https://doi.org/10.56452/2022-11-002

ISSN: 0974-5823 Vol. 7 No. 07 November, 2022

International Journal of Mechanical Engineering

Analysing And Modelling Dissolved Oxygen Concentration Using Deep Learning Architectures

Jitha P Nair*
Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India, [0000-0002-0962-8393]

Vijaya M S
Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India, **2[0000-0002-4623-5572]**

ABSTRACT
Accurate dissolved oxygen concentration assessment is vital for a variety of environmental applications like water quality prediction. The complicated connections between many processes that impact dissolved oxygen (DO) concentration in flowing water and the challenge of applying process-based water quality models, build modelling DO concentration in running water challenging. This study employs deep learning algorithms like Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) an effective prediction model for forecasting DO levels in river water. The models are developed and validated using the river water quality data collected from eleven sampling stations during the year 2016 to 2020. Parameters such as water temperature, specific conductance, pH, biological oxygen demand (BOD) and chemical oxygen demand (COD), are used as independent input variables for training the model wherein the top predictors are water temperature and pH. The deep learning learning-based on model is developed using LSTM and RNN to predict dissolved oxygen concentration. The performance of the LSTM and RNN based DO prediction models are compared with traditional machine learning approaches such as support vector regressor, random forest, linear regression, and MLP regressor. The experimental results show the best prediction accuracy for recurrent neural networks-based DO prediction models than other algorithms.

Keywords: DO concentration, Deep learning, Machine learning, Water quality index, Prediction. LSTM, RNN.

1. INTRODUCTION

In addition to being essential for human survival, water is vital to the expansion of human culture through supporting economic growth. The availability of clean water is intricately connected to the health of the environment and the pursuit of other activities. Every living thing on Earth requires access to water and oxygen in order to exist. Several water parameters are used to predict DO in the present investigation.

The deterioration of surface water quality due to an increase in pollution loads impacts aquatic ecosystems globally [1,2]. An adequate level of dissolved oxygen, i.e., 5 mg/L, is required for the existence of diverse aquatic life. Dissolved oxygen is the most essential measure for analysing the effects of contaminants on river water [3,4]. Primarily in rivers, dissolved oxygen is primarily caused by the interaction of air and water in the atmosphere reaeration, photosynthesis of aquatic plants [5] and denitrification. Degradation of organic matter, nitrification, and aquatic respiration are examples of dissolved oxygen sinks [6]. Natural factors such as ambient water temperature, salt content, and river water level are human effects such as agricultural activity and urban sprawl, and determine the amount of dissolved oxygen in the water [7]. The interaction of natural and human effects on DO makes it difficult for researchers and water resource

managers to identify and quantify individual DO-related processes in rivers.

Reaeration and deoxygenation processes were coupled in a simple prediction equation in the 1920s, and field data from the Ohio River was used to validate it [8]. Since then, a lot of work has gone into building several water quality models that include DO as an inherent component [9,10]. These mechanistic models typically make use of modified versions of the original Streeter-Phelps equation and are based on mass-balance theory, which includes the DO interacting processes [11]. The majority of these models require input data for specific DO-affecting processes in order to undertake simulations and validations.
This approach typically necessitates a certain level of comprehension of the DO-related interactions and is computationally demanding. The model's application is constrained because the relevant data sets are frequently unavailable. Despite the development of numerous DO models for aquatic systems, proper model implementation was susceptible to change due to site-specific natural and human factors, as well as a lack of suitable data sets [12].

Data-driven modelling, a distinct method, has been dubbed the fourth paradigm in scientific discovery [13]. In contrast to process-based models, a data-driven model was created by feeding relevant data from the modelled system into machine learning algorithms [14].

Data-driven methods have gained traction in recent years, according to a review of the literature on water quality prediction [15]. Numerous studies on water quality modelling have utilized ANNs and variants of them, including the prediction of DO in reservoirs and rivers. Regression and principal component analysis were also utilized in conjunction with diffused oxygen modelling and prediction [16,17].

Using neuro-fuzzy techniques and fuzzy logic, another data-driven modelling method has been used to predict DO in riverine systems [18]. Using neural networks, hybrid models were also used to predict DO in reservoirs [19,20]. Predicting DO in rivers using water quality variables and a support vector machine (SVM) was another method of data-driven modelling [21]. A recurrent neural network is a type of artificial neural network (ANN) used to forecast sequential or time-series data. A type of RNN called Long Short-Term Memory can learn long sequences. The advantage of LSTM is that it can handle enormous datasets in a short amount of time. The main difference between RNN and LSTM is in the input and output layer workings, with LSTM executing quicker than RNN for large datasets.

Dissolved oxygen concentrations are employed as a measure of water health since they are linked to high production and low pollution. One of the most crucial water quality indicators is dissolved oxygen. To remain viable, fish and other aquatic creatures require dissolved oxygen, b Because of the aerating function, winds dissolve oxygen in surface water. Aquatic plants release oxygen into the water as a result of photosynthesis. The dissolved oxygen concentration is critical when predicting water quality.

