DOI: 10.24996/ijs.2025.66.6.30

ISSN: 0067-2904

Water Quality Prediction and Classification using AFSO based Long Short-Term Model with Data Transformation Manuscript

Divyajyothi M G[1]*, Rachappa Jopate[1], Piyush Kumar Pareek [2], Anwar Al Daeri[1]

[1] Department of IT, University of Technology and Applied Sciences- Al Mussanah, Muscat, Sultanate of Oman

[2]Professor, Department of AI / ML, NITTE Meenakshi Institute of Technology, Bangalore, India

Received: 29/9/2024 Accepted: 26/11/2024 Published: 30/6/2025

Abstract

Water is a precious, essential, and dwindling resource in both developing and developed nations. As a vital nutrition for human beings, it easily takes the cake as the planet's most valuable natural resource. Various wastes, including municipal, industrial, agricultural (including pesticides and fertilizers), medical, etc., contribute to geo-environmental contamination and render water unfit for human or animal use. Therefore, it is crucial to establish effective means to automate water suitability checking. In this investigation, the variables included are pH, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, and turbidity. These measurements serve as a feature vector to represent the state of the water. The article employed an enhanced deep learning model (IDL) to predict the water quality class. Normalization, spitting, and transformation are three of the models used in preliminary data processing. Long Short-Term Memory (LSTM) models take in preprocessed data as input, and their weights are ideally chosen using Artificial Fish Swarm Optimization (AFSO). Using a dataset obtained from Kaggle, tests were conducted with several levels of granularity. The findings showed that the LSTM classifier is superior to the rest. The results show that deep learning methods may effectively forecast the viability of water quality.

Keywords: Improved deep learning; Artificial Fish Swarm Optimization; Long Memory; Water Quality Prediction; Kaggle Dataset.

1. Introduction

Continued global population growth and socioeconomic development have accelerated the depletion of natural resources, particularly the polluting of aquatic environments. The growing discharge of physical, chemical, and biological pollutants in water sources has harmed and continues to harm aquatic life [1] [2]. Hydro-morphological sources include those associated with natural processes and human activities, such as water abstraction, while point sources include industries and direct effluent disposal. Non-point sources include surface runoff and airborne outflows. Given the numerous potential risks to aquatic life and public health posed by the various sources of toxins in aquatic ecosystems, it is crucial to devise a management approach to mitigate these issues. Water's suitability for any given use (drinking, irrigation, recreation, industrial activities, etc.) is contingent on factors like the amount and kind of pollutants present in the water.

*Email: divya@act.edu.com

Water quality indices (WQIs) often include contamination criteria [3] to express these pollutants. Water quality prediction is useful for both the dynamic control of water quality and the preparation for unforeseen situations. It is a must-do in the realms of water management and pollution prevention [4]. Management practices of yesteryear emphasized prior-prevention over inferior-disposal. Water indicators are influenced by several factors, not all of which are well-structured [5]. Furthermore, the nondeterministic and nonlinear characteristic determines the complexity of water quality prediction. There are a variety of national and international tools available for predicting water quality [6]. Most of these methods are based on mathematical statistics, grey theory, chaos theory, and neural networks. Mathematical statistics is an effective modeling technique, but with imperfect predictions [7]; nonlinear functions are too complex to be approximated by grey theory. The chaos theory approach works well [8] when the training data is very abundant. The nonlinearity, concept, and learning capacities of classical neural networks make them well-suited for processing non-linear, random input. Due to the nature of time series data, traditional neural networks are unable to process it [9].

Time series data used to evaluate water quality indicators show clear seasonal fluctuation. Period series forecasting [10] includes the study of water quality. This study presents a neural network-based method for predicting water quality, making it possible to handle time series data. Deep learning is a tactic that may be used to learn the characteristics of a dataset [11]. Scientists have been trying to develop deep-learning-based algorithms to solve the time series prediction problem in recent years. In the past, water quality monitoring was performed manually, requiring the collection of water samples and their subsequent shipment to labs for examination. Using these techniques, you cannot obtain timely information [12]. The suggested water quality monitoring system uses wireless technologies to allow for constant data delivery.

The water environment is complex, dynamic, and non-linear [13] due to the interplay of several water quality factors such as temperature, dissolved oxygen, pH value, ammonia nitrogen, nitrites, and nitrates. Recent progress has been achieved using models to forecast water quality time series. Most earlier techniques, for instance [14], relied on the combined model to enhance the traditional BP neural network and support vector machine approach. This demonstrates that the integrated model can produce reliable water quality forecasts. A WQI is a method for characterizing water as a whole by reducing a large amount of data on water quality to a single, standardized numerical number. Since WQIs may be used to compare water quality over time and space, they are useful for making predictions about water quality. Specifically, WQIs are used to (i) increase the knowledge of general water quality and communicate information regarding water quality in a plain manner to the public, policymakers, and non-water professionals.

