International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Hybrid Intelligent Technique between Supervised and Unsupervised Machine Learning to Predict Water Quality

Hanan Anas Aldabagh[1] and Ruba Talal Ibrahim[2]

[1] *Department of Computer Science,The General Directorate of Education in Nineveh Governorate,Mosul,Iraq*
[2] *Department of Computer Science,University of Mosul, Mosul,Iraq*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: Water is the secret of life and makes up almost 70% of the Earth's surface. It has become necessary to protect the water resources around us from pollution and neglect, which can result in the loss of life and health. Artificial intelligence (AI) has the potential to improve water quality analysis, forecasting, and monitoring systems for sustainable and environmentally friendly water resource management. As a result, this work focuses on the prediction of accurate and sustainable water quality prediction model using hybridization between supervised and unsupervised machine learning techniques. A set of multi-model learning features was used to represent the state of the water and determine its suitability category (i.e., safe or unsafe). This is done by building a hybrid model between supervised algorithms (LGBM) and unsupervised algorithms (COPOD, IForest, and CBLOF) after fusing their outliers, and the proposed model is called (HLGBM+Fusion CIC). Also, the Gamel herd swarm optimization algorithm was applied to find the optimum hyper-parameters. The models were evaluated with or without class balancing and compared in terms of accuracy, recall, precision, f1 score, and area under the curve (AUC). The results showed that the proposed model (HLGBM+Fusion CIC) outperformed other models by 99.2% in accuracy, AUC, and f1-score. Also, it achieved 99% precision and 99.3% recall. Finally, this paper presented a framework for researchers using hybrid machine learning to forecast water quality.

Keywords: Prediction, Water Quality, Artificial Intelligent, Machine Learning, supervised, unsupervised, Camel Herd Algorithm.

1. INTRODUCTION

Water is the most important vital resource for sustaining the life of living organisms, as clean water is used in various aspects of life, such as drinking, agriculture, energy generation, and entertainment[1]. But at the same time, as a result of progress and development of modern technology, aquatic ecosystems have become a threat to the existence of living beings. As most countries in the world use and sell chemicals, this exposes the water to toxic substances and makes it unfit for consumption by living organisms in addition to agriculture. In the context of the new green economy, monitoring and analyzing water quality is critical for the sustainability of all living creatures[2]. Because of the existence of precise water quality standards, traditional chemical monitoring methods are unable to assess the complex interactions and impacts of many stressors on microorganisms in water[3]. Many organizations employ manual techniques to monitor water quality and assess complicated interactions, calculating the water quality index equation after collecting samples and analyzing them in a laboratory, which has proven to be costly and time-consuming. Recently, many artificial intelligence studies have demonstrated the possibilities of employing machine learning technology and sensors to handle the problem of forecasting water quality ,consumption and automating their monitoring, as well as the ability to gather data in real time[4],[5]. Machine learning, a branch of artificial intelligence, enables a system to automatically learn and train data in order to recognize trends and update itself without the need for explicit programming[6]. Machine learning opens up new prospects for predicting the WQ index in water body investigations by giving photo-sensors that rely on determining the wavelength of a given color or variations in amplitude values, which may be utilized to detect various dissolved water contaminants[7],[8]. The outputs of these sensors can generate data that is processed using machine learning techniques with high accuracy and performance. Machine learning models may successfully mimic hydrological processes and pollution transport when big datasets are available[9].

Researchers in the field of artificial intelligence have preferred to use traditional machine learning approaches to forecast water quality, with excellent outcomes. However, in this work, a new hybrid technique combining supervised machine learning methods and unsupervised machine learning methods was applied and outperformed previous studies, and the following contributions were made:

1) The dataset was processed by using normalization and

E-mail address: hanan_aldabagh1985@nan.epedu.gov.iq, rubatalal@uomosul.edu.iq

oversampling to balance it.
2) The (LGBM) hyperparameter's were tuned using the Gamel herd method.
3) Water quality was predicted using existing features in the dataset, as well as new features created by combining supervised (LGBM) and unsupervised (COPOD, IForest, and CBLOF) machine learning methods after fused their outliers.
4) The performance of the proposed models was evaluated using a number of performance metrics (accuracy, precision, recall, F1 score, AUC-ROC).
5) A comparison was done between the traditional (LGBM) technique after balancing the dataset and the hybridized (LGBM) with unsupervised (COPOD, IForest, and CBLOF) machine learning approaches
6) Finally, the proposed model was compared to the previous studies.
The rest of work organized as follow: section 2 will discuss related work, section 3 and 4 will present structure of supervised and unsupervised ML, section 5 will discuss Gamel herd algorithm, section 6 will present description and analysis of the dataset, section 7 will present correlation analysis. Finally, section 8 will present Research Methodology and results discussion followed by section 9, which is conclusion.

