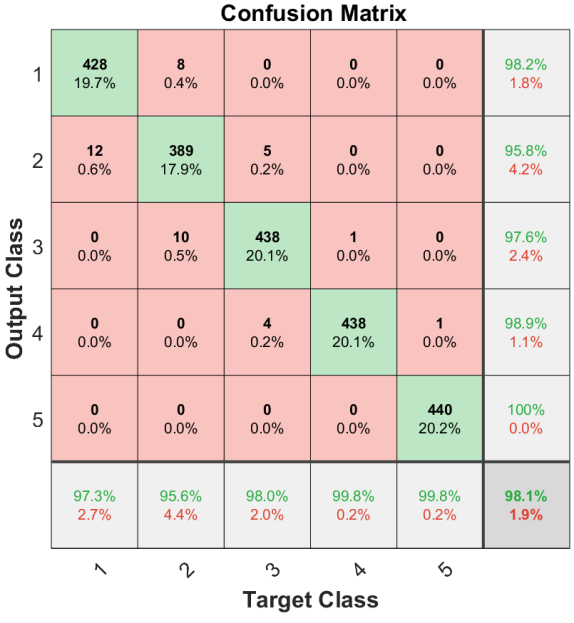
Figure 6 shows the convergence curve of the modified sand cat optimisation algorithm. The convergence curve is one well-known metric used to evaluate the stability of an algorithm. It shows the optimal score obtained throughout the optimisation process. As can be observed from this figure, the proposed SCOA-based hyperparameter optimisation algorithm converges at almost the 7th iteration. Additionally, it can be observed from the result that the modified version of SCOA can find the approximate or optimal solution in a reasonable time.

6.3 Classification and result interpretation results

This experiment aims to evaluate the efficiency and reliability of the overall proposed model. Additionally, it aims to compare the obtained results of the proposed model with other state-of-the-art models previously proposed in the literature. Figure 7 shows the confusion matrix of the proposed water quality classification model based on optimised ANN using the modified version of the sand cat optimisation algorithm, where the number of true negatives, true positives, false negatives, and false positives are reported. The y-axis values represent the output class from the proposed model, and the x-values represent the target class. Additionally, annotation 1 is used to denote the excellent class, 2 denotes the good class, 3 denotes the poor class, 4 denotes the very poor class, and finally, 5 denotes the unsuitable drinking class. As can be observed, only a few samples are misclassified. Furthermore, the proposed water quality classification model identified a total of 12 incorrectly classified samples out of 406 samples. The model incorrectly classified several samples as class 2, when they should have been class 1. Another noteworthy observation is that the proposed model excels at identifying class 5. This means that it can efficiently identify the samples that are unsuitable for drinking based on the reported value of biological oxygen demand, faecal coliform, dissolved oxygen, total coliform, pH level, conductivity, and nitrate.

To further evaluate the performance of the proposed water quality classification model, receiver operating characteristic shortly ROC is used. ROC is one of the commonly used methods to evaluate the performance of a binary classifier. It is used to visualise the confusion metric properties, such as false-positive true positives. Moreover, it is used to calculate the area under the curve value, which is used for summarising the performance of an algorithm with a single value. Figure 8 displays the ROC curve of the proposed model using a real dataset, where x-axis values denote the false positive rate and y-axis values denote the true positive rate. As it can be observed, the misclassification rate is very low and the detection rate is very high. Furthermore, the overall area under the curve (AUC) produced by the proposed water quality classification model is practically perfect, approaching one. This indicates the model's ability to differentiate between different classes is remarkable, with few misclassification instances.

