International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Water Quality Prediction of Ganga River using Time-series Models

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

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

***Corresponding Author:-** Chunnu Lal
* PH.D. SCHOLAR, Quantum University, Roorkee

ABSTRACT
Life of Living organism have present on the earth depends on Water. Water Quality is also equally important as Water. Ganga river is fulfilling the needs of water of a large population of India. Being a citizen of India it's our responsibility to keep the Ganga River neat & clean. A large number of governments funded base stations available for forecasting the Water Quality of ganga river. But there is a need of low-cost prediction techniques of waterquality based on data available from these base stations. It can help the government to take the necessary decisions to cure the water quality of Ganga River & save the lives of many livings' organism depends on Ganga River. Monitoring & forecasting of water quality of Ganga River is most important because ganga river is the main source of drinking water of a large population of India. In this paper two time series-based models such asAuto-Regressive Integrated Moving Average (ARIMA), SeasonalARIMA (SARIMA) have been used to predict the water quality of Ganga River. The models are developed on water quality data available of 10 base stations on the Uttarakhand Pollution Control Board's official website. Four water quality parameters-Temp, pH, DO, BOD data is used for models training & calculating WQI (Water Quality Index). The result of experiment shows that SARIMA model predict the water quality parameters as well as Water Quality Index (WQI) more accurately.

Index Terms—Ganga River, time series models, Water Quality Prediction, Water Quality Index (WQI), ARIMA model, SARIMA model,

I. INTRODUCTION

More than 30% population of the world face the scarcity of good quality water [1]. More than 80% open water is not suitable for human being in India [2]. Ganga Water Quality decreasing day by day due to sewage, industrial waste, fertilizers used in agriculture land [12]. Population depends on Ganga River water are more prone to various types of disease like cancer due to killer pollutants available in Water. Water Quality depends on water quality parameters such as temperature, pH (potential of Hydrogen), DO (Dissolved Oxygen), BOD (Biochemical Oxygen Demand) etc. Each quality parameter has their unique importance in Water quality [12]. To shows the importance of every individual parameter in Water Quality, WQI (Water Quality Index) is calculated [3]. The traditional techniques available for measuring the water quality is very costly, time consuming & prone to errors. The Machine Learning gives the power to computer systems that they can learn the rule from the available data & forecast the future data [1].

Advanced technologies such as time series analysis methods can be used to analyse & forecasting water quality. Based on latest research, various time series models such as Prophet, ARIMA, SARIMA used for forecasting two parameters such as DO & BOD of Water & water quality index of river ganga in Uttar Pradesh. Comparative study has done to find the best model based on values of performance metrices such as MAE (Mean Absolute Error) & RMSE (Square root of Mean Squared Error) [2]. An efficient hybrid deep learning model named as CNN-BiLSTM-SVR is implemented for forecasting water quality parameters such as DO & BOD of river ganga in Uttar Pradesh. Implemented model has been compared based on various performance metrices such as MSE & RMSE with existing deep learning models such LSTM, CNN-LSTM, BiLSTM [3]. Unsupervised machine learning techniques such as principal component analysis, cluster analysis & correlation is used to find the Spatial & temporal changes in water quality of river ganga in Uttar Pradesh. The study result shows that two quality parameters used in our study pH, DO had correlation with season [4]. Machine Learning based framework implemented to extract concentrations of different optical & non-optical parameters from Landsat-8 satellite images [1]. Study shows the relations between Advanced Space borne Thermal Emission and Reflection Radiation (ASTER) data and observed water quality parameters using regression analysis. Parameters calculated from ASTER reflectance bands have the same values of parameters as predicted through regression analysis [5]. Water quality analysis based on statistical methods have done on data available of various sampling locations of Semenyih River [6]. In our study, we have used the time series models such as ARIMA & SARIMA for analysis of Water Quality parameters data available on Uttarakhand pollution control board official website. We have used four parameters' data such as temperature, pH, DO & BOD in our study.

We have implemented two frequently used time- series models such as ARIMA & SARIMA to predict WQI of the river Ganga (Consider only 10 sampling locations installed by Uttarakhand government from Rishikesh to Haridwar) with high accuracy. The paper is organized as follows. The Motivation for research is discussed in Section II. The methodology is discussed in Section III. The discussion on result is done in Section IV. The comparative analysis of models used in the study is done in Section V. The paper is concluded in Section VI.