2. Related Work

Several previous studies have validated the use of artificial intelligence algorithms for water quality prediction and analysis. Here's a summary of these studies:
Furqan Rustam et al.[10] reviewed machine learning techniques to improve the prediction of water consumption and quality, using two types of unbalanced datasets, the first from the Gaggle website to predict water quality and the second from GitHub to predict water consumption. The limitation of this study was unbalanced datasets. After tuning the hyperparameters, paper employed a variety of machine learning methods. This study improved ANN Model after adding ReLU activation function followed by dropout layer with 50% dropout rate to reduce complexity and prevent overfitting. The ANN model was constructed up of three layers: the first and second layers each had 256 nodes, while the final layer had two nodes to predict water quality and one node to predict water consumption. The findings revealed that this study obtained an accuracy range of 90% to 99%, with an enhanced ANN outperforming the other models with an accuracy of 96% for forecasting water quality and a 99% R2 score for water usage. At the same time, Nida Nasir et al.[11]introduced study which involved a variety of machine learning algorithms, including SVM, RF, LR, DT, CATBoost, XGBoost, and (MLP), as well as an ensemble of all models. To estimate water quality, the paper analyzed data obtained from different Indian towns. The CatBoost method was considered the most dependable by the study, achieving 94.5% accuracy and producing 100% accuracy after ensemble the models. Duie Tien Bui et al.[12]assessed the efficacy of four standalone (RF, REPT, M5P, and RT) and 12 hybrid data-mining algorithms(hybrids of standalones with CVPS, bagging, and RFC) in predicting WQI. The study relied on a dataset collected from northern Iran. The modeling procedure found that fecal coliform content was the most critical factor influencing WQI. The findings showed that the performance of the separate and hybrid models varied based on the differences in the input features (water samples). The features with the highest correlation coefficient had the most predictive power, and vice versa. The hybrid (BART) approach outperformed the other hybrid or standalone models, with an R2 score of 0.94%, although it may not perform as well in different datasets and environments. Mohamed Torky et al.[13]presented machine learning techniques to predict whether drinking water samples are safe or dangerous, in addition to predicting the Water Quality Index (WQI). In the field of classification, nine machine learning models were applied to classify water samples, and the results showed the overcome of the Random Forest (RF) and Light Gradient Boosting Machine (Light GBM) models over other models with an accuracy of 0.96% and 0.97%. As for regression, six models were used to predict Water Quality Index (WQI), with superiority LGBM regression models and Extra tree regression models with an accuracy of 95.5% on the rest. Fitore Muharemi et al.[14]employed time series data gathered by the General Water Company of Germany as a challenge to estimate water quality. The study used a variety of machine learning methods (logistic regression, support vector machines (SVM), linear discriminant analysis, recurrent neural network (RNN), artificial neural network (ANN), deep neural network (DNN), and long short-term memory (LSTM)), and the findings revealed that imbalanced data has a significant impact on the performance of machine learning algorithms and makes them vulnerable. As a result, the paper did not produce satisfactory findings, particularly when applying time series algorithms (DNA, RNN, and LSTM). Meanwhile, Umair Ahmed et al.[15]classified water quality (WQC) and predicted the water quality index (WQI) by applying a set of machine learning algorithms. The researchers collected dataset from several different sources for Lake Rawal in the city of Pakistan. The research relied on a number of important parameters after performing a number of preprocessing on them, such as temperature, pH, and others. The results demonstrated that gradient boosting and polynomial regression achieved best accuracy for predicting the water quality index while in water quality classification, the MLP model overcame the rest models with an accuracy of 85%. To forecast water quality, Md. Mehedi Hassan et al.[16], applied several of supervised machine learning models in India. The research relied on a dataset collected from Kaggle consisting of a number of important biometric features that indicate water quality and purity. The findings showed that MLR outperformed the other models with about 99% accuracy. At the same year, M. H. Al-Adhaileh, and F. W. Alsaade[17]employed two approaches. The first approach was to use the created Adaptive Neural Fuzzy Inference System (ANFIS) algorithm to estimate the Water Quality Index (WQI). The second is to use feed-forward neural networks (FFNN)

and K-nearest neighbors to classify water quality. The analysis was based on seven major features. Following evaluation with a variety of performance and statistical indicators, the two models produced the best results. Saber Kouadri[18]proposed two scenarios: in the first scenario, all parameters were utilized as inputs and tried to shorten the time required for WQI computation. In the second scenario, all inputs were decreased based on sensitivity analysis and aimed to illustrate the fluctuation in water quality in crucial instances where the required assessments are not available. The study employed eight artificial intelligence algorithms to forecast water quality indicators in an arid desert setting using 114 samples taken at various time intervals from six aquifers in Illizi Province, southeastern Algeria. The findings revealed that the MLR model had the highest accuracy. Afaq Juna et al.[19]predicted water quality based on data at the kaggle website after processing it, such as eliminating missing values using KNN imputer or manually. The work applied a number of traditional machine learning methods, in addition to improving the MLP model, which consists of nine layers, with 256 nodes in each layer. The model was implemented over 20 epochs and used the loss function(binary_crossentropy) with Adam Optimizer. The results showed that the improved model with KNN imputer achieved the best results with an accuracy of 0.99%. Table I summarized Related Work.

3. Supervised Machine Learning

In machine learning an algorithm is trained using data that is labeled. Each data point includes input characteristics and their corresponding output labels.The LightGBM algorithm is an example of supervised machine learning that was applied in this work.

The algorithm was presented by Ke et al.[20] and based on Decision Tree algorithm. In comparison to traditional techniques, the algorithm's design, which combines gradient-based one-sided sampling (GOSS) and exclusive feature pooling (EFB), offers high efficiency, accuracy, and regression in data classification[21]. GOSS relies on high gradients and leaves out features with low gradients. In order to minimize the amount of features, mutually incompatible features are bundled together using EFB[22]. It is characterized by:

1) It is called light because of its speed in training data.
2) Less memory consumption.
3) Reaching the best accuracy.
4) Dealing with big data.
5) Followed parallel learning
6) It reduces the cost of loss because it relies on dividing the tree into leaves and not at the depth level that used in previous Boosting algorithms.

4. Unsupervised Machine Learning

Unsupervised machine learning includes algorithms that recognize patterns and structures, within data without needing labels. Below the (COPOD, IForest, CBLOF) algorithms are an examples of unsupervised machine learning that was applied in this work.

A. *Copula-Based Outlier Detection (COPOD)*

It was introduced by Zheng Li[23], who characterized it as being motivated by copulas for modeling multivariate distributions. Copulas are mathematical functions that allow the COPOD model to distinguish marginal distributions from a random data. This offers COPOD the ability to be employed in high-dimensional datasets[24]. The method creates an empirical copula to estimate the tail probability of each data point and identify its ”extreme” level.The working steps are[25]:

1) The dataset is collected, and then preprocessed to deal with missing and abnormal values.
2)The variables in the dataset are treated as having a uniform distribution using marginal distribution functions. Use the copula function like (Gaussian, Clayton etc.) to represent the dependence structure between the converted variables.
3) The parameters of the chosen copula function are determined using maximum probability estimation or other fitting approaches.
4) The copula function constructs synthetic data points that reflect the dependence structure of the original dataset.
5) Data points that deviate significantly from the expected adoption structure are considered outliers.
6) Statistical analysis is utilized to describe the features of outliers and the causes for their anomaly.

