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

Forecasting Water Quality Index of the Ganga River Using CCL Hybrid Deep Neural Network

Chunnu Lal[1*], Dr. Satender Kumar[2]

Submitted:26/01/2024 **Revised**: 04/03/2024 **Accepted**: 20/03/2024

Abstract— In this paper, Convolutional Neural Network-Convolutional Neural Network- Long Short-Term Memory (CNN-CNN_LSTM) hybrid deep learning neural network is developed to forecast the water quality of the river Ganga. Various deep learning models like LSTM, CNN, CNN_LSTM have been designed as baseline models to compare the outcome to the proposed model. Water Quality parameters data collected from ten base stations stablished by Uttarakhand Pollution Control Board is used for training & testing of the model developed. Water Quality Index is calculated using basic four Water Quality Parameters like BOD (Biochemical Oxygen Demand), pH (potential of Hydrogen), DO (Dissolved Oxygen), Temperature. The proposed CNN-CNN_LSTM(CCL) model provides better forecasting results for Water Quality Index (WQI).

Index Terms— *Ensemble Learning, Ganga River, Water Quality (WQ), Water Quality Index (WQI), Deep Learning Models*

I. Introduction

More than 30% population in the world have the shortage of Good Quality Water [1]. 80% water available on earth is not suitable for humans [2]. A large population of India can suffer from disease like cancer depends on Ganga River water due to killer pollutants available in Ganga River water. Water quality parameters like BOD (Biochemical Oxygen Demand), pH (potential of Hydrogen), DO (Dissolved Oxygen), temperature etc. are used to find the Water Quality. WQI is calculated to shows the contribution of each parameter in Water Quality [3]. The methods available in literature for finding the water quality is very costly & not gives the accurate data. The Machine & Deep Learning algorithms can easily read the dependency of different parameters available & forecast the future data based on trained data. Forecasting water quality can be done using Time series analysis methods [17]. Time series analysis methods like Prophet, ARIMA, SARIMA is discussed for forecasting the water quality parameters like DO and BOD of Water & Water Quality Index of the river ganga in Uttar Pradesh. Best model is finding by doing comparative study based on values of performance metrices such as square root of Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [2]. CNN-BiLSTM-SVR is developed, an efficient deep learning model to forecast DO and BOD of Ganga River in Uttar Pradesh (UP). Developed model has been compared to other models like LSTM, Bi-LSTM, CNN-LSTM based on performance parameters like MSE & RMSE [3].

To find the temporal and spatial variations in water quality parameters of the river ganga in UP.

Unsupervised machine learning methods like cluster analysis, PCA (Principal Component Analysis) & correlation is introduced. The study finds that pH, DO water quality parameters had correlation with season [4]. In this paper, we have used the deep learning models like LSTM, CNN, CNN_LSTM for training & testing of Water Quality parameters data available on UPCB (Uttarakhand Pollution Control Board) website. Study used total four parameter's data such as pH, DO, temperature & BOD in our study.

We have implemented CCL Hybrid deep neural network to predict accurate WQI of the River Ganga. The paper is divided in five sections. Section II discusses the motivation of the research. Section III is discussed the methodology used to develop the efficient deep learning model. The discussion on result is done in Section IV. The model comparisons based on performance metrices such as MSE, RMSE & R2 score is done in Section V. Section VI contains the conclusion of the research.

II. Motivation For Research

Water pollution is a major problem at various places in Uttarakhand. Future water quality prediction using deep learning neural networks can helps the individuals and the government to take necessary actions on time. The research implemented deep learning models using ensemble learning for analysis and forecasting of Water Quality Index.

III.Steps For Methodology

Methodology used for developing the efficient deep learning model is given below-

[1*]*Ph.D. Scholar, Quantum University, Roorkee*
[2]*Dean Academics, Quantum University, Roorkee*

A. Study Area

Uttarakhand state is suffered from water pollution due to industrial area from Rishikesh to Roorkee. This paper has chosen the area of Uttarakhand for research. Fig.1 shows the geographical location of the Ganga River in Uttarakhand used for study.