This study is aimed to build a DO concentration prediction model and to assess the efficiency of deep learning algorithms such as long short-term memory, and recurrent neural networks, in predicting dissolved oxygen concentration in river water. Data collected from eleven different monitoring stations of the Bhavani River which flows into two states, Kerala and Tamilnadu have been used to train the predictive model. The deep learning-based predictive models are evaluated for their efficiency in predicting the dissolved oxygen concentration.

2. COLLECTION OF DATA AND EXPLORATORY DATA ANALYSIS

The Bhavani River is taken as the field of study for predicting dissolved oxygen concentration. The river flows through Tamil Nadu and Kerala, it begins at Nilgiri Hills and enters the Silent Valley National Park in Kerala, and then flows through Tamil Nadu. The Bhavani River is 217 km, filled by both the southwestern and north-eastern monsoons. The basin is a 620,000-hectare channel that provides water supply to the states in which the river flows. Fig. 1. shows the map of the Bhavani River, which mainly flows through the Attappady Plateau in the Palakkad district and then into the Coimbatore and Erode districts of Tamilnadu. Irrigation of crops with river water accounts for roughly 90% of its supply. River water data are gathered from eleven sampling stations in Kottathara, Chalayur, Cheerakuzhi, Thavalam, Elachivazhi, Karathur, Sirumugai, Cheerakuzhi, Bhavani, Bhavanisagar, Badrakaliamman kovil, and Sathyamangalam.

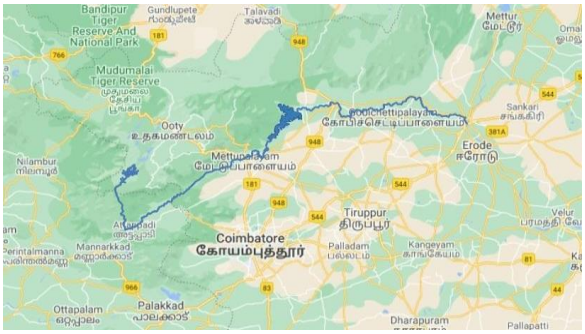


Fig.1. Bhavani River Map

Eleven sampling stations along the Bhavani River were used to collect data for the period of January 1st, 2016, to December 31st, 2020. Temperature, turbidity, conductivity, pH, biological oxygen demand, ammonia, chemical oxygen demand, nitrate, wind speed, date and wind speed are the primary parameters used to determine river water dissolved oxygen concentration.

Physicochemical parameters such as turbidity, conductivity, COD, temperature, BOD, ammonia, pH, and nitrate, and meteorological parameters such as wind speed are used in analysing and predicting dissolved oxygen concentration. Temperature is directly related to dissolved oxygen, when the temperature of the water is low then DO concentration will be high and vice versa. Compared to hot water, cold water can hold onto more dissolved oxygen. In the winter and early spring, the concentration of dissolved oxygen increases when the water temperature is low. In the summer and autumn, high water temperatures frequently result in low dissolved oxygen levels. pH and dissolved oxygen are not directly correlated, but the indirect relationship between external factors might increase algal growth. When the pH level is too low then the water will be acidic then the aquatic life affects badly to absorb the DO, that is pH decreases then hydrogen ions react with DO also decreases.

Conductivity plays a crucial role in determining salinity and total dissolved solids (TDS), which have an impact on river water quality and aquatic life. The dissolved oxygen concentration of water decreases when the salinity level rises. Turbidity in water negatively affects aquatic life as well as humans. Dissolved oxygen and turbidity are inversely related, higher turbidity and lower dissolved oxygen are considered to be unhygienic water. The amount of oxygen consumed when the organic matter that can be oxidised is exposed to powerful oxidants is measured by the chemical oxygen demand (COD). The presence of all kinds of organic matter, both biodegradable and non-biodegradable, as well as the degree of water pollution, are reflected in high COD levels.

Ammonia is a toxic material found in waste. When ammonia is oxidised, oxygen is consumed and oxygen levels in the water decrease and increase ammonia levels by inhibiting nitrification. Nitrate directly mixes with water through effluent from fertilisers containing nitrate in agricultural fields. Nitrogen is essential for all living organisms, but high nitrate content in drinking water can be harmful to health, especially for babies and pregnant women. The more turbulent water in streams and rivers, such as waterfalls and rapids, the more oxygen it absorbs. Furthermore, wind turbulence on the surface of bodies of water tends to raise dissolved oxygen levels. For this study, 10560 instances were collected from monitoring stations, and the sample data for the above parameters are shown in Table I.

