Global NEST Journal, Vol 26, No X, pp 1-13

Water quality prediction and classification using attention based deep differential recurflownet with logistic giant armadillo optimization

Venkatraman M.[1], Surendran R.[2], Srinivasulu S.[3] and Vijayakumar K.[4]

[1,2]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, 602105, Tamil Nadu, India

[3]Department of Computer Science and Engineering, School of Computing, Sathyabama Institute of science and Technology, Chennai, 600119, Tamil Nadu, India

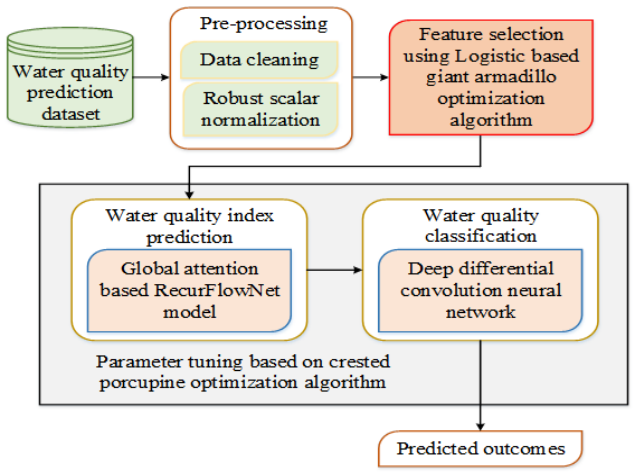
[4]Department of Information Technology, St. Joseph's Institute of Technology, OMR, Chennai, Tamil Nadu, India

Received: 11/09/2024, Accepted: 21/11/2024, Available online: 25/11/2024

*to whom all correspondence should be addressed: e-mail: surendran.phd.it@gmail.com

https://doi.org/10.30955/gnj.006799

Graphical abstract



Abstract

Water quality prediction and classification plays a crucial role in ecosystem sustainability, agriculture, aquaculture and environmental monitoring. The nonlinearity and nonstationarity of water quality are challenging for traditional prediction techniques to adequately capture. The rapid advancement of deep learning in recent decades has made it a hot topic for predicting water quality and classification. In this paper, a new Optimization driven Deep Differential RecurFlowNet (ODD-RecurFlowNet) model with feature selection is proposed for predicting and categorizing the water quality. Preprocessing methods are utilized to evaluate the collected data to predict the water quality class and water quality index. Before deploying feature selection algorithm, preprocessing procedures such as data cleaning and robust scalar normalization are carried out. A logistic based giant armadillo optimization algorithm (GArO) algorithm is used for optimal feature selection. Next, the water quality index is predicted using global attention (GA) based RecurFlowNet model. Subsequently, a deep differential convolution neural network (DDiff-CNN) model is employed for the classification of different levels of water quality. In addition, the hyper-parameters of ODD-RecurFlowNet is tuned using the crested porcupine optimization algorithm (CPoOA). For simulation, python platform is used and the standard water quality dataset from Kaggle library is used to validate the experiment. The finding shows that the proposed ODD-RecurFlowNet model obtains the overall accuracy of 98.01% and RMSE value of 0.039. Thus, the obtained results prove the superiority of proposed model to the existing methods.

Keywords: Water quality index; feature selection; giant armadillo optimization; recurflownet; differential convolution; crested porcupine optimizer; water quality classification

1. Introduction

Water is the fundamental resource that all living things on Earth share, including animals, plants, and humans Gavrilescu, M. (2021). It is required for every kind of creature to survive. Water makes up the majority of the earth's surface and is vital to the existence of every animal and human being Mishra, R.K. (2023). Around 326 cubic meters, equivalent to 71% of the planet's entire area are made up of water, and the remaining 97% is ocean. Only half of the world's potable water is usable; the other two-thirds either reside deep below the planet's surface beyond human access or are locked in icebergs, ice caps on the polar regions, the atmosphere, the ground, or other places Varotsos, C.A *et al*. (2023). Therefore, one of the biggest ecological issues is water quality contamination. To control water pollution and alert consumers when low water quality is detected Venkatraman, M. and Surendran, R. *et al.* (2023), it is imperative to develop a technique for estimating water quality Patel, P.S *et al*. (2023). A balanced

existence on Earth and sustainability are dependent on the forecast of water quality on a worldwide scale Koranga, M. *et al.* 2022.

Conventional methods of water quality estimation mostly depend on a multitude of manual tasks, such as choosing the river's monitoring locations to be examined and periodically gathering data to be sent to a lab for detection and evaluation Ahmed, U. *et al.* (2022). Yet, the conventional methods are error-causing, and early detection is not achievable. Therefore, in recent times, the water quality evaluation requirements have been raised by the expansion of artificial intelligence (AI) and computer technology Hmoud Al-Adhaileh, M. *et al.* (2021). The introduction of Internet of Things (IoT) equipment has made the computations and measurements regarding the overall condition of the water simpler to compute and more accurate Shin, H. *et al*. (2024). Consequently, AI is only a supporting tool for computerizing water quality assessments with the incorporation of IoT devices Wiryasaputra, R. *et al.* (2024). Water consumers can lessen the adverse effects of water quality pollution by becoming aware of abrupt occurrences of pollution through an accurate forecast of metrics for water quality. Recent research has shown that deep learning techniques, with their significant capacities for tracing highly nonlinear connections at an acceptable computation speed, are being extensively used for the prediction of water quality Chhipi-Shrestha, G. *et al.* (2023).

Moreover, time series data-based prediction methods like Long Short-Term Memory (LSTM) Manocha, A. *et al.* (2023). neural networks Islam, M.M. *et al.* (2023). encoder-decoders Jatoi, G.M. *et al.* (2023). and hybrid models have been extensively applied in recent works. But as the amount of information on water quality rises, it becomes more unstable and extremely nonlinear, making good prediction more difficult Baek, S.S. *et al.* (2020). Furthermore, it is maintained that the inability of physical mechanics to adjust for the predicted outcomes of variations in water quality limits the use of deep learning techniques Prasad, D.V.V. *et al.* (2022). There is an absence of information in the justification of the deep learning results for the forecasts of water quality Dodig, A. *et al.* (2024). Therefore, a novel model to water quality prediction and classification is developed for various regions of water resources. The data cleaning and normalization technique is used by the proposed model as a pre-processing step. The metaheuristic optimization algorrithn is employed to select the optimal features and reduce the dimensionality issues Mokarram, M. *et al.* (2024). In addition, optimizer-based hyperparameter tuning is utilized for water quality parameter/index forecast and classify water quality. The key contribution of this proposed model is articulated as: To propose a novel optimization driven deep differential RecurFlowNet (ODD-RecurFlowNet) model to achieve highly accurate water quality metrics prediction and classification Zhao, Z. *et al.* (2024).

To select the optimal features using logistic based giant armadillo optimization algorithm (GArO) algorithm, which uses the advantages chaotic mapping and metaheuristic optimizer for minimizing the dimensionality issues Bi, J., Lin, Y. *et al.* (2021). To introduce a combined global attention (GA) based RecurFlowNet and deep differential convolution neural network (DDiff-CNN) model for predicting the water quality indices and categorizing the levels of water with maximum performance Barzegar, R. *et al.* (2020). To tune the hyper-parameters of the prediction and classification model using crested porcupine optimization algorithm (CPoOA) Ewuzie, U. *et al.* (2022).

Several criteria are crucial for the prediction and classification of water quality include physical parameters includes temperature, turbidity, color, pH value, electrical conductivity. Chemical parameters include dissolved oxygen, biochemical oxygen demand, total dissolved solids, alkalinity, and Chloride (Cl-), Sulfate (SO4 2-), Fluoride (F-). Biological parameters such as total organic carbon, algal biomass, and contamination level Periasamy, S. *et al.* (2024). Evaluating water quality is essential for safeguarding human health and promoting environmental sustainability. Conventional techniques for water quality assessment are often laborious, time-consuming, and susceptible to human error. Motivation for this study Precise and timely water quality evaluation to create a model capable of precisely predicting and classifying water quality in real-time, facilitating prompt intervention and decision-making. Scope of the research aims to design and assess a new deep learning model, the ODD-RecurFlowNet, for predicting and classifying water quality. The key contributions of this research are Novel Model Architecture that proposed ODD-RecurFlowNet architecture effectively captures complex patterns in water quality data, leading to improved prediction and classification accuracy.