B. *Isolation Forest (IForest)*

In 2008, Zhou Zhihua produced unsupervised IForest algorithm. It is an efficient and ensemble learning method that identifies outliers throughout the full sample space[26]. This method provides a good level of accuracy and execution efficiency. It may identify anomalous data by isolating data points that are sparse and dispersed from high density clusters. The principle of its work is[26],[27]:

1) A subset of the training data is chosen randomly.
2) iteratively creates binary trees, each branch of which is called an isolation tree(itree).
3) Each time, the feature and partition value(p) are selected at random, with the condition that the partition value (p) is within the feature value range. If feature ¡ p, then put it in the left tree otherwise put in the right tree.
4) The stopping condition for the algorithm is to reach the deepest node in the tree or isolate a single feature in the leaf node.
5) The final form is to reach an isolated forest of features.
6) Find the average path length h(d) for each feature in the isolation forest, where d is the dataset. Equation 1 is showed that.

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \tag{1}$$

Where C(n) denoted of the average of h(d) and n is the number of leaves, H(t) is the harmonic number that calculated using ln(t) + (Euler's constant= 0.5772156649) , and the anomaly score can be computed by equation 2:

$$s(d,n) = 2^{\frac{-E(h(d))}{c(n)}} \tag{2}$$

TABLE I. Summarization of Related Work

Papers	Year	Methods	Dataset	Best Results
[10]	2022	DT, RF, Extra Tree, LR, AdaBoost, CNN, LSTM, Gated Recurrent unit and improved ANN	https://www.kaggle.com/datasets/ adityakadiwal/water-potability	96% forecasting water quality 99% water Consumption
[11]	2022	SVM, LR, RF, DT, XGBoost CATBoost, and (MLP)	https://kaggle.com/anbarivan/ indian-water-quality-data	Cat boost 95% 100% Meta decision tree, Meta MLP, Meta CATBoost
[12]	2020	M5P, RF, RT, REPT (reduced error pruning tree), BA(bagging)-M5P, BA-RF, BA-RT, BA-REPT, CVPS(CV parameter selection)-M5P, CVPS-RT, CVPS-REPT, RFC-RF, RFC-M5P, RFC-RT, RFC-REPT	Private	94% BA-RT
[13]	2023	XGBoost, LightGBM, MLP, Decision Tree, ETC Classifier, GBC, RF, SVM, ANN	https://www.kaggle.com/datasets/ mssmartypants/water-quality	Classification 0.96% RF 0.97% LGBM Regression 95.5% LGBM and DT
[14]	2019	LR, linear discriminant analysis, SVM, ANN, recurrent neural network (RNN), deep neural network (DNN), LSTM	Private	0.36% SVM
[15]	2019	Multiple Linear Regression, Ridge Regression, Polynomial Regression, Lasso Regression, Elastic Net Regression, RF, SVM, Gaussian Naïve Bayes, MLP, LR, Stochastic gradient descent, K Nearest Neighbor, DT, Bagging Classifier	http://www.pcrwr.gov.pk/	Classification (MLP) 85% accuracy Regression Gradient Boosting 7.2011 MSE polynomial regression 12.7307 MSE
[16]	2021	ANN, SVM, bagged tree (BT) models, RF, multinomial logistic regression (MLR)	Indian dataset pollution https://www.kaggle.com/code/ anbarivan/indian-water-quality-data	MLR 100%
[17]	2021	Adaptive Neural Fuzzy Inference System (ANFIS), feed-forward neural networks (FFNN), K-nearest neighbors	Indian water quality data (kaggle.com)	ANFIS 92.39%accuracy FFNN 100% accuracy KNN 80.63% accuracy
[18]	2021	Multi linear regression (MLR), ANN, SVM, M5P tree, Random subspace(RSS), RF, Additive regression (AR), and Locally weighted linear regression (LWLR)	Private	MLR 100%
[19]	2022	LR,SVC, DT, RF,KNN,Stochastic Gradient Decent Classifier (SGDC), and XGBoost, MLP-9	https://www.kaggle.com/datasets /adityakadiwal/water-potability	MLP-9 0.9990% accuracy

Where E(h(d)) is the average of all h(d).
7) If the value of S(x, n) is near to one, it indicates that the data is more probable to be anomalous; If S(x, n) is near to zero, it indicates normal data.

C. *Cluster-based Local Outlier Factor (CBLOF)*

CBLOF was suggested by He, Xu, and Deng[28]. It describes anomalies as a result of local distances to neighboring clusters and the overall size of the specific clusters where the data point belongs. It first divides data points into large and small clusters. Data points within a small cluster close to a nearby larger cluster are recognized as outliers. The local outliers could not represent a single point, instead being a tiny group of separated points. CBLOF considers both the distance between a data point and the closest cluster as well as the size of the cluster to which a data point belongs.The steps of CBLOF procedure are[29]:
1) A data point is given to exactly one cluster using K-means, which is a good clustering algorithm.
2) Clusters are ranked from large to small based on their size, and over time, data counts are calculated. The ”large” clusters keep up to 90% of the data, while the ”small” clusters keep the remaining 10%.
3) Finds a data point’s distance towards the centroid and outlier score using two rules.Firstly,the distance between data points in a large cluster is measured from the cluster’s centroid. The distance is multiplied by the number of data points in the cluster to determine the outlier score.The second rule ,if a data point is in the smallest cluster, the distance is calculated using the centroid of the next large cluster. The calculation of the outlier score involves multiplying the distance by the amount of data present in the small cluster containing the corresponding data point.

5. Camel Herd Algorithm (CHA)

It is an optimal intelligent algorithm that relies on the collective behavior of camels’ herd in the desert to solve complex problems. Its goal is to reach various solutions by exploring multiple paths and starting from different points. It also avoids falling into local optimum and reaching the global one. The basic idea behind how it operates is that, based on the amount of humidity in the air, the leader of each camel herd directs the group toward water and food. In order to find the best multiple solutions and the fastest convergence, the herd leader leads the herd as they investigate the solution regions utilizing neighborhood approaches and humidity levels[30]. The algorithm takes into account the food and humidity content, as well as the herd’s number.