Table I - Sample Data Collected from Sampling Station

Temp	pH	Conductivity	Turbidity	COD	Ammonia	BOD	Nitrate-N	Wind speed	Date
25	7.15	340	2	4	0.25	0.89	1.1	16.3	01/01/2016
24	7.46	339	2	3.9	0.25	0.87	1.1	14.4	02/05/2016
25	7.5	339	2	4	0.25	0.89	1.1	13.1	13/01/2017
25	7.18	340	2	3.9	0.25	0.88	1	15.4	24/10/2017
25	7.45	340	2	4	0.25	0.85	1.2	14	15/01/2018
24	7.05	342	2	4	0.25	0.87	1	18.7	29/11/2018
24	7.4	341	2	4	0.25	0.82	1.2	40.2	19/01/2019
25	7.38	339	2	3.9	0.25	0.81	1.2	13.6	28/08/2019
25	7.56	340	2	3.9	0.25	0.88	1.2	14.4	09/01/2020
25	7.1	340	2	4	0.25	0.82	1.2	14.9	16/06/2020
24	7.27	339	2	4	0.25	0.89	1.1	14	31/12/2020

Critical examination of the data distribution, outliers, and anomalies in order to gain a better understanding of it. Through exploratory data analysis, researchers can learn about the characteristics of primary data in relation to various statistical measures. EDA is a critical first step after data collection in which data is simply visualised, plotted, and manipulated without any assumptions to aid in determining data quality, performing data pre-processing, and selecting

features. EDA aims to assist in the development of models and the recognition of natural patterns by demonstrating potential connections between exposure and outcome variables.

In exploratory data analysis of experimental data, heatmaps, box plots, and pair plots are utilised to determine the causes of water quality fluctuation. Using the Pearson correlation function, which is commonly employed to determine the relationship between variables, the correlation matrix can be shown as a heatmap. A positive correlation shows that both variables are going in the same direction and that the coefficient of correlation is larger than zero. When the correlation coefficient is negative, the two variables are going in opposite directions. A box plot is utilised to illustrate the distribution of quantitative data in a manner that facilitates the comparison of variables. The box depicts the quartiles of the data set, while the whiskers depict the remaining distribution. Box plots are a standard approach for depicting the distribution of data in five categories: minimum, first quartile, median, third quartile, and maximum. A pair plot is employed to illustrate the distribution of single variables and their relationships.

The attribute distributions and correlations are investigated, using univariate and multivariate analysis and performed on the river water data which includes ten water quality parameters and 10560 instances. Heat map analysis showed that the water temperature is negatively correlated with pH, turbidity and nitrate, whereas all other parameters are positively correlated. Similarly, wind speed is negatively correlated with conductivity, COD, turbidity, and ammonia whereas all other parameters are positively correlated. The pair plot analysis on river water data depicted the correlation of temperature with pH, conductivity, ammonia, wind speed, and other attributes. The correlation is lower if the plot is scattered. The temperatures range from 22 to 33 according to a box plot analysis, with the majority falling between 25 and 28. The values of conductivity range from 1 to 1200, but the majority are between 60 and 210. The majority of biological oxygen demand values fall between 0.8 and 1.8, ranging from 0 to 2. The various analytical results of exploratory data analysis on river water quality data are visualised in Table 2, Fig.2a and Fig 2b. A sample box plot result is depicted in Fig. 2c.

Table 2: Correlation Between Water Quality Parameters

	Temp	pH	Conductivity	Turbidity	COD	Ammonia	BOD	Nitrate	Wind speed
Temp	1	-0.151	0.161468	-0.036	0.1985	0.1984	0.1824	-0.050	0.00799
pH	-0.151	1	0.051478	-0.299	-0.290	-0.227	-0.066	0.1372	0.00207
Conductivity	0.1614	0.0514	1	0.0395	0.4037	0.4087	0.2537	0.081	-0.2993
Turbidity	-0.036	-0.299	0.03956	1	0.2879	0.2457	0.1408	-0.144	-0.1781
COD	0.1985	-0.290	0.403798	0.2879	1	0.7849	0.3873	-0.207	-0.2065
Ammonia	0.1984	-0.227	0.408742	0.2457	0.7849	1	0.3076	-0.186	-0.2026
BOD	0.1824	-0.067	0.253743	0.1408	0.3873	0.3076	1	-0.051	-0.083
Nitrate	-0.050	0.1372	0.08111	-0.144	-0.207	-0.186	-0.051	1	0.01504
Wind speed	0.0079	0.0020	-0.29934	-0.178	-0.206	-0.202	-0.083	0.0150	1

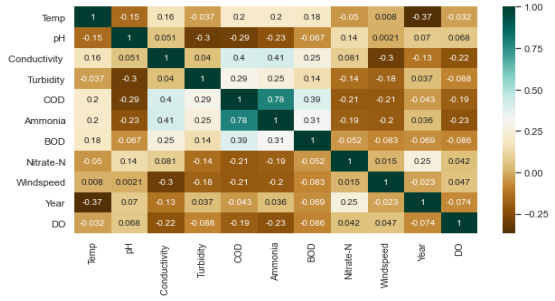


Fig 2a. Heat Map Results Illustrating Correlation Between Parameters

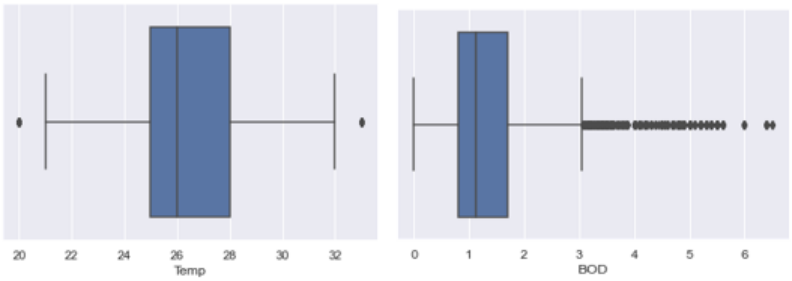


Fig.2b. Sample Box Plot Analysis

https://doi.org/10.56452/2022-11-002

Fig. 2c. Pair plot Analysis of the Water Parameters

Various analytical charts and graphs have been used to distinguish between the factors of variation in water data during exploratory data analysis of experimental data. The attribute distribution and correlation explored above through EDA have offered suitable solutions for data preparation and data modelling requirements.

