



Multi-ahead electrical conductivity forecasting of surface water based on machine learning algorithms

Deepak Kumar¹ · Vijay Kumar Singh² · Salwan Ali Abed³ · Vinod Kumar Tripathi¹ · Shivam Gupta⁴ · Nadhir Al-Ansari⁵ · Dinesh Kumar Vishwakarma⁶ · Ahmed Z. Dewidar^{7,8} · Ahmed A. Al-Othman⁸ · Mohamed A. Mattar^{7,8,9}

Received: 27 April 2023 / Accepted: 13 August 2023 / Published online: 8 September 2023
© The Author(s) 2023

Abstract

The present research work focused on predicting the electrical conductivity (EC) of surface water in the Upper Ganga basin using four machine learning algorithms: multilayer perceptron (MLP), co-adaptive neuro-fuzzy inference system (CANFIS), random forest (RF), and decision tree (DT). The study also utilized the gamma test for selecting appropriate input and output combinations. The results of the gamma test revealed that total hardness (TH), magnesium (Mg), and chloride (Cl) parameters were suitable input variables for EC prediction. The performance of the models was evaluated using statistical indices such as Percent Bias (PBIAS), correlation coefficient (R), Willmott's index of agreement (WI), Index of Agreement (PI), root mean square error (RMSE) and Legate-McCabe Index (LMI). Comparing the results of the EC models using these statistical indices, it was observed that the RF model outperformed the other algorithms. During the training period, the RF algorithm has a small positive bias (PBIAS = 0.11) and achieves a high correlation with the observed values ($R = 0.956$). Additionally, it shows a low RMSE value (360.42), a relatively good coefficient of efficiency (CE = 0.932), PI (0.083), WI (0.908) and LMI (0.083). However, during the testing period, the algorithm's performance shows a small negative bias (PBIAS = -0.46) and a good correlation ($R = 0.929$). The RMSE value decreases significantly (26.57), indicating better accuracy, the coefficient of efficiency remains high (CE = 0.915), PI (0.033), WI (0.965) and LMI (-0.028). Similarly, the performance of the RF algorithm during the training and testing periods in Prayagraj. During the training period, the RF algorithm shows a PBIAS of 0.50, indicating a small positive bias. It achieves an RMSE of 368.3, R of 0.909, CE of 0.872, PI of 0.015, WI of 0.921, and LMI of 0.083. During the testing period, the RF algorithm demonstrates a slight negative bias with a PBIAS of -0.06. The RMSE reduces significantly to 24.1, indicating improved accuracy. The algorithm maintains a high correlation ($R = 0.903$) and a good coefficient of efficiency (CE = 0.878). The index of agreement (PI) increases to 0.035, suggesting a better fit. The WI is 0.960, indicating high accuracy compared to the mean value, while the LMI decreases slightly to -0.038. Based on the comparative results of the machine learning algorithms, it was concluded that RF performed better than DT, CANFIS, and MLP. The study recommended using the current month's total hardness (TH), magnesium (Mg), and chloride (Cl) parameters as input variables for multi-ahead forecasting of electrical conductivity (EC_{t+1} , EC_{t+2} , and EC_{t+3}) in future studies in the Upper Ganga basin. The findings also indicated that RF and DT models had superior performance compared to MLP and CANFIS models. These models can be applied for multi-ahead forecasting of monthly electrical conductivity at both Varanasi and Prayagraj stations in the Upper Ganga basin.

Keywords Decision tree · Multilayer perceptron · Random forest · Co-adaptive neuro-fuzzy inference system · Electrical conductivity

Introduction

Clean and safe water is essential for human consumption, hygiene, and sanitation. Contaminated water harbors harmful microorganisms, pathogens, and pollutants that cause

waterborne diseases, such as diarrhea, cholera, typhoid, and hepatitis (Nemčić-Jurec et al. 2022). Water quality plays a crucial role in maintaining the health of aquatic ecosystems, including rivers, lakes, wetlands, and oceans. Aquatic organisms, including fish, plants, and other wildlife, depend on clean water to survive and thrive. Polluted water, containing chemicals, toxins, or excessive nutrients, leads to habitat

Extended author information available on the last page of the article

degradation, species loss, and disruptions in the ecological balance (Bhat et al. 2022). Water quality is critical for agriculture and food production. Contaminated water used for irrigation negatively impacts crop health and productivity. It also leads to the accumulation of harmful substances in food, posing risks to human health. Maintaining good water quality in irrigation systems is essential for sustainable agriculture and safe food production (Mensah-Akutteh et al. 2022). Water quality is crucial for human health, ecosystem functioning, sustainable development, and various economic activities. It is essential to monitor, manage, and maintain water quality to safeguard human well-being, protect the environment, and ensure the sustainable use of this precious resource (Giri et al. 2022).