II.MOTIVATION FOR RESEARCH

Water quality of Ganga River is degraded due to high pollution at various places in Uttarakhand. It is the requirement to design & develop a Robust Machine & deep learning model that can predict the future water quality at various places with high accuracy in Uttarakhand. The Paper implemented two time series models for analysis and future prediction of Water Quality Parameters. To find the best model for Water Quality prediction, comparative analysis based on performance metrices is done.

III.METHODOLOGY

The steps of methodology adopted for doing research is explained below-

A. Area of Study

Large number of studies have done in literature in UP area. Water pollution is also a problem in Uttarakhand. So, we have selected the area of Uttarakhand for our study. Hence, the study area is geographically located in Uttarakhand as highlighted in Fig. 1.

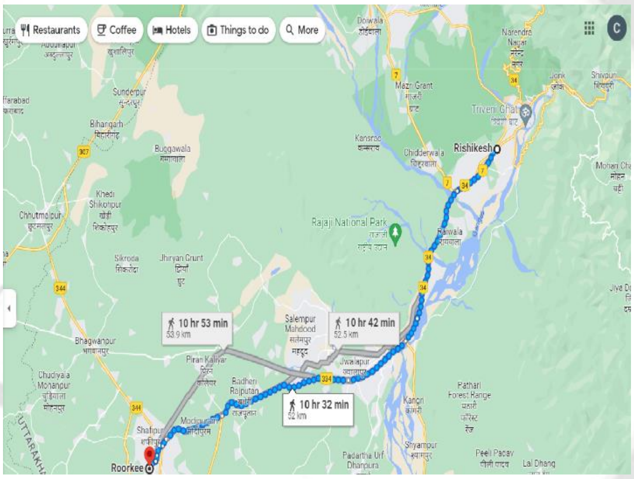


Fig. 1: Study area (Part of the river Ganga that flows through Uttarakhand).

B. Water Quality Data collection & preprocessing

The data set was collected from Uttarakhand Pollution Control Board (UPCB) official website. The water samples were collected from ten sampling locations such as Upper Ganga Canal D/S Roorkee Haridwar, Upper Ganga Canal D/S Harkipouri Haridwar, Upper Ganga Canal D/S Harkipouri Haridwar(Damkothi), Upper Ganga Canal Rishikul Bridge D/S Harkipouri Haridwar, Upper Ganga Canal Lalita Rao Bridge, Haridwar, Upper Ganga Canal D/S Balkumari Mandir, Ajeetpur, Haridwar, Ganga U/S Bindughat Dudhiyabad, Haridwar, River Ganga D/S Raiwala Dehradun, River Ganga U/S Lakshmanjhula Rishikesh, Ganga at D/S of Rishikesh at Bairaj Near Pashulok, Uttarakhand .So, the data contains 1440 samples collected monthly corresponding to these ten locations of the river Ganga collected from 2011 to 2022 for four water quality parameters temperature, pH, DO & BOD. Temperature shows water chemistry, biological activities and also have impact on water quality [8]. pH shows how acidic or basic the water & shows the percentage of hydrogen ion concentration [10]. DO quantity can harm the living being & affects water quality. BOD quantity increased in water when sewage & industrial waste (rich of organic matter) get mixed with water [4]. The original data contains some missing values. To fill up all missing values the backward fill method is used. To scale the water quality parameters on same range (0-1 range), min-max normalization techniques is used. After scaling the data is divided in training & test set.

C. Calculation of Water Quality Index [3]

WQI have been calculated to assign a unique value to overall water quality & it is calculated using different water quality parameters that represent the actual quality of water individually [7]. In our study four parameters namely temperature, pH, DO and BOD are used to calculate WQI. The *q* value represents the normalized value for individual parameters from the range 0-100. The WQI can be calculated by using Eq. 1 [11],

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

In equation (1), *n* shows total number of water quality parameters, Q_i represents *q*-value associate with i^{th} water quality parameters, W_i represents weight factor of i^{th} water quality parameters. The parameters with its corresponding weight factor used in WQI calculation is given in Table I.

TABLE I: Weight associated with parameters for WQI Calculation

Parameters	**Weight Factor**
Temp	0.10
pH	0.11
DO (mg/L)	0.17
BOD (mg/L)	0.11

D. Test of Stationarity

The time series does not show any trend or seasonal effect over time can be consider as stationary. To check the stationarity of each parameter an augmented dicky-fuller test has done. Difference method is used for converting non-stationary time series to stationary.