3. PROPOSED DO CONCENTRATION PREDICTION MODEL

The proposed DO concentration prediction models explore the deep learning framework for predicting dissolved oxygen concentration using the nine physicochemical water parameters. The data collected from Bhavani River sampling stations are used here for data modelling and learning the trends in water parameters. The DO concentration prediction modelling consists of four phases, (i) time series data collection (ii) EDA and pre-processing (iii) DO concentration prediction model building (iv) model evaluation. The proposed DO concentration prediction model is illustrated in Fig.3.

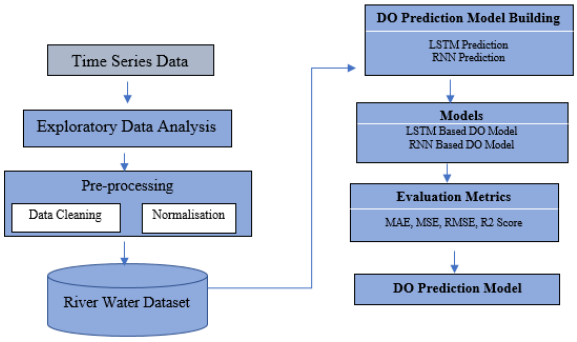


Fig 3. An Overview of the Work Plan

Creation of Dataset

The river water data is gathered from eleven monitoring stations along the Bhavani River in Kerala and Tamil Nadu from January 2016 to December 2020. Nine physicochemical parameters and one meteorological parameter are described in section 2 with 10560 instances and are used to find the DO concentration in water.

Finally, a river water dataset has been developed with 10560 instances comprising nine physicochemical and meteorological parameters as independent attributes and dissolved oxygen as the target variable. Exploratory data analysis is performed on the river water dataset to understand the data distributions and correlations to distinguish between the sources of variation in water parameters which are visualised using box plots, heat maps and pair plots. The exploratory data analysis carried out in this study provided a comprehensive understanding of the data and revealed that the dataset contains duplicates, parameters that are negatively correlated, and a wide range of values with some parameters such as conductivity.

Data Pre-processing

Data pre-processing enhances the quality and efficiency of the data. Data cleaning is the process of removing redundant, inaccurate, or incomplete data from a dataset. Twelve duplicate instances are removed from the river water quality dataset. Since the parameter conductivity ranges from a minimum value of 1 and the maximum value is 1200, Z score normalisation has been applied to

standardise the parameter values using the following formula,

$v' = (v - \bar{A})/\sigma A$ (1)

The DO concentration prediction model is constructed using 80% of the instances for training and 20% for testing from the river water dataset.

DO concentration prediction model building

In the proposed work, a deep learning approach has been employed to build an accurate DO concentration prediction model. Deep learning is a method of machine learning that makes use of numerous hidden layers in the network and is based on the idea of a neural network. More information is extracted from the raw input data, the deep learning architecture makes use of an infinite number of hidden layers that are limited in size. The complexity of the training data determines the number of hidden layers. To efficiently deliver the correct results, more hidden layers are required for more complex data. The recurrent neural network and long short-term memory are the deep learning architectures that were used in the research to create predictive models.

Another application of artificial neural networks that can learn features from sequence data is the recurrent neural network (RNN). RNN is made up of many layers, each of which has its own bias and weight. RNN makes it possible to identify temporal dynamic behaviour by sequentially running the relationships between nodes in a direct cycle graph. By providing a recurrent hidden state that identifies relationships across time scales, it can deal with temporal sequences and uses internal memory to store sequence information from earlier inputs, making it useful in several different areas.

A type of recurrent neural network (RNN) known as Long Short-Term Memory (LSTM) deals with dependencies that last a long time. It learns the complete data sequence through feedback connections and has been used in a variety of time-series data domains, including categorization, processing, and prediction. Input, forget, and output gates make up the LSTM architecture. The cell state is long-term memory in the LSTM cell that recalls and stores data from prior intervals. The input gate determines which values should be entered into the cell state. Using a sigmoid function with a range of [0, 1], the forget gate can determine which information should be forgotten. The information from the current time that should be considered in the subsequent step is selected by the output gate.

