International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Water Quality Prediction Using Combined Model of Convolutional Neural Network and Long Short-Term Memory

Raju Amireddy*[1], Dileep Pulugu[2]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: In recent decades, the quality of water is affected due to contamination and pollution of water bodies. The existing techniques face issues related to poor water quality prediction with less accuracy. This research focusses on an effective water quality classification framework by predicting it as safe or unsafe. Initially, the data is acquisitioned from Kaggle and it is subjected to the stage of pre-processing using standard scalar. The pre-processed output is provided for feature selection takes place using dynamic Particle Swarm Optimization (PSO). After this, the classification is performed using combined method of Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM). The CNN acts as front end of model which processes the input features based on non-linear characteristics and LSTM acts as the back end which receives the abstracted data that helps in predicting the water quality as safe or unsafe. The outcome through the experimental validation shows that the suggested framework achieves prediction accuracy of 99.99% which is comparably higher than ensemble model with classification accuracy of 98.1%.

Keywords: *Classification, Combined model, Dynamic particle swarm optimization, Prediction, Water quality*

1. Introduction

Water is considered as the significant natural resources which plays a vital role in day-to-day life of living organisms. Moreover, water is widely utilized in drinking, cooking, washing and irrigation and so on [1-3]. The water with a good quality refers to the quality which is consumable by humans, animals and suited for agricultural purposes. There are numerous studies developed for this in the recent years to enhance water quality by means of physical, chemical and biological processes [4,5]. The quality of the water is accessed based on Total Dissolved Solids (TDS) and the TDS is comprised with various type of inorganic salts such as calcium, magnesium, potassium, nitrate, sodium and chloride. The water released from industries and factories are mixed with the natural water sources such as river, lake, pond [6]. However, taking an appropriate action by building a partition helps to prevent the water from being affected. In case of the waste water from the industries are fed into the ground, it affects the land surface as well as the ground level water [7,8]. Moreover, the urban waste water based on dye and fabrics contributes to a great extent in polluting the source of waters. So, the water needs to meet specified criteria to classify itself as potable water. More number of processes have been followed by organized international bodies and government to diminish the efficiency reckless water utilization and protecting it from being polluted.

[1] *Department of Computer Science and Engineering (CSE), Malla Reddy University, Hyderabad, India*
[2] *Department of Computer Science and Engineering (CSE), Malla Reddy College of Engineering and Technology, Hyderabad, India*
ORCID ID : 0000-0003-2519-7113
** Corresponding Author Email: rajucse531@gmail.com*

The Water Quality Index (WQI) is a subjective method, widely utilized to access the suitability of the surface and ground water. The assessment by means of subjective approach assign weights to the parameters for evaluating WQI score [9,10]. The evaluation of water quality by means of subjective approach considers physiochemical parameters. However, measuring the quality of groundwater sometimes gives inappropriate results due to its contamination [11]. The traditional methods based on water quality prediction are time consuming because it needs an expert to decide the quality of water. So, an automated process helps in adopting a data driven approach using Machine Learning (ML) or Deep Learning (DL) algorithms. The process of monitoring and assessing possible concerns using ML and DL approaches acts as an efficient modelling approach to estimate WQI and WQC. However, ML techniques are least suitable for predicting the non-linear water quality parameters [12,13]. So, the DL techniques are vastly utilized in predicting the water quality due to its flexibility in capturing non-linear parameters related to water quality prediction. Nonetheless, the existing prediction approaches are prone to adapt themselves with the varying environmental conditions that hinder the effect of assessing water quality [14,15]. So, this research develops an advanced deep learning approach to enhance the accuracy of predicting the water quality. The major findings of this study are as follows:

1. The pre-processed output using standard scalar is fed into feature selection which is based on dynamic PSO. The dynamic PSO is developed based on exponential decay and particle elimination that selects the best feature sets.

2. This research develops a combined model integrating CNN and LSTM in which, CNN acts as the frontend that processes the input features based on non-linear characteristics, while the LSTM acts as the back end which receives the abstracted data that helps in predicting the water quality as safe or unsafe.

The rest of manuscript is structured in the following manner: The recent researches based on predicting the water quality is presented in Section 2. The proposed framework to enhance the accuracy of water quality prediction is delved in Section 3, while Section 4 presents the experimental outcomes achieved while evaluating the suggested approach, and the concluding phase of the research is described in Section 5 of the manuscript.

2. Related works

This section provides recent studies on basis of water quality prediction using different approaches along with its advantages and limitations.