The LSTM model, with its weights ideally chosen by AFSO, is employed in this study to forecast future water quality. Normalization, data partitioning, and data transformation are three of the approaches used for further analysis. The rest of the paper follows this structure: we review the significant literature in Section 2, while Section 3 offers an overview of the proposed model. Section 4 presents the validation analysis along with accompanying remarks. Finally, Section 5 concludes with a summary of the study.

2. Related Work

To identify water's fitness for drinking or other applications, Dritsas and Trigka [15] present a supervised learning strategy to develop feasible prediction models using a labeled training dataset. To classify the water as safe or unsafe, they propose using a collection of physicochemical and microbiological attributes as input features to effectively capture the current state of the system. These characteristics are selected for their ability to provide a comprehensive representation of the system's conditions. The challenge is approached as a binary classification problem, where we evaluate multiple machine learning algorithms to assess their performance metrics, including accuracy, recall, precision, and area under the curve. This evaluation is conducted both with and without the application of class balancing techniques, specifically utilizing the Synthetic Minority Oversampling Technique (SMOTE). Our findings indicate that the Stacking classification model, when combined with SMOTE and 10-fold cross-validation, demonstrates superior performance, achieving remarkable metrics of 98.1% accuracy and a perfect recall rate of 100%. Precision (100%) and Area Under the Curve (AUC) (99.9%). In conclusion, this study presents a paradigm that might aid researchers in their pursuit of water quality prediction using ML.

Using seven widely used WQI models and three entirely new and recently proposed models, in his research, Uddin [16] investigated several machine learning classification techniques to determine the best approach for assessing water quality categories. The study considered a variety of algorithms, including support vector machines (SVM), Naive Bayes (NB), random forest (RF), k-nearest neighbors (KNN), and gradient boosting (XGBoost). For seven different WQI models, the KNN (100% correct and 0% incorrect) and XGBoost (99.9% correct and 0.1% wrong) algorithms failed to provide reliable water quality predictions. Model validation results show that the XGBoost classifier excels in predicting the proper categorization of water quality, with an accuracy of 1, precision of 0, sensitivity of 0, specificity of 1, and F1 score of 0.99. In addition, the weighted quadratic mean (WQM) and unweighted WQI models demonstrated improved prediction accuracy, precision, sensitivity, specificity, and F1 score compared to WQI models for each class. The results of this investigation indicated that the WQM and RMS models have the potential for accurate assessment of coastal water quality. Therefore, this study may aid in supplying researchers, policymakers, and water research staff with precise water quality information for monitoring and utilising the WQI model.

A farming forecast for cold water and warm water fish, plants, and bacteria is provided by Nemade and Shah [16] to enhance the aquaponics farming industry. At its outset, the suggested system gathers information from the Internet of Things sensors. The next step is data cleaning, which involves erasing any outliers or missing data. The next step is to extract characteristics that are relevant to the sensed data and eliminate any that aren't. Then, to solve the issue of unequal classes, they provide a new M-SMOTE method. As a final step, the multi-model categorization is used in the suggested method for the aquaponic environment. The suggested technique employs a voting mechanism for optimum prediction to compare the efficacy of six different classifiers. Based on a voting approach, the suggested technique selects XGBoost and random forest as the best classifiers. The testing outcomes demonstrate that the suggested technique provides a novel, cutting-edge prediction model for aquaponics farming with an accuracy of 99.13%.

Using data from monitoring stations in the Pearl River Estuary in Guangdong, China, Yan et al. [17] offer a Bayesian-optimized machine learning approach to analyzing pollution levels. Accuracy (0.992) and Kappa (0.987) were both maximized in the optimized stacked generalization (SG-op) model. The relevance of features in the prediction model was in line with conventional measures of air pollution. The significance of the fluctuation patterns between several pollution indicators was calculated using the Spearman rank correlation coefficient. The study found an increasing trend in many major pollution markers, including total petroleum (PET). This framework may be used to make accurate predictions about water quality in the future and offer technical assistance in the event of an emergency involving pollution.

To optimize a long short-term memory (LSTM) neural network, Wang et al. [18] offer a method that uses variational mode decomposition (VMD) and an enhanced version of the grasshopper optimization algorithm (IGOA). To make the water quality data more stable and hence easier to forecast, VMD was first used to break it down into a series of less volatile components; these were then fed into the IGOA-LSTM model. The projected values were then calculated by adding each individual component. The data utilized for training and prediction in this work came from the Ganjiang River's Dayangzhou Station and Shengmi Station monitors. Compared to other models, such as the integrated model of Ensemble Empirical Mode Decomposition (EEMD), the integrated model of Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), the Nonlinear Autoregressive Network with Exogenous Inputs (NARX), the Recurrent Neural Network (RNN), and others, the proposed VMD-IGOA-LSTM model performed better in short-term prediction, as shown by the experimental results. This work can give a solid method for predicting water quality in other regions.