In this work, LSTM, and RNN architectures have been implemented on the river water training dataset to train and build the DO concentration predictive models by learning the self-extracted patterns from the input parameters. A variety of hyperparameters such as hidden layers, dense, optimizer, epoch, batch size, and dropout are used to fine-tune the best DO concentration prediction model. Hidden layers are the layers that exist between the input and output layers. The mini-batch gradient descent method is used to update network parameters. The epoch shows the number of times the network runs the entire training dataset. Important regularisation technique dropouts are used to reduce the effects of overfitting. The predictive models are built using the sparse category cross-entropy loss function for training. The learning rate determines the speed at which a deep model replaces an already learned concept with a new one. The Adam Optimizer minimises the error function. The DO concentration predictive models have been developed with predefined hyperparameters. The independent predictive models have been tested for their efficiency using the test dataset.

Model Evaluation

The performance of the LSTM and RNN, DO concentration prediction models are evaluated with the test dataset using mean squared error, root mean squared error, mean absolute error, and R2 score. If the mean squared error, root mean squared error, and mean absolute error, value is less, then the model fits perfectly with the dataset and is more accurate in DO concentration prediction.

The mean squared error of an estimator measures the average of error squares i.e., the average squared difference between the predicted value Y_b and actual value Y_a. The mean squared error is calculated using the following formula,

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_a - Y_b)^2 \quad (2)$$

The mean absolute error calculates the average difference between the predicted value Y_b and actual values Y_a. The mean absolute error is calculated using the following formula,

$$MAE = abs(Y_a - Y_b) \quad (3)$$

The root mean square error is a standard way to measure the error of a model in predicting quantitative data. The error is a measure that how

well a regression line fits the data points and is calculated using the following formula.

$$RMSE = \frac{\sqrt{(Y_a - Y_b)^2}}{n} \quad (4)$$

where Y_a and Y_b are the actual response and the predicted value, respectively, and n is the total number of instances.

The R2 score is a performance evaluation measure for regression models and is also known as the coefficient of determination. If the value of the R2 score is 1, the model is considered to be perfect, and if it is 0, the model will perform badly in predicting the target value.

The R2 score is calculated using the following formula,

R2 Score = 1 – (SS/TSS) (5)

where SS is the sum of squares of residuals and TSS is the total sum of squares.

4. EXPERIMENTAL RESULTS

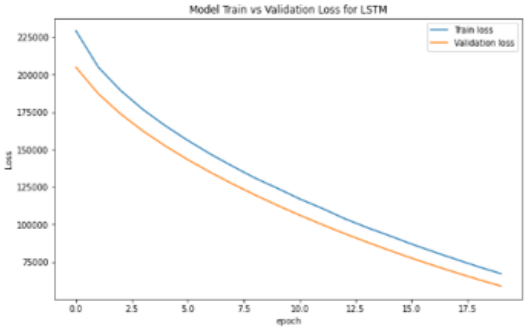
An effective model for predicting the DO concentration using the Bhavani River water dataset has been developed through several experiments. A training dataset with 7081 instances and a testing dataset with 3485 instances have been separated from the 10560 instances and 10 attributes of the dataset. Using Python libraries, deep learning algorithms such as LSTM and RNN are implemented using training datasets to develop the DO concentration prediction models. Varying epochs of 20, 50, 100, 150, and 200 have been experimented with both LSTM and RNN and the epoch size of 200 is fixed for the accurate prediction model.

The LSTM and RNN algorithms employ hyperparameters such as hidden layers, dense layers, optimizer, epoch, batch size and drop out to fine-tune the deep network. The experiments were conducted with the parameter settings as depicted in Table 3 and DO concentration models have been built.

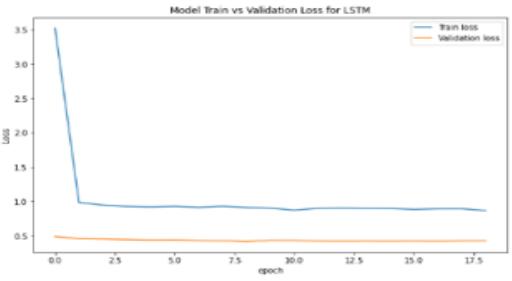
Table 3: Hyperparameters Setting for LSTM and RNN

Hyperparameter	Values
Optimizer	Adam
Epoch	200
Batch size	64
Dropout	30
Learning rate	0.01

The performance of the LSTM and RNN based DO concentration predictive models have been tested with a test dataset with respect to various performance metrics such as mean absolute error, mean squared error, root mean squared error and R2 score. The results of the experiments for various epoch sizes are illustrated in Table 4. The difference between the model train and validation loss of the DO concentration when predicted using LSTM and RNN based prediction models is depicted in Fig .4.