Electrical conductivity (EC) is an important parameter used to assess and monitor water quality. It is a measure of the ability of water to conduct an electric current and is influenced by the concentration of dissolved ions in the water (Naiel et al. 2022). Electrical conductivity provides an estimate of the TDS concentration in water. TDS refers to the total concentration of inorganic and organic substances dissolved in water, including salts, minerals, metals, and other dissolved solids. High EC values generally indicate higher TDS levels, which can affect the taste, odor, and overall palatability of water (Ezea et al. 2022). Electrical conductivity is commonly used as a measure of salinity in water bodies. Salinity refers to the concentration of salts, particularly sodium chloride, in water. High EC values indicate higher salinity levels, which have adverse effects on aquatic life, irrigation practices, and the suitability of water for various uses (Liu et al. 2023). Changes in electrical conductivity indicates the presence of pollutants or contaminants in water. Elevated EC levels may suggest the presence of dissolved substances like heavy metals, industrial chemicals, agricultural runoff, or sewage contamination. Monitoring EC can help identify potential pollution sources and assess the overall water quality (Lakrouit et al. 2022). Electrical conductivity is also used in agriculture to assess soil and irrigation water quality. High EC in irrigation water leads to increased soil salinity, negatively affecting soil fertility and crop growth. Monitoring EC helps farmers and agricultural professionals make informed decisions about irrigation practices and nutrient management to maintain soil health and crop productivity (Gomaa 2020). Electrical conductivity plays a significant role in assessing water quality by providing information about TDS concentration, salinity, pollution levels, and suitability for various uses. Monitoring EC helps identify potential issues, implement appropriate water management practices, and ensure compliance with regulatory standards to protect both human health and the environment.

EC measurement is primarily sensitive to the presence of ions in the water. It may not detect non-ionic substances or organic contaminants that do not significantly contribute

to the electrical conductivity. Therefore, EC measurements alone may not provide a comprehensive assessment of water quality in terms of organic pollutants or certain non-ionic substances (Igboama et al. 2022). EC is influenced by temperature variations. Changes in temperature affects the accuracy of EC measurements. To overcome this limitation, temperature compensation techniques or standardized temperature conditions are often employed, but variations in temperature still introduce uncertainties in the measurement (Ghorbani et al. 2017). EC measurements are typically point measurements taken at a specific location in the water body. This limited spatial resolution may not capture the spatial heterogeneity of water quality parameters, especially in large or complex water systems. To obtain a more comprehensive understanding of water quality, additional sampling points or continuous monitoring techniques may be required (Mahanta et al. 2022). While changes in EC indicates the presence of contaminants, it does not provide information on the specific sources or types of contaminants. Additional testing and analysis are often needed to identify and quantify specific pollutants in the water (Maansi et al. 2022). It is important to consider these limitations when interpreting EC measurements and to complement them with other water quality parameters and analyses to obtain a more comprehensive understanding of water quality conditions and potential pollutant sources. The accurate measurements of EC in laboratory with considering these limitation is very difficult, time consuming, require large amount of assets, require high skill man power and also need large amount of economy. The limitations of electrical conductivity (EC) measurements pose challenges in obtaining accurate and efficient results, especially when considering factors like time, cost, and expertise. Machine learning algorithms offer a promising approach to address these challenges and enhance the accuracy and efficiency of EC prediction (Taşan 2023). By training machine learning models using historical data that includes both EC measurements and associated water quality parameters, it is possible to establish relationships and patterns that are used to predict EC values in the future. Machine learning models have the potential to capture complex relationships and nonlinear dependencies between EC and other water quality parameters, allowing for more accurate predictions (Boubakri et al. 2023). Using machine learning algorithms also helps overcome the limitations of EC measurements by incorporating additional variables and parameters that may contribute to water quality. By considering a wider range of inputs, such as temperature, pH, dissolved oxygen, and specific ion concentrations, machine learning models provides a more comprehensive understanding of water quality conditions and potential pollutant sources. Additionally, machine learning models automate the prediction process, reducing the need for extensive manual

labor and time-consuming laboratory measurements. Once the model is trained, it quickly processes new data and provides near real-time predictions of EC, making it a more efficient and cost-effective solution compared to traditional laboratory measurements.