E. Development using Time Series Analysis Models

Statistical technique like time series analysis that deals with variables having a trend or seasonal effects with respect to time. There are two types of time series models available in literature that is Univariate and Multivariate time series models [9]. The difference between univariate & multivariate time series models is that univariate uses single variable & multivariate uses multiple variables over an equal interval time. In the proposed methodology, the univariate time series forecasting is performed by taking monthly WQI index of 10 different locations using ARIMA model and SARIMA model.

1) ARIMA Model [2]:

The general form of ARIMA (p, d, q) have the following non negative integer parameters such as Auto Regressive (AR) order is denoted by p, number of differencing required to make the time series stationary is denoted by d and Moving Average (MA) order is denoted by q.

Mathematically ARIMA model can be represented using the Eq.2 given below.

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1\text{-}L)^d\, y_{t = c +}\left(1 + \sum_{j=1}^{q} \theta_j L^j\right) \epsilon_t \quad (2)$$

Where the Lag operator denoted using L, moving average term is denoted by ϕ_i, and ε_t denotes the error term. The p, d, q refer as order of ARIMA model.

2) *SARIMA Model [2]:* ARIMA with seasonal effect is called SARIMA. The SARIMA model can be represented as ARIMA (p, d, q) (P, D, Q, m). Here, model parameter p represents non-seasonal AR term, d represents non-seasonal differencing, and q represents non-seasonal MA term respectively. While P, D, Q, and m represents the seasonal AR term, seasonal differencing, seasonal MA term and time span of repetitive seasonal effect respectively. Mathematically SARIMA can be represented using Eq.3 given below.

$$\emptyset_p\,(B)\ \phi_p(B^s)(1-B^s)^D(1-B)^d Z^t = \theta_q(B)\Theta_Q(B^s)\epsilon_t \quad (3)$$

Where, forecast variable is represented using Z^t, $\phi_p(B)$ shows AR polynomial of order p, $\Phi_p(B^s)$ represents seasonal AR polynomial of order P, $\theta_q(B)$ is MA polynomial of order q, $\Theta_Q(B^s)$ represents seasonal MA polynomial of order Q and ε_t is white noise process. The nonseasonal and seasonal differencing operators represented using $(1-B)^d$ and $(1-B^s)^D$ respectively. Whereas d, D and s are order of difference, seasonal order of difference and seasonal length, respectively. The order of ARIMA and SARIMA model is determined through grid search for WQI.

F. Performance matrices [4]

G. Time series forecasting models' accuracy can be determined using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE shows the deviation between the original values and predicted values whereas Mean Squared Error (MSE) or Mean Squared Deviation (MSD) is the average of squares of errors and RMSE value is the square root of MSE. Here, deviation between actual and estimated values e_i, for i = 0,1,2,3...n is called error. The model which have lower MAE & RMSE is considered as best model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad (5)$$

$$\text{RMSE}=\sqrt{MAE} \qquad (6)$$

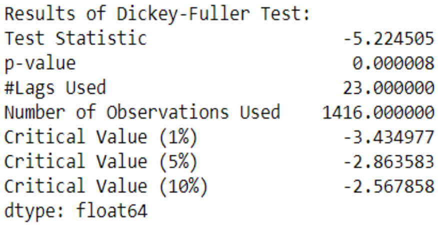
IV.DISCUSSION ON RESULTS

The important results of the developed time-series models for the forecasting WQI is discussed in this section.

A. Stationary Test [2]

The stationary test is performed on the pre-processed data at the individual variable level by using the augmented Dicky-Fuller test. The rolling statistics is the plot of the mean and standard deviation values of the time series data. The rolling statistics and Dicky-Fuller test results of the WQI time-series data are shown in Fig. 3. The dicky-fuller test results showed that the test statistics of WQI has the p-value < 0.05, and hence, the WQI time series is considered as stationary.

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*



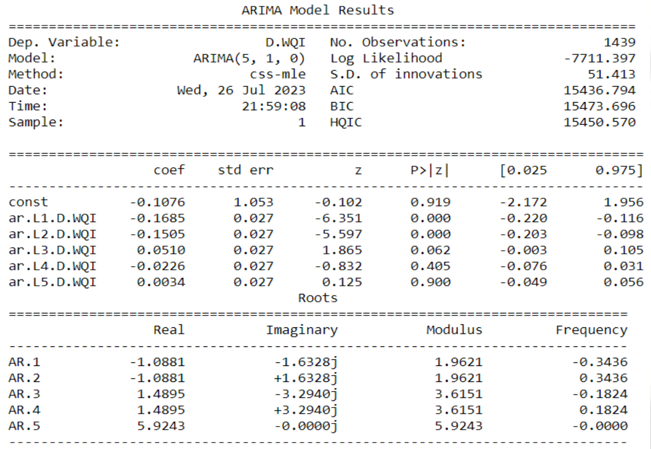
```
Results of Dickey-Fuller Test:
Test Statistic                   -5.224505
p-value                           0.000008
#Lags Used                       23.000000
Number of Observations Used    1416.000000
Critical Value (1%)              -3.434977
Critical Value (5%)              -2.863583
Critical Value (10%)             -2.567858
dtype: float64
```