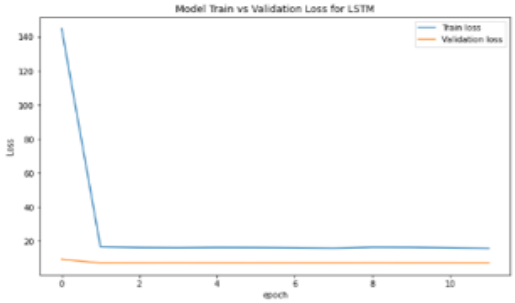


a) 20 epochs



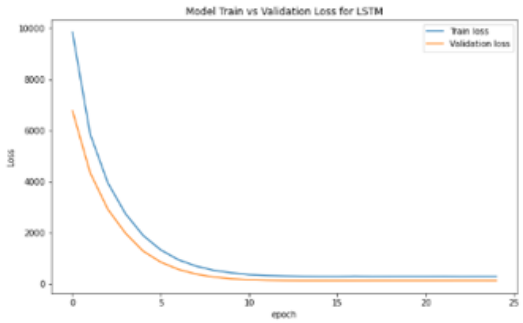
b) 50 epochs

c) 100 epochs



d) 150 epochs

https://doi.org/10.56452/2022-11-002



e) 200 epochs

Fig. 4. Model Train vs Validation Loss of LSTM and RNN for Various Epoch Sizes

Table 4: Results of LSTM and RNN based DO concentration Prediction Models

	Epoch 20		Epoch 50		Epoch 100		Epoch 150		Epoch 200	
	LSTM	RNN	LSTM	RNN	LSTM	RNN	LSTM	RNN	LSTM	RNN
MAE	0.6	0.56	0.6	0.5	0.543	0.4774	0.5664	0.483	0.675	0.4645
RMSE	0.8	0.7526	0.8	0.7151	0.823	0.6602	0.7436	0.667	0.72	0.647
R-2	0.832	0.847	0.838	0.862	0.837	0.88244	0.847	0.8801	0.848	0.887

The experimental results were observed for various epoch sizes and the values of performance metrics were found to be stable for epoch size 200. The prediction results showed that the RNN prediction model had a mean absolute error of 0.4645 and the LSTM prediction model had a mean absolute error of 0.675 for the 200 epochs. Similarly, it was found that the root mean squared error value of the LSTM prediction model with 0.62, and for RNN based model with 0. 647. Whereas the mean squared error value of the LSTM prediction model is 0.592 and for the RNN prediction model 0.462. The R2 score value of the RNN prediction model gives 0.887 with epoch 200 whereas the LSTM prediction model is 0.834 with epoch 200. Thus, a high error rate is produced by the LSTM based prediction model as compared to the lower rate for the RNN prediction model, also RNN yielded better accuracy than LSTM. The prediction results of deep network algorithms trained with epoch size 200 for various performance metrics such as mean absolute error, mean squared error, root mean squared error and R2 score are illustrated in Fig.5a, Fig.5b, Fig 5c and Fig.5d respectively.