The issue of water salinity is especially vital in arid and semiarid, mostly waterfront zones, where overuse of groundwater to cover temperate crop water needs, particularly in summer, may prompt ocean water interruption, in this manner falling apart irrigation water quality, lastly, bringing about the supposed optional soil salinization (Szabolcs 1986). High salt concentrations in waste effluents can build up the salinity of the water, which may result in unfriendly natural impacts on amphibian biota (Dunlop et al. 2005). Electrical conductivity can be utilized for evaluating the impact of various particles on chemical balance, physiological consequences for plants or creatures, and erosion rates (Kisi and Ay 2014). Environmentalists and ecologists are applying electrical conductivity to follow the movement of the water system in order to foresee and take care of issues with marine life (Chapman 1996). An effective investigation of salinity and irrigation water quality needs electrical conductivity data, which is measured only at specific stations of the central water commission in India. The large number of relationships developed with other water quality parameters for estimating the electrical conductivity of water. However, no forecasting model was developed for the prediction of the electrical conductivity of water.

In recently years, the large quantity of research was completed based on machine-learning in several fields such as hydrology (Mirbagheri et al. 2010; Noori et al. 2011; Kisi and Shiri 2012; Atieh et al. 2015; Singh et al. 2016a, b, 2017, 2018a, b, 2022a, b; Kumar et al. 2019; Esfahani and Keshavarzi 2020; Elbeltagi et al. 2022, 2023a, b, c; Markuna et al. 2023; Achite et al. 2023; Saroughi et al. 2023; Mirzania et al. 2023), ground water level forecasting (Samantarray et al. 2022; Saroughi et al. 2023), soil permeability and hydraulic conductivity (Sihag et al. 2019a, b; Singh et al. 2022b), soil moisture (Gu et al. 2021), sprinkler irrigation's wind drift and evaporation losses (Al-Ghobari et al. 2018; Mattar et al. 2022), soil temperature (Tabari et al. 2011; Napagoda and Tilakaratne 2012; Singh et al. 2018b), evapotranspiration (Mattar 2018; Roy et al. 2022), irrigation water infiltration rate (Mattar et al. 2017), meteorological droughts and standardized precipitation index (SPI) (Bouaziz et al. 2021; Elbeltagi et al. 2023b), global solar radiation (Hassan et al. 2017; Voyant et al. 2017; Cornejo-Bueno et al. 2019; Feng et al. 2019; Ağbulut et al. 2021), water quality (Maier et al. 2010; Barzegar et al. 2018; Deng et al. 2021; Singha et al. 2021), medical (Chen et al. 2015; Venkatesan and Velmurugan 2015; Podolsky et al. 2016). Tutmez et al. (2006) applied the adaptive neuro-fuzzy inference system (ANFIS) algorithm to predict the EC of groundwater using

ions present in the water. The performance of the ANFIS algorithm was better than that of traditional approaches. Wen et al. (2013) applied artificial intelligence techniques to develop a reliable prediction model for the dissolved oxygen (DO) of surface water in the Heihe River. Also, most input variables influence on output was examined by sensitivity analysis. Nemati et al. (2015) investigated the usability of machine learning and regression algorithms in the simulation of the DO. It is the most effective physical properties of quality water to study aquatic subsidies, health care, water quality, and aquatic ecosystems. Eight different physiochemical parameters were used for the prediction of DO in two different rivers. Ghorbani et al. (2017) forecasted the EC concentration of surface water in the Asi River based on a machine learning algorithm to assess the water quality based on physiochemical parameters of surface water. In addition, outcomes were compared with regression-based techniques. Khadr and Elshemy (2017) studied the abilities of the ANFIS model for forecasting different water quality factors based on electrical conductivity, turbidity, discharge, water temperature, dissolved oxygen, total suspended solids, total dissolved solids, and pH. Najah Ahmed et al. (2019) applied different artificial intelligence algorithms for the prediction of suspended solids, ammoniacal nitrogen, and pH of surface water in the Johor River Basin. The results of all algorithms were found satisfactory. Ekemen Keskin et al. (2020) used machine learning and regression-based techniques to predict groundwater EC. Various hydro-chemical parameters were applied as input variables, which affect the EC. Mokhtar et al. (2022) demonstrated the potential of using multiple regression models and the developed machine learning methods for predicting the irrigation water quality index. The models were served as rapid decision tools for modeling irrigation water quality, assisting in water resource management strategies. Masoudi et al. (2023) demonstrated that the RF model outperformed the other models, albeit with only a slight difference in accuracy. Additionally, the analysis of relative importance highlighted that the sodium adsorption ratio (SAR) and pH had the most significant impact in explaining groundwater quality classes.