Fig. 2: Dicky-Fuller test for WQI

B. Model Recognition

To estimate and evaluate the best order of the non-seasonal ARIMA and seasonal ARIMA model, the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are the time series statistical approaches. Table II, Fig.2, Fig.3 shows the obtained best order and its lower AIC, BIC values for ARIMA and SARIMA models for this study.

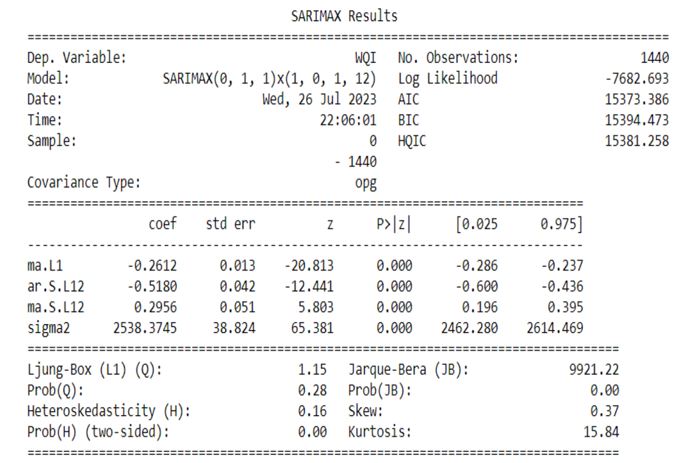
Model Name	Water Quality Parameters	WQI
ARIMA	Order (p, d, q)	(5,1,0)
	AIC	15436.794
	BIC	15473.696
SARIMA	Order (p, d, q)	(0,1,1)
	Seasonal Order (P, D, Q, m)	(1,0,1,12)
	AIC	15373.386
	BIC	15394.473



```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                  D.WQI   No. Observations:                 1439
Model:                 ARIMA(5, 1, 0)   Log Likelihood               -7711.397
Method:                       css-mle   S.D. of innovations             51.413
Date:                Wed, 26 Jul 2023   AIC                          15436.794
Time:                        21:59:08   BIC                          15473.696
Sample:                             1   HQIC                         15450.570

===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -0.1076      1.053     -0.102      0.919      -2.172       1.956
ar.L1.D.WQI    -0.1685      0.027     -6.351      0.000      -0.220      -0.116
ar.L2.D.WQI    -0.1505      0.027     -5.597      0.000      -0.203      -0.098
ar.L3.D.WQI     0.0510      0.027      1.865      0.062      -0.003       0.105
ar.L4.D.WQI    -0.0226      0.027     -0.832      0.405      -0.076       0.031
ar.L5.D.WQI     0.0034      0.027      0.125      0.900      -0.049       0.056
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.0881           -1.6328j            1.9621           -0.3436
AR.2           -1.0881           +1.6328j            1.9621            0.3436
AR.3            1.4895           -3.2940j            3.6151           -0.1824
AR.4            1.4895           +3.2940j            3.6151            0.1824
AR.5            5.9243           -0.0000j            5.9243           -0.0000
-----------------------------------------------------------------------------
```