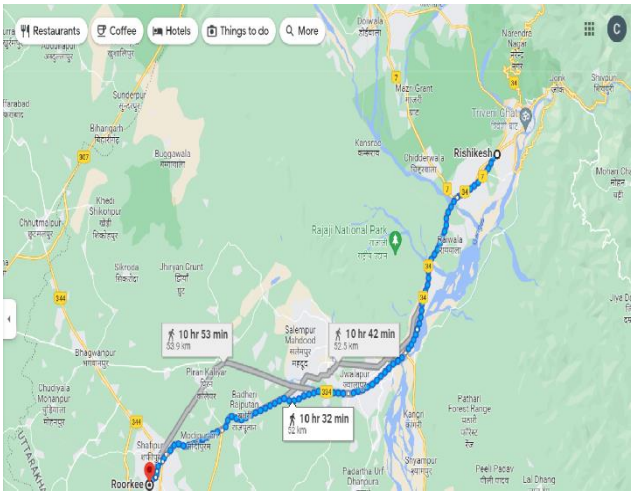

Fig. 1: Part of the Ganga River flows from Rishikesh to Roorkee

B. Preprocessing of Water Quality Parameters

Water Quality data set is used for research have been taken from Uttarakhand Pollution Control Board website. Data set available from 10 base stations is used for study. The points of base stations are- P1: Rishikesh at Bairaj Near Pashulok, Uttarakhand, P2: Lakshmanjhula Rishikesh, P3: Raiwala Dehradun, P4: Bindughat Dudhiyabad, Haridwar, P5: Balkumari Mandir, Ajeetpur, Haridwar, P6: Lalita Rao Bridge, Haridwar, P7: Rishikul Bridge D/S Harkipouri Haridwar, P8: Harkipouri Haridwar (Damkothi), P9: Harkipouri Haridwar, P10: Roorkee Haridwar. Total 1440 samples of four water quality parameters like BOD, Temperature, DO, pH from year 2011 to 2022 is used for training & testing models.

C. Formula of WQI Calculation [3][9]

WQI is calculated using four parameters like pH, DO, temperature and BOD. The value of *q* shows the value of individual parameters in the range 0-100. Eq.1 shows the calculation of WQI [11].

$$WQI = \sum_{i=1}^{n} Wi * Qi \qquad (1)$$

In Eq.1 W_i defines assigned weight to i^{th} water quality parameter, *n* defines count of water quality parameters, Q_i defines *q*-value associate with i^{th} water quality parameter. Weight assigned to each parameter for WQI calculation is shown in Table I.

TABLE I: Contribution of various parameters in WQI Calculation [3]

WQ Parameters	Weight of Individual Parameter
Temperature	0.10
BOD (mg/L)	0.11
DO (mg/L)	0.17
pH	0.11

D. Proposed Hybrid Model

Proposed deep hybrid model is designed using ensemble modelling technique to forecast the univariate time series data. The flow diagram of the developed model is shown in Fig. 2.

CNN Model: This Feedforwards Neural Network can forecast the time series using spatial features available in the time series data. [5]-[9].

CNN_LSTM Model: Forecasting the water quality parameters is a tedious task. Individual deep learning models is not sufficient to read the seasonal and time variations of data. It can be done using hybrid models easily. Here, CNN_LSTM Hybrid deep learning model is designed by combining CNN and LSTM neural networks in a special manner. First CNN is constructed using ReLU activation function and kernel size is equal to 1. Second a MaxPooling layer is connected and pool size is set to 2. Now the features extracted from second layer is forwarded to the LSTM layer to find the predicted values of WQI after adding a flatten layer to CNN. Next, the dense layer is used to find the output.