It also identifies the herd leader, whose responsibility it is to guide the herd down numerous routes in search of the best solution, beginning at various positions[31]. The camel herd procedure is revealed in Algorithm 1[31]

```
Algorithm1: Pseudocode of Camel Herd Algorithm
Input: No. of camel (M), no. of herds (H), max_Humidity (maxH)
Output: best short path
Begin
For i = 1 to $H_i$ Do
  //Choose leader ($LH_i$ )from the herd by using selection approach
End for
Repeat
        For i := 1 to $H_i$ Do
        b := 1
        Initialize (Humidity)
        For j := 1 to length ($LH_i$)
              For each solution Do
                    Establish random neighbors (RN) of $LH_i$ // RN denote no. of
                  camel except leader
             For z := 1 to RN Do
                  (best neighbors) $BN_Z$= $BN_Z$ * 1\ Humidity
                   $BN_Z$= $LH_i$ - $BN_{Z\backslash}$ dis ($LH_i$ ,$BN_Z$ )
             End for
           $LH_i$ [j] [b+1] = $LH_i$ [j] [b] + $BN_Z$
             End for
      Update Humidity
      End for
   End for
Until achieve goal or maximum Humidity
End
```

Figure 1. Pseudocode of (CHA)

6. Description and Analysis of the Dataset

The datasets for this study were acquired from well-known site such as Kaggle[32]. This dataset has 8000 items and 21 features. All of the features in the water quality dataset are real numbers except target class which is integer. Table II has a thorough overview of the dataset's attributes. This study aimed to clarify the distribution of 20 features utilized in water quality prediction. Figure 2 depicts the various distributions of features after cleaning and deleting missing data.

TABLE II. Explanation of the water quality dataset's features

No.	Attribute	Explanation	Range per Liter
1	aluminum	Water is dangerous if higher than 2.8	0–5.05
2	ammonia	Water is dangerous if higher than 32.5	0.08–29.8
3	arsenic	Water is dangerous if higher than 0.01	0–1.05
4	barium	Water is dangerous if higher than 2	0–4.94
5	cadmium	Water is dangerous if higher than 0.005	0–0.13
6	chloramine	Water is dangerous if higher than 4	0–8.68
7	chromium	Water is dangerous if higher than 0.1	0–0.9
8	copper	Water is dangerous if higher than 1.3	0–2
9	flouride	Water is dangerous if higher than 1.5	0–1.5
10	bacteria	Water is dangerous if higher than 0	0–1
11	viruses	Water is dangerous if higher than 0	0–1
12	Lead	Water is dangerous if higher than 0.015	0–0.2
13	nitrates	Water is dangerous if higher than 10	0–19.8
14	nitrites	Water is dangerous if higher than 1	0–2.93
15	mercury	Water is dangerous if higher than 0.002	0- 0.1
16	perchlorate	Water is dangerous if higher than 56	0 – 60
17	radium	Water is dangerous if higher than 5	0–7.99
18	selenium	Water is dangerous if higher than 0.5	0 – 0.1
19	silver	Water is dangerous if higher than 0.1	0–0.5
20	uranium	Water is dangerous if higher than 0.3	0–0.9
21	is _safe	Target Class	not safe=0 , safe=1

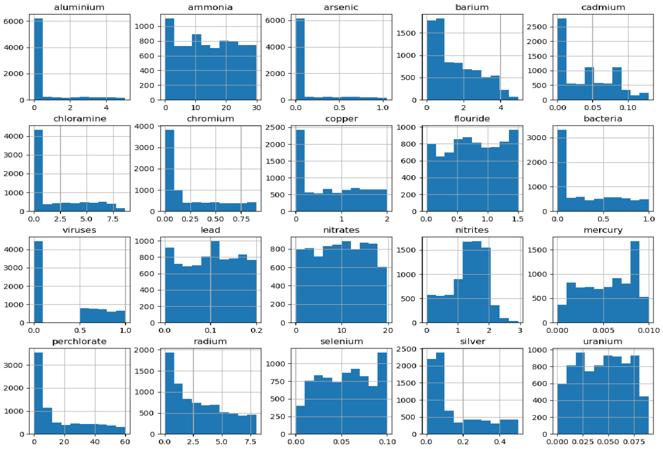


Figure 2. Distribution of water including chemicals in the dataset

In addition, Table III displays a range of statistical values for the input features in the dataset. Also,it illustrated that the count of parameters in the dataset is equal (7996.000000). The minimum value is (-0.08000), which belongs to ammonia. perchlorate also achieved the maximum value and height standard deviation of (60.010000) and (17.688827) respectively.

7. Correlation Analysis (CA)

A table that shows the correlation coefficients for several attributes is called a correlation matrix. Every conceivable value pair in the table is represented by the matrix[16]. The dataset's attribute correlation matrix with the output and each other is shown in Figure3. The graph reveals that aluminum and chloramine have a greater influence on forecasting water quality, but cadmium and arsenic have the least.

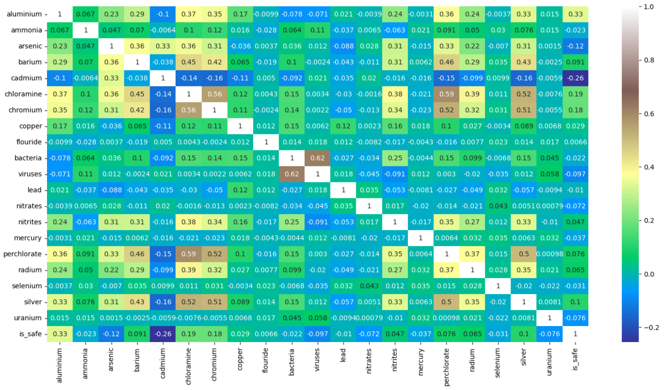


Figure 3. Correlation Matrix of Dataset

8. Research Methodology and Approach

A. *Research Requirement*

1) Environmental Requirement.