After reviewing the importance of fresh water, the role of electrical conductivity for the assessment of water quality, limitations during measurement of electrical conductivity, potential and accuracy of machine learning, and its wide range of applications in different fields. The authors applied four different machine learning algorithms namely decision tree (DT), multilayer perceptron (MLP), random forest (RF), and co-adaptive neuro-fuzzy inference system (CANFIS) for the prediction of the electrical conductivity of surface water in the Upper Ganga basin. The study aimed to investigate the optimal input–output arrangements and the length of training and testing datasets using cross-correlation and gamma testing. Additionally, the researchers aimed to develop an

efficient machine learning model for multi-ahead forecasting of electrical conductivity in the Upper Ganga basin and compare the performance of the DT, MLP, RF, and CANFIS algorithms for electrical conductivity prediction.

Materials and methods

Data collection and study area

The Ganga River originates in the Uttarakhand state of India, beginning with the Gangotri glacier. The present study was concentrated on Varanasi and Prayagraj sites switched in the main branch of the Upper Ganga basin, as presented in Fig. 1. The total area of 87,787 km² is covered by the Upper Ganga Basin. The high-altitude difference present in the Upper Ganga basin ranges from 7500 m in the upper mountain region to 100 m in the lower plains. The study site is presented in the lower plane of the Upper Ganga basin. The yearly normal precipitation varies from 550 to 2500 mm in mountainous to plain areas of the Upper Ganga basin. The monthly water sample was picked up from Prayagraj and Varanasi sites of the Upper Ganga River basin and analyzed by the Central Water Commission (CWC), New Delhi, India. The water sample was collected in the first week of the month. The total of 37 years from 1980 to 2016 of water quality data was acquired from the divisional office of the Central Water Commission, Lucknow, Uttar Pradesh, India.

The total data was divided into two different sets: the training data set (362), used for the development of the model, and the testing data set (82) used for multi-ahead validation of the model.