Figure 7 Confusion matrix of the proposed water quality classification model (see online version for colours)



To further interpretation of the proposed water quality classification model, SHAP analysis is utilised. The influence of each feature on the overall performance of the model is ordered for the SHAP plot, with the feature with the greatest impact shown first. The test sample is indicated by the coloured plotted point in the SHAP plot. The real value of the test sample data is shown by the coloured point, which ranges from blue (low values) to red (high values). Figure 9 shows the SHAP summary plot for the fifth class (unsuitable for drinking). As can be observed, the pH level is the most important parameter that affects detecting whether the water is suitable or unsuitable for drinking. This parameter indicates the acidity of the water. Additionally, it can be observed that biological oxygen demand, nitrate, and faecal coliform count, shown as red dots, move toward positive SHAP values, but the pH level, shown as red dots, moves toward negative SHAP values. These apparent patterns demonstrate the significance of the pH level, biological oxygen demand, nitrate, and faecal coliform count features in influencing the water quality.

The ability to apply SHAP analysis to individual test instances is another benefit when utilising it for the model's interpretation. This allows for the observation of the features that have a significant impact on the performance of the proposed model in a single test case. Figure 10 shows the SHAP force plot for a single test instance per class. As can be observed, pH level has the highest impact on the prediction of water quality level. The biological oxygen demand is in second place. The acidity of the water (pH) feature, biological oxygen demand, and nitrate are essential for thoroughly comprehending and assessing water quality since they provide in-depth information on the composition of chemicals and physical characteristics of the water. They are a component of a larger framework that takes into account several factors to offer a

comprehensive evaluation of water quality. This obtained result is consistent with the obtained result in Figure 9.

Figure 8 ROC curve for the proposed water quality classification model (see online version for colours)

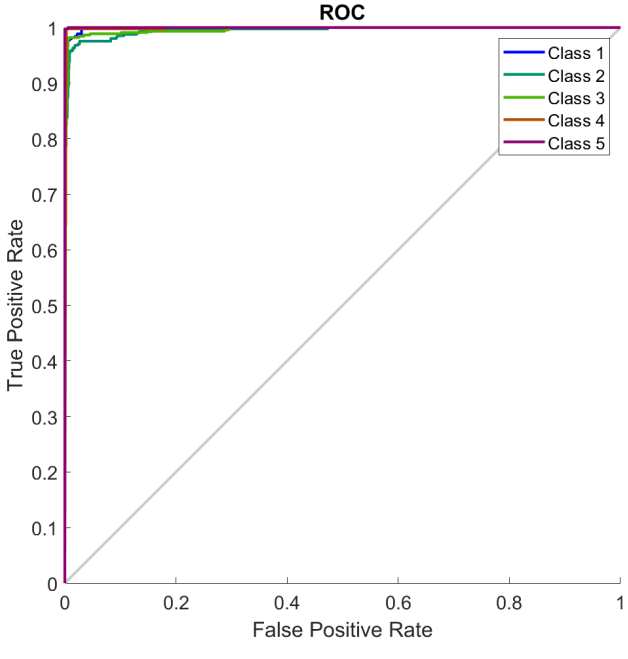


Figure 9 SHAP summary plot (see online version for colours)

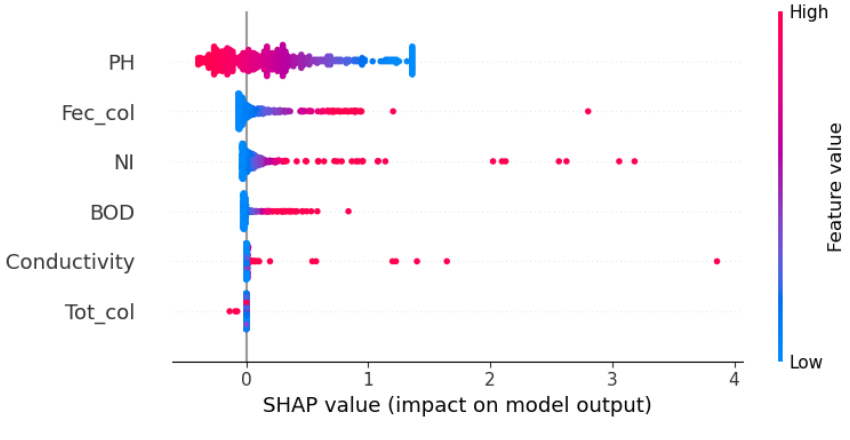


Figure 10 SHAP force plot for a single test instance for each class in the adopted dataset (see online version for colours)