TABLE III. Statistical Metric on Dataset

Material	count	mean	Standard deviation	min	25%	50%	75%	max
aluminium	7996.000000	0.666396	1.265323	0.0000	0.040000	0.070000	0.280000	5.050000
ammonia	7996.000000	14.278212	8.878930	-0.08000	6.577500	14.130000	22.132500	29.840000
arsenic	7996.000000	0.181477	0.252832	0.00000	0.030000	0.050000	0.100000	1.050000
barium	7996.000000	1.567928	1.216227	0.00000	0.560000	1.190000	2.482500	4.940000
cadmium	7996.000000	0.042803	0.036049	0.0000	0.008000	0.040000	0.070000	0.130000
chloramine	7996.000000	2.177589	2.567210	0.000000	0.100000	0.530000	4.240000	8.680000
chromium	7996.000000	0.247300	0.270663	0.000000	0.050000	0.000000	0.440000	0.900000
copper	7996.000000	0.805940	0.653595	0.000000	0.090000	0.750000	1.390000	2.000000
flouride	7996.000000	0 0.771648	0.435423	0.000000	0.407500	0.770000	1.160000	1.500000
bacteria	7996.000000	0.319714	0.329497	0.000000	0.000000	0.220000	0.610000	1.000000
viruses	7996.000000	0 0.328706	0.378113	0.000000	0.002000	0.008000	0.700000	1.000000
lead	7996.000000	0.099431	0.058169	0.000000	0.048000	0.102000	0.151000	0.200000
nitrates	7996.000000	9.819250	5.541977	0.000000	5.000000	9.930000	14.610000	19.830000
nitrites	7996.000000	1.329846	0.573271	0.000000	1.000000	1.420000	1.760000	2.930000
mercury	7996.000000	0.005193	0.002967	0.000000	0.003000	0.005000	0.008000	0.010000
perchlorate	7996.000000	16.465266	17.688827	0.000000	2.170000	7.745000	29.487500	60.010000
radium	7996.000000	2.920106	2.322805	0.000000	0.820000	2.410000	4.670000	7.990000
selenium	7996.000000	0.049684	0.028773	0.000000	0.020000	0.050000	0.070000	0.100000
silver	7996.000000	0.147811	0.143569	0.000000	0.040000	0.080000	0.240000	0.500000
Uranium	7996.000000	0.044672	0.026906	0.000000	0.020000	0.050000	0.070000	0.090000
Is_safe	7996.000000	0.114057	0.317900	0.000000	0.000000	0.000000	0.000000	1.000000

- windows OS.
- Anaconda includes Jupyter notebook tools for the Python programming language.
  2) Functional Requirement.
- A group of libraries was used in the Python language to implement the desired goals, which are sklearn, pandas, NumPy, matplotlib, lightgbm, niapy and seaborn.
- To achieve paper goals, a hybridization of supervised (LGBM) and unsupervised (COPOD, IForest, CBLOF) machine learning algorithms was done.

B. Proposed Methodology

The proposed methodology of the water quality prediction model consists of five phases: Preprocessing phase, Unsupervised ML phase, Tuning phase, Prediction using supervised ML phase and finally Performance evaluation phase. The framework of the proposed methodology can be simply described in Figure 4.

1) Pre-processing Phase

Pre-processing is essential for improving the quality of data analysis. It refers to the act of acquiring and manipulating numerous data components in order to produce usable and relevant information. Pre-processing phase included Data cleaning, Data normalization, Data splitting, and lastly resampling training data.

- Data Cleaning
  Cleaning data was performed by deleting records that contained incomplete data.
- Data Normalization
  Normalization is a technique for standardizing attribute values in a dataset by placing data in a predefined range between 0 and 1 without affecting the underlying distribution. It ensures that the data keeps its original shape when scaled to a defined range. Equation (3) calculated the feature's normalization on a scale of 0 to 1 [33].

  $$f_{scaled} = \frac{f - f_{min}}{f_{max} - f_{min}} \tag{3}$$

  Where f$_{max}$ represents the feature's maximum value and f$_{min}$ refers the minimum value. This is done by using MinMaxScaler function (f scaled).
- Data Splitting
  The data was divided into two groups, with 70% going to training and 30% going to the testing technique.
- Resampling Training Dataset
  Unbalanced datasets have unequal categories, one with more samples than the other. Classifiers may perform effectively in the majority class but poorly on the minority due to their greater effect. Unbalanced datasets often need to be resampled to achieve a more even distribution of class states[34],[35]. SMOTE sampling, an adaptive oversampling approach, has been applied to process the raw dataset for guaranteeing high accuracy of the training data. The SMOTE approach efficiently motivates the minority class to become broader. Oversampling the minority class is a technique for dealing with unbalanced datasets. Duplicating samples in the minority class is the simplest solution, but these examples add no new information to the model[36].

2) Unsupervised ML Phase

At this phase, three models of unsupervised machine learning were used, and each model utilized features without labels (outcome). The function of models was to discover anomaly score for each data point(outlier), which is added and fused as an additional feature for using in the prediction algorithm by LGBM. Finally, a hybrid model was pro-

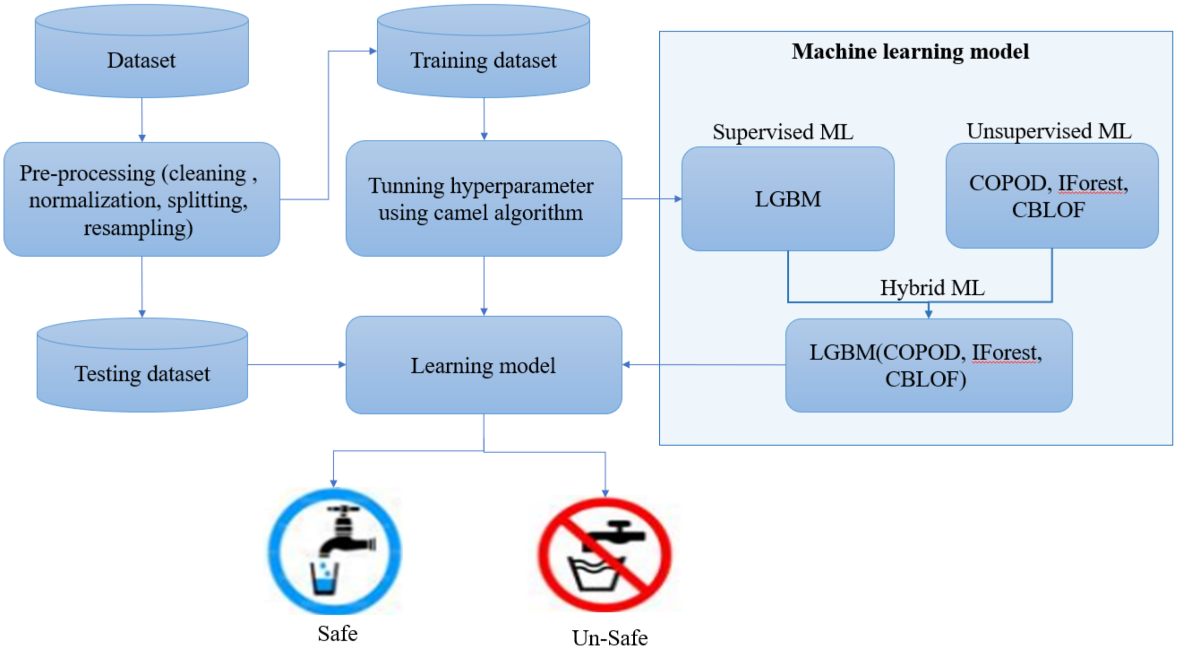


Figure 4. Methodology Framework

posed called (HLGBM+Fusion CIC) that combines supervised ML(LGBM) and unsupervised ML(COPOD, IForest, CBLOF) after fusing their outliers. that used the results of fusion unsupervised algorithms (COPOD, IForest, and CBLOF) as input for the LGBM algorithm.