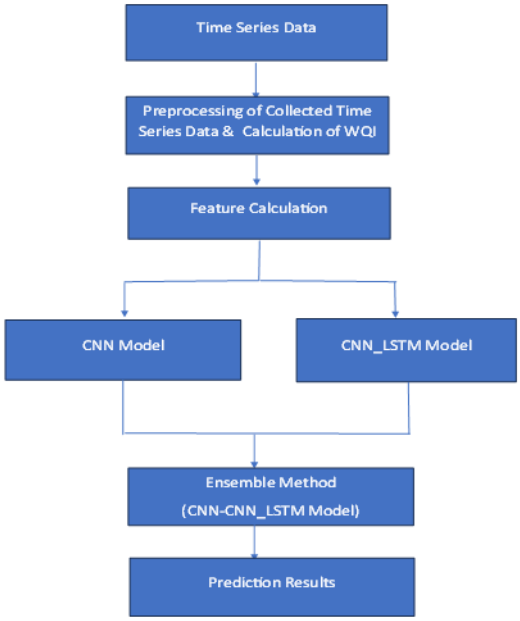

Fig.2: Flow diagram of the CCL Hybrid Deep Neural Network

E. Calculation of Performance matrices [4][9]

We can easily find the accuracy of deep learning models using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) performance metrices. MSE can be calculated by finding the deviation between the predicted and the original values. RMSE is equal to square root of MSE. Error is the deviation between actual and predicted values e_i, for i = 0,1,2,3...n. The model which has smallest MSE & RMSE values will be worked as best model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (2)$$

$$\text{RMSE}=\sqrt{MSE} \quad (3)$$

***F.* Model parameter**

For designing the deep learning models, parameter values of various hyper parameters are set as per detail given in Table II.

TABLE II: Parameter of CCL Hybrid Deep Neural Network

Parameter	Value
Count of Convolution layer filters	64
Loss function	MSE
Padding of Convolution layer	Same
Size of MaxPooling layer	2
Optimizer	Adam
Padding used at MaxPooling layer	Same
Activation Function	ReLU
Batch size	64
Epochs	100
Kernel size	1

IV. The Discussion Of Results

Three baseline models like LSTM, CNN, CNN-LSTM has been implemented in this paper. The proposed CNN-CNN-LSTM model is compared with the baseline models to shows that it is the best model. Data set is divided in training & test data set. 80% of data set have been used for training of models & 20% of data set is used for testing. Prediction have been done on test data set. Performance matrices such as MSE, RMSE and R2 Score is used to decide the best model.

A. Baseline Models Development

The system configuration used for developing all the models is as follows: Operating system: Windows 64-bit Operating System, x64-based processor, CPU-Intel® Core™ i5-10210U @ 1.60GHz 2.11 GHz, RAM-4 GB. Deep learning packages like pandas, NumPy, Keras, TensorFlow, and matplotlib is used for developing all models. Epochs is set to 100. Adam optimizer is used whereas the batch size set to 64. Loss and activation functions are used as MSE and ReLU (Rectified Linear Unit) respectively. Development steps of the baseline models is given below.

1) LSTM Model- Forecasting of univariate time series can be done using LSTM model. Next value in the series based on past observations can be easily predicted using LSTM model. LSTM is accepting a three-dimensional input and generate a two-dimensional output based on feature extraction done from the sequence. The LSTM model try to find a function by which output observations can easily be find by using the past observations of the input. In this paper, LSTM is designed using LSTM layer with dimension of output vector=16, input shape= (1,1) with ReLU activation. The next layer is a dense layer is used in this model to get the output.

2) CNN Model- A Convolutional Neural Network (CNN) is a neural network used for working with two-dimensional image data. Extracting features from univariate time series data can be done easily. In this, CNN is designed using an input_shape= (1,1), Convolution layer with ReLU activation function & the kernel size is set to 1. Next the MaxPooling layer is added with pool size of 2. Then, the extracted features are inputted to the flatten layer. Now dense layer with dimension of output vector=50, is added with ReLU activation. Lastly, the dense layer with dimension of output vector=1, is added to get the output.