The SVABEG method proposed by Bi et al. [19] combines a Savitzky-Golay (SG) filter, Variational Mode Decomposition (VMD), cleans up the original time series, and copes with any nonlinearities. SVABEG uses the SG filter and VMD. Then, SVABEG combines BiLSTM, the ED structure, and the attention mechanism to simultaneously capture long-term, bidirectional correlations, realize dimensionality reduction and extract crucial information. In addition, SVABEG uses GSPSO for hyperparameter optimization. The proposed SVABEG improves upon the accuracy of predictions made by the state-of-the-art algorithms, as shown by experiments conducted on real-world datasets.

3. Proposed System

3.1 Predicting water portability dataset

Safe drinking water is an important part of any health protection program and a vital human right. This matters for national, regional, and local health and development. Investments in water supply and sanitation have been found to be economically beneficial in some areas since the savings made on healthcare and medical expenses due to improved health are more than the upfront expenditures of these initiatives. There are 3276 individual water bodies represented by quality measurements in the Predicting Water Portability dataset. Retrieved from "https://www.kaggle.com/adityakadiwal/water-potability," the dataset may be found here. The following table lists the values for the dataset's parameters.

Table 1: The parameters used in the dataset.

S. No	Parameters
1	Hardness
2	Solids (Total dissolved solids - TDS)
3	Organic_carbon
4	pH value
5	Chloramines
6	Sulfate
7	Conductivity
8	Trihalomethanes
9	Turbidity
10	Potability (0 = Not potable , 1 = Potable)

3.2 Pre-processing

3.2.1 Data Normalization

Neural computing techniques need transformations to standardize raw data on both the independent and dependent variables. The transformation modifies input variables so their distributions are consistent with the estimated results. Reasons for scaling data samples include initial parity in the importance of variables and enhanced interpretability of network weights. To ensure that all inputs and outputs are given equal weight during training and to speed up network convergence, neural networks are often decomposed and linearly transformed once they have been initialized. Data on water quality is normalized using Equation (1).

$$X_i = a \times \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

The original data values, a and B, are represented in the normalized data set X_i by the constants A and B, respectively.
Z-score transformation also works well to enhance data samples. Following this formula will allow you to calculate your Z-score. This is a statistical technique for measuring the dispersion around the mean of a given data value. This equation depicts the mean and standard deviation of the original data, with x standing in for the answer.

$$X_i = \frac{x_i - \bar{x}}{\sigma} \tag{2}$$

3.2.2 Data division (data splitting)

To further strengthen their predictions, water quality models need to be validated by incorporating previously unseen data. So that the model would almost always yield the same results across different data sets. This idea is widely discussed in published works. The split between training and validation data substantially affects the accuracy of predictive models.

3.2.3 Data Transformation

Data transformation is used in the construction sector to convert numerical data into categorical data for use with our proposed techniques. Equal-width and equal-frequency methods are broadly used because of the ease of implementation. The equal-width method allows you to divide the possible values for a variable into several narrower bands. Typically, the user determines the interval size based on his or her level of familiarity with the topic at hand [20]. Multiple data intervals of about the same length can be created by using the same frequency threshold. The five numerical variables were classified into three groups using the equal frequency approach. The equal-frequency method makes outliers less of an issue.

It is possible to simplify the process of making predictions by converting category variables into numerical ones using various methods. For categorical variables with L levels, several researchers have used a method called "one-hot encoding," which generates matrices with just L-1 columns. Data with many dimensions is possible if there are several category variables. Embedding networks and other deep learning methods may be utilized to create dense representations of categorical variables [21]. Using data transformation to study massive amounts of time series data can further reduce computing expenses. The symbolic aggregate approximation (SAX) technique [22] [23] allows for the symbolic representation of simplified numerical time series in construction data. By doing so, we may reduce the size of the starting data while keeping the amount of information lost during transformation to a minimum.

4. Water Quality Prediction-Deep Learning Based LSTM Model

The LSTM can't exist without the regular RNN framework. The issue of Recurrent ANN's inability to handle long distance dependency is solved by applying various models for computing the hidden state. Despite the advantages of the newly constructed model, these models do not learn to their full potential. This LSTM method uses a collection of memory units, each of which consists of 3 gates and varies in performance. The following are values of the particular conditions of the LSTM unit of the tth word, where the text feature S vector is used as input, and the tth word is used as a sample. The following is a processing phrase in which the sigmoid function () and the dot multiplication operator () are defined.
The f_t implies a forget gate:

$$f_t = \sigma\left(W_f w_t + U_f h_{t-1} + b_f\right) \tag{1}$$

The i_t signifies an input gate:

$$i_t = \sigma(W_i w_i + U_i h_{t-1} + b_i) \tag{2}$$

The $\tilde{c}_t$ Recent ingestion of the candidate memory cell criterion, whereas tanh is the hyperbolic definition of the tangent;

$$\tilde{c}_t = \tanh\,(W_c w_t + U_i h_{t-1} + b_c) \tag{3}$$

Recent time in the memory cell is represented by c_t, which is the sum of the f_t and i_t series values between 0 and 1. Data from the candidate unit c _t is written to c_t during the processing of i_tc_t. Data conservation is undesirable for memory, as indicated by the function f_tc_t1. c_{t-1}.