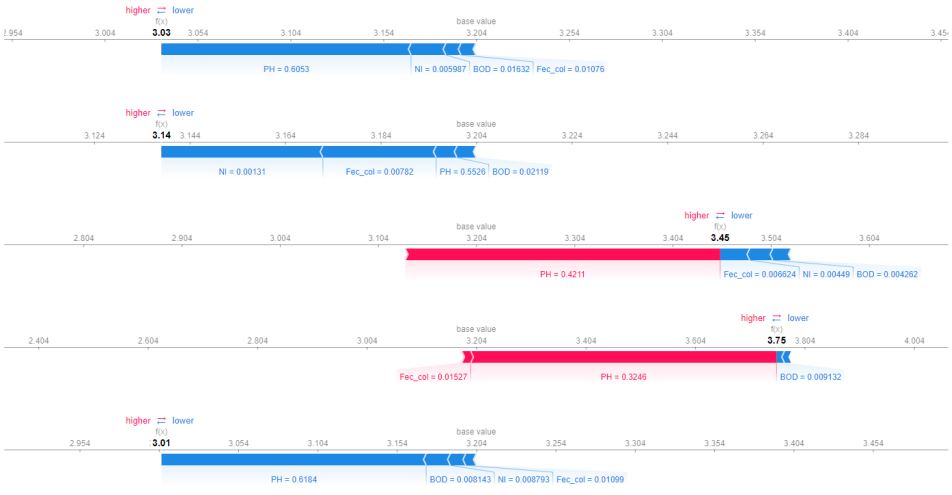


Table 7 Comparison of the obtained results of the proposed water quality classification model based on optimised ANN with other state-of-the-art models

	#classes	*Acc. (%)*	*Spec. (%)*	*Sens. (%)*	*FSc. (%)*	*Year*
Aldhyani et al. (2020)	5	97	97	**99**	98	2020
Radhakrishnan and Pillai (2020)	5	98	–	–	–	2020
Hassan et al. (2021)	3	99	–	–	–	2021
Al-Adhaileh and Alsaade (2021)	4	**100**	**99**	**99**	**100**	2021
Al-Adhaileh and Alsaade (2021)	4	**100**	**99**	**99**	–	2021
Nasir et al. (2022)	4	94	–	94	94	2022
Relangi et al. (2023)	4	99	**99**	**99**	–	2023
Shams et al. (2023)	3	99	–	**99**	99	2023
Proposed model	5	98	**99**	98	98	2024

Table 7 compares the obtained results of the proposed water quality classification model with other proposed models in the literature in terms of accuracy (Acc.), specificity (Spec.), sensitivity (Sens.), and f-score (FSc.). It should be mentioned that to make a fair comparison with all competitive models, all of these models employed the same dataset. As can be observed, the proposed model obtained very competitive results. Another finding, however, the model in Al-Adhaileh and Alsaade (2021) obtained the highest accuracy, but the authors categorised the value of the water quality index into four classes, not five classes as in this paper. These classes are excellent, good, poor, and very poor. The same finding is for Al-Adhaileh and Alsaade (2021) and Nasir et al. (2022), where only four classes are considered. Additionally, for the proposed model in Hassan et al. (2021) and Shams et al. (2023), only three classes are considered. These classes are good, poor, and unsuitable for drinking. Thus the performance of these models can not be guaranteed for handling five classes. Therefore, the reported accuracy results from those models can be decreased. Another finding, although the model in Radhakrishnan and Pillai (2020) achieved identical accuracy findings, it is important to

highlight that reporting only accuracy might be misleading, especially when dealing with imbalanced datasets, as in the adopted dataset. In such cases, where there is a severe class imbalance, additional measures such as f-score must be considered to provide a more comprehensive evaluation of the model's performance.