3) Tuning Hyperparameter Phase

Choosing the correct hyperparameters has a significant impact on the effectiveness of the prediction model, and also allows for a more optimal solution with a better level of accuracy, but it is a difficult matter to achieve. So, swarm intelligent algorithms have demonstrated their capacity to perform such jobs[34]. The camel herd algorithm was applied for tuning the hyperparameters of LGBM algorithm, and the Table IV showed the best one.

TABLE IV. Hyper-Parameter of Models

Models	Hyperparameters
LGBM	num_leaves=141, n_estimators=196
LGBM + COPOD	num_leaves=46, n_estimators=352
LGBM + IForest	num_leaves=44, n_estimators=333
LGBM + CBLOF	num_leaves=81, n_estimators=242
HLGBM+Fusion CIC	num_leaves=46, n_estimators=352

4) Prediction using Supervised ML Phase

At this stage, the water quality prediction process is carried out after pre-processing the dataset and tuning the hyperparameters of the LGBM algorithm.

5) Performance Evaluation

After designing the model, its performance was evaluated using multiple metrics, including ROC AUC, precision, recall, f1 score, and accuracy. AUC-ROC is a classification metric that measures how effectively a classifier can distinguish between classes at different thresholds. AUC-ROC illustrates the trade-off within specificity and sensitivity in tests that produce numerical results rather than a binary positive or negative outcome. The AUC-ROC (decision thresholds) determines the optimum cut-off for both sensitivity and specificity. Accuracy represents categorization task performance and counts the number of accurately estimated examples across all data samples. Furthermore, Recall is an appropriate statistic for identifying model faults as well as how accurately the model recognizes actual ”safe” and ”non safe” occurrences. Precision refers to the percentage of positively (either ”safe” or ”non safe”) identifies that have been correct. Precision measures quality, whereas recall measures quantity. F1 score is a statistic that aims to find a balance between precision and recall. These metrics are defined in 4,5,6,7 equations as follows :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

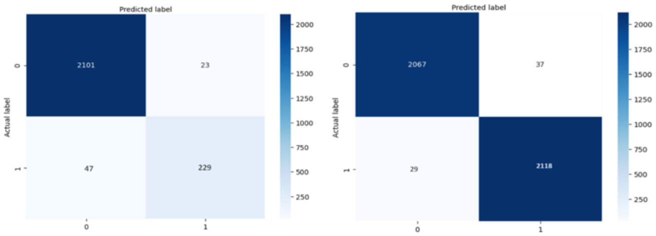
TP, FP, TN, and FN represent True Positive, False Positive, True Negative, and False Negative, respectively. They range from zero to one and used to determine the ML model that performs better to identify ”safe” and ”nonsafe” instances[37].

C. Result and Discussion

The results of LGBM model were evaluated before and after the SMOTE process, as shown in Table V. The previous table shows that the results of applying LGBM model on data after SMOTE are better than before applying SMOTE, because balanced data ensures that the LGBM model is not biased. Figure 5 displays the confusion matrix for the LGBM model before and after SMOTE, and Figure 6 shows the height of the AUC curve after oversampling in comparison with before.

TABLE V. Results of LGBM Model before and after the SMOTE

Evaluation metrics	Before SMOTE	After SMOTE
AUC	0.909	0.984
Precision	0.909	0.983
Recall	0.830	0.986
F1-score	0.867	0.985
Accuracy	0.971	0.984

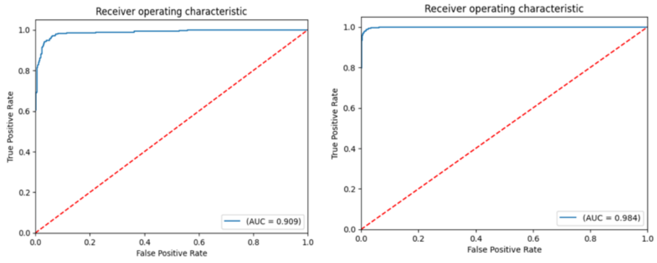


Before SMOTE After SMOTE

Figure 5. Confusion Matrix of LGBM Model Before and After SMOTE

After applying the SMOTE algorithm to the original training dataset, tuning of the hyperparameters of the LGBM algorithm was performed using Gamel herd algorithm. Next, the LGBM algorithm was hybridized with the outliers generated by the COPOD algorithm. Similarly, the same procedure was repeated independently once on the IForest algorithm and once on the CBLOF algorithm. Finally, the results of the three unsupervised algorithms (COPOD, IForest, and CBLOF) were fused as input to the LGBM algorithm.

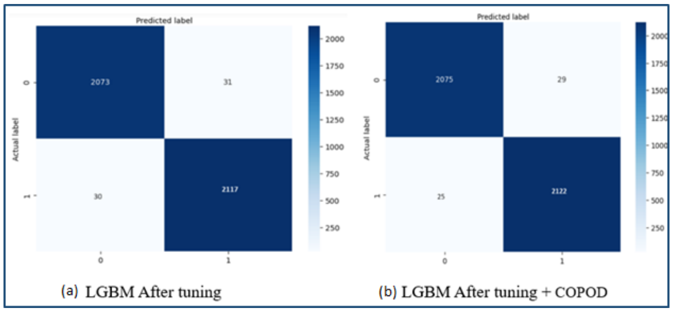
Table VI represents the performance evaluation results for all previous models, which shows that the(LGBM+IForest) model overcome the (LGBM+COPOD) model and which shows that the (LGBM+CBLOF) model overcome the (LGBM+IForest) model, finally proposed model (HLGBM+Fusion CIC) superior on the three previous models (COPOD, IForest, CBLOF).