3) CNN_LSTM Model-

Forecasting the water quality parameters is not an easy task. Individual models such as CNN & LSTM is not sufficient to read the seasonal or temporal information from a time series data. This can be done easily using hybrid models. A hybrid mode of CNN model with an LSTM backend where the CNN is used to interpret hidden features of input & output of CNN model is provided to LSTM model to interpret. This hybrid model is called a CNN_LSTM model. Here, CNN model is designed with a one-dimensional ReLU activated convolution layer and kernel size is set to 1. Next layer is MaxPooling with a pool size of 2. Next, flatten layer is added & the extracted features are inputted to LSTM layer to get the forecasted values of WQI. Next, the dense layer is used to get the output.

B. Comparative Analysis

The values of performance metrices like MSE, RMSE & R2 score is compared with other baseline models in Table III. Designed deep learning model have predicted the values with lower values of MSE & RMSE and higher values of R2 Score. The values of performance metrices given in Table III clearly shows that CCL model performing well compared to other models.

TABLE III: Parameter of CCL Hybrid Deep Neural Network

Model	**Performance Metrices**	**WQI**
LSTM	MSE	0.039
	RMSE	0.198
	R2 score	0.717
CNN	MSE	0.037
	RMSE	0.193
	R2 score	0.734
CNN_LSTM	MSE	0.036
	RMSE	0.191
	R2 score	0.739
CCL(CNN-CNN_LSTM)	**MSE**	**0.035**
	RMSE	**0.188**
	R2 score	**0.747**

The values of WQI index forecasted by different models for the next 12 months is compared in Table IV given below. Forecasted values using CCL deep learning model is closer than the other baseline deep learning models. Forecasting values of WQI using CCL model clearly shows that this model is the best model compared to other deep learning models. The graphical representation of the forecasted values is represented in the Fig.3.

Forecasted values using CCL is represented using yellow bar which is closer to actual values of WQI represented using blue bar.

LSTM	CNN	CNN_LSTM	CNN-CNN_LSTM(CCL)	Actual Values
78.30547	78.43514	76.48007	76.25723778	85.23682962
82.13594	83.070915	79.98531	91.86679323	98.60148711
84.33764	86.2733	82.5501	94.95396298	106.1784528
85.59395	88.485504	84.41666	103.203607	112.1831989
86.307884	90.01368	85.7697	99.64309445	108.6226863
86.71478	91.06937	86.74768	110.2895112	114.7793071
86.94639	91.79861	87.45308	113.4799873	121.3371302
87.07814	92.30237	87.961105	101.7263091	110.705901
87.15304	92.650375	88.3266	98.32415757	101.6915045
87.19562	92.890755	88.58932	111.4364816	117.0487265
87.21981	93.05683	88.77808	110.3140326	117.0487265
87.23356	93.171555	88.91361	87.36687233	89.61177029