Shams [16] introduced water quality prediction using machine learning based on grid search method. The raw data was pre-processed with the help of data manipulation and normalization. The grid search was utilized in the process of optimizing and parameter tuning for four regression models. The evaluation was performed using cross validation approach and the water quality assessment was performed to evaluate the quality of water as good or bad. However, the introduced method required prediction for a time period.

Dritsas and Trigka [17] introduced an efficient data driven machine learning models to predict the quality of water. A supervised machine learning method was developed from a labelled training data to detect the suitability of water. The physiochemical and the microbiological parameters were provided as input features which helped to present the quality of water as safe or unsafe. The efficiency of the machine learning models was assessed with and without balancing the classes using Synthetic Minority Oversampling Technique (SMOTE). However, the use of machine learning was only applicable with smaller datasets.

Shah [18] developed an environmental assessment-based underground water quality prediction using consistent bigdata through hyperparameters optimized machine learning models. The PCA was utilized to choose the significant input parameter set and removed the least influential for exhibiting total dissolved solids and oxygen levels in water. For better prediction of the particle swarm optimization was integrated with gene expression programming and feed-forward neural network for structure formation and hyperparameter tuning. This developed model was utilized for effectively solving the hyperparameter setting problems. Due to the drawback of conventional techniques, the developed model handled only linear and stationary datasets.

Bhardwaj [19] implemented a machine learning and IoT based framework for water quality calculation and device component monitoring. The ML techniques investigated the water quality, device management and monitoring approaches. The developed model was involved to perform proactive monitoring, alongside warning of devices and systems. The developed model had advantages like efficiency, cost effective, low time consuming and effort. Nevertheless, this developed model struggled to capture long-term patterns in water quality data which are often crucial for accurate predictions.

Singha [20] introduced a prediction of groundwater quality by effective machine learning techniques like RF, XGBoost and ANN. A total of 226 samples were composed and several physicochemical parameters were estimated to calculate the entropy weight-based water quantity in agriculture intensive area. RF, XGBoost, and ANN processed large-scale data without feature selection, offered high training speed, noise immunity, and accuracy in quantity prediction. The developed model enhanced the performance by decreasing the error and computational rate in the execution process. Nevertheless, the suggested model does not consider the feature selection phase which utilized inappropriate features for classification.

Mehedi Hassam. Md [21] developed the Machine Learning (ML) methods like Random Forest (RF), Neural Network (NN), Multinominal Logistic Regression (MLP), Support Vector Machine (SVM) and Bagged Tree Model (BTM) for classifying the water quality in different locations. The water quality was represented through features like Total Coliform (TC), Dissolved Oxygen (DO), Nitrate pH, Biological Oxygen Demand (BOD) and Electric Conductivity (EC). The developed method effectively selected the feature importance of model. However, the developed method has not normalized the data.

3. Proposed methodology

This research introduces an effective framework to predict the quality of water. The major objective of this research is to utilize advanced deep learning algorithm to analyse quality of water as safe and unsafe. The proposed framework undergoes four stages such as data collection, pre-processing, feature selection and classification. The block diagram for the workflow involved in proposed algorithm is presented in Fig. 1.

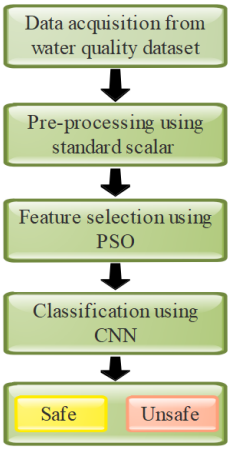

Fig. 1. Workflow of the process involved in water quality prediction

3.1. Data acquisition

The collection of raw data is the initial stage in water quality prediction which takes place using the data from water quality dataset [22] in Kaggle. The data is a set of data obtained from urban environment and is comprised with attributes of numeric variables. There are about 21 attributes present in the dataset, the first 20 attributes are based on the ingredients present in the water along with the ranges present in water. The final attribute is used to predict the classes as 0 or 1. The class 1 refers to the water quality is safe and the class 0 refers to the water quality being unsafe. The list of attributes along with their ranges are present in the water quality dataset have Aluminium (2.8), Ammonia (32.5), Arsenic (0.01), Barium (2), Cadmium (0.05), Chloramine (4), Chromium (0.1), copper (1.3), Fluoride (1.5), Bacteria (0), Virus (0), Lead (0.015), Nitrates (10), Nitrates (10), Nitrites (1), Mercury (0.002), Perchlorate (56), Radium (5), Selenium (0.5), Silver (0.1), Uranium (0.3). If the range of elements exceeds the fore mentioned ranges, then it is considered as dangerous.