Fig.3: ARIMA Model Results



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                WQI   No. Observations:                 1440
Model:             SARIMAX(0, 1, 1)x(1, 0, 1, 12)   Log Likelihood               -7682.693
Date:                            Wed, 26 Jul 2023   AIC                          15373.386
Time:                                    22:06:01   BIC                          15394.473
Sample:                                         0   HQIC                         15381.258
                                           - 1440
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.2612      0.013    -20.813      0.000      -0.286      -0.237
ar.S.L12      -0.5180      0.042    -12.441      0.000      -0.600      -0.436
ma.S.L12       0.2956      0.051      5.803      0.000       0.196       0.395
sigma2      2538.3745     38.824     65.381      0.000    2462.280    2614.469
===================================================================================
Ljung-Box (L1) (Q):                   1.15   Jarque-Bera (JB):              9921.22
Prob(Q):                              0.28   Prob(JB):                         0.00
Heteroskedasticity (H):               0.16   Skew:                             0.37
Prob(H) (two-sided):                  0.00   Kurtosis:                        15.84
===================================================================================
```

Fig.4: SARIMA Model Results

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

C. Prediction using Univariate Models

In this study, after performing the stationary test, two univariate time-series models such as ARIMA, SARIMA are used to train water quality time series over the period of Six years from 2011 to 2016. Prediction have been done for six-year data from 2017 to 2022. The results for prediction of individual water quality parameters such as WQI values using ARIMA, SARIMA are shown in Fig. 5 - Fig. 10.

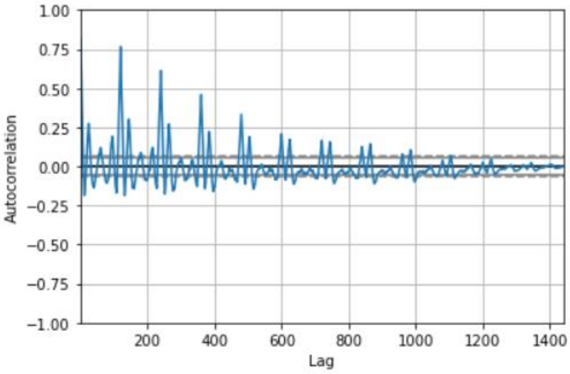


Fig.5: Autocorrelation of WQI

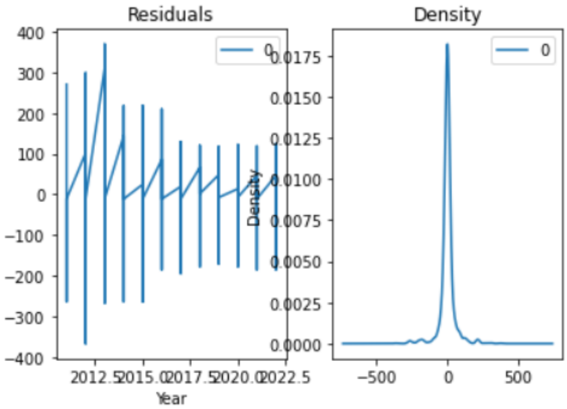


Fig.6: Residuals & Density of WQI

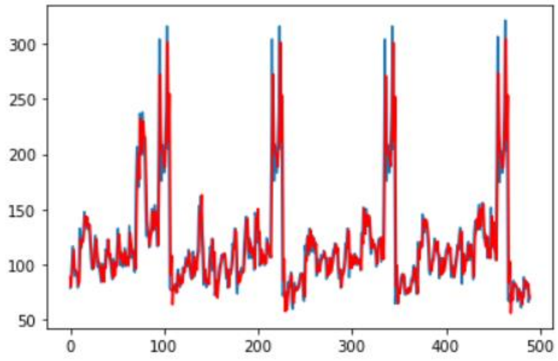
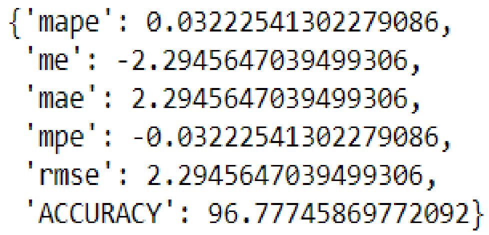


Fig.7: Prediction results of ARIMA Model represented using Red Color



```
{'mape': 0.03222541302279086,
 'me': -2.2945647039499306,
 'mae': 2.2945647039499306,
 'mpe': -0.03222541302279086,
 'rmse': 2.2945647039499306,
 'ACCURACY': 96.77745869772092}
```

Fig.8: Performance matrices & accuracy using ARIMA model

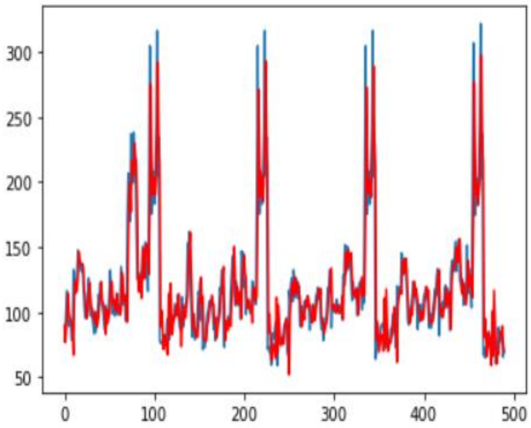
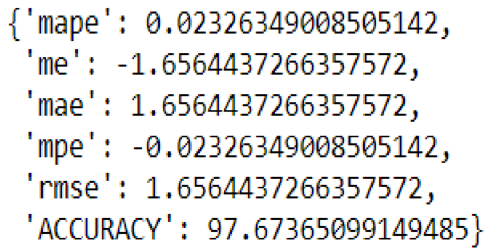


Fig.9: Prediction results of SARIMA Model represented using Red Color