The primary objective of this project is to create a reliable and precise deep learning model, the ODD-RecurFlowNet, for the effective prediction and classification of water quality. Utilize sophisticated methodologies such as attention mechanisms, differential recurrent neural networks, and optimization algorithms to achieve the model's objectives. Enhance Water Quality Forecast Precision: To precisely forecast water quality indicators, including pH, temperature, turbidity, and dissolved oxygen levels, using both historical and real-time data.

Improve Water Quality Categorization: To precisely categorize water quality into distinct classifications (e.g., excellent, good, bad, extremely poor) based on forecasted indicators and other pertinent criteria.

A variety of measures, comprising accuracy, recall, mean square error (MSE), R-squared (R2) and others, are utilized to assess the performance of designed model Prasad, D.V.V. *et al.* (2022). The rest of this article is summarized as follows: The recent research works related to this article are explained in Section 2. The proposed methodology and algorithm steps are detailed in Section 3. The research findings, discussion and comparison are provided in Section 4. In Section 5, the article is terminated with the conclusion part.

2. Related works

A brief overview of the latest studies on this issue is given in this section. The reliability and efficiency of deep learning techniques in water quality calculations has recently been demonstrated by researchers. A Deep Neural Network (DNN) framework was built to anticipate water quality index depending on variables chosen for both wet and dry periods across the year, according to research by Bi, J., Lin, Y. *et al.* (2021). The Principal Component Analysis/Factor Analysis (PCA/FA) modelling along with Hierarchical Cluster Analysis (HCA) were used to analyse seasonal variations and the origins of the springs. According to the analysis's findings, the designed DNN model has a high accuracy, low Mean Square Error metric, and a high R-Squared (R2) value El-Shebli. *et al.* (2024). Still, its computational complexity is high. The combined multivariate long and short-term memory-based network (LSTM) was implemented by Wang, J. *et al*. (2024) in semi-arid river basins to anticipate the primary pollutants used in successful water quality monitoring and prediction analytic methodologies. When the forecast period was one day, the best outcomes were attained. Prediction accuracy consistently exceeded 85–89%. The accuracy of this model was very low over the other measures.

A unique gated graph neural network (GGNN) model was proposed by Li, Z. *et al*. (2024) for real-time water quality forecasting in WDNs. To describe the structure and dynamics of the system, the GGNN algorithm incorporates hydraulic flow routes and WQ data. The model was trained using a masking operation for improving the prediction accuracy. But when it comes to increasing forecast accuracy, the sensor's position contributes more than its amount. The water quality time series were predicted using a hybrid prediction technique termed VBAED, according to Bi, J. *et al.* (2024). VBAED was defined as an encoder-decoder framework that combines a bidirectional attention mechanism with BiLSTM and employs VMD as mode decomposition. By lowering the input information nonlinearity and instability, VBAED increases prediction accuracy. However, processing times vary depending on the volume of data.

Zheng, H. *et al*. (2023). developed an accessible deep learning architecture to forecast the spatiotemporal fluctuations of water quality metrics in a significant geographic area of China. The incorporation of socioeconomic as well as land-use indicators with hydrological data might enhance the model's forecast accuracy. The R2 values for the prediction approach were close to 0.80, indicating that the deep learning algorithm performed satisfactorily in the case area when it came to these parameters' predictions Selvanarayanan, R. *et al.* (2024). The Savitzky-Golay (SG) filter method, Variational Mode Decomposition (VMD) model, an Attention mechanism with BiLSTM, an ED framework, and a hybrid algorithm known as Genetic Simulated annealing-based Particle Swarm Optimization (GSPSO) were all combined in the hybrid water quality prediction technique known as SVABEG, which was proposed by Bi, J. *et al*. (2023). The SVABEG attained better accuracy in predicting than the state-of-the-art techniques as shown by experimental findings using real-world datasets. Unbalanced datasets have a higher potential for inaccuracy.

Khullar, S. and Singh, N. (2022). presented a deep learning methods of Bi-LSTM approach (DLBL-WQA) to predict the Yamuna River, India, water quality variables. The proposed approach demonstrated a unique technique that applies optimal loss function to minimalize training error, creates feature maps for a Bi-LSTM design to enhance learning, and adds missing value imputation. Consequently, the suggested model lowers error rates and increases predicting accuracy. However, this research was data dependent, lacks real-time execution, and has a significant computational cost. An improved deep learning method for predicting the WQ index (WQI), which is essential for evaluating the health of water bodies, was presented in the research of Ehteram *et al*. (2024). By efficiently detecting intricate patterns of water quality, this model combines Convolutional Neural Networks (CNN), clockwork Recurrent Neural Networks (CRNN), and M5 Tree methods to improve prediction accuracy. The superiority of the CNN-CRNN-M5T model for both temporal and spatial WQI forecasts in Malaysia in terms of lowering MAE and raising efficiency. Although deep learning advances for predicting water quality are encouraging, real-time evaluations still face difficulties, especially when certain parameters are missing.

A CNN-BiLSTM model was used in the work by Geetha *et al*. (2024). to provide a unique method of river water quality monitoring over the Kaveri River. This approach addresses the rising demand for effective environmental control by using deep learning methods to improve the reliability and efficacy of water quality evaluations. However, there were still issues with handling environmental unpredictability and sensor imperfections, which would distress the accuracy of data and the performance of models. Zamani *et al*. (2024) highlighted the significance of spatiotemporal aspects in water quality monitoring by presenting a hybrid WT-CNN-GRU framework for evaluating reservoir water quality parameters. To improve prediction accuracy, this novel method combines gated recurrent units (GRU) and CNN with wavelet transform (WT). The model efficacy was validated using statistical measures including the correlation coefficient along with Nash-Sutcliffe efficiency, which showed improved performance in both the training as well as testing stages. But in terms of combining various data sources and guaranteeing model resilience in the face of changing environmental circumstances. A unique deep learning ensemble approach to predicting WQ is introduced in the study of Liu *et al*. (2023) highlighting the significance of choosing features and optimization. This strategy is in line with current developments in machine learning methods for environmental monitoring, especially when it comes to evaluating water quality. Hybrid models, includes LSTM networks have shown potential in modelling changes in water quality over 70% accuracy in important metrics. Subsequent studies suggest to concentrate on improving these models for wider use in other environmental scenarios.

Advanced methods for forecasting water quality characteristics utilizing network models and deep learning were investigated in the study of Zamani *et al*. (2023). For efficient management of water resources and pollution control, this strategy was essential. The subsequent segments accentuate pivotal perspectives from important studies that strengthen and elaborate on their results. Models based on deep learning can be less accurate in dynamic situations because of their tendency to have difficulties with irregular data circumstances. This means that models would need to be continuously improved and adjusted. Han, Su, and Na. *et al.* (2023) work employed an enhanced deep learning technique that includes spatiotemporal characteristics to forecast the water quality in the Tanghe Reservoir. In order to progress the accuracy, this method emphasized the use of sophisticated algorithms, particularly a deep network that combines a generative adversarial network layer together with a backpropagation (BP) based neural network layer. The results indicated that the enhanced deep learning technique not only tackles the intricacies of evaluating water quality but also advances better control of water resources, especially in areas where pollution is a problem Santhanaraj. *et al.* (2024).

Problem statement: In recent times, several methods have been employed to predict and classify the water quality, However, the efficiency of existing techniques for assessing and forecasting water quality is hampered by a several issues. The usage of conventional statistical methods, which might not be adequate to capture the intricate, non-linear correlations exist in environmental data, is one of the primary challenges. The deep learning techniques have problems with overfitting or underfitting and this lead to minimize the accuracy. Besides, many existing models produced erroneous forecasts owing to their failure to account for temporal variations in water quality. Moreover, the completeness and quality of the input data can have a big impact on how well these models function since the missing or noisy data might generate biased findings. Thereby, a new deep learning based method needs to be focused to address the limitations of the existing methods by capturing complex patterns in water quality.

3. The proposed model

In this section, a novel ODD-RecurFlowNet model is discussed to detect water quality index and effectively classify the water quality into distinct classes with maximum accuracy. The steps involved in ODD-RecurFlowNet model are data collection, pre-processing, feature selection, water quality index forecast and water quality classification. At first, the data collection stage collects the input data from different sensors. The collected raw input data normally desires to be pre-processed to enhance the data representation using data cleaning and robust scalar normalization. Next, a feature selection method based on logistic based GArO algorithm is utilized to select the features representing the water quality. Then, the selected features are fed to GA based RecurFlowNet model to achieve enhanced prediction. Further, the predicted water quality parameters are used for classifying the water quality using DDiff-CNN model. The hyper-parameters are tuned using an efficient optimization algorithm called CPoOA. The block diagram of ODD-RecurFlowNet model is presented in **Figure 1**.