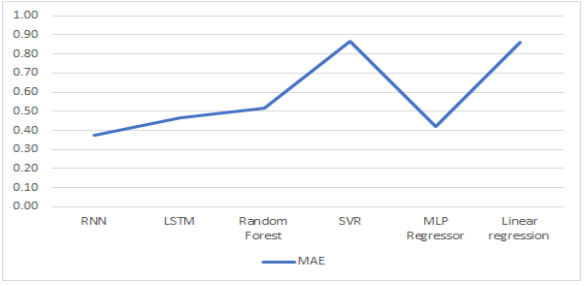


Fig.5a. Mean Absolute Error for Prediction Models

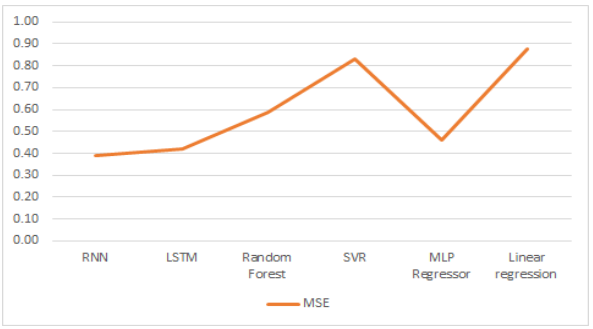


Fig 5b. Mean Squared error for prediction model

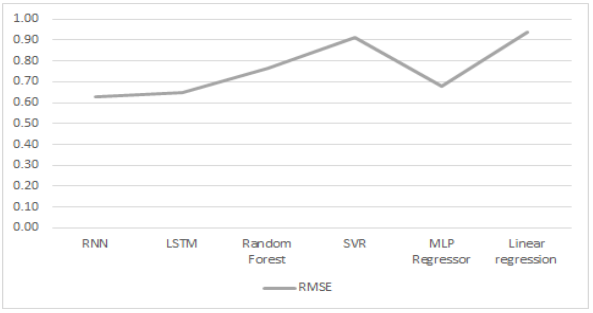


Fig 5c. Root Mean Squared Error Prediction Models

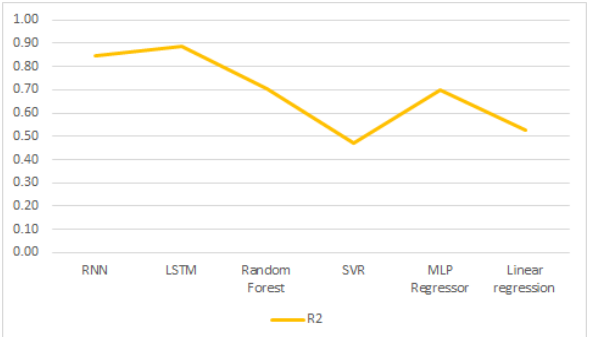


Fig 5d. R- Squared Values of prediction models

The overall performance evaluation results of the LSTM and RNN based DO concentration prediction model are compared with results of traditional machine learning algorithms such as linear regression, support vector regressor, MLP regressor and random forest. The comparative results with respect to performance metrics mean absolute error, mean squared error, root mean squared error and R2 score of deep learning algorithms such as LSTM, RNN and traditional machine learning algorithm linear regression, support vector regressor, MLP regressor, random forest is given in Table 5 and illustrated in Fig.6. The mean absolute error value of the support vector regressor is compared to other prediction models whereas RNN yields the least error rate. The mean squared error value of the support vector regressor is higher than other prediction models while RNN yields the minimal error rate. The root mean squared error value of the support vector regressor and is higher when evaluating with other prediction models, but RNN yields a less error rate. RNN prediction model yields high accuracy of 88.7% and LSTM based model gives 84.88%, whereas linear regression, MLP regressor, support vector regressor and random forest yield 52.61%, 69.86%, 47.037% and 70.66%.

Table 5: Comparative Performance Analysis of DO Concentration Prediction Models

	RNN	LSTM	Random Forest	SVR	MLP Regressor	Linear regression
MAE	0.375	0.4645	0.51602	0.867	0.4194	0.86241
MSE	0.392	0.421	0.587	0.831	0.462	0.876
RMSE	0.626	0.649	0.766	0.91211	0.680	0.936
R2	0.848	0.887	0.70668	0.4703	0.69862	0.5261

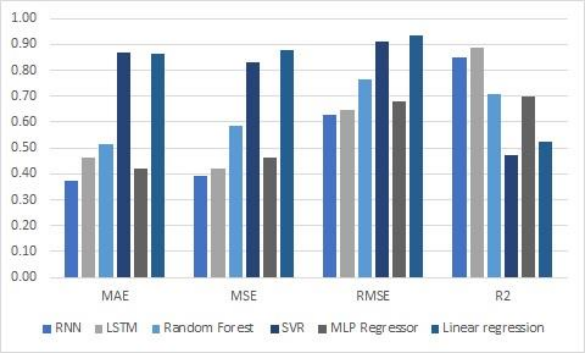


Fig 6. Comparative performance evaluation of DO concentration prediction models

The predicted and actual values of the DO concentration prediction models are analysed using the box plot analysis. The DO concentration prediction model predicts the target variable with independent input parameters given to the input layer. The actual value is the value of the target variable supplied to the model for learning the patterns from the input parameters, whereas the predicted values are the future predictions of the DO concentration, predicted by the deep learning-based DO concentration prediction model after self-analysing the patterns. The box plot analysis of actual and predicted values of the LSTM and RNN based prediction models depicts that the RNN based models suit well in predicting DO concentration. The actual values of the LSTM based DO concentration prediction model is between 0 and 1, whereas the predicted value is in the range -0.2 and 0.9. The RNN based DO concentration prediction model produced the actual values as 0.1 to 0.88, and the predicted values lie between 0.08 and 0.85. The box plot analysis of DO concentration prediction is illustrated in Fig.7a and Fig.7b.

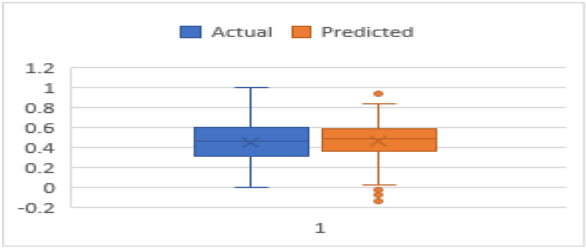


Fig.7a. Boxplot analysis of LSTM predicted values.

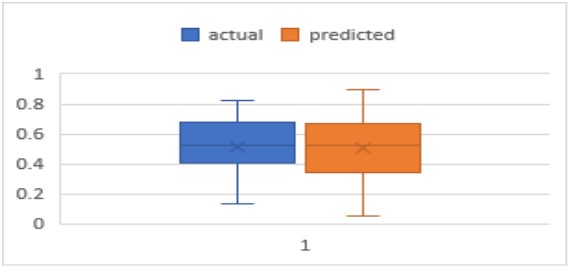


Fig.7b. Boxplot analysis of RNN with predicted values

Thus, the experimental results confirm that the developed deep learning models are able to predict the DO concentration accurately. The performance evaluation of the models was done using the performance metrics like mean absolute error,

mean squared error, root mean squared error and R2 score. The LSTM based DO concentration prediction model yields a high error rate as compared to RNN based prediction model with respect to epoch size 200. The recurrent neural network DO concentration prediction model gives high accuracy as compared to LSTM prediction model and traditional machine learning models such as linear regression, support vector regressor, MLP regressor and random forest. Hence it is concluded that the RNN prediction model produces more promising results than the LSTM model and other machine learning models considered here, in predicting DO concentration using river water data.