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \tag{4}$$

The o_t indicates output gate:

$$o_t = \sigma(W_0 w_t + U_0 h_{t-1} + b_0) \tag{5}$$

h_t implies hidden layer state at time t:

$$h_t = o_t \odot \tanh\,(c_t) \tag{6}$$

Inadequate training data can be found in the LSTM's series of past observations. If it uses future information, it can serve several purposes. The bidirectional LSTM consists of two LSTM layers, one facing forward and one facing backward [24][26]. In the following, we present the r function: The forward layer utilizes past data from a series, whereas the reverse layer takes in future data. Then, the resulting layers are combined into a single one. One great thing about this approach is that it takes a series of background data into account. At time t, the output of the forward hidden unit is h_(t-1), and the output of the backward hidden unit in the simulation is h _(t+1), with the word embedding w_t serving as the input of time. When all is said and done, the result at time t for both backward and concealed units is the same as indicated below.:

$$\vec{h}_t = L\left(w_t, \vec{h}_{t-1}, c_{t-1}\right) \tag{7}$$

$$\overleftarrow{h}_t = L\left(w_t, \overleftarrow{h}_{t+1}, c_{t+1}\right) \tag{8}$$

Where $L(.)$ mentions the hidden layer of the LSTM hidden layer A forward subsequent vector is$\vec{h}_t \in R^{1\times H}$ as well as backward output vector is $\overleftarrow{h}_t \in R^{1\times H}$, which must be joined to achieve the text feature. It is labeled that shows H the amount of hidden layer cells:

$$H_t = \vec{h}_t || \overleftarrow{h}_t \tag{9}$$

5. Parameter Tuning Process – Artificial fish Swarm Optimization (AFSO)

While rapid progress is being made during training, the learning rate (_t) is constrained using this strategy. Increases in the poundage are achieved by training a larger step size, or "learning rate." The learning rate is a positive but small number modeling hyper-parameter utilised in NN training. The route towards fixing these problems is well-balanced, thanks to the learning rate. For slower learning rates, longer training epochs that provide incremental weight changes are necessary. The training epochs are short, and the learning rates are high, so significant improvements may be made. Learning rate simulation is a difficult and time-consuming process. If the learning rate is set too high, the training process becomes divergent, while setting it too low slows down convergence. Training using a range of activatable learning rates yields the best results.

Fish in the wild may either undertake their own research to determine where the best food is, or they can follow the path of other fish that have already found it. The solution space and the territory of other AFs make up an AF's environment. What it does next depends on how things are right now and the surrounding environment. In addition, three other forms of updating are considered in this AFS improvement, which are discussed below.

5.1 Prey behavior

This crucial biological activity helps provide sustenance. U_i is assertively selecting a state U_j within its detecting range based on the states of simulated fish. The separation between the simulated fish is $d_{i,j} = ||U_i - U_j||$; here, i and j are random fish

$$U_j = U_i + visual.r \tag{10}$$

$$U_i^{(t+1)} = U_i^{(t)} + \frac{U_j - U_i^{(t)}}{||U_j - U_i^{(t)}||}.step.rand, \tag{11}$$

where rand generates random numbers between 0 and 1, and step is the largest possible AF step size. Visual distance is the shortest path between an object in the viewer's field of view to the observer's retina.

5.2 Swarm behavior

Suppose the current state of AF is $U_i(d_{i,j} < Visual)_{\mathrm{a}}$; number of AF is n_f if $(n_f < \delta)$ means there's more food and less competition for it with your partner; if it's higher than F_i, you should go towards the partnership's center.

$$U_i^{(t+1)} = U_i^{(t)} + \frac{U_s - U_i^{(t)}}{||U_s - U_i^{(t)}||}.step.r \tag{12}$$

5.3 Follow behavior

Assume the AF is in state L_i; using information about its near-by neighbors and the number of its partners, determine its ideal state L_max. $L_{\max}$ is if $(n_f < \delta)$ Assume the AF is in state L_i; using information about its near-by neighbors and the number of its partners, determine its ideal state L_max.

6. Results and Discussion

This study's code was created in Python using the Keras framework, and it has been tested and evaluated using the Waikato Environment for Knowledge Analysis software package. The suggested method is implemented on a machine with 12 GB of DDR3 RAM, a 2.5 GHz Intel Core i5 7200U CPU, and a host of other high-end components.