Machine learning algorithms

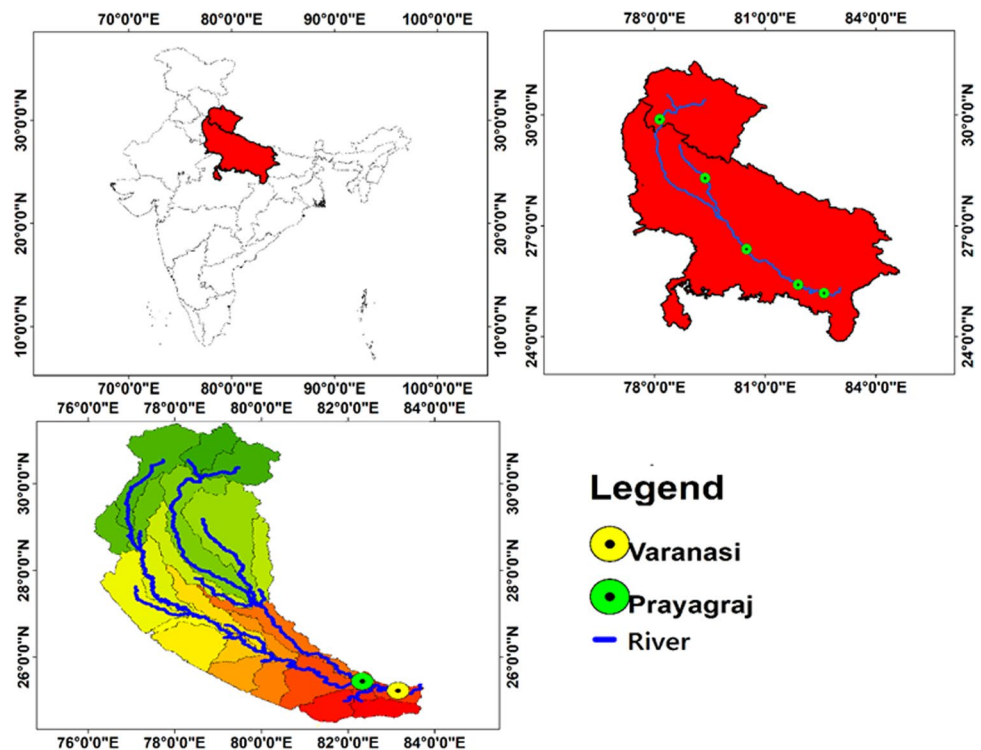
Multilayer perceptron (MLP) algorithm

MLP is a system architecture of ANN that consists of one or more hidden layers between the independent vectors (input vectors) and dependent vector (output vector). The independent layer nodes obtain signal permits to depended layer by hidden layer (Singh et al. 2018a). The independent vectors are multiplied over and done with the initial weight value; then, produce every neuron input vector of the next layer. Then, acquire dependent of hidden (dependent vector) layer neurons through activation function. The neurons in each node were found through a large number of trails. The literature on MPL can be found in Fu (1994); Srinivasulu and Jain (2006); Rezaeianzadeh et al. (2014).

Co-adaptive neuro-fuzzy inference system (CANFIS) algorithm

A standout among the most prominent incorporated frameworks is CANFIS, which has indicated promising outcomes in displaying nonlinear arrangements. In CANFIS, a Takagi–Sugeno-type fuzzy induction framework is utilized. The

Fig. 1 Location information about the present study area of Upper Ganga basin



independent yield of each standard can be a direct composite of independent input vectors in addition to a consistent term. The data space of CANFIS is disconnected into three one-of-a-kind parts to store the data (Malik and Kumar 2015; Singh et al. 2016b; Gholami et al. 2021). The methodologies are named scatter, grid, and art. Fragmentary examination is normally utilized as a piece of control world in light of its rightness (Aytek 2009). Fuzzy inference frameworks are also refreshing as they relate the illustrative idea of tenets (membership function) with the direction of “black-box” neural systems. The optimal parameters of CANFIS algorithms were defined by a large number of trails. The elementary information about CANFIS algorithms was found in Mizutani and Jang (1995); Aytek (2009); Tahmasebi and Hezarkhani (2011).

Random forest (RF) regression algorithm

The RF algorithm is a characterization and regression strategy consisting of a mix of tree indicators, where each tree is produced utilizing an arbitrary vector examined freely from the information vector. In regression, the RF regression algorithm is a troupe strategy fit for performing both relapse and order undertakings with the utilization of different choice (decision) trees and a method called Bootstrap and Aggregation, ordinarily known as bagging (Breiman 1999, 2001). The essential thought behind this is to join numerous choice trees in deciding the last yield as opposed to depending on singular choice trees (Hastie et al. 2001; Karimi et al. 2018). The RF algorithm has numerous choice trees as base learning models. We arbitrarily perform line examining and highlight inspecting from the dataset, shaping example datasets for each model. This part is called Bootstrap (Pal and Mather 2003; Chen et al. 2017). Bootstrap alludes to irregular inspection with substitution. Bootstrap permits us to more likely comprehend the predisposition and the difference in the dataset. Bootstrap includes the arbitrary examination of a small subset of information from the dataset. It is a general strategy that can be utilized to lessen the fluctuation for those calculations that have high differences, commonly called choice trees. Sacking makes each model run freely and afterward totals the yields toward the end without inclination to any model (Quinlan 1992; Adams and Sterling 1992).

Decision tree (DT) algorithm

The DT algorithm is one of the outstanding classification strategies. DT constructs relapse or characterization models as a tree structure. It separates a dataset into smaller and smaller subsets while simultaneously a related DT is gradually evolved. The conclusive outcome is a tree with choice (decision) hubs and leaf hubs (Fakhari and Moghadam 2013). A choice hub (e.g., Outlook) has at least two

branches, each speaking to values for the property tried (Nasridinov et al. 2013; Nagalla et al. 2017). The leaf hub speaks to a choice on the numerical objective. The highest-choice hub in a tree, which compares to the best indicator called the root hub. Choice trees can deal with both all-out and numerical information (Lee et al. 2013; Loh 2014). The facts of the DT algorithm can be assessed from (Harb et al. 2009; Fakhari and Moghadam 2013; Czajkowski and Kretowski 2016). The flow diagram of methodology is presented in Fig. 2.