7 Discussion

From the obtained results in Table 6 and Figures 5 and 6, it can be revealed the ability of the proposed SCOA-based hyperparameters optimisation algorithm to boost the classification performance of the whole model. Additionally, it can be observed that the proposed SCOA-based hyperparameters optimisation algorithm can remarkably find the optimal weights and bias coefficients of the ANN architecture in a reasonable time. Moreover, it can be observed from the conducted experiment in Table 6 that employing the optimised ANN using the modified version of the SCOA can significantly boost the performance of the traditional ANN. However, it should be mentioned that employing swarm optimisation such as SCOA to tune the hyperparameters of ANN introduces challenges inherent in their nature which are mainly based on stochastic behaviour. Because swarm optimisation techniques are inherently stochastic, their parameters contain randomness and lack deterministic guarantees. As a result, the behaviour of these algorithms cannot be predicted exactly. The hyperparameter optimisation procedure in this paper is based on the trial-and-error methodology. Because swarm algorithms are stochastic, numerous runs may be required to thoroughly explore the hyperparameter space. In each run, the initial population is randomly generated. The ideal values for ANN parameters are identified iteratively, with each iteration being a probabilistic attempt to identify configurations that improve the model's performance on the given data.

Another challenge is that employing swarm intelligence algorithms can introduce a potential challenge in terms of computational complexity as shown in Figure 5. In particular, the goal of this paper is to find the best weight and bias values for the ANN with ten hidden layers. However, it is critical to acknowledge that as the number of hidden layers increases, so will the computational time. Furthermore, when dealing with larger datasets, the complexity can be increased as well. The complex ANN architecture and large dataset can significantly affect the convergence curve as well. In this paper, the proposed SCOA-based hyperparameters optimisation only takes seven iterations to reach the optimal solution as in Figure 6. However, this behaviour can't be guaranteed with either complex ANN architecture or a large dataset, as it may take more iterations greater than seven to reach the optimal solution.

Several critical indicators, namely pH level, biological oxygen demand (BOD), nitrate concentration, and faecal coliform count have a significant impact on water quality classification, especially pH level as can be observed from Figure 10. The pH level is one of these characteristics that serves as an indicator of acidity or alkalinity, providing important information about the chemical balance. The amount of dissolved oxygen required by microorganisms to break down organic materials in water is represented by biological oxygen demand. Nitrate concentrations are a good indicator of nutrient levels, and they are frequently related to agricultural runoff and possible contamination. The faecal coliform count indicates microbial contamination and can provide insight into potential health issues. The importance of these measures in

assessing the environmental health and potential pollutants in aquatic environments is highlighted by their major impact on water quality classification. The interaction of pH, BOD, faecal coliform count, and nitrate, provides a more nuanced understanding of water quality, assisting in better monitoring, analysis, and management methods for long-term water resource use.

Moreover, it can be revealed from the conducted experiments that the proposed water quality classification model can be further used to monitor the level and quality of drinking water with a high detection rate. The reliability and robustness of the proposed model can be used by authorised governments to find a suitable strategy that serves the demands of the community. Thus, the proposed model can be further used to the serve Sustainable Development Goal that aims to guarantee access to clean water for all.

8 Conclusions and future work

A vital concern for health is clean water. This is due to that they lower negative health consequences and medical expenses more than they cost to execute, investments in water supply and sanitation have been demonstrated to generate a net economic gain in some locations. However several contaminants are ruining the purity of drinking water. Additionally, detecting the level and quality of drinking water is considered a very important task for the environment's protection. This paper introduces a new model for classifying the quality level of drinking water from different locations in Indian states. Additionally, a modified version of SCOA is applied to ANN to find the optimal values of weights and bias parameters. Moreover, the obtained results of the proposed model are explained and interpreted using one of the XAI techniques, namely SHAP. Employing SHAP can remarkably measure how each feature can impact the final decision. The experimental results revealed that the proposed model is promising and can effectively be further used as a smart water quality monitoring system. Additionally, the results demonstrated that the pH level feature has the highest impact on the prediction of water quality. Further studies will be done to apply the proposed model to more complex and real-life datasets. Additionally, the proposed modified version of SCOA can be further applied for hyper-parameter optimisation of other deep-learning architectures.