6.1 Performances metrics

The performance of the classification algorithms was compared according to the criteria of accuracy, precision, and F-score that is shown in Eq. (13-16).

$$Accuracy\ (AC) = \frac{A+B}{A+B+C+D} \tag{13}$$

$$Precision\ (P) = \frac{A}{A+B} \tag{14}$$

$$Sensitivity\ (S) = \frac{A}{A+D} \tag{15}$$

$$F - Measure\ (F - M) = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} \tag{16}$$

Table 2 : RESULTS OF VARIOUS DEEP LEARNING MODELS.

	Model	F-M (%)	S (%)	AC (%)	P (%)
With AFSO	DBN	83.50	86.20	86.23	82.80
	CNN	84.20	84.40	84.74	83.90
	RNN	90.80	92.10	92.06	89.60
	Proposed	96.50	96.50	96.49	96.50
Without AFSO	DBN	84.30	86.90	86.89	83.70
	CNN	84.50	84.90	84.93	84.20
	RNN	90.90	92.10	92.12	89.70
	Proposed	95.80	95.90	95.86	95.80

In the above Table 2 signifies the results of various deep learning models. In the analysis of AFSO, the DBN model attained the F-M score of 83.50 and the sensitivity calculation of 86.20 and accuracy range of 86.23, and the precision range of 82.80 respectively. Then the CNN model attained an F-M score of 84.20 and the sensitivity calculation of 84.40 and accuracy range of 84.74 and the precision range of 83.90, respectively. Then the RNN model attained an F-M score of 90.80 and the sensitivity calculation of 92.10 and the accuracy range of 92.06 and the precision range of 89.60, respectively. Then the Proposed model attained the F-M score of 96.50 and the sensitivity calculation of 96.50 and the precision range of 96.50, respectively. Then without AFSO ratio, the DBN model attained the F-M score of 84.30 and the sensitivity calculation of 86.90 and the accuracy range of 86.89 and the precision range of 83.70, respectively. Then the CNN model attained the F-M score as 84.50 and the sensitivity calculation of 84.90, and the accuracy range of 84.93, and the precision range of 84.20, respectively. Then the RNN model attained the F-M score of 90.90 and the sensitivity calculation as 92.10, and the accuracy range of 92.12, and the precision range of 89.70, respectively. Then the Proposed model attained the F-M score of 95.80 and the sensitivity calculation of 95.90 and the accuracy range of 95.86 and the precision range of 95.80, respectively.