3.2. Pre-processing

Pre-processing is performed to make the data suitable for further process involved in water quality prediction. Next to data acquisition, data pre-processing is accomplished using standard scalar. Basically, the standard scalar is utilized to standardize the raw data which helps the machine learning techniques to perform better by distributing the features in a same scale. For a reliable approach, the data should be standardized within the range of 0-1. The evaluation of standard scalar takes place based on the (1).

$$z = (x - \mu)/s \tag{1}$$

Where, the mean value is denoted as μ and the standard deviation is represented as s. The standard scalar helps to arrange the negative data to a normally distributed function. Moreover, it is utilized when the classification is more significant than regression. Next to the process of pre-processing, the feature selection proceeds to a place with the help of dynamic Particle Swarm Algorithm (PSO).

3.3. Feature selection using dynamic PSO

Next to pre-processing, the feature selection takes place using the dynamic PSO which is the improved version of PSO with exponential decay and elimination of particles. The pre-processed output is comprised with 21 features; in that, dynamic PSO selects the best 10 features for classification. The fitness function is defined based on the accuracy of K-Nearest Neighbors (KNN) classifier, which serves as a proxy for feature relevance. In PSO, each individual population is referred to as the particle noted as the candidate solution. The process takes place in PSO optimization that attains the global optimal solution. The position of the particle is accompanied with four vectors including the best position in previous iteration $p_i^t = (p_{i,1}^t, p_{i,2}^t, \dots, p_{i,D}^t)$, position is $x_i^t = (x_{i,1}^t, x_{i,2}^t, \dots, x_{i,D}^t)$, velocity is $v_i^t = (v_{i,1}^t, v_{i,2}^t, \dots, v_{i,D}^t)$ and global best position is $g^t = (g_{i,1}^t, g_{i,2}^t, \dots, g_{i,D}^t)$. The count of iteration in the present state is denoted as t and variables are denoted as D. The particle's position is on the basis of (2) and (3).

$$x_{i,j}^{t+1} = x_{i,j}^t + v_{i,j}^{t+1} \tag{2}$$

$$v_{i,j}^{t+1} = v_{i,j}^t + c_1 \times r_1 \times \left(p_{i,j}^t - x_{i,j}^t\right) + c_2 \times r_2 \times \left(g_j^t - x_{i,j}^t\right) \tag{3}$$

Where, the index of the particle is represented as i, the size of the population is denoted as N, and index for each variable in the population is j. The local acceleration co-efficient is represented as c_1, the global acceleration co-efficient is represented as c_2, and the global acceleration co-efficient is represented as r_1 and r_2. Based on the aforementioned Eqs. (2) and (3), the ideology of PSO is presented in the following way. Every individual particle is attracted towards the best and global best position by velocity adjustment. Moreover, the velocity of the particle at position i is represented based on (4).

$$x_{i,j}^t = l_j + r_3 \times (u_j - l_j) \tag{4}$$

Where, the number dispersed in a randomized manner is represented as r_3 which lies between 0 and 1. Similarly, upper and lower boundaries of jth dimension are denoted as l_j and u_j, respectively. The number of populations is determined as 50 and fitness of each population is evaluated based on the threshold value. The particles vary by adopting themselves towards the individual in the best positions. The particles which do not exceed the predetermined range after the speed update is utilized by retaining the actual speed. The maximal velocity is allotted to the particle which does not exceed the predefined range continue to retain in the actual population. The cyclic process of eliminating particles is repeated until the particle population is 8. The final solution is selected as 8 to select the optimal feature

sets from the total number of features. When the number of iteration achieve the required range, the iteration is terminated and the optimal solution is achieved.

3.4. Classification

Next to the stage of selecting the features, the classification takes place with help of combined model based on Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM). The best selected 10 features using the dynamic PSO is utilized in the process of classification.

3.4.1. Convolutional Neural Network (CNN)

The architecture of CNN is general kind of artificial neural network which is known for its effective classification. In common, the architecture of CNN is comprised with one or more convolutional layers and a fully connected layer along with weights and pooling layer.