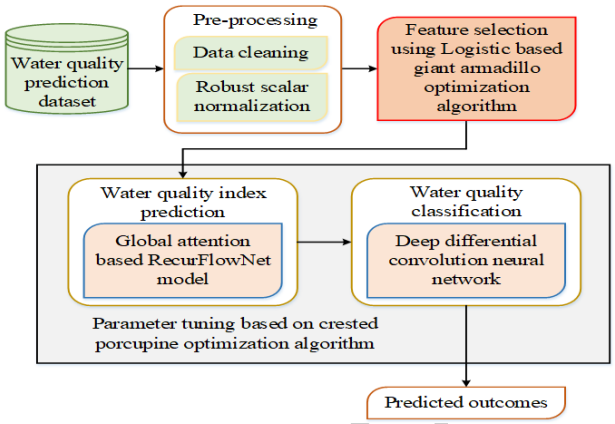


Figure 1. Block diagram of ODD-RecurFlowNet model

3.1. Data pre-processing

Data pre-processing is regarded as a crucial step in the framework of water quality prediction and classification since it supports in preparing the raw data for analysis and modelling. In the proposed ODD-RecurFlowNet method, data pre-processing encompasses data cleaning and robust scaler normalization Sadoune, Hadjer. *et al.* (2023).

Data cleaning: To deal with outliners and missing numbers, the proposed ODD-RecurFlowNet model employs a data cleaning procedure. As the dataset has a marginally greater percentage of missing values, the proposed ODD-RecurFlowNet model preserves the dataset's instances by using the replacement strategy. In order to replace the missing value, the average of five nearest samples present, previous, and next to the missing values is calculated.

Robust scaler normalization: The robust scaler method of normalizing data is same as the min-max normalization method. It scales the data depending on the quintile range, which is the only difference. The below equation resembles the robust scaler, with *y* resembling for the scalar values as well as Q1 and Q3 for the 25th and 75th quantiles, considerably.

$$y = \frac{y_j - Q_1(y)}{Q_3(y) - Q_1(y)} \tag{1}$$

After performing pre-processing, the significant features are selected using the optimization strategy.

3.2. Feature selection

Feature selection is the process of recognizing the ideal subset of features. Developing high-performance models and reducing computational complexity require feature selection strategy. In the proposed ODD-RecurFlowNet model, the Logistic based giant armadillo optimization algorithm (GArO) algorithm is employed to select the optimal features. GArO 38. Alsayyed, Omar. *et al.* (2023) is a bioinspired metaheuristic algorithm designed to mimic the actions of giant, wild armadillos. The main inspiration came from the way giant armadillos hunt, visiting specific regions and excavating termite mounds. There are two

stages to GArO's mathematical modelling and theory: exploration and exploitation. In the exploration stage, the giant armadillos are simulated to be moving in the direction of termite mounts, and in the exploitation stage, they are simulated to be digging in order to discover and destroy termite mounds. The GArO algorithm exhibit strong exploration, exploitation and balancing abilities at the search process. Thereby, the optimization issues can be effectively handled by GArO algorithm. Owing to these benefits, the proposed water quality prediction technique chose logistic based GArO for key selection. However, sometimes the GArO fell into local optimal issues. Therefore, the chaotic logistic map Abdel-Salam. *et al.* (2024) is incorporated with GArO. Giant armadillos are referred to as features in this context. The steps involved in logistic based GArO for optimal feature selection are as follows:

Initialization: Giant armadillos are in charge of populating the GArO algorithm. The values that each member of GArO chooses for the problem's decision variables establish where the problem falls in the problem-solving space. Thereby, the giant armadillo in the population describes a possible solution for the issue characterized by a vector. Using the following equation, the primary position of giant armadillo is randomly initialized at the commencement of algorithm execution.

$$z_{k,d} = LoBo_d + T.(UpBo_d - LoBo_d) \tag{2}$$

where, $z_{j,\,d}$ characterizes the d^{th} dimension of k^{th} GArO member in search space (decision variable), $LoBo_d$ and $UpBo_d$ specifies the lower and upper bounds of d^{th} decision variable, and T designates a random number between [0, 1].

Fitness function:

As each giant armadillo's position in the problem-solving space resembles a potential solution, it is possible to compute each giant armadillo's fitness function value. The typical dimensions of a dataset are $R_f \times R_S$, where R_f indicates the number of features and R_S symbolizes the total number of samples. To accomplish its task, initially the feature selection process splits the entire feature set into smaller subsets (V) whose sum of dimensions is lesser than R_f. The below equation provides the expression of fitness function based on the member of GArO for selecting the feature subsets Surendram R. *et al.* (2023).

$$G_k = \psi \times \varpi_V + (1-\psi) \times \left(\frac{|V|}{R_f}\right) \tag{3}$$

where, G_k characterizes the computed fitness function depending on k^{th} GArO member, ψ has chosen in the interval [0−1] and is applied to balance between $\left(\frac{|V|}{R_f}\right)$ and ϖ_V. $|V|$ indicates the picked features, while ϖ_V exemplifies the classification error. The highest value determined for the fitness function indicates the best candidate solution (optimal member), while the worst value obtained for the fitness function resembles the worst candidate solution (worst member).

Exploration phase (Termite mound attack): In the exploration phase of hunting, the population members' position is updated depending on the giant armadillo's simulated attack on termite mounds. As it gets nearer to the termite mounds, the giant armadillo moves to update the location of population members. The molding of this attack process tends to location change of giant armadillo, which improves the exploration potentiality. For every individual in the population, the set of possible termite mounds is characterized as follows:

$$\begin{aligned} &TMou_k = \{Y_m : G_m < G_k \ and \ m \neq k\}, \\ &Where\, k = 1,2,3..,Q and\, m \in \{1,2,3..,Q\} \end{aligned} \tag{4}$$

where, $TMou_k$ characterizes the set of candidate termite mound positions for k^{th} giant armadillo, Z_m signifies the population member with maximum fitness value than k^{th} giant armadillo, G_m denotes its fitness value. For each population member, the new location is calculated based on the migration of giant armadillo in the direction of termite mounds using the below equation.

$$z_{k,J}^{R1} = z_{k,J} + T_{k,J}.(CTMou_{k,J} - K_{k,J}.z_{k,J}) \tag{5}$$

Then, the previous position of resultant member is interchanged based on the following equation if this new position maximizes the fitness as resembled by the above equation.

$$Z_k = \begin{cases} Z_k^{R1}, & G_k^{R1} \le G_k \\ Z_k, & else \end{cases} \tag{6}$$

where, $CTMou_k$ designates the preferred termite mound for k^{th} giant armadillo, Y_j^{Q1} resembles the new position designed for k^{th} giant armadillo using the attacking phase, $Z_{k,\,l}^{R1}$ represents its k^{th} dimension, Z_k^{R1} characterizes its fitness value, $T_{k,\,l}$ designates the random numbers between [0, 1], and $K_{k,\,l}$ implies the numbers that are randomly picked as 1 or 2

In GArO, the chaotic mapping is used to update the parameter $T_{k,\,l}$. A relatively frequent phenomenon is recognized as chaos in nonlinear systems. One traditional mapping 1D maps is called as logistic mapping Demir, Fahrettin Burak. *et al.* (2024) and it is elected as follows:

$$Z_{p+1} = \mu.Z_p.(1-Z_p) \tag{7}$$

where, μ designates the chaotic factor, $\mu \in (0, 4]$. Now, the expression for parameter updation based on logistic map in GArO is offered below as follows:

$$T_{k,J(u+1)} = 4T_{k,J(u)}.(1-T_{k,J(u)}) \tag{8}$$

here, μ is set to 4.

Exploitation phase (Digging in termite mounds): In the exploitation step, a replication of a giant armadillo breaking into termite mounds to feed on termites is employed to update the population members' location. This minimizes fluctuation in giant armadillo location and maximizes the capability to influence local search. By simulating giant armadillo's ability to excavate in termite mounds, the following equation is utilized to determine a new location.

$$z_{k,l}^{R2} = z_{k,l} + (1-2T_{k,l}).\frac{UpBo_l - LoBo_l}{v} \quad (9)$$

If the value of fitness is enhanced with the expression below, then this new location interchanges the earlier one.

$$Z_k = \begin{cases} Z_k^{R2}, & G_k^{R2} \le G_k \\ Z_k, & else \end{cases} \quad (10)$$

where, Z_k^{R2} characterizes the new location calculated for k^{th} giant armadillo using the digging stage, $Z_{k,\,l}^{R2}$ indicates its l^{th}dimension, G_k^{R2} denotes the value of its fitness function, and v suggests the iteration counter. The pseudocode of GArO is presented in Algorithm 1.