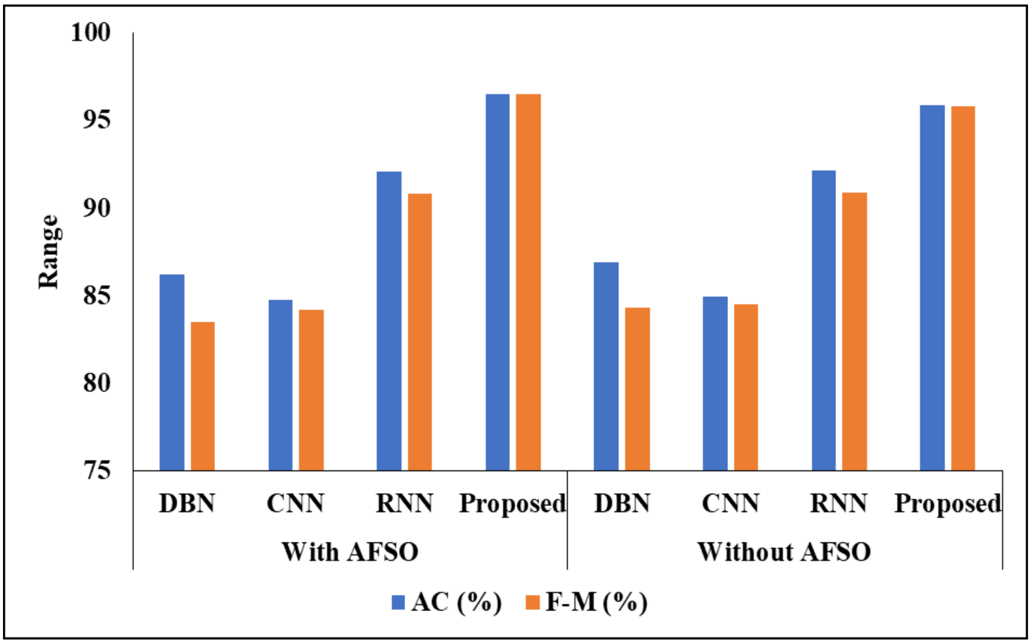

Figure 1: Analysis of various DL models

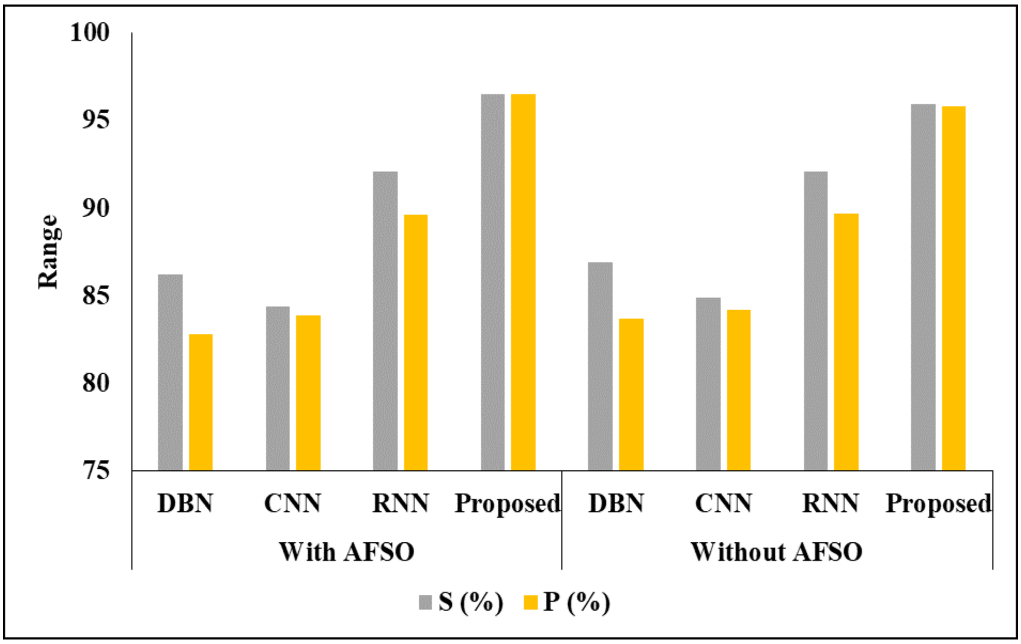

Figure 2: Comparative study of Proposed model with existing techniques.

Table 3: NUMBER OF INSTANCE BASE PROPOSED MODEL EVALUATION.

Sensitivity (%)				
Number of Instance	DBN	CNN	RNN	Proposed
100	87.90	92.60	93.30	94.80
200	84.60	88.40	92.30	95.20
300	86.40	93.20	93.60	96.30
400	88.60	92.40	96.90	97.60
500	89.10	93.60	96.00	98.00
Specificity (%)				
100	83.40	84.20	92.60	94.70
200	83.60	86.10	91.20	96.80
300	86.90	87.30	92.40	95.00
400	82.10	88.30	88.60	91.20
500	86.40	89.30	90.40	93.80
Accuracy (%)				
100	76.80	89.40	91.60	95.10
200	78.60	91.30	92.40	95.90
300	77.80	87.60	90.40	95.30
400	80.10	86.40	93.20	97.10
500	82.40	86.30	92.80	97.40

Table.3 signifies that the Number of instance base proposed model evaluation. In the Sensitivity analysis, in 100 instances, the DBN model reached the sensitivity of 87.90 and the CNN model reached as 92.60, and the RNN model reached as 93.30 and finally, the proposed model reached 94.80 respectively. Then 200 instances, DBN model reached the sensitivity of 84.60 and the CNN model reached 88.40, and the RNN model reached as 92.30 and finally, the proposed model reached 95.20, respectively. Then 300 instances, DBN model reached the sensitivity as 86.40 and the CNN model reached 93.20 and the RNN model reached 93.60 and finally the proposed model reached 96.30, respectively. Then 400 instances, DBN model reached the sensitivity 88.60, and the CNN model reached 92.40, the RNN model reached 96.90, and finally the proposed model reached 97.60 respectively. Then 500 instances, the DBN model reached the sensitivity of 89.10 and the CNN model reached 93.60, and RNN model reached 96.00 and finally the proposed model reached 98.00 respectively. Then Specificity analysis, 100 instances, the DBN model reached the sensitivity of 83.40 and the CNN model reached as 84.20 and the RNN model reached as 92.60 and finally the proposed model reached as 94.70 respectively. Then 200 instances, DBN model reached the sensitivity of 83.60 and the CNN model reached 86.10 and the RNN model reached 91.20, and finally the proposed model reached 96.80 respectively. Then 300 instances, the DBN model reached the sensitivity of 86.90 and the CNN model reached as 87.30 and the RNN model reached 92.40, and finally the proposed model reached 95.00, respectively. Then 400 instances, DBN model reached the sensitivity as 82.10 88.30, the RNN model reached as 88.60, and finally the proposed model reached as 91.20 respectively. Then 500 instances, DBN model reached the sensitivity as 86.40 and the CNN model reached 89.30, and the RNN model reached as 90.40 the 93.80 respectively. Then The Accuracy analysis, in 100 instances, the DBN model reached the sensitivity of 76.80, the CNN model reached 89.40, and the RNN model reached 91.60 and finally the proposed model reached 95.10 respectively. Then 200 instances, the DBN model reached the sensitivity as 78.60, and the CNN model reached 91.30, and the RNN model reached 92.40 and finally the proposed model reached 95.90 respectively. Then 300 instances, the DBN model reached the sensitivity of 77.80, the CNN model reached 87.60 and the RNN model reached as 90.40, and finally, the proposed model reached 95.30 respectively. Then 400 instances, the DBN model reached the sensitivity of 80.10, the CNN model reached 86.40, the RNN model reached 93.20, and finally, the proposed model reached 97.10, respectively. Then 500 instances, DBN model reached the sensitivity as 82.40 and the CNN model reached 86.30, the RNN model reached 92.80, and finally, the proposed model reached 97.40, respectively.