3.4.1.1. Convolutional layer

In this layer, the convolution of each sub regions in the input data with kernel is evaluated based on biasing and an activation function to produce feature map in the next layer. The samples of input features are considered as $x_i^0 = [x_1, x_2, \ldots x_n]$, where the total samples are represented as n and the output is evaluated based on the (5).

$$c_i^{l,j} = h(b_j + \sum_{m=1}^{M} w_m^j x_{i+m-1}^{0j}) \quad (5)$$

Where, the index of the layer is represented as l and the activation function is h that is utilized to produce non-linearity. The bias term of jth feature map is represented as b and the kernel size is M. The weight of the jth feature map is represented as w_m^j.

3.4.1.2. Average pooling layer

This layer is known as subsampling layer followed by convolutional layer to diminish the size of feature. Moreover, it considers small rectangular data blocks and provides a single output for each block. This research utilizes average pooling layer utilized in evaluating the average values in the input set. The process of pooling the feature map in the layer is represented in (6).

$$p_i^{l,j} = max_{i \times T+r}^{c^{l,j}} \quad (6)$$

Where, the pooling window size is represented as R and the stride is represented as T. After convolutional and the pooling layers, the features are transferred to a single dimensional vector. Finally, the classification takes place in the fully connected layer of CNN.

3.4.2. Long Short-Term Memory

LSTM is kind of recurrent neural network which is comprised with a cell, input gate, output gate and a forget gate. The forget gate in memory block structure is regulated with help of one layered neural network and this gate's activation is evaluated based on the (7).

$$f_t = \sigma(W[x_t, h_{t-1}, C_{t-1}] + b_f) \quad (7)$$

Where, the input sequence is represented as x_t and the output of the previous block is represented as h_{t-1}. The block memory of previous LSTM block is represented as C_{t-1}, the biasing vector and the sigmoid function is represented as b_f and σ, respectively.

The new memory is generated in the input gate of the cell with $tanh$ activation function and the memory block in the previous state is based on (8) and (9).

$$i_t = \sigma(W[x_t, h_{t-1}, C_{t-1}] + b_i \quad (8)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tanh([x_t, h_{t-1}, C_{t-1}]) + b_c \quad (9)$$

The output gate is generated based on the output of LSTM which is generated based on the (10) and (11).

$$\sigma_t = \sigma(W[x_t, h_{t-1}, C_{t-1}] + b_0 \quad (10)$$

$$h_t = o_t . \tanh(C_t) \quad (11)$$

The connections between the time steps is generated an internal feedback which permits the state to understand the concept of time and temporal dynamics. The error propagation in memory cell permits LSTM to bind the time lags and helps in effective prediction.

3.4.3. Unified model of CNN-LSTM for water quality prediction

The proposed methodology of the suggested framework is based on CNN as the front end which helps in processing the non-linear characteristics of input feature, and the LSTM acts as the back end which receives the abstracted data and helps in classification. The architectural diagram of proposed model is presented in Fig. 2.

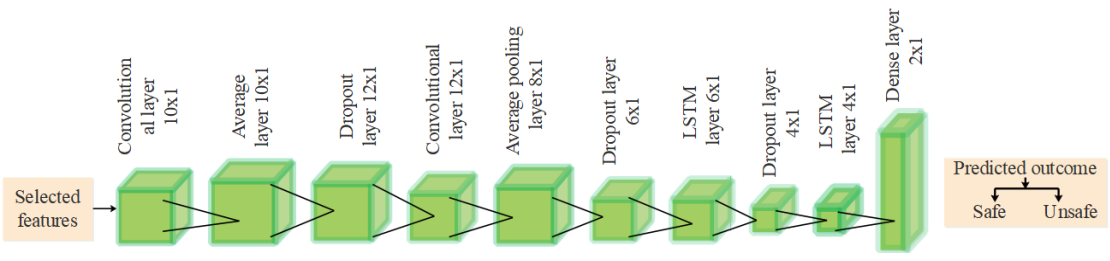

Fig. 2. Architectural presentation for CNN-LSTM

The best selected 10 features are provided as input into the convolutional layer of CNN. The specified structure of the

network is on the basis of numerous training times. In the suggested approach, depth of the neural network and the parameters utilized in each layer helps to improvise the prediction accuracy. The input layer of the CNN is convolved to produce the first layer and the average pooling layer is applied on each feature map at dropout layer. The CNN network is provided at an interval sequence with similar structure with two layers which utilize average pooling layer. The output of dense layer with similar shape, and merged as a new input at the LSTM and the output is achieved at the fully connected layer which helps to predict the outcome. The weights and biases are based on (12) and (13).