```
{'mape': 0.02326349008505142,
 'me': -1.6564437266357572,
 'mae': 1.6564437266357572,
 'mpe': -0.02326349008505142,
 'rmse': 1.6564437266357572,
 'ACCURACY': 97.67365099149485}
```

Fig.10: Performance matrices & accuracy using ARIMA model

V. COMPARATIVE ANALYSIS

In this section, the implemented models have been compared based on performance matrices such as MAE, RMSE & Accuracy. The model with lower values of MAE and RMSE is considered as the best model for forecasting. Table III shows the values of performance matrices calculated using prediction results done by ARIMA & SARIMA models. Here, SARIMA model provides the lowest MSE and RMSE values & high accuracy and hence, is found to be more effective in prediction.

Model Name	Performance metric	WQI
ARIMA	MAE	2.294564
	RMSE	2.294564
	ACCURACY	96.777458
SARIMA	MAE	1.656443
	RMSE	1.656443
	ACCURACY	97.673650

VI. CONCLUSION

In this study, the prediction of WQI is performed using two most widely used time series forecasting models, such as ARIMA & SARIMA. The comparative analysis of models is performed to find the best suitable model having the least values of MAE, RMSE and high accuracy for WQI prediction. Here, SARIMA model performed best with the least values of MAE and RMSE. Further study can be done using hybrid and ANN models to achieve higher accuracy in predicting and classifying the WQI.

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] K.W. Abdelmalik, “Role of statistical remote sensing for Inland water quality parameters prediction”, The Egyptian Journal of Remote Sensing and Space Science, Volume 21, Issue 2,2018, Pages 193-200, ISSN 1110-9823, https://doi.org/10.1016/j.ejrs.2016.12.002.

[6] F. Al-Badaii, M. Shuhaimi-Othman, M. Barzani Gasim, "Water Quality Assessment of the Semenyih River, Selangor, Malaysia", Journal of Chemistry, vol. 2013, Article ID 871056, 10 pages, 2013. https://doi.org/10.1155/2013/871056.

[7] A. K. Bisht, R. Singh, R. Bhutiani, A. Bhatt, "Artificial Neural Network Based Water Quality Forecasting Model for Ganga River", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-8 Issue-6, August 2019.

[8] B. Srivastava, S. S. Sandha, V. Raychoudhury, S. Randhawa, V. Kapoor and A. Agrawal, "Building an Open, Multi-Sensor, Dataset of Water Pollution of Ganga Basin and Application to Assess Impact of Large Religious Gatherings," 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 2020, pp. 1-6, doi: 10.1109/PerComWorkshops 48775.2020.9156149.

[9] P. Kumar, R. K. Kaushal, A. K. Nigam, "Assessment and Management of Ganga River Water Quality Using Multivariate Statistical Techniques in India", Asian Journal of Water, Environment and Pollution, vol. 12, no. 4, pp. 61-69, 2015.

[10] A. K. Shukla, C. S. P. Ojha and R. D. Garg, "Surface water quality assessment of Ganga River Basin, India using index mapping," 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 2017, pp. 5609-5612, doi: 10.1109/IGARSS. 2017. 8128277.

[11] G. Tripathi, A. Chandra Pandey and B. Ranjan Parida, "Spatio- Temporal Analysis of Turbidity in Ganga River in Patna, Bihar Using Sentinel-2 Satellite Data Linked with Covid-19 Pandemic," 2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS), Ahmedabad, India, 2020, pp. 29-32, doi: 10.1109/InGARSS48198. 2020.9358965.

[12] S. Shakhari, A. K. Verma and I. Banerjee, "Remote Location Water Quality Prediction of the Indian River Ganga: Regression and Error Analysis," 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 2019, pp. 1-5, doi: 10.1109/ICTKE47035.2019.8966796.