Algorithm 1: Pseudocode of GArO algorithm for feature selection

Start

Initialize the population size (Q),fitness function, dimension, maximum iteration (V)

Determine the initial random population matrix

$z_{k,d} = LoBo_d + T.(UpBo_d - LoBo_d)$

Express the fitness function

For v = 1 to V

For k = 1 to Q

Stage 1: Exploration

Calculate the termite mounds' set for k^{th} GArO member

$TMou_k = \{Y_m : G_m < G_k \ and m \neq k\},$

Select the termite mounds randomly for k^{th} GArO member

Determine new location of k^{th} GArO member using

$z_{k,l}^{R1} = z_{k,l} + T_{k,l}.(CTMou_{k,l} - K_{k,l}.z_{k,l})$

$$Z_k = \begin{cases} Z_k^{R1}, & G_k^{R1} \le G_k \\ Z_k, & else \end{cases}$$

Update k^{th} member of GArO using

Stage 2: Exploitation

Calculate new position of k^{th} GArO member

$z_{k,l}^{R2} = z_{k,l} + (1-2T_{k,l}).\frac{UpBo_l - LoBo_l}{v}$

$$Z_k = \begin{cases} Z_k^{R2}, & G_k^{R2} \le G_k \\ Z_k, & else \end{cases}$$

Update k^{th} member of GArO using

End

Store the best solution acquired so far

End

Output the best solution (optimal features)

End

3.3. Water quality prediction and classification

Water quality prediction encompasses predicting the future values of water quality parameters depending on historical data. The proposed ODD-RecurFlowNet model has employed global attention (GA) based RecurFlowNet to effectively predict the water quality parameters for agricultural purposes Zhao, Jinghua *et al.* (2021). The water quality parameters are significant to evaluate the water suitability for irrigation and ensuring the crops health and soil. After prediction process, the classification process is performed to categorize the water quality into predefined classes. A deep differential convolution neural network (DDiff-CNN) model is used in the proposed ODD-RecurFlowNet model for water quality classification. In addition, the hyper-parameter of the prediction and classification model is tuned using crested porcupine optimization algorithm (CPoOA). The design of water quality prediction and classification is given in **Figure 2**.

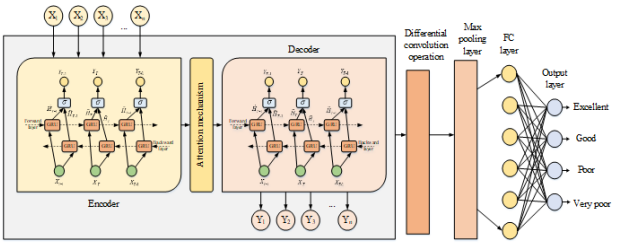


Figure 2. Architecture of water quality prediction and classification

3.3.1. Attention based recurflownet

The attention based RecurFlowNet is a novel architecture that incorporates GA and bidirectional gated recurrent unit (BiGRU) with an encoder-decoder structure. It is designed to effectively acquire the and model sequential data for predicting the water quality parameters. Besides, the RecurFlowNet redirects the recurrent nature of BiGRU She, Daoming. *et al.* (2021) and smooth information flow between encoder and decoder. The ability of the model to process sequential data in backward and forward directions permit it to capture complex temporal patterns in water quality.

The RecurFlowNet architecture is currently based on a learning model being utilized as the most advanced sequence prediction architecture. Variant-length sequences are read and generated by two learning networks, termed as the encoder and the decoder. An encoder–decoder structure is the building block of an attention mechanism Niu, Zhaoyang. *et al.* (2021). From the input, the encoder creates an attention vector, which it then fed to the decoder. The encoder's output is fed into the decoder, which creates a hidden state. The components of encoder and decoder used in RecurFlowNet architecture are BiGRU. Based on the preceding historical data and the temporal context, future data are predicted by the encoder component. The BiGRU encoder's responsibility is to encode the input data such that the context vector (CV) can be generated. A CV is a vector with a fixed length that signifies the incoming data's temporal representation. After decoding the CV, the BiGRU accomplishes prediction. To calculate the probability of prediction sequence, the formula below is employed:

$$P(\omega_1,\omega_2,\omega_3,.....,\omega_\Delta \mid x_1,x_2,x_3,...,x_t) = \prod_{T=1}^{\Delta} P(F_T \mid c_t,\omega_1,\omega_2,\omega_3,.....,\omega_{T-1}) \quad (11)$$

where (x_1, x_2, x_3,…, x_t) represents the input features. The encoder component of encoder-decoder structure compresses every historical information's hidden representations into a CV Feltane, Amal. (2016). The encoding and decoding stages are interfaced by means of the temporal attention layer. By using BiGRU as an encoder, it preserves the hidden internal state of H by

accepting the time series (X_1, X_2, X_3,…, X_t) as input. The GRU reads X_T and updates the hidden mode H_T for each step T in the following ways:

$$H_T = BiGRU(X_T, H_{T-1}) \tag{12}$$

Next, the BiGRU outcome is utilized to construct the temporal context vector c_j in j^{th} decoding phase, which is the weighted sum of encoder network's hidden states and signifies the overall weight of hidden state. The hidden representation of best encoder that map the decoder's attention to these representations are chosen using these vectors. The temporal context attention vector c_j is calculated as follows:

$$c_j = \sum_{T=1}^{t} \beta_{jT} H_T \tag{13}$$

The weight β_{jT} of each hidden state H_T is computed as:

$$\beta_{jT} = \frac{\exp(\phi_{jT})}{\sum_{l=1}^{t} \exp(\phi_{jl})} \tag{14}$$

The below equation designates the general GA computation between s_{j-1} and H_k,

$$\phi_{jT} = A(s_{j-1}, H_k) \tag{15}$$

where, the hidden states of encoder layer and decoder layer are designated by H_k and s_{j-1}, considerably. The architecture of GA based RecurFlowNet is offered in **Figure 3**.

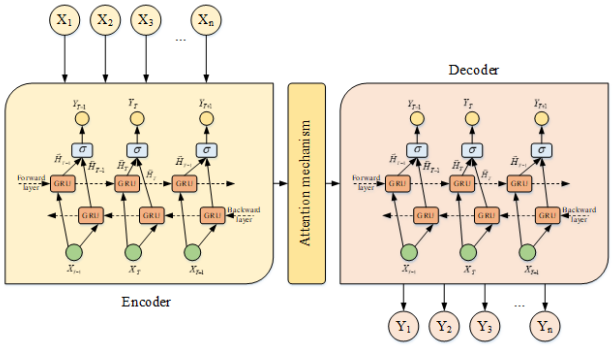


Figure 3. Architecture of GA based RecurFlowNet model

The relationship of the input value at position T and the output value at position j is designated by φ_{jT}, which also specifies the scoring function that is employed to calculate the correlation value. The scoring function in RecurFlowNet is general GA, which is established using the below equation Lei, Baiying. *et al.* (2018).

$$\phi_{jT} = s_{j-1} wm_A H_T \tag{16}$$

where, the scoring function's weight matrix is characterized by wm_A. In order to predict the output at time j, attention weights are designed using Equation (13), which is associated to the time series at time T. Over the input time series, the vector E_{jT}, whose length is t, is utilized as the attention mask. The attention layer's final state is considered as c_j. The E_{jT} vector normalization is accomplished using the softmax function. Together, the decoder and encoder are trained to optimize the sequence of water quality parameter outputs, which are:

$$\underset{\phi}{Arg\min} c = -\sum_{j=1}^{N} \log P(\omega_j \mid x_j; \phi) + \gamma \left\| \phi^2 \right\| \tag{17}$$

where, φ implies the model parameter, comprising weight ω and bias A of each layer, N represents the training dataset size, and γ specifies the regularity of loss function or the significance of penalty. The MSE is employed by the model as the loss function of training process. To prevent overfitting, the early stop is used to end the training procedure if the validation loss stops reducing.

3.3.2. Deep differential convolution network

CNN is a neural network architecture that outperforms the traditional neural networks with its fast and precise method in detection and classification tasks. It has been used to enhance the classification accuracy of several standard databases. Even with improved accuracy, the complexity of convolution network remains challenging. Thereby, the proposed ODD-RecurFlowNet model has employed deep differential convolution neural network (DDiff-CNN) model to classify the water quality. The DDiff-CNN model contains several convolution layers that can analyze water samples' data representation for determining the patters associated to the quality indices. The feature maps are produced in DDiff-CNN by applying a differential factor and pre-defined hyperactivity values. To extract more detail for water classification, the DDiff-CNN utilizes more differential features maps instead of adding more convolution layers or parameters. In compliance with the need of computing approaches, the DDiff-CNN model reduced the convolutional network structures' complexity without losing the values.