$$\Delta W_l(t+1) = -\frac{x_\lambda}{r} W_1 - \frac{x}{n}\frac{\partial c}{\partial W_l} + m\Delta W_l(t) \quad (12)$$

$$\Delta B_l(t+1) = -\frac{x}{n}\frac{\partial c}{\partial B_l} + m\Delta B_l(t) \quad (13)$$

Where, the weight, bias, number of layers, regularization parameter, rate of learning, training samples, cost function, momentum and updated step are denoted as $W, B, l, \lambda, x, n, C, m$ and t, respectively. The Rectified Linear Unit (ReLU) is utilized as activation function for convolutional layers. The network is employed with the dropout layer which removes the randomized partitions of the network and prohibits the neurons from opting training data. Moreover, the batch normalization is performed to enhance the model's convergence. The pseudocode of the proposed architecture is presented as follows:

Pseudo code for CNN-LSTM

```
Input: selected feature subset from Dynamic PSO

Output: Water Quality (safe, unsafe)


Initialize model
model input = input layer


Add Convolutional layers
for i in range(Number of convolutional layers):
   convolutional layer = Conv1D(filters=number of filters,
kernel size=kernel size, activation='ReLU',
padding='same')(model input)
   pool layer = MaxPooling1D(pool size=pool
size)(convolutional layer)
   model input = pool layer

end for


Add LSTM layers

LSTM output = LSTM (units=number of units, return
sequences=False) (model input)

LSTM output = LSTM(units=num_units, return
sequences=False)( lstm_output_1 )
Add skip connection (residual connection)
skip connection = Add()([LSTM output, input layer])


Add Dense layers for classification
dense layer = Dense(100, activation='ReLU')(skip
connection)
output layer = Dense(num_classes,
activation='softmax')(dense layer)


Define and compile the model
model = Model(inputs=input layer, outputs=output layer)
model. Compile(optimizer='Adam',
loss='categorical_crossentropy', metrics=['accuracy'])

Train the model
for epoch in range(number of epochs):
   for batch in range(number of batches):
      X batch, Y batch = get next batch()

 Get the next batch of training data
      Train the model on _batch(X batch, Y batch)

   end for

end for

Evaluate the model
loss, accuracy = Evaluate the model (X test, Y test)
```

4. Results and analysis

In this section, the experimental validation of results based on existing techniques. The evaluation of the proposed framework is performed in Python 3.7, Windows 10 OS, intel core i7 processor and 16 GB random access memory. The parameter setting of CNN-LSTM and dynamic PSO is presented in table 1 and table 2.

Table 1. Parameter setting of CNN-LSTM

Parameter	Value
Batch size	128
Optimizer	Adam
Activation function	Softmax
Loss function	Categorical cross entropy
Epoch	100

Table 2. Parameter setting of dynamic PSO

Parameter	Value
Population size	50
Inertia weight	0.5

The efficiency of the suggested approach is evaluated by considering the performance metrics like accuracy, precision, recall and F1 score, whose formulas are presented in (14-17).

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \quad (14)$$

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

$$\text{F1 score} = 2 \times \frac{Precision*Recall}{Precision+Recall} \quad (17)$$

Where, TN signifies true negative, TP signifies true positive, FP signifies false positive and FN signifies false negative.

5. Performance analysis

The efficiency of the dynamic PSO utilized in feature selection and the classifier is evaluated in this section. The outcomes are validated based on the efficiency with state of art techniques. Initially, the evaluation is performed based on different feature selection techniques and the evaluation is performed based on the efficiency of the classifier, with and without feature selection, as presented in Section 4.1.1 and 4.1.2 respectively.

5.1. Evaluation based on different feature selection

This section presents the evaluation of outcome on different feature selection techniques such as Recursive Feature Elimination (RFE), Mutual Information (MI), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and PSO. Table 4 presents the outcomes on the different feature selection techniques.

The outcomes displayed in table 3 exhibits that the proposed dynamic PSO achieves superior outcome when compared to the state of art techniques. The dynamic PSO utilized in this research achieves a classification accuracy of 99.99%, which is comparably higher than the state of art feature selection techniques. The exponential decay and elimination of particles helps to attains better results for selecting the best features with best fitness.