Compliance with ethical standards

Data availability

The dataset used during the current paper are obtained by the Indian Government and available in the Kaggle repository, https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data.

References

Al-Adhaileh, M.H. and Alsaade, F.W. (2021) 'Modelling and prediction of water quality by using artificial intelligence', *Sustainability*, Vol. 13, No. 8, pp.1–18, DOI: 10.3390/su13084259.

Ahmadzadeh, E., Lee, J. and Moon, I. (2017) 'Optimized neural network weights and biases using particle swarm optimization algorithm for prediction applications', *Journal of Korea Multimedia Society*, Vol. 20, No. 8, pp.1406–1420, DOI: 10.9717/kmms.2017.20.8.1406.

Aldhyani, T., Al-Yaari, M., Alkahtani, H., Maashi, M. (2020) 'Water quality prediction using artificial intelligence algorithms', *Applied Bionics and Biomechanics*, Vol. 2020, pp.1–12, DOI: 10.1155/2020/6659314.

Anbarivan, N. (2018) *Indian Water Quality Data* [online] https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data (accessed 1 September 2023).

Barua, S., Islam, M. and Murase, K. (2011) 'A novel synthetic minority oversampling technique for imbalanced data set Learning', *Neural Information Processing: 18th International Conference, ICONIP 2011*, Shanghai, China, pp.735–744.

Ben-Daoud, M., El Mahrad, B., Moroșanu, G.A., Elhassnaoui, I., Moumen, A., El Mezouary, L., ELbouhaddioui, M., Essahlaoui, A. and Eljaafari, S. (2023) 'Stakeholders' interaction in water management system: insights from a MACTOR analysis in the R'Dom Sub-basin, Morocco', *Environmental Management*, Vol. 71, No. 6, pp.1129–1144, DOI: 10.1007/s00267-022-01773-x.

Berthet, A., Vincent, A. and Fleury, P. (2021) 'Water quality issues and agriculture: an international review of innovative policy schemes', *Land Use Policy*, Vol. 109, p.105654, https://doi.org/10.1016/j.landusepol.2021.105654.

Bourouis, S. (2022) 'Detection of coronavirus disease using texture analysis and machine learning methods', *International Journal of Intelligent Engineering Informatics*, Vol. 10, No. 3, pp.196–211, DOI: 10.1504/IJIEI.2022.128448.

Bui, D., Khosravi, K., Tiefenbacher, J., Nguyen, H. and Kazakis, N. (2020) 'Improving prediction of water quality indices using novel hybrid machine-learning algorithms', *Science of the Total Environment*, Vol. 721, p.137612, https://doi.org/10.1016/j.scitotenv.2020.137612.

Dilmi, S. and Ladjal, M. (2021) 'A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques', *Chemometrics and Intelligent Laboratory Systems*, Vol. 214, p.104329, https://doi.org/10.1016/j.chemolab.2021.104329.

Elshaboury, N., Abdelkader, E., Al-Sakkaf, A. and Alfalah, G. (2021) 'Teaching-learning-based optimization of neural networks for water supply pipe condition prediction', *Water*, Vol. 13, No. 24, pp.1–20, DOI: 10.3390/w13243546.

Gaya, M., Zango, M.U., Yusuf, L., Mustapha, M., Muhammad, B., Sani, A., Tijjani, A., Wahab, N. and Khairi, M. (2017) 'Estimation of turbidity in water treatment plant using Hammerstein-Wiener and neural network technique', *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 5, No. 3, pp.666–672, DOI: 0.11591/ijeecs.v5.i3.pp666-672.