Convolution layer: Several pooling layers, convolutional layers, and fully connected (FC) layers are present in basic CNN model. Identifying the local connection features of existing layer is the function of convolution layer. The below expression describes the formula for describing a single output matrix a:

$$B_k = g\left(\sum_{j=1}^{P} J_j * L_{j,k} + Bias_k \right) \tag{18}$$

where, J resembles the input vector, $Bias_k$ indicates the bias value, and L specifies the resultant convolution kernel with a size of $Bias_k \times$ p (p< input size). Next, the sum of every convoluted matrices is calculated. A bias value $Bias_k$ is added to each element of resultant matrix. To create the output matrix B, a non-linear activation function g, operates on each element in the preceding matrix.

Activation function: In order to evaluate the learning rate and CNN's classification performance, a polished linear function is used as activation function. The equation below defines the formula as follows:

$$g(y) = \max(0, y) (\mathrm{Re}LU) \tag{19}$$

In order to minimize the feature maps' fidelity, the pooling layer integrates linguistically related features.

Deep differential convolutional feature map: Convolution is deliberated as the key component of the deep learning architecture, wherein the input predicted water quality indices are passed through several filters. Nevertheless, when there are more feature maps present in the model's feature extraction layers, more number of features are classified. In standard CNN, the feature maps are produced by transferred knowledge or random initialization. By employing pre-defined hyperactive values as well as the differential operator Sarıgül, Mehmet. *et al.* (2019) traditional convolution feature maps are employed to create the feature maps in DDiff-CNN. An addition variation is computed using differential convolution maps to examine the patterns related to water quality indices and their neighbourhood areas. The difference between the data is calculated through mathematical differentiation to account for the sequence change. The difference is computed by using each feature map, as shown in **Figure 4**.

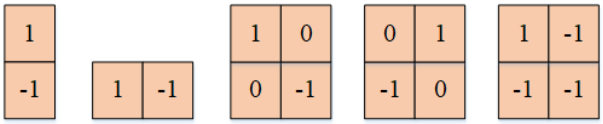


Figure 4. Structure of predefined filters

The difference in one direction is counted for every feature map. Further, the additional feature maps with variations in various directions are obtained. Conversely, in order to extract more features for water quality classification, one static filter is added to the original algorithm. Here, a fixed filter Abdel-Basset. *et al.* (2024) is added and the feature maps are added directly. Consider, the initial feature map produced using classical neural networks is g_1. The five feature maps that occurred by applying the differential operator are g_2, g_3, g_4, g_5 and g_6. Using the below equations, the neurons in these maps are computed.

$$g_{2,j,k} = g_{1,j,k} - g_{1,j+1,k} \tag{20}$$

$$g_{3,j,k} = g_{1,j,k} - g_{1,j,k+1} \tag{21}$$

$$g_{4,j,k} = g_{1,j,k} - g_{1,j+1,k+1} \tag{22}$$

$$g_{5,j,k} = g_{1,j+1,k} - g_{1,j,k+1} \tag{23}$$

$$g_{6,j,k} = g_{1,j+1,k+1} - g_{1,j,k+1} \tag{24}$$

where, j and k indicate the neuron coordinates in the feature maps of convolution. Consider that the size of g_1 is m × n, and size of g_2, g_3, g_4, g_5 and g_6 are $(m-1)\times n$, $m\times(n-1)$, $(m-1)\times(n-1)$, $(m-1)\times(n-1)$, $(m-1)\times(n-1)$ considerably. Using the differential operators, the differential convolutional feature maps are computed from the initial feature map as soon as the first feature map is produced by classic convolution feature map. The feature maps of differential convolution are employed to determine the variations in data. From the above derivation procedure, DDiff-CNN extracts additional details from the data representation of water samples without adding more convolution layers by using more differential feature maps. Therefore, the complexity of convolution network structure is reduced by the DDiff-CNN, which lowers the computing requirements.

3.3.3. Hyper-parameter tuning

The proposed ODD-RecurFlowNet model has used CPoOA to tume the hyper-parameters of the prediction and classification model. CPoOA Islam, Nazrul. *et al.* (2022) is one of a meta-heuristic optimization algorithm inspired by nature that has established to precisely solve a range of optimization problems. The defense tactic of crested porcupine's (CPoP) encourages CPoOA. From least to most aggressive, the crowned porcupine employs sight, odor, sound, and physical attack as its four main defense strategies. The first and second defensive tactics that mimic the CPoP's exploring activity are sight and sound. On the other hand, the third and fourth defense tactics that mimic the exploitative actions of CPoP are the smell and physical assault. Experiments conducted on various test suites demonstrate that the CPoOA can perform competitively and has unique stability qualities on high-dimensional benchmarks and real-world problems. In addition, the CPoOA boosts the efficiency of optimization computations and promotes the continuous advancement and expansion of artificial intelligence applications. It also progresses into a powerful tool for handling difficult problems in the actual world. Thereby, the proposed ODD-RecurFlowNet model has selected CPoOA to tune the hyperparameters of the prediction and classification model. Here, the crested porcupine is considered as the tunable parameter Surendran R. *et al.* (2023) and the physical attack behavior of the exploitation phase is imitated to update the optimal hyperparameter values as follows:

$$Z_j^{u+1} = (Z_{CPoP}^u + (\varphi(1-\upsilon)+\upsilon)\times(\lambda\times Z_{CPoP}^u - Z_j^u) - \upsilon\times\lambda\times\delta_u\times H_j^u \tag{25}$$

where, Z_j^u characterizes the location of j^{th} individual at iteration u (resembles the predator position), υ designates a random value between [0,1], ϕ states a convergence speed factor, λ implies the parameter employed to manage the search direction, H_j^u states the average force of CPoP that affected the j^{th} predator and $Z_{CPoP}{}^u$ suggests the best-attained solution and describes the CPoP.

4. Results and discussion

The experimental evaluation to compare the results of proposed ODD-RecurFlowNet model with other methods is covered in this section. The graphical representation of the simulated results, the performance evaluation, and the comparison are provided in the subsections below. The propose method's performances are examined using the PYTHON platform. The proposed ODD-RecurFlowNet model has used water quality dataset acquired from Kaggle repository for experimentation. For accomplishing the research, the data collected have been derived from various locations in India, which included 1679 samples taken from 666 lake and river foundations. Data for the dataset has gathered from the year 2005 to 2014. Eight significant variables such as nitrate, pH, temperature, DO, fecal coliform, total coliform, and conductivity are

comprised in the dataset. The class distribution of the dataset used in proposed ODD-RecurFlowNet model is given in **Table 1**.

Table 1. Investigation of class distribution

Name of class	Total samples
Good	726
Poor	273
Excellent	324
Very poor	355
Total count	1678

Further, the collected data from dataset is divided into 70% and 30% for training and testing. The ODD-RecurFlowNet is set up with an Intel CoreI i5-4760S CPU @ 3.10 GHz processor with 16.00 GB of main memory running 64-bit Windows 10 operating system. **Table 2** shows the proposed model's parameter configuration. In this case, the proposed ODD-RecurFlowNet model uses the size of the epoch, which is set to 25, and the hyper-parameters are efficiently tuned using crested porcupine optimization algorithm (CPoOA). The hyper-parameter setup of the proposed ODD-RecurFlowNet model is offered in **Table 2**.

Table 2. Hyper-parameter setup of ODD-RecurFlowNet model

Parameter	Values
Maximum epoch	25
Drop out	0.2
Learning rate	0.01
Optimizer	CPoOA

4.1. Performance indicators

The proposed ODD-RecurFlowNet model has used deep learning based regression and classification model for the prediction and classification of water quality. In this subsection, the performance metrics like MSE, root MSE (RMSE), R-squared error (R2), precision, accuracy, recall, and f1-score, are considered to assess the performance of proposed ODD-RecurFlowNet model.

Mean square error: The average squared difference between the predicted and actual values is dignified by the MSE metric. It is more widely utilized and highlights greater errors. The equation MSE is given below:

$$MSE = \frac{1}{p}\sum_{j=1}^{L}(z_j - \tilde{z}_j)^2 \tag{26}$$

where, L represents the number of samples, z_j indicates the actual value of j^{th} sample, and $\hat{Z}_j$ resembles the expected value of j^{th} sample.