Table 3. Evaluation based on different feature selection

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RFE	67.45	62.09	59.07	60.54
MI	70.32	67.67	62.46	64.96
PCA	75.20	71.90	68.50	70.15
LDA	72.80	73.20	67.90	71.50
PSO	94.51	93.12	93.67	93.45
Dynamic PSO	99.99	99.99	99.99	99.99

5.2. Evaluation based on different classifiers

The section presents the outcomes on different classifiers with and without feature selection. Table 4 presents the outcome on various classifiers without feature selection. The state of art classifiers such as Recurrent Neural Network (RNN), CNN, Gated Recurrent Unit (GRU), Bidirectional Encoder Representations from Transformers (BERT).

Table 4. Evaluation of classifiers without feature selection

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RNN	69.05	66.03	64.72	65.36
CNN	67.78	63.45	61.32	62.36
GRU	70.56	67.53	62.73	65.04
BERT	72.30	69.80	65.50	67.60
CNN LSTM	93.53	93.80	93.00	93.50

Table 5. Evaluation of classifiers with feature selection

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RNN	85.32	78.65	73.56	76.01
CNN	90.78	85.05	84.67	84.85
GRU	78.67	73.03	71.42	72.21
BERT	80.50	79.90	79.25	79.50
CNN-LSTM	99.99	99.99	99.99	99.99

The results achieved from table 4 and table 5 exhibits that the proposed CNN-LSTM attains better outcomes when compared to existing techniques, in both the cases related to presence and absence of feature selection. For instance, the accuracy of CNN-LSTM without feature selection is 93.53% and accuracy with feature selection is 99.99%. Thus, the suggested framework attains better outcome in both the terms of presence and absence of feature selection.

The commendable outcomes are due to CNN as the front end which helps in processing the non-linear characteristics of input feature, and the LSTM acts as the back end which receives the abstracted data and helps in classification. Fig. 3 presents the graphical depiction of results on various classifiers with feature selection.

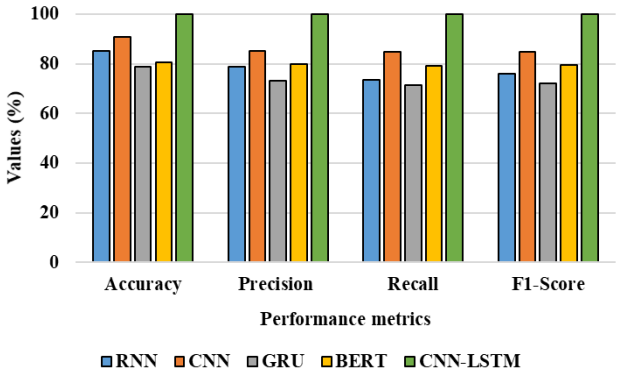

Fig. 3. Graphical depiction of results based on different classifiers with feature selection

The confusion matrix based on the performance of classifier in predicting the outcome based on water quality is depicted in Fig. 4.

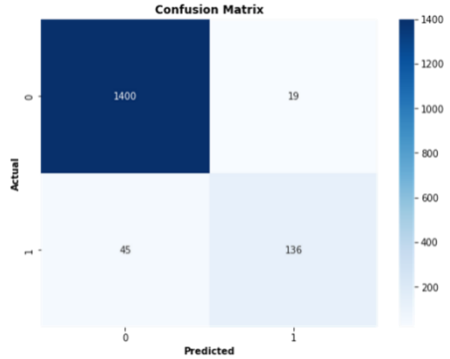

Fig. 4. Confusion matrix for the performance of the classifier

5.3. Comparative analysis

This section presents the comparative evaluation of outcomes with the existing techniques. Table 6 presents the outcomes of the proposed classifier with the existing methods. The evaluation is performed based on accuracy, precision, recall and F1 score.

The outcome from table 6 exhibits that the CNN-LSTM attains an accuracy of 99.99% which is more preferable than the existing ensemble method with the accuracy of 98.1%. The exponential decay and the elimination of particles using dynamic PSO helps in selecting the features with the best fitness. Similarly, the combination of CNN-LSTM helps in processing the non-linear characteristics of input feature, and the LSTM acts as the back end which receives the abstracted data that helps in predicting the water quality as safe or unsafe.