Before SMOTE After SMOTE

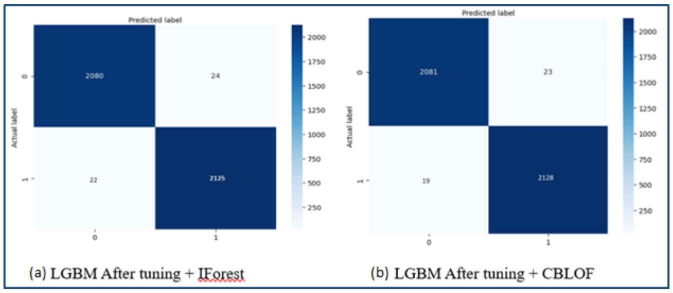
Figure 6. AUC of LGBM Model Before and After SMOTE

Figure 7,8,9 depicts the results of the confusion matrix for all applied models, demonstrating that the proposed model (HLGBM+Fusion CIC) overcome the other models.



(a) LGBM After tuning (b) LGBM After tuning + COPOD

Figure 7. (a) confusion matrix of LGBM After tuning. (b)LGBM After(tuning+COPOD) Model



(a) LGBM After tuning + IForest (b) LGBM After tuning + CBLOF

Figure 8. (a)confusion matrix of LGBM After tuning+IForest. (b)LGBM After tuning+COPODn

The results of proposed model (HLGBM+Fusion CIC) was

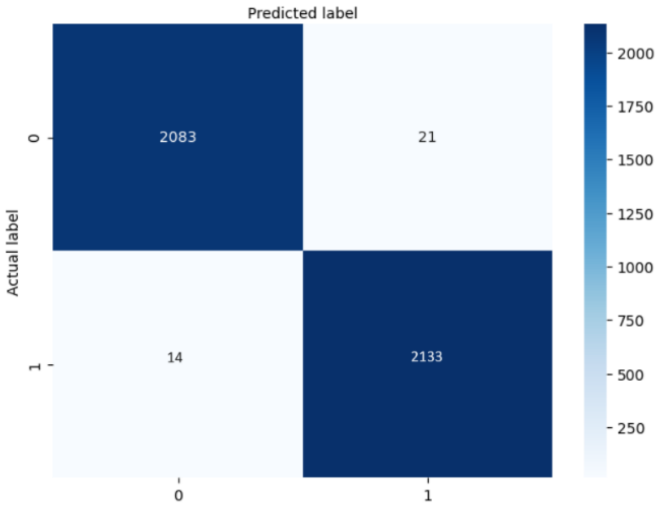


Figure 9. confusion matrix of (HLGBM+Fusion CIC) Model

TABLE VI. Performance Evaluation Results

Evaluation metrics	LGBM Before tuning	LGBM After tuning	After Tuning Hybrid (LGBM+ COPOD)	Hybrid (LGBM+ IForest)	Hybrid (LGBM+ CBLOF)	HLGBM +Fusion CIC
AUC	0.984	0.986	0.987	0.989	0.990	0.992
Precision	0.983	0.986	0.987	0.990	0.989	0.990
Recall	0.986	0.986	0.988	0.990	0.991	0.993
F1-score	0.985	0.986	0.987	0.989	0.990	0.992
Accuracy	0.984	0.986	0.987	0.989	0.990	0.992

compared to the highest result mentioned in a related work conducted by Furqan Rustam et [10]. Table 7 shows that the result obtained from the proposed model was better compared to Furqan Rustam et [10].

TABLE VII. Comparison with Related Work[10]

Paper	Accuracy	Precision	Recall	F1 score
Furqan Rustam et. [10]	0.96	0.91	0.87	0.89
HLGBM+Fusion CIC	**0.99**	**0.99**	**0.99**	**0.99**

9. Conclusion

Precise monitoring of changes in water quality is crucial for delivering drinking water. Conventional techniques like computing the water quality index (WQI) can be time intensive and prone to mistakes. The global issues of water scarcity and pollution underscore the need to automate water suitability assessments. Artificial Intelligence (AI) presents opportunities, for enhancing the analysis and forecasting of water quality.AI approaches can cut down expenses, help ensure adherence to water quality regulations and establishing monitoring systems is essential, for sustainable friendly water resource management. This study focused on predicting water quality. In order to achieve this, an evaluation and comparison of different models was conducted to hybridize supervised learning with unsupervised learning after using the SMOTE process and using swarm optimization to develop the model and obtain the best prediction results. The results of show that the model after SMOTE, tuning training dataset, fusion unsupervised algorithms (COPOD, IForest, CBLOF) and hybridizing them with the LGMB algorithm overcome the other models with accuracy, AUC and f1 score are 99.2%, a precision is 99% and a recall is 99.3%.

Furthermore, these results have important implications for learning how to develop a new model that combines the features of supervised and unsupervised learning algorithms to achieve multi-model learning and high-representation prediction, including whether they are suitable for human consumption, agricultural irrigation, or other industrial or environmental applications.

References

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

[2] J. Wolfram, S. Stehle, S. Bub, L. L. Petschick, and R. Schulz, "Water quality and ecological risks in european surface waters–monitoring improves while water quality decreases," *Environment International*, vol. 152, p. 106479, 2021.

[3] N. Kedia, "Water quality monitoring for rural areas-a sensor cloud based economical project," in *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, 2015, pp. 50–54.

[4] K. Abirami, P. C. Radhakrishna, and M. A. Venkatesan, "Water quality analysis and prediction using machine learning," in *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2023, pp. 241–245.

[5] O. Alshaltone, N. Nasir, F. Barneih, E. A. Majali, and A. Al-Shammaa, "Multi sensing platform for real time water monitoring using electromagnetic sensor," in *2021 14th international conference on developments in eSystems engineering (DeSE)*. IEEE, 2021, pp. 174–179.

[6] A. Y. Sun and B. R. Scanlon, "How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions," *Environmental Research Letters*, vol. 14, no. 7, p. 073001, 2019.

[7] N. Nasir, O. Al Bashier, A. A. Murad, and M. Al Ahmad, "Optical detection of dissolved solids in water samples," in *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. IEEE, 2018, pp. 1–6.