Root mean square error: RMSE takes the square root of RMSE in order to obtain an understandable metric in the same unit as the dependent variable. The RMSE equation is given below as follows:

$$MSE = \sqrt{\frac{1}{p}\sum_{j=1}^{L}(z_j - \tilde{z}_j)^2} \tag{27}$$

R-squared error: The coefficient of determination or R2 is deliberated as a statistical measure that computes the variance portion in the dependent variable described by the independent variables in a regression model. It gives a suggestion of how well the data fit the regression model. The expression of R2 is given below as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{L}(z_j - \bar{z}_j)^2}{\sum_{j=1}^{L}(z_j - \tilde{z}_j)^2} \tag{28}$$

where, $\tilde{Z}_j$ signifies the mean of actual values.

Accuracy: One metric that's generally utilized for classification tasks is accuracy. It evaluates the percentage of accurate predictions among all the predictions made by a model. This formula is used to calculate accuracy is given below as follows:

$$Accuracy = \frac{No.of\ correct\ prediction}{Total\ no.of\ prediction} \tag{29}$$

Precision: The ratio of true positive (TP) estimates to every positive prediction a model makes is termed as precision. By computing the ratio of anticipated positive values that are really positive, it assesses the accuracy of positive predictions. The equation of precision is given below as follows:

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{30}$$

where, *FP* indicates false positive.

Recall: The proportion of TP predictions to every actual positive values is measured by a performance measure called recall. It evaluates how well the proposed classification model can

distinguish among all of the actually positive cases. The expression of recall is given below as follows:

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{31}$$

where, *FN* specifies false negative.

F1-score: By computing the harmonic mean of recall and precision, the F1 score delivers a fair assessment of model's performance. It is mainly useful if there is an unequal distribution between the negative and positive classes since this statistic takes both the model's capability to capture all positive instances (recall) and the proficiency to reliably identify positive instances (precision). The expression of F1-score is given below as follows:

$$F1 - score = 2 \times \frac{\mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \tag{32}$$

4.2. Analysis in terms of MSE, RMSE and R2 metrics

In this section, the results of ODD-RecurFlowNet are assessed to determine the efficacy of watesr quality index prediction. A proportionate results analysis of ODD-RecurFlowNet model with existing models in terms of RMSE is exposed in **Figure 5**. The existing modes such as feedforward neural network (FFNN), artificial neural network (ANN), random forest (RF), logistic regression (LR), polynomial regression (PR), support vector machine (SVM), gradient boosting (GB), and optimal stacked BiLSTM (OSBiGRU) [48]. According to the figure, the proposed ODD-RecurFlowNet model produced effective results with the lowest RMSE value of 0.039, while the ANN, FFNN, and OSBiGRU models recognized somewhat enhanced RMSE of 0.7158, 0.5967, and 0.0436, respectively. These values

guaranteed the proposed ODD-RecurFlowNet model's effective outcomes on water quality index prediction.

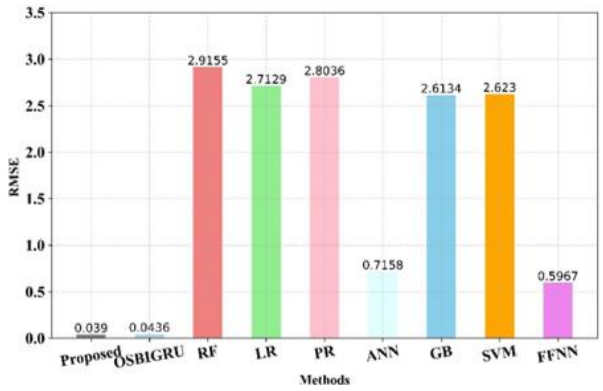


Figure 5. Analysis of RMSE for the proposed and existing methods

A comparative analysis of ODD-RecurFlowNet with other models in terms of R2 is shown in **Figure 6**. From the graphical representation, it is perceived that the proposed ODD-RecurFlowNet model has a maximum R2 value of 98.3% showing better performance than the ANN, FFNN, PR, LR, GB, RF, SVM models, and OSBiGRU, which exhibit slightly lower R2 of 89.87%, 83.66%, 82.50%, 91.17%, 86.44%, 84%, 80.93%, and 96.48%, respectively. Among the existing methods, OSBiGRU has achieved close to the proposed ODD-RecurFlowNet model. The usage of enhanced architecture in the proposed ODD-RecurFlowNet model effectively capture the temporal dependencies in water quality data. Besides, the use of CPoOA for hyper-parameter tuning permits the model to boost the performance.

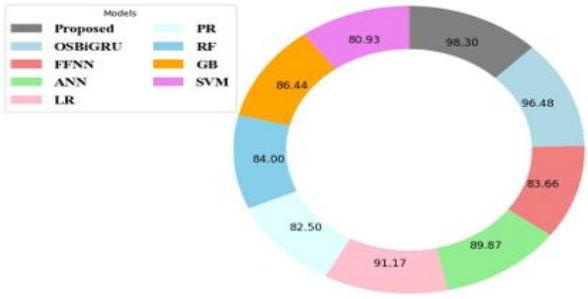


Figure 6. Analysis of RMSE for the proposed and existing methods

Similarly, when considering the MSE metric, the proposed ODD-RecurFlowNet model has demonstrated superior performance against the existing methods. The MSE value achieved by the proposed ODD-RecurFlowNet model is 0.0014 and it significantly outperforms the state-of-the-art methods such as FFNN, ANN, LR, RF, PR, SVM, GB, and OSBiGRU. The existing OSBiGRU records lower MSE value of 0.0019, which is close to the proposed ODD-RecurFlowNet model. The traditional ANN and FFNN attained 0.5123 and 0.356. SVM and LR exhibited maximum MSE values of 6.88 and 7.36. The RF, PR and GB have attained 8.5, 7.86 and 6.83 of MSE value. Other existing models including RF, PR and GB have attained 8.5, 7.86 and 6.83 of MSE value, demonstrating that they are less effective in predicting the quality index of water. The significantly lower MSE of proposed ODD-RecurFlowNet model highlights its efficacy in acquiring the underlying pattern in data, creating it more reliable for evaluating the water quality over the existing methods. This performance underscores the benefits of advanced deep learning methods in environmental monitoring applications.

4.3. Analysis in terms of accuracy, precision, recall and f1-score

In this section, the performance of the classification model is accessed in terms of accuracy, recall, precision, and f1-score. The existing methods such as multi-layer perceptron (MLP), stochastic gradient descent (SGD), k-nearest neighbor (KNN), and decision tree (DT) gaussian naïve bayes (GNB), and artificial ecosystem optimization with improved elman neural network (AEO-IENN) are used for comparison. A confusion matrix is deliberated as a statistic utilized to assess the accuracy of classification model. In contrast to the actual outcomes, a visual representation of the model's prediction is provided in the confusion matrix. This permits to analyze the model performance across different classes. Typically, the confusion matrix comprises of four components such as TP, FP, true negative (TN), and FN.

Figure 7. Analysis of confusion matrix

Figure 7 provides the confusion matrix of proposed ODD-RecurFlowNet model. The confusion matrix in the graphical representation reveals how the samples are accurately classified into good, excellent, poor and very poor. Here, the correctly predicted values are presented along the diagonal of matrix, and other values indicates the misclassification, where the samples are wrongly dispensed to various category. On the total of 99 excellent samples, one is misclassified as poor and very poor. Considering the 220 good samples, 217 are correctly classified and only three is misclassified as very poor. The poor class has a total of 84 samples among that one is misclassified as poor. In the same way, the very poor class has a total of 108 samples among that 106 samples are correctly identified as very poor class and two is misclassified as good.

The comparative analysis of ODD-RecurFlowNet with the dominant models in terms of accuracy is exposed in **Figure**

8. The graphical depiction suggests that the existing methods of the SGD, LR, and DT representations has been reported to be lower, with respective accuracy rates of 83.44%, 84.57%, and 84.35%. Followed by this, KNN has established a marginally superior outcome with an accuracy of 86.81%. The GNB and MLP models produced results with respectable accuracy of 90.98% and 90.46%, respectively. The proposed ODD-RecurFlowNet model has demonstrated superior performance, with a maximum accuracy of 98.01%. When compared to other existing models, AEO-IENN has attained accuracy value of 97.42%, which is near to the proposed ODD-RecurFlowNet model.