7. Results and Discussion

Water is a scarce and extremely precious resource in both emerging and developed nations. In recent years, international organizations and national governments have taken steps to curb wasteful water usage and limit pollution that renders water unusable. In this study, an LSTM model is used to forecast water quality, with the LSTM's weight chosen using an AFSO optimization procedure. The dataset, known as Kaggle, is publicly available, and three models are utilized for pre-processing. The findings show that the suggested model is superior to the existing methods in terms of accuracy (96%), precision (96%), recall (96%), and F-measure (96%). In the future, we hope to enhance the DL framework by applying swin transform and assess them using real-world data on the metrics to determine if a feature is useful for water quality prediction.

References

[1] Kumar, N., Nandihal, P., Madhumala, R. B., Pareek, P. K., Nikshepa, T., & Sowmya, S. R. (2022, December). A Novel Machine Learning-Based Artificial Voice Box. In *2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)* (pp. 1-7). IEEE.

[2] P. Nandihal, V. Shetty, T. Guha, and P. K. Pareek, "Glioma detection using improved artificial neural network in MRI images," in *Proc. 2022 IEEE 2nd Mysore Subsection Int. Conf. (MysuruCon)*, Oct. 2022, pp. 1-9.

[3] C. Subbalakshmi, P. K. Pareek, and M. V. Narayana, "A gravitational search algorithm study on text summarization using NLP," in *Proc. Int. Conf. Artif. Intell. Data Sci.*, Dec. 2021, pp. 144-159.

[4] S. Prasath Alais Surendhar, G. Ramkumar, R. Prasad, P. K. Pareek, R. Subbiah, A. A. Alarfaj, ..., and R. Raju, "[Retracted] Prediction of Escherichia coli bacterial and coliforms on plants through artificial neural network," *Adv. Mater. Sci. Eng.*, vol. 2022, no. 1, Art. no. 9793790, 2022.

[5] R. Jopate, P. K. Pareek, and A. S. Z. J. Al Hasani, "Prediction of thyroid classes using feature selection of AEHOA based CNN model for healthy lifestyle," *Baghdad Sci. J.*, vol. 21, no. 5 (SI), p. 1786, 2024.

[6] M. G. Divyajyothi, R. Jopate, and R. A. A. Albalushi, "AI precision for irrigation, crop management, and pest control for sustainable agriculture in Oman," in *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1401, no. 1, Art. no. 012005, Oct. 2024.

[7] F. I. Ezema, M. Anusuya, and A. C. Nwanya, Eds., *Materials for Sustainable Energy Storage at the Nanoscale*. CRC Press, 2023.

[8] S. M. Deepa, N. Revathi, K. Sivakami, and P. K. Pareek, "Machine learning based education system with sentiment analysis for students," in *Proc. 2022 IEEE 2nd Mysore Subsection Int. Conf. (MysuruCon)*, Oct. 2022, pp. 1-6.

[9] T. Goswami, D. M. Ganapathi, and P. Goswami, "Soil classification and crop prediction using machine learning techniques," in *Intelligent Robots and Drones for Precision Agriculture*, Cham: Springer Nature Switzerland, 2024, pp. 101-118.

[10] M. G. DivyaJyothi, R. Jopate, R. A. A. Albaushi, and H. A. N. N. Alabri, "Augmented reality based assisted healthcare for enhancing medical rescue and doctor-patient consultations with AR-headset," *Southeast Eur. J. Soft Comput.*, vol. 13, no. 1, pp. 79-85, 2024.

[11] S. Sivakumar, S. Saminathan, R. Ranjana, M. Mohan, and P. K. Pareek, "Malware detection using the machine learning based modified partial swarm optimization approach," in *Proc. 2023 Int. Conf. Appl. Intell. Sustain. Comput. (ICAISC)*, Jun. 2023, pp. 1-5.