Gu, X., Angelov, P. and Soares, E. (2020) 'A self-adaptive synthetic oversampling technique for imbalanced classification', *International Journal of Intelligent Systems*, Vol. 35, pp.923–943, DOI: 10.1002/int.22230.

Hassan, M., Mehedi, H., Mahedi, M., Akter, L., Rahman, M.M., Zaman, S., Hasib, K.M., Jahan, N., Smrity, R.N., Farhana, J. and Raihan, M. (2021) 'Efficient prediction of water quality index (WQI) using machine learning algorithms', *Human-Centric Intelligent Systems*, Vol. 1, pp.86–97, DOI: 10.2991/hcis.k.211203.001.

Iraji, A., Karimi, J., Keawsawasvong, S. and Nehdi, M. (2022) 'Minimum safety factor evaluation of slopes using hybrid chaotic sand cat and pattern search approach', *Sustainability*, Vol. 14, No. 13., pp.1–24, DOI: 10.3390/su14138097.

Jeyanthi, S., Venkatakrishnaiah, R. and Raju, K. (2023) 'Utilising recurrent neural network technique for predicting strand settlement on brittle sand and geocell', *International Journal of Intelligent Engineering Informatics*, Vol. 11, No. 2, pp.122–137, DOI: 10.1504/IJIEI.2023.132699.

Kaddoura, S. (2022) 'Evaluation of machine learning algorithm on drinkingwater quality for better sustainability', *Sustainability*, Vol. 14, No. 18, pp.1–17, DOI: 10.3390/su141811478.

Kamali, M., Appels, L., Yu, X., Aminabhavi, T. and Dewil, R. (2021) 'Artificial intelligence as a sustainable tool in wastewater treatment using membrane bioreactors', *Chemical Engineering Journal* Vol. 417, p. 128070, https://doi.org/10.1016/j.cej.2020.128070.

Koech, R. and Langat, P. (2018) 'Improving irrigation water use efficiency: a review of advances, challenges and opportunities in the Australian context', *Water*, Vol. 10, No. 12, pp.1–17, DOI: 10.3390/w10121771.

Krishnan, S.R., Nallakaruppan, M., Chengoden, R., Koppu, S., Iyapparaja, M., Sadhasivam, J. and Sethuraman, S. (2022) 'Smart water resource management using artificial intelligence: a review', *Sustainability*, Vol. 14, No. 20, pp.1–28, DOI: 10.3390/su142013384.

Kumar, M. and Sharma, A. (2023) 'Progressive global best artificial bee colony algorithm for wind farm layout optimisation problem', *International Journal of Intelligent Engineering Informatics*, Vol. 11, No. 3, pp.272–297, DOI: 10.1504/IJIEI.2023.133075.

Letzgus, S., Wagner, P., Lederer, J., Samek, W., Müller, K. and Montavon, G. (2022) 'Toward explainable artificial intelligence for regression models: a methodological perspective', *IEEE Signal Processing Magazine*, Vol. 39, No. 4, pp.40–58, DOI: 10.1109/MSP.2022.3153277.

Malviya, A. and Jaspal, D. (2021) 'Artificial intelligence as an upcoming technology in wastewater treatment: a comprehensive review', *Environmental Technology Reviews*, Vol. 10, No. 1, pp.177–187, DOI: 10.1080/21622515.2021.1913242.

Mehta, A., Jangid, J., Saxena, A., Shekhawat, S. and Kumar, R. (2022) 'Harmonics estimator design with Trigonometric function inspired grey wolf optimiser', *International Journal of Intelligent Engineering Informatics*, Vol. 10, No. 3, pp.212–241, DOI: 10.1504/IJIEI.2022.128447.

Muhammad, S.Y., Makhtar, M., Rozaimee, A., Aziz, A. and Jamal, 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, DOI: 10.14257/ijseia.2015.9.6.05.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A. and Al-Shamma'a, A. (2022) 'Water quality classification using machine learning algorithms', *Journal of Water Process Engineering*, Vol. 48, p.102920, https://doi.org/10.1016/j.jwpe.2022.102920.