5. CONCLUSION

This research work has been proposed to develop an efficient DO concentration prediction model using a river water dataset. The implementation of a deep learning approach for developing prediction models has been elucidated in this paper. The river water data for various contributively properties were collected from eleven different sampling stations situated across the Bhavani River, and employed in building the DO concentration models. Deep learning architectures such as LSTM and RNN have been implemented with fine-tuned hyperparameters to build DO concentration prediction models with improved efficiency of the prediction. The performance comparison of deep learning-based DO concentration prediction models with traditional machine learning-based models proved that RNN deep neural architecture produced more accurate DO concentration prediction models. The prediction of dissolved oxygen is highly imperative as it is correlated with many water quality attributes in defining the water quality index. Hence, water quality index prediction can be explored further with accurate DO concentration prediction.

REFERENCE

1. Ji, X., Shang, X., Dahlgren, R.A., Zhang, M., 2017. Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machines: a case study of Wen-Rui Tang River, China. Environ. Sci. Pollut. Res. 24, 16062–16076.
2. Vorosmarty, ¨ C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global M.H. Ahmed and L.-S. Lin Journal of Hydrology 597 (2021) 126213 12 threats to human water security and river biodiversity. Nature 467, 555–561.
3. Basant, N., Gupta, S., Malik, A., Singh, K.P., 2010. Linear and nonlinear modelling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water -A case study. Chemometr. Intellig. Lab. Syst.104,172–180.
4. Wen, X., Fang, J., Diao, M., Zhang, C., 2013. Artificial neural network modelling of dissolved oxygen in the Heihe River, North western China. Environ. Monit. Assess. 185, 4361–4371.
5. Chapra, S., 2008. Surface water-quality modelling. McGraw-Hill Companies Inc, Newyork.
6. Loucks, D.P., van Beek, E., 2017. Water Quality Modelling and Prediction. In: Water Resource Systems Planning and Management. Springer International Publishing, Cham, pp. 417–467.
7. Sanchez, E., Colmenarejo, M.F., Vicente, J., Rubio, A., García, M.G., Travieso, L., Borja, R., 2007. Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. Ecol. Ind. 7, 315–328. https://doi.org/10.1016/j. ecolind.2006.02.005
8. Streeter, H.W., Phelps, E.B., 1925. A Study of the Pollution and Natural Purification of the Ohio Rivers. Public Health Service Bulletin, U.S.
9. Cox, B.A., 2003. A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. Sci. Total Environ. 314–316, 335–377.
10. Kannel, P.R., Lee, S., Lee, Y.S., Kanel, S.R., Khan, S.P., 2007. Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment. Environ. Monit. Assess. 132, 93–110.
11. Li, G., 2006. Stream temperature and dissolved oxygen modelling in the Lower Flint River Basin. PhD Dissertation. University of Georgia, Athens, GA.
12. Rankovi´c, V., Radulovi´c, J., Radojevi´c, I., Ostoji´c, A., Comi ˇ ´c, L., 2010. Neural network modelling of dissolved oxygen in the Gruˇza reservoir, Serbia. Ecol. Model. 221, 1239–1244.
13. Hey, T., 2012. The Fourth Paradigm – Data-Intensive Scientific Discovery, in: E-Science and Information Management. Presented at the

International Symposium on Information Management in a Changing World, Springer, Berlin, Heidelberg, pp. 1–1. 10.1007/978-3-642-33299-9_1.
14. Herzog, S., Worg ¨ otter, ¨ F., Parlitz, U., 2018. Data-Driven Modeling and Prediction of Complex Spatio-Temporal Dynamics in Excitable Media. Front. Appl. Math. Stat. 4
15. Tung Tiyasha, T.M., Yaseen, Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 585, 124670.
16. Heddam, S., 2014. Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. Environmental Technology (United Kingdom) 35, 1650–1657. 10.1080/ 09593330.2013. 878396.
17. Zhang, Y.-F., Fitch, P., Thorburn, P.J., 2020. Predicting the Trend of Dissolved Oxygen Based on the kPCA-RNN Model. Water 12, 585.
18. Ay, M., Kis¸i, O., ¨ 2017. Estimation of dissolved oxygen by using neural networks and neuro-fuzzy computing techniques. KSCE J. Civ. Eng. 21, 1631–1639.
19. Chen, W.-B., Liu, W.-C., 2014. Artificial neural network modelling of dissolved oxygen in the reservoir. Environ Monit Assess 186, 1203–1217.
20. Nemati, S., Fazelifard, M.H., Terzi, O., ¨ Ghorbani, M.A., 2015. Estimation of dissolved oxygen using data-driven techniques in the Tai Po River, Hong Kong. Environ Earth Sci 74, 4065–4073.
21. Heddam, S., Kisi, O., 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J. Hydrol. 559, 499–509.
22. Jitha P Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 1747-1753.
23. Jitha P Nair, and M. S. Vijaya. "River Water Quality Prediction and index classification using Machine Learning." Journal of Physics: Conference Series. Vol. 2325. No. 1. IOP Publishing, 2022.
24. Jitha P Nair, Vijaya, M.S. Exploratory Data Analysis of Bhavani River Water Quality Index Data. In: Kumar, S., Hiranwal, S., Purohit, S.D., Prasad, M. (eds) Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-19-3951-8_74