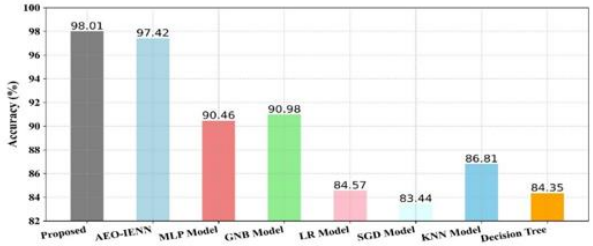


Figure 8. Analysis of accuracy for the proposed and existing methods

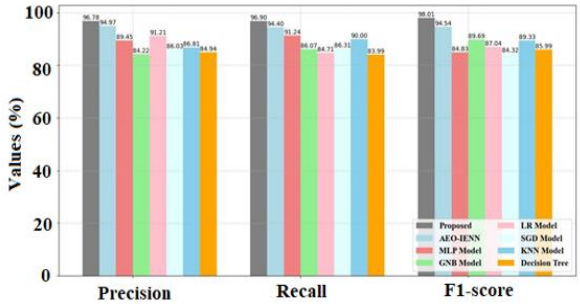
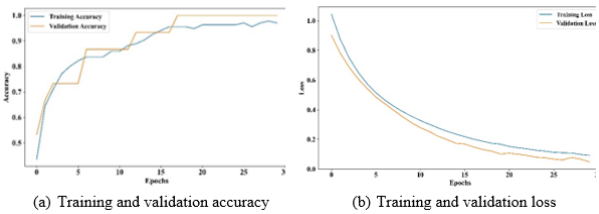


Figure 9. Analysis of precision, recall and f1-score for the proposed and existing methods

Figure 9 portrays a thorough analysis of the proposed and most recent models in terms of several metrics such as recall, precision, and f1-score. In comparison to other models, the implication is that the SGD, MLP, and DT representations have revealed the least values. Besides, there has been a minor improvement in the classifier outcomes of LR model. In addition, there has been a moderate improvement in classification performance between the KNN and GNB models. On the other hand, the proposed ODD-RecurFlowNet model has demonstrated significantly greater values with a maximum recall of 96.9%, precision of 96.78%, and an f1-score of 98.01%. The existing AEO-IENN also achieved better performance with recall of 94.4%, precision of 94.97%, and f1-score of 94.54%, which are slightly lower to the proposed ODD-RecurFlowNet model. Overall, when compared to alternative methods, the proposed ODD-RecurFlowNet model has achieved favorable results with regard to the classification of water quality.

4.4. Analysis in terms of accuracy-loss curve

Initially, the performance obtained by the proposed ODD-RecurFlowNet for water quality prediction and classification is compared with various existing studies in terms of difference evaluation metrics. In this section, the detailed analysis of experimental outcome in terms of accuracy and loss based on training and validation data is designated. **Figure 10** exemplifies the accuracy-loss curve of proposed ODD-RecurFlowNet model. In the accuracy curve, the analysis is performed with training and validation data, as exposed in Figures. The findings ensured that accuracy increases with increasing epochs. Furthermore, it appears that training accuracy is superior to testing accuracy. Considering the loss curve, loss percentage of proposed ODD-RecurFlowNet model is highly minimized, and if the number of epochs is maximized, the resultant value is very low.



(a) Training and validation accuracy (b) Training and validation loss

Figure 10. Analysis of accuracy and loss

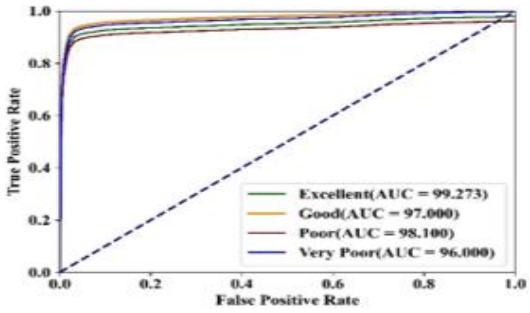


Figure 11. Analysis of ROC

4.5. ROC analysis

The receiver operating characteristic analysis or ROC analysis is employed in the proposed ODD-RecurFlowNet model to determine the performance of classification by plotting TP rate against the FP rate. This performance is evaluated by the area under the curve (AUC), which has values between 0 and 1. An AUC of 0.5 implies no discrimination while an AUC of 1 describes perfect classification. The ROC evaluation of proposed ODD-RecurFlowNet model for water quality classification is exposed in **Figure 11**. The graphical depiction suggested that the samples has been effectively classified into four distinct classes with maximum ROC values by the proposed classification model due to the use of differential features with optimal parameter tuning.

5. Conclusion

A unique ODD-RecurFlowNet model for accurately classifying water quality by using predicted water quality index is proposed in this paper. The ODD-RecurFlowNet model automatically and successfully recognizes water quality by selecting the best features and adjusting classifier hyperparameters. The ODD-RecurFlowNet use logistic basedGArO to select important features. While the GA based RecurFlowNet network handles water quality

index prediction, DDiff-CNN classifies the water quality as good, excellent, poor and very poor. Furthermore, CPoOA is applied to optimize the model's hyperparameters. The Python platform is used to train and assess the model using the publicly available dataset from kaggle repository. On the water quality prediction dataset, the ODD-RecurFlowNet model's overall accuracy is 98.01%, and R2 of 98.3% respectively. Despite achieving a high performance and maximum accuracy, the ODD-RecurFlowNet model necessitates greater computational resources because of its complexity during both the training and testing stages. Besides, the diversity of real-world scenarios cannot be adequately reflected in the datasets used to test the model's performance. Future direction focus on hybrid feature selection techniques that integrate filter-based, wrapper-based, and embedding approaches to identify the most informative features. Gather statistical information from various geographical regions and environmental contexts to enhance model predictability.

References

Abdel-Basset M., Mohamed R. and Abouhawwash M. (2024). Crested Porcupine Optimizer: A new nature-inspired metaheuristic. *Knowledge-Based Systems*, **284** 111257.

Abdel-Salam M., Hu G., Çelik E., Gharehchopogh F.S. and El-Hasnony I.M. (2024). Chaotic RIME optimization algorithm with adaptive mutualism for feature selection problems. *Computers in Biology and Medicine*, **179**, 108803.

Ahmed U., Mumtaz R., Anwar H., Mumtaz S. and Qamar A.M. (2020). Water quality monitoring: from conventional to emerging technologies. *Water Supply*, **20**(1), 28–45.

Alsayyed O., Hamadneh T., Al-Tarawneh H., Alqudah M., Gochhait S., Leonova I., Malik O.P. and Dehghani M. (2023). Giant Armadillo optimization: A new bio-inspired metaheuristic algorithm for solving optimization problems. *Biomimetics*, **8**, 619.

Baek S.S., Pyo J. and Chun J.A. (2020). Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water*, **12**(12), 3399.

Barzegar R., Aalami M.T. and Adamowski J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*, **34**(2), 415–433.

Bi J., Chen Z., Yuan H. and Zhang J. (2024). Accurate water quality prediction with attention-based bidirectional LSTM and encoder–decoder. *Expert Systems with Applications*, **238**, 121807.

Bi J., Lin Y., Dong Q., Yuan H. and Zhou M. (2021). Large-scale water quality prediction with integrated deep neural network. *Information Sciences*, **571**, 191–205.

Bi J., Lin Y., Dong Q., Yuan H. and Zhou M. (2021). Large-scale water quality prediction with integrated deep neural network. *Information Sciences*, **571**, 191–205.

Bi J., Zhang L., Yuan H. and Zhang J. (2023). Multi-indicator water quality prediction with attention-assisted bidirectional LSTM and encoder-decoder. *Information Sciences*, **625**, 65–80.

Chhipi-Shrestha G., Mian H.R., Mohammadiun S., Rodriguez M., Hewage K. and Sadiq R. (2023). Digital water: artificial intelligence and soft computing applications for drinking water quality assessment. *Clean Technologies and Environmental Policy*, **25**(5), 1409–1438.

Demir F.B., Tuncer T. and Kocamaz A.F. (2020). A chaotic optimization method based on logistic-sine map for numerical function optimization. *Neural Computing and Applications*, **32**, 14227–14239.

Dodig A., Ricci E., Kvascev G. and Stojkovic M. (2024). A novel machine learning-based framework for the water quality parameters prediction using hybrid long short-term memory and locally weighted scatterplot smoothing methods. *Journal of Hydroinformatics*, **26**(5), 1059–1079.

Ehteram M., Ahmed A.N., Sherif M. and El-Shafie A. (2024). An advanced deep learning model for predicting water quality index. E*cological Indicators*, **160**, 111806.