Patel, V., Bhatti, D., Ganatra, A. and Tailor, J. (2022) 'An automated approach for electroencephalography based seizure detection using machine learning algorithms', *International Journal of Intelligent Engineering Informatics*, Vol. 10, No. 4, pp.332–358, DOI: 10.1504/IJIEI.2022.128890.

Priya, S. and Khanaa, V. (2022) 'An intelligent fuzzy and IoT-aware air quality prediction and monitoring system using CRF and Bi-LSTM', *International Journal of Intelligent Engineering Informatics*, Vol. 10, No. 5, pp.379–396, DOI: 10.1504/IJIEI.2022.129095.

Pruss-Ustun, A. (2008) *Safer Water, Better Health: Costs, Benefits and Sustainability of Interventions to Protect and Promote Health*, World Health Organization.

Radhakrishnan, N. and Pillai, A. (2020) 'Comparison of water quality classification models using machine learning', *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, pp.1183–1188.

Relangi, N., Chaparala, A. and Sajja, R. (2023) 'Effective groundwater quality classification using enhanced whale optimization algorithm with ensemble classifier', *International Journal of Intelligent Engineering and Systems*, Vol. 16, No. 1, pp.214–223, DOI: 10.22266/ijies2023.0228.19.

Sayed, G., Soliman, M. and Hassanien, A. (2021) 'A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization', *Computers in Biology and Medicine*, Vol. 136, p.104712, https://doi.org/10.1016/j.compbiomed.2021.104712.

Sayed, G. and Hassanein, A. (2023) 'Air pollutants classification using optimized neural network based on war strategy optimization algorithm', *Automatic Control and Computer Sciences*, Vol. 57, No. 6, pp.600–607, DOI: 10.3103/S0146411623060081.

Sayed, G., Soliman, M. and Hassanien, A. (2018) 'Modified optimal foraging algorithm for parameters optimization of support vector machine', *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, Springer, Egypt, pp.23–32.

Seyyedabbasi, A. and Kiani, F. (2023) 'Sand cat swarm optimization: a natureinspired algorithm to solve global optimization problems', *Engineering with Computers*, Vol. 39, No. 4, pp.2627–2651, DOI: 10.1007/s00366-022-01604-x.

Shams, M.Y., Elshewey, A.M., El-kenawy, E.S.M., Ibrahim, A., Talaat, F. and Tarek, Z. (2023) 'Water quality prediction using machine learning models based on grid search method', *Multimedia Tools and Applications*, pp.1–28, https://doi.org/10.1007/s11042-023-16737-4.

Sillberg, C., Kullavanijaya, P. and Chavalparit, O. (2021) 'Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River', *Journal of Ecological Engineering*, Vol. 22, No. 9, pp.70–86, DOI: 10.12911/22998993/141364.

Singh, C. and Vigila, S. (2023) 'An investigation of machine learning-based intrusion detection system in mobile ad hoc network', *International Journal of Intelligent Engineering Informatics*, Vol. 11, No. 1, pp.54–70, DOI: 10.1504/IJIEI.2023.130704.

Vadood, M., Semnani, D. and Morshed, M. (2011) 'Optimization of acrylic dry spinning production line by using artificial neural network and genetic algorithm', *Journal of Applied Polymer Science*, Vol. 120, No. 2, pp.735–744, DOI: 10.1002/app.33252.

Wu, D., Rao, H., Wen, C., Jia, H., Liu, Q. and Abualigah, L. (2022) 'Modified sand cat swarm optimization algorithm for solving constrained engineering optimization problems', *Mathematics*, Vol. 10, No. 22, pp.1–41, DOI: 10.3390/math10224350.

Zhao, L., Dai, T., Qiao, Z., Sun, P., Hao, J. and Yang, Y. (2020) 'Application of artificial intelligence to wastewater treatment: a bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse', *Process Safety and Environmental Protection*, Vol. 133, pp.169–182, DOI: 10.1016/j.psep.2019.11.014.