Performance evaluation

It is very important to clearly identify the criteria that are used to judge the performance of the model in order to evaluate its effectiveness. Throughout the literature, it has been clearly mentioned that the performance of the model in this study was measured according to a set of criteria. All of the

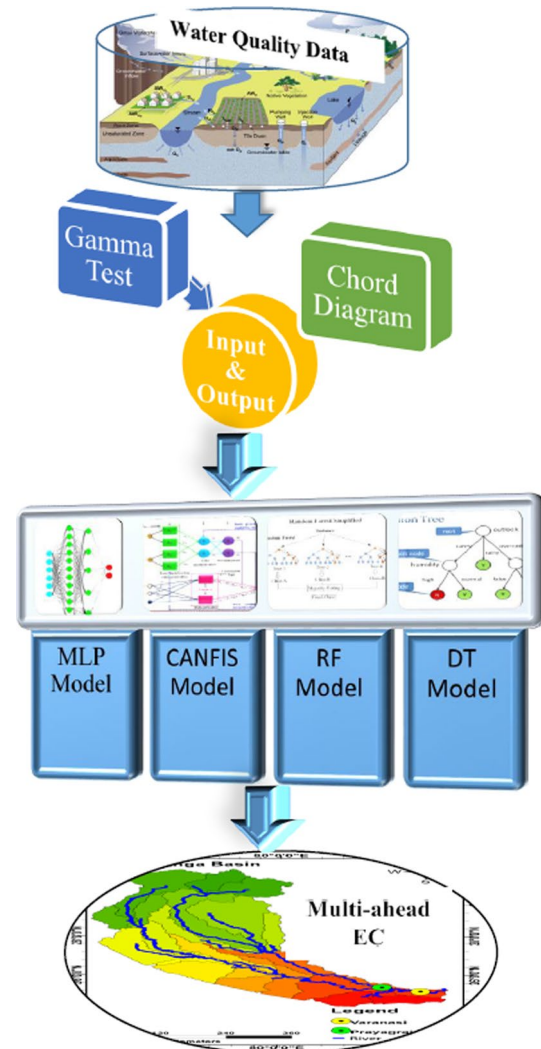


Fig. 2 Flowchart of complete process

forementioned decision tree (DT), multilayer perceptron (MLP), random forest (RF), and co-adaptive neuro-fuzzy inference system (CANFIS) models were evaluated on the basis of a variety of measures of goodness of fit in order to assess their predictive power. A variety of statistical indices were used in this analysis, including Percent Bias (PBIAS), Root Mean Square Error (RMSE), Correlation Coefficient (R), Coefficient of Efficiency (CE), Index of Agreement (PI), Willmott's Index (WI) and Legate-McCabe Index (LMI). The following equations were used to calculate these indices (Vishwakarma et al. 2018; Kumar et al. 2022; Singh et al. 2022a, b; Khatti and Grover 2023):

$$\text{PBIAS}(\%) = \frac{\sum_{i=1}^N (EC_p - EC_t)}{\sum_{i=1}^N EC_t} \times 100 \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (EC_t - EC_p)^2}{n}} \quad (2)$$

$$R = \left[\frac{\sum_{i=1}^n \{ (EC_t - \overline{EC_t})(EC_p - \overline{EC_p}) \}}{\sqrt{\sum_{i=1}^n (EC_t - \overline{EC_t})^2} \sqrt{\sum_{i=1}^n (EC_p - \overline{EC_p})^2}} \right] \quad (3)$$

$$\text{CE} = 1 - \frac{\sum_{i=1}^n (EC_t - EC_p)^2}{\sum_{i=1}^n (EC_t - \overline{EC_t})^2} \quad (4)$$

$$\text{PI} = 1 - \frac{\sum_{i=1}^n (EC_t - EC_p)^2}{\sum_{i=1}^n (|EC_p - \overline{EC_t}| + |EC_t - \overline{EC_p}|)^2} \quad (5)$$

$$\text{WI} = 1 - \frac{\sum_{i=1}^n (EC_t - EC_p)^2}{\sum_{i=1}^n (|EC_p - \overline{EC_t}| + |EC_t - \overline{EC_p}|)^2} \quad (6)$$

$$\text{LMI} = 1 - \left[\frac{\sum_{i=1}^n |EC_p - \overline{EC_p}|}{\sum_{i=1}^n |EC_t - \overline{EC_t}|} \right] \quad (7)$$

where EC_t is measured electrical conductivity data series, EC_p is modeled electrical conductivity data series, n is number of observations, $\overline{EC_t}$ is average of the measured electrical conductivity data series and $\overline{EC_p}$ is average of modeled electrical conductivity data series.

R measures how well the model fits the data, but RMSE measures how well the model fits the data in an absolute sense, and a lower value of RMSE indicates that the model fits better. The index of agreement (d) was developed by

Willmott (1981) as a standardized measure of the degree of model prediction error and varies between 0 and 1. A value of 1 indicates a perfect match, and 0 indicates no agreement at all. The model that has the highest index of agreement (d), r , and R^2 , and the least error (RSME), will be the best model (Kushwaha et al. 2021; Vishwakarma et al. 2022a, b).