El-Shebli M., Sharrab Y. and Al-Fraihat D. (2024). Prediction and modeling of water quality using deep neural networks. Environment, *Development and Sustainability*, **26**(5), 11397–11430.

Ewuzie U., Bolade O.P. and Egbedina A.O. (2022). Application of deep learning and machine learning methods in water quality modeling and prediction: a review. *Current trends and advances in computer-aided intelligent environmental data engineering*, 185–218.

Feltane A. (2016). Time-frequency based methods for nonstationary signal analysis with application to EEG signals. *University of Rhode Island*., **8**, 1–27.

Gavrilescu M. (2021). Water, soil, and plants interactions in a threatened environfment. *Water*, **13**(19), 2746.

Geetha T.S., Chellaswamy C., Raja E. and Venkatachalam K. (2024). Deep learning for river water quality monitoring: a CNN-BiLSTM approach along the Kaveri River. *Sustainable Water Resources Management*, **10**(3), 1–15.

Han M., Su Z. and Na X. (2023). Predict water quality using an improved deep learning method based on spatiotemporal feature correlated: a case study of the Tanghe Reservoir in China. *Stochastic Environmental Research and Risk Assessment*, **37**(7), 2563–2575.

Hmoud Al-Adhaileh M. and Waselallah Alsaade F. (2021). Modelling and prediction of water quality by using artificial intelligence. *Sustainability*, **13**(8), .4259.

Islam M.M., Kashem M.A., Alyami S.A. and Moni M.A. (2023). Monitoring water quality metrics of ponds with IoT sensors and machine learning to predict fish species survival. *Microprocessors and Microsystems*, **102**, 104930.

Islam N. and Irshad K. (2022). Artificial ecosystem optimization with deep learning enabled water quality prediction and classification model. *Chemosphere*, **309**, 136615.

Jatoi, G.M., Rahu, M.A., Karim, S., Ali, S.M. and Sohu, N., (2023). Water Quality Monitoring in Agriculture: Applications, Challenges and Future Prospectus with IoT and Machine Learning. *Journal of Applied Engineering & Technology* (JAET), **7**(2), 46–54.

Khullar S. and Singh N. (2022). Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. *Environmental Science and Pollution Research*, **29**(9), 12875–12889.

Koranga M., Pant P., Kumar T., Pant D., Bhatt A.K. and Pant R.P. (2022). Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand. *Materials today: proceedings*, **57**, 1706–1712.

Kumarasamy S.R., Rajendran S., Romero C.A.T. and Murugaraj S.S. (2023). Internet of Things Enabled Energy Aware Metaheuristic Clustering for Real Time Disaster Management. *Comput. Syst. Sci. Eng*. **45**(2) 1561–1576.

Lei B., Huang S., Li R., Bian C., Li H., Chou Y.H. and Cheng J.Z. (2018). Segmentation of breast anatomy for automated whole breast ultrasound images with boundary regularized convolutional encoder–decoder network. *Neurocomputing*, **321**, 178–186.

Li Z., Liu H., Zhang C. and Fu G. (2024). Real-time water quality prediction in water distribution networks using graph neural networks with sparse monitoring data. *Water Research*, **250**, 121018.

Liu W., Liu T., Liu Z., Luo H. and Pei H. (2023). A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction. *Environmental Research*, **224**, 115560.

Manocha A., Sood S.K. and Bhatia M. (2023). Artificial intelligence-assisted water quality index determination for healthcare. *Artificial Intelligence Review*, **56**(Suppl 2), 2893–2915.

Mishra R.K. (2023). Fresh water availability and its global challenge. *British Journal of Multidisciplinary and Advanced Studies*, **4**(3), 1–78.

Mokarram M., Pourghasemi H.R. and Pham T.M. (2024). Enhancing water quality monitoring through the integration of deep learning neural networks and fuzzy method. *Marine Pollution Bulletin*, **206**, 116698.

Niu Z., Zhong G. and Yu H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, **452**, 48–62.

Patel P.S., Pandya D.M. and Shah M., (2023). A systematic and comparative study of Water Quality Index (WQI) for groundwater quality analysis and assessment. *Environmental Science and Pollution Research*, **30**(19), 54303–54323.

Periasamy S., Subramanian P. and Rajendran S. (2024). An intelligent air quality monitoring system using quality indicators and transfer learning based lightweight recurrent network with skip connection. *Global Nest*, **1**.18.

Prasad D.V.V., Venkataramana L.Y., Kumar P.S., Prasannamedha G., Harshana S., Srividya S.J., Harrinei K. and Indraganti S. (2022). Analysis and prediction of water quality using deep learning and auto deep learning techniques. *Science of the Total Environment*, **821**, 153311.

Prasad D.V.V., Venkataramana L.Y., Kumar P.S., Prasannamedha G., Harshana S., Srividya S.J., Harrinei K. and Indraganti S. (2022). Analysis and prediction of water quality using deep learning and auto deep learning techniques. *Science of the Total Environment*, **821**, 153311.

Sadoune H., Rihani R. and Marra F.S. (2023). DNN model development of biogas production from an anaerobic wastewater treatment plant using Bayesian hyperparameter optimization. *Chemical Engineering Journal*, **471**, 144671.

Sarıgül M., Ozyildirim B.M. and Avci M. (2019). Differential convolutional neural network. *Neural Networks*, **116**, 279–287.

Selvanarayanan R., Rajendran S., Algburi S., Ibrahim Khalaf O. and Hamam H. (2024). Empowering coffee farming using counterfactual recommendation based RNN driven IoT integrated soil quality command system. *Scientific Reports*, **14**(1), 6269.

She D. and Jia M. (2021). A BiGRU method for remaining useful life prediction of machinery. *Measurement*, **167**, 108277.

Shin H., Byun Y., Kang S., Shim H., Oak S., Ryu Y., Kim H. and Jung N. (2024). Development of water quality prediction model for water treatment plant using artificial intelligence algorithms. *Environmental Engineering Research*, **29**(2).

Surendran R., Alotaibi Y. and Subahi A.F. (2023). Lens-Oppositional Wild Geese Optimization Based Clustering Scheme for Wireless Sensor Networks Assists Real Time Disaster Management. Comput. Syst. Sci. Eng., **46**(1), 835–851.

Surendran R., Alotaibi Y. and Subahi A.F. (2023). Wind Speed Prediction Using Chicken Swarm Optimization with Deep Learning Model. *Computer Systems Science & Engineering* **46**(3), 1–19.

Varotsos C.A., Krapivin V.F., Mkrtchyan F.A. and Xue Y. (2023). Global Water Balance and Pollution of Water Reservoirs. In Constructive processing of microwave and optical data for hydrogeochemical applications (119–161). Cham: Springer International Publishing.

Venkatraman M. and Surendran R. (2023). Aquaponics and Smart Hydroponics Systems Water Recirculation Using Machine Learning. In 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), 998–1004.

Wang J., Xue B., Wang Y., Yinglan A., Wang G. and Han D. (2024). Identification of pollution source and prediction of water quality based on deep learning techniques. *Journal of Contaminant Hydrology*, **261**, 104287.

Wiryasaputra R., Huang C.Y., Lin Y.J. and Yang C.T. (2024). An IoT Real-Time Potable Water Quality Monitoring and Prediction Model Based on Cloud Computing Architecture. *Sensors*, **24**(4), 1180.

Zamani M.G., Nikoo M.R., Al-Rawas G., Nazari R., Rastad D. and Gandomi A.H. (2024). Hybrid WT–CNN–GRU-based model for the estimation of reservoir water quality variables considering spatio-temporal features. *Journal of Environmental Management*, **358**, 120756.

Zamani M.G., Nikoo M.R., Jahanshahi S., Barzegar R. and Meydani A. (2023). Forecasting water quality variable using deep learning and weighted averaging ensemble models. *Environmental Science and Pollution Research*, **30**(59), 124316–124340.

Zhao J., Zeng D., Liang S., Kang H. and Liu Q. (2021). Prediction model for stock price trend based on recurrent neural network. *Journal of Ambient Intelligence and Humanized Computing*, **12**, 745–753.

Zhao Z., Fan B. and Zhou Y. (2024). An Efficient Water Quality Prediction and Assessment Method Based on the Improved Deep Belief Network—Long Short-Term Memory Model. *Water*, **16**(10), 1362.

Zheng H., Liu Y., Wan W., Zhao J. and Xie G. (2023). Large-scale prediction of stream water quality using an interpretable deep learning approach. *Journal of environmental management*, **331**, 117309.