Table 6. Comparative results

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Ensemble Method [17]	98.1	100	98.1	-
ML [21]	99.83	-	-	-
CNN-LSTM	99.99	99.99	99.99	99.99

5.4. Discussion

This section presents the overall validation of numerical results when evaluated with the state of art techniques and the existing methods. The efficiency of different feature selection techniques is evaluated in Section 4.1.1 and the efficiency of dynamic PSO is validated with RFE, MI, PCA, LDA and PSO. Among the aforementioned existing techniques, the dynamic PSO achieves better accuracy of 99.99%. The process of exponential decay and the particle elimination facilitates in achieving better outcomes, in contrast to the state of art techniques. Similarly, the efficiency of the classifier is assessed in Section 4.1.2 based on presence and absence of feature selection. The proposed classifier achieves accuracies of 93.53% and 99.99% based on the presence and absence of feature selection. Moreover, the classification accuracy of suggested framework is validated with existing ensemble model; in that, the suggested approach attains classification accuracy of 99.99%, whereas the ensemble model accomplishes a classification accuracy of 98.1%. The CNN helps in processing the non-linear characteristics of input feature, and the LSTM acts as the back end to predict the water quality as safe or unsafe.

6. Conclusion

Water is considered as a valuable source of resource which is known for its various needful applications. This research focuses on predicting the water quality as safe and unsafe, further aiding in predicting the water quality as fit and unfit, based on its contents. The data collected from water quality dataset is pre-processed using standard scalar and the features are selected with the help of dynamic PSO. After selecting the optimal feature sets, the classification is performed using the combined model of CNN-LSTM. The CNN processes the input features from dynamic PSO based on the non-linear characteristics, and the LSTM at the back end of the model obtains the abstracted data to analyze the quality of water as safe and unsafe. The experimental validation exhibits that the proposed classifier accomplishes a classification accuracy of 99.99% which is comparably superior to the existing ensemble model with a classification accuracy of 98.1%. The future research will focus on using advanced deep learning models with a recent optimization technique to enhance the performance of water quality prediction.

Author contributions

Raju Amireddy: Conceptualization, Methodology, Software, Field study, Data curation, Writing-Original draft preparation **Dileep Pulugu:** Software, Validation, Field study, Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] N. H. A. Malek, W. W. F. Yaacob, S. A. Md Nasir, and N. Shaadan, "Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques," *Water*, vol. 14, p. 1067, Mar. 2022, https://doi.org/10.3390/w14071067.

[2] Q. B. Pham, R. Mohammadpour, N. T. T. Linh, M. Mohajane, A. Pourjasem, S. S. Sammen, D. T. Anh, and V. T. Nam, "Application of soft computing to predict water quality in wetland," *Environ. Sci. Pollut. Res*., vol. 28, no. 1, pp. 185-200, Jan. 2021, https://doi.org/10.1007/s11356-020-10344-8.

[3] S. Park, S. Jung, H. Lee, J. Kim, and J. -H. Kim, "Large-Scale Water Quality Prediction Using Federated Sensing and Learning: A Case Study with Real-World Sensing Big-Data," *Sensors*, vol. 21, no. 4, p. 1462, Jan. 2021, https://doi.org/10.3390/s21041462.

[4] B. Sakaa, A. Elbeltagi, S. Boudibi, H. Chaffaï, A. R. M. T. Islam, L. C. Kulimushi, P. Choudhari, A. Hani, Y. Brouziyne, and Y. J. Wong, "Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin," *Environmental Science and Pollution Research*, vol. 29, no. 32, pp. 48491-48508, Jul. 2022, https://doi.org/10.1007/s11356-022-18644-x.

[5] T. P. Latchoumi, K. Raja, Y. Jyothi, K. Balamurugan, and R. Arul, "Mine safety and risk prediction mechanism through nanocomposite and heuristic optimization algorithm," *Meas.: Sens*., vol. 23, p. 100390, Oct. 2022, https://doi.org/10.1016/j.measen.2022.100390.

[6] P. Chen, B. Wang, Y. Wu, Q. Wang, Z. Huang, and C. Wang, "Urban river water quality monitoring based on self-optimizing machine learning method using multi-source remote sensing data," *Ecol. Indic*., vol. 146, p. 109750, Feb. 2023, https://doi.org/10.1016/j.ecolind.2022.109750.

[7] J. Zhang, T. Zou, and Y. Lai, "Novel method for industrial sewage outfall detection: Water pollution monitoring based on web crawler and remote sensing interpretation techniques," *J. Cleaner Prod*., vol. 312, p. 127640, Aug. 2021, https://doi.org/10.1016/j.jclepro.2021.127640.

[8] R. Bogdan, C. Paliuc, M. Crisan-Vida, S. Nimara, and D. Barmayoun, "Low-Cost Internet-of-Things Water-Quality Monitoring System for Rural Areas," *Sensors*, vol. 23, p. 3919, Apr. 2023, https://doi.org/10.3390/s23083919.