In addition to the statistical parameters stated in Eqs. (1–7), the correctness of the investigated models was validated using a scatter plot, Violin diagram, a discrepancy ratio (DR), and a Taylor diagram (TD) (Taylor 2001), among other. A simplified definition of the Taylor diagram is a thorough depiction of the observed and expected data (Taylor 2001). Taylor delivered a single demonstration demonstrating how to simultaneously show several assessment metrics in real-time. Correlation coefficients and standard deviation values between expected and observed values might be shown in this diagram to aid in the detection of changes between the two values (Taylor 2001; Elbeltagi et al. 2023b; Saroughi et al. 2023; Vishwakarma et al. 2023).

Results

In this study, the capability of four different modeling algorithms for forecasting the electrical conductivity of surface water was evaluated based on performance evaluation criteria (PBIAS, RMSE, R, CE, PI, WI and LMI) obtained from observed and modeled datasets during training (calibration) and testing periods. The obtained outcomes from four various algorithms were presented in terms of a scatter diagram and a time series diagram between measured and modeled electrical conductivity of surface water for the duration of testing. Four different algorithms, specifically MLP, RF, CANFIS, and DT, were applied for the prediction of the EC of surface water in the Upper Ganga basin at Varanasi and Prayagraj sites and developed models to compare with each other. Table 1 illustrates the parameters of these four different algorithms applied for prediction. The modeling capability of MLP, RF, CANFIS, and DT algorithms is influenced by their parameters. The optimal parameters of different algorithms were defined through a large number of trials and compared with different performance indices. Similarly, Table 2 illustrates the performance of the developed model during the training and testing periods at the Varanasi and Prayagraj sites. The CANFIS and MLP algorithm-based electrical conductivity forecasting models were established via NeuroSolution 5.0 and the R Studio 3.5.1 version used for the establishment of RF and DT models.

Input vector selection based on gamma test

Gamma test has a decent capability of selecting appropriate input vectors from a large number of input vectors

Table 1 Input and outputs combinations for EC prediction

Model	Output/Inputs	Varanasi		Prayagraj	
		(<i>I</i>)	SE	(<i>I</i>)	SE
E₁	EC_t = f (TH, Mg, Cl)	0.0154	0.0345	0.0148	0.0245
E ₂	EC _t = f (TH, Na, Ca, Har.)	0.5248	0.1846	0.2484	0.1546
E ₃	EC _t = f (Na, Cl, Ca, Har.)	0.1247	0.1521	0.2487	0.1457
E ₄	EC _t = f (Mg, Na, TH, Cl)	0.2144	0.1741	0.2477	0.3851
E ₁	EC _t = f (TH, Na, Mg, Cl)	0.1481	0.2487	0.1944	0.1844
E ₂	EC _t = f (Mg, HCO ₃ , SO ₄)	0.2547	0.3744	0.1942	0.2542
E ₃	EC _t = f (Mg, Na, TH, Cl, Har.)	0.531	0.384	0.315	0.345
E ₄	EC _t = f (Ca, Mg, Na, TH, Cl, Har.)	0.245	0.184	0.471	0.345

*Bold color show best input output combination

Table 2 Performance of different machine learning algorithms at Varanasi.