[12] J. Lenin, M. D. Jyothi, N. S. H. Alzadjali, and S. A. Azeem, "IoT based emergency handling communication system for medical and traffic rescue teams," in *J. Phys.: Conf. Ser.*, vol. 1964, no. 4, Art. no. 042053, Jul. 2021.

[13] A. Alqahtani, N. Alqahtani, A. A. Alsulami, S. Ojo, P. K. Shukla, S. V. Pandit, ..., and H. S. Khalifa, "Classifying electroencephalogram signals using an innovative and effective machine learning method based on chaotic elephant herding optimum," *Expert Syst.*, Art. no. e13383, 2023.

[14] S. Rani, P. K. Pareek, J. Kaur, M. Chauhan, and P. Bhambri, "Quantum machine learning in healthcare: Developments and challenges," in *Proc. 2023 IEEE Int. Conf. Integrated Circuits Commun. Syst. (ICICACS)*, Feb. 2023, pp. 1-7.

[15] C. Chethana and P. K. Pareek, "Analysis of credit card fraud data using various machine learning methods," in *Big Data, Cloud Computing and IoT*, Chapman and Hall/CRC, 2023, pp. 103-116.

[16] B. Nemade and D. Shah, "An IoT-based efficient water quality prediction system for aquaponics farming," in *Computational Intelligence: Select Proceedings of InCITe 2022*, Singapore: Springer Nature Singapore, 2023, pp. 311-323.

[17] N. S. H. Alzadjali, M. S. Jereesha, C. Savarimuthu, and M. G. Divyajyothi, "A recommender system for Alzheimer patients in Sultanate of Oman using neutrosophic logic," in *Proc. 2020 Int. Conf. Emerging Trends Inf. Technol. Eng. (ic-ETITE)*, Feb. 2020, pp. 1-5.

[18] S. Johri, M. G. Divyajyothi, S. Anitha, M. S. Rani, T. Murari, and N. Shirisha, "A novel deep learning approach for capturing time series dependencies and improving short-term weather forecasting," in *Proc. 2023 Seventh Int. Conf. Image Inf. Process. (ICIIP)*, Nov. 2023, pp. 357-362.

[19] B. Prasad, R. Jopate, P. Savita, A. B. Reddy, B. P. Shankar, and M. S. Arunkumar, "Enhancing IoT-edge computation with data forwarding based decentralized deep neural networks," in *Proc. 2023 Seventh Int. Conf. Image Inf. Process. (ICIIP)*, Nov. 2023, pp. 417-423.

[20] G. Li, Y. Hu, H. Chen, H. Li, M. Hu, Y. Guo, ..., and M. Sun, "Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions," *Appl. Energy*, vol. 185, pp. 846-861, 2017.

[21] C. Fan, Y. Sun, Y. Zhao, M. Song, and J. Wang, "Deep learning-based feature engineering methods for improved building energy prediction," *Appl. Energy*, vol. 240, pp. 35-45, 2019.

[22] M. S. Piscitelli, S. Brandi, A. Capozzoli, and F. Xiao, "A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings," *Building Simulation*, vol. 14, pp. 131-147, Feb. 2021.

[23] M. S. Piscitelli, D. M. Mazzarelli, and A. Capozzoli, "Enhancing operational performance of AHUs through an advanced fault detection and diagnosis process based on temporal association and decision rules," *Energy Build.*, vol. 226, Art. no. 110369, 2020.

[24] L. P. Joseph, R. C. Deo, R. Prasad, S. Salcedo-Sanz, N. Raj, and J. Soar, "Near real-time wind speed forecast model with bidirectional LSTM networks," *Renewable Energy*, vol. 204, pp. 39-58, 2023.

[25] A. M. Kadim and W. R. Saleh, "Morphological and optical properties of CdS quantum dots synthesized with different pH values," *Iraqi J. Sci.*, pp. 1207-1213, 2017.

[26] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712-731, 2007.

[27] S. Romano, J. Bailey, V. Nguyen, and K. Verspoor, "Standardized mutual information for clustering comparisons: one step further in adjustment for chance," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2014, pp. 1143-1151.

[28] J. Brownlee, *Clever Algorithms: Nature-Inspired Programming Recipes*, Jason Brownlee, 2011.

[29] Home Office, *Code of Practice for the Housing and Care of Animals Used in Scientific Procedures*, 1989.

[30] J. Kratochvíl, L. Plch, and E. Koriťáková, "Compliance with ethical rules for scientific publishing in biomedical Open Access journals indexed in Journal Citation Reports," *Vnitřní Lékařství*, vol. 65, no. 5, pp. 338-347, 2019.

[31] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "Performance analysis of the water quality index model for predicting water state using machine learning techniques," *Process Saf. Environ. Prot.*, vol. 169, pp. 808-828, 2023.