TABLE IV: Forecasted Values for next one year by different models

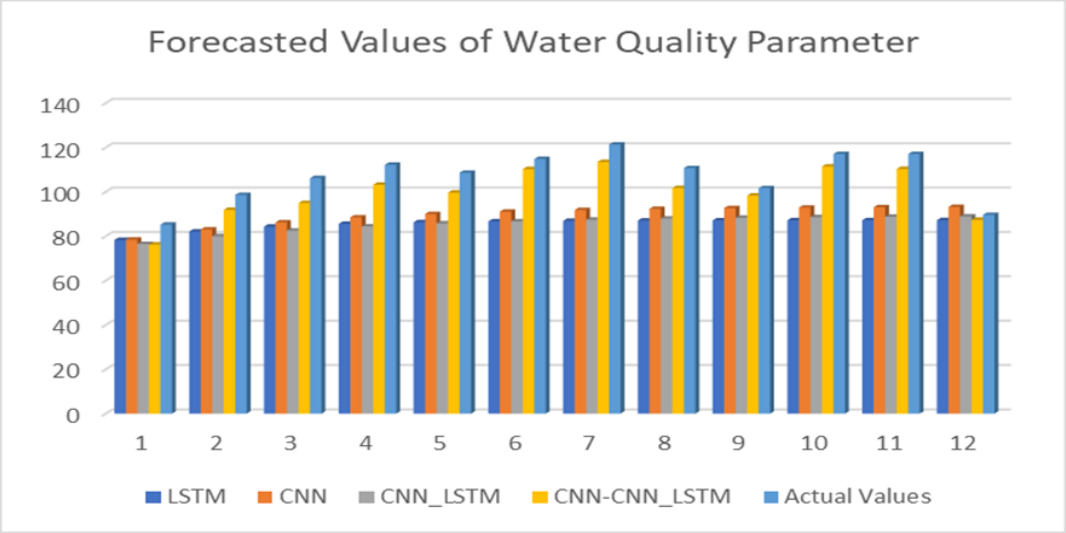

Fig.3: Comparison of forecasted Values for next one year by different models

V. Conclusions

In this paper, Ensemble technique is used to develop an efficient CNN-CNN_LSTM(CCL) deep hybrid model. Forecasted values using CCL model is closer to the Actual values of Water Quality Index. Using the proposed model water quality index is predicted with lower values of MSE, RMSE & higher value of R2 Score, which shows that the proposed model is the best model when compared with the baseline models. Further, the research study may be extended by hybrid the different deep learning models using various ensemble techniques.

References

[1] A. Krishnaraj, R. Honnasiddaiah, “Remote sensing and machine learning based framework for the assessment of spatio-temporal water quality in the Middle Ganga Basin”, Environ Sci Pollut Res 29, 64939–64958, https://doi.org/10.1007/s11356-022-20386-9, 2022.

[2] A. P. Kogekar, R. Nayak and U. C. Pati, "Forecasting of Water Quality for the River Ganga using Univariate Time-series Models," 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2021, pp. 52-57, doi: 10.1109/ICSCC51209.2021.9528216.

[3] A. P. Kogekar, R. Nayak and U. C. Pati, "A CNN-BiLSTM-SVR based Deep Hybrid Model for Water Quality Forecasting of the River Ganga," 2021 IEEE 18th India Council International Conference (INDICON), Guwahati, India, 2021, pp. 1-6, doi: 10.1109/INDICON52576.2021.9691532.

[4] A. Krishnaraj, P.C. Deka, Spatial and temporal variations in river water quality of the Middle Ganga Basin using unsupervised machine learning techniques. Environ Monit Assess 192, 744 (2020). https://doi.org/10.1007/s10661-020-08624-4

[5] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, “A cnn-lstm-based model to forecast stock prices,” Complexity, vol. 2020, 2020.

[6] K. Wu, J. Wu, L. Feng, B. Yang, R. Liang, S. Yang, and R. Zhao, “An attention-based cnn-lstm-bilstm model for short-term electric load forecasting in integrated energy system,” Int. Trans. on Electrical Energy Systems, vol. 31, no. 1, p. e12637, 2021.

[7] E. Hoseinzade and S. Haratizadeh, “Cnnpred: Cnn-based stock market prediction using a diverse set of variables,” Expert Systems with Applications, vol. 129, pp. 273–285, 2019.

[8] Chunnu Lal, et al. (2023). Water Quality Prediction of Ganga River using Time-series Models.

International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 4845–4850. https://doi.org/10.17762/ijritcc.v11i9.10080

[9] C. Lal and S. Kumar, "Ganga River Water Assessment Using Deep Neural Network: A Study," 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP), Uttarakhand, India, 2022, pp. 184-186, doi: 10.1109/ICFIRTP56122.2022.10063185.