[9] T. Selmane, M. Dougha, S. Djerbouai, Djamaleddine djemiat, and N. Lemouari, "Groundwater quality evaluation based on water quality indices (WQI) using GIS: Maadher plain of Hodna, Northern Algeria," *Environ. Sci. Pollut. Res*., vol. 30, no. 11, pp. 30087-30106, Mar. 2023, https://doi.org/10.1007/s11356-022-24338-1.

[10] M. A. K. Fasaee, E. Berglund, K.J. Pieper, E. Ling, B. Benham, and M. Edwards, "Developing a framework for classifying water lead levels at private drinking water systems: A Bayesian Belief Network approach," *Water. Res*., vol. 189, p. 116641, Feb. 2021, https://doi.org/10.1016/j.watres.2020.116641.

[11] Z. Wang, Q. Wang, and T. Wu, "A novel hybrid model for water quality prediction based on VMD and IGOA optimized for LSTM," *Front. Environ. Sci. Eng*., vol. 17, no. 7, p. 88, Feb. 2023, https://doi.org/10.1007/s11783-023-1688-y.

[12] R. Tan, Z. Wang, T. Wu, and J. Wu, "A data-driven model for water quality prediction in Tai Lake, China, using secondary modal decomposition with multidimensional external features," *J. Hydrol.: Reg. Stud*., vol. 47, p. 101435, Jun. 2023, https://doi.org/10.1016/j.ejrh.2023.101435.

[13] S. Yang, Shaojun, S. Zhong, and K. Chen, "W-WaveNet: A multi-site water quality prediction model incorporating adaptive graph convolution and CNN-LSTM," *Plos one*, vol. 19, no. 3, p. e0276155, Mar. 2024, https://doi.org/10.1371/journal.pone.0276155.

[14] H. Ghosh, M. A. Tusher, I. S. Rahat, S. Khasim, and S. N. Mohanty, "Water Quality Assessment Through Predictive Machine Learning," in International Conference on Intelligent Computing and Networking, Proceedings of IC-ICN 2023, Springer Nature, Singapore, 2023, pp. 77-88, https://doi.org/10.1007/978-981-99-3177-4_6.

[15] L. Chen, T. Wu, Z. Wang, X. Lin, and Y. Cai, "A novel hybrid BPNN model based on adaptive evolutionary Artificial Bee Colony Algorithm for water quality index prediction," *Ecol. Indic*., vol. 146, p. 109882, Feb. 2023, https://doi.org/10.1016/j.ecolind.2023.109882.

[16] M. Y. Shams, A. M. Elshewey, E. -S. M. El-kenawy,

A. Ibrahim, F. M. Talaat, and Z. Tarek, “Water quality prediction using machine learning models based on grid search method,” *Multimedia Tools Appl*., vol. 83, no. 12, pp. 35307-35334, Apr. 2024, https://doi.org/10.1007/s11042-023-16737-4.

[17] E. Dritsas and M. Trigka, “Efficient Data-Driven Machine Learning Models for Water Quality Prediction,” *Sensors*, vol. 23, p. 1161, Jan. 2023, https://doi.org/10.3390/computation11020016.

[18] M. I. Shah, M. F. Javed, A. Alqahtani, and A. Aldrees, “Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data,” *Process Saf. Environ. Prot*., vol. 151, pp. 324-340, Jul. 2021, https://doi.org/10.1016/j.psep.2021.05.026.

[19] A. Bhardwaj, V. Dagar, M. O. Khan, A. Aggarwal, R. Alvarado, M. Kumar, M. Irfan, and R. Proshad, “Smart IoT and machine learning-based framework for water quality assessment and device component monitoring”, *Environ. Sci. Pollut. Res*., vol. 29, no. 30, pp. 46018-46036, Jun. 2022, https://doi.org/10.1007/s11356-022-19014-3.

[20] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, “Prediction of groundwater quality using efficient machine learning technique,” *Chemosphere*, vol. 276, p. 130265, Aug. 2021, https://doi.org/10.1016/j.chemosphere.2021.130265.

[21] 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, and S. Mollick, “Efficient prediction of water quality index (WQI) using machine learning algorithms”, *Human-Centric Intelligent Systems*, vol. 1, no. 3, pp. 86-97, Dec. 2021, https://doi.org/10.2991/hcis.k.211203.001.

[22] Link for dataset: https://www.kaggle.com/datasets/mssmartypants/water-quality