[8] N. Nasir, M. Al Ahmad, and A. A. Murad, "Capacitive detection and quantification of water suspended solids," in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*. IEEE, 2019, pp. 1–5.

[9] M. Ehteram, S. Ghotbi, O. Kisi, A. Najah Ahmed, G. Hayder, C. Ming Fai, M. Krishnan, H. Abdulmohsin Afan, and A. EL-Shafie, "Investigation on the potential to integrate different artificial intelligence models with metaheuristic algorithms for improving river suspended sediment predictions," *Applied Sciences*, vol. 9, no. 19, p. 4149, 2019.

[10] F. Rustam, A. Ishaq, S. T. Kokab, I. de la Torre Diez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "An artificial neural network model for water quality and water consumption prediction," *Water*, vol. 14, no. 21, p. 3359, 2022.

[11] N. Nasir, A. Kansal, O. Alshaltone, F. Barneih, M. Sameer, A. Shanableh, and A. Al-Shamma'a, "Water quality classification using machine learning algorithms," *Journal of Water Process Engineering*, vol. 48, p. 102920, 2022.

[12] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, "Improving prediction of water quality indices using novel hybrid

machine-learning algorithms," *Science of the Total Environment*, vol. 721, p. 137612, 2020.

[13] M. Torky, A. Bakhiet, M. Bakrey, A. A. Ismail, and A. I. E. Seddawy, "Recognizing safe drinking water and predicting water quality index using machine learning framework," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023.

[14] F. Muharemi, D. Logofătu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set," *Journal of information and telecommunication*, vol. 3, no. 3, pp. 294–307, 2019.

[15] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, p. 2210, 2019.

[16] M. M. Hassan, M. M. Hassan, L. Akter, M. M. Rahman, S. Zaman, K. M. Hasib, N. Jahan, R. N. Smrity, J. Farhana, M. Raihan *et al.*, "Efficient prediction of water quality index (wqi) using machine learning algorithms," *Human-Centric Intelligent Systems*, vol. 1, no. 3, pp. 86–97, 2021.

[17] M. Hmoud Al-Adhaileh and F. Waselallah Alsaade, "Modelling and prediction of water quality by using artificial intelligence," *Sustainability*, vol. 13, no. 8, p. 4259, 2021.

[18] S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S. Kateb, "Performance of machine learning methods in predicting water quality index based on irregular data set: application on illizi region (algerian southeast)," *Applied Water Science*, vol. 11, no. 12, p. 190, 2021.

[19] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. Eshmawi, A. Mohamed, and I. Ashraf, "Water quality prediction using knn imputer and multilayer perceptron," *Water*, vol. 14, no. 17, p. 2592, 2022.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[21] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, lightgbm, and xgboost regression," *Automation in Construction*, vol. 129, p. 103827, 2021.

[22] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (lightgbm)," *Diagnostics*, vol. 11, no. 9, p. 1714, 2021.

[23] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "Copod: copula-based outlier detection," in *2020 IEEE international conference on data mining (ICDM)*. IEEE, 2020, pp. 1118–1123.

[24] R. K. Kennedy, Z. Salekshahrezaee, F. Villanustre, and T. M. Khoshgoftaar, "Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning," *Journal of Big Data*, vol. 10, no. 1, p. 106, 2023.

[25] X. Sun, Y. Wang, and Z. Shi, "Insider threat detection using an unsupervised learning method: Copod," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*. IEEE, 2021, pp. 749–754.

[26] W. Zhang and H. Fan, "Application of isolated forest algorithm in deep learning change detection of high resolution remote sensing image," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2020, pp. 753–756.

[27] D. Xu, Y. Wang, Y. Meng, and Z. Zhang, "An improved data anomaly detection method based on isolation forest," in *2017 10th international symposium on computational intelligence and design (ISCID)*, vol. 2. IEEE, 2017, pp. 287–291.

[28] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern recognition letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[29] P. Kasture and J. Gadge, "Cluster based outlier detection," *International Journal of Computer Applications*, vol. 58, no. 10, 2012.

[30] A. T. S. Al-Obaidi, H. S. Abdullah *et al.*, "Camel herds algorithm: A new swarm intelligent algorithm to solve optimization problems," *International Journal on Perceptive and Cognitive Computing*, vol. 3, no. 1, 2017.

[31] Z. O. Ahmed, A. T. Sadiq, and H. S. Abdullah, "Solving the traveling salesman's problem using camels herd algorithm," in *2019 2nd Scientific Conference of Computer Sciences (SCCS)*. IEEE, 2019, pp. 1–5.

[32] A. Kadiwal, "Kaggle dataset," Accessed: Feb. 20, 2024. [Online]. Available: https://www.kaggle.com/datasets/mssmartypants/water-quality/data.

[33] A. Gökhan, C. O. Güzeller, and M. T. Eser, "The effect of the normalization method used in different sample sizes on the success of artificial neural network model," *International Journal of Assessment Tools in Education*, vol. 6, no. 2, pp. 170–192, 2019.

[34] G. A. Al-Talib *et al.*, "Prediction of covid-19 effect on patients during six month after recovery, by using ai algorithm," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 17, no. 1, pp. 53–61, 2023.

[35] H. A. Aldabagh and G. A. Altalib, "Predicting the effect of covid-19 on physical activity of survivors using gso and hybrid intelligent model," in *2022 2nd International Conference on Advances in Engineering Science and Technology (AEST)*. IEEE, 2022, pp. 739–745.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[37] J. Tohka and M. Van Gils, "Evaluation of machine learning algorithms for health and wellness applications: A tutorial," *Computers in Biology and Medicine*, vol. 132, p. 104324, 2021.

Hanan Anas Al-Dabbagh She is an employee at the General Directorate of Education in Nineveh Governorate, Mosul, Iraq. She obtained a master's degree in computer science in 2014 from the University of Mosul/Iraq, after which she obtained a doctorate degree in computer science in the field of artificial intelligence from the University of Mosul/Iraq in 2023.

Ruba T. Ibrahim She is a faculty member in the Department of Computer Science, University of Mosul, Iraq. She obtained a master's degree and PhD. degree in computer science in the field of artificial intelligence from the University of Mosul/Iraq in 2012 and 2023, respectively. Her current research area covers digital image processing, computer vision and artificial intelligence.