Output	Algorithms	Architecture	Training							Testing						
			PBIAS	RMSE	r	CE	PI	WI	LMI	PBIAS	RMSE	r	CE	PI	WI	LMI
EC _n	MLP	3-29-1	1.55	372.39	0.932	0.867	0.102	0.964	0.055	-10.77	48.02	0.744	0.876	0.081	0.605	0.305
EC _n	CANFIS	MFs 4	0.24	419.13	0.799	0.638	0.096	0.856	0.109	-7.90	54.19	0.708	0.738	0.082	0.791	0.060
EC _n	RF	Tree=300, mtry = 8	0.11	360.42	0.956	0.932	0.083	0.908	0.083	-0.46	26.57	0.929	0.915	0.033	0.965	-0.028
EC _n	DT	Size of tree = 6	0.12	376.15	0.842	0.709	0.086	0.896	0.083	-1.38	35.23	0.890	0.858	0.044	0.946	-0.093
EC _{n+1}	MLP	3-29-1	1.21	423.39	0.922	0.867	0.121	0.954	0.065	-11.99	48.52	0.701	0.867	0.101	0.585	0.185
EC _{n+1}	CANFIS	MFs 4	0.39	489.13	0.789	0.618	0.099	0.846	0.119	-8.95	51.45	0.695	0.628	0.091	0.788	0.090
EC _{n+1}	RF	Tree=300, mtry = 8	0.17	390.42	0.948	0.912	0.088	0.918	0.085	-1.15	29.56	0.914	0.922	0.060	0.949	0.029
EC _{n+1}	DT	Size of tree = 6	0.79	406.15	0.818	0.686	0.087	0.896	0.085	-2.01	38.58	0.874	0.697	0.066	0.932	-0.004
EC _{n+2}	MLP	3-29-1	1.21	423.39	0.896	0.825	0.121	0.915	0.065	-11.25	49.15	0.662	0.846	0.101	0.525	0.185
EC _{n+2}	CANFIS	MFs 4	0.39	489.13	0.761	0.600	0.166	0.860	0.119	-9.00	58.59	0.763	0.619	0.124	0.661	0.090
EC _{n+2}	RF	Tree=300, mtry = 8	0.17	330.19	0.914	0.900	0.091	0.943	0.085	-1.26	38.47	0.874	0.916	0.047	0.914	0.031
EC _{n+2}	DT	Size of tree = 6	0.79	381.13	0.801	0.667	0.094	0.902	0.085	-2.37	50.14	0.781	0.688	0.069	0.863	-0.004
EC _{n+3}	MLP	3-29-1	3.21	467.47	0.872	0.820	0.156	0.925	0.078	-13.98	77.67	0.642	0.843	0.118	0.511	0.192
EC _{n+3}	CANFIS	MFs 4	3.81	506.86	0.746	0.600	0.195	0.851	0.125	-9.50	76.26	0.673	0.619	0.138	0.571	0.093
EC _{n+3}	RF	Tree=300, mtry = 8	1.72	397.96	0.903	0.897	0.096	0.953	0.087	-2.26	52.66	0.718	0.915	0.049	0.629	0.030
EC _{n+3}	DT	Size of tree = 6	1.99	469.94	0.788	0.658	0.084	0.916	0.086	-3.75	52.59	0.706	0.683	0.064	0.634	-0.004

*Bold and blue color show best input output combination.

*Bold and blue color show best input output combination

(Stefánsson et al. 1997; Shukla et al. 2021; Markuna et al. 2023). A large number of input vectors has a high probability of overfitting the model and increases its complexity (Bajirao et al. 2021a, b). In the present study, we analyzed the twenty-one biological, chemical, and physical parameters of surface water at Varanasi and Prayagraj stations of the Upper Ganga basin using a correlation chord diagram, which is presented in Fig. 3. It shows from the correlation diagram that calcium (Ca), magnesium (Mg), sodium (Na), total hardness (TH), chloride (Cl), and hardness due to MgCO₃ parameters of water quality were correlated with the EC of surface water at both stations. Therefore, these parameters were applied to develop the input–output arrangement of the model based on the gamma test, which is presented in Table 1. Eight different combinations were explored to determine the most efficient input–output arrangements based on the minimum values of gamma and standard error. It was therefore the gamma test applied for electing relevant input and output arrangements. After analyzing the results of the gamma test, it was finalized that total hardness (TH), magnesium (Mg), and chloride (Cl) inputs combinations have a minimum value of gamma (0.0154) and standard error (0.0345) for Varanasi station; gamma (0.0148) and standard error (0.0245) for Prayagraj station. The basic literature can be

assessed in several articles (Stefánsson et al. 1997; Kakaei Lafdani et al. 2013; Czajkowski and Kretowski 2016). Singh et al. (2018a) compared the potential of gamma test and correlation function for sorting appropriate input–output combination. The gamma test was found to be effective in capturing both linear and nonlinear relationships between input and output variables. By comparing its performance across several datasets, Markuna et al. (2023) evaluated the gamma test's robustness and generalization capacity. It was determined that the machine learning model with the greatest potential and efficiency for long-term rainfall was obtained at minimal values for gamma, SE, and V-ratio. According to Kumar et al. (2022), the gamma test (GT) examines the relationship between the information vector, which consists of the days of anthesis, days of silking, days to senescence, plant height, ear length, number of kernels, and gross yield per plant, and the yield vector, represented by the drought-tolerance index (DTI) and stress-tolerance index (STI), and both. The goal of the research was to determine the configuration that produced the lowest gradient, standard error (SE), and gamma values. While the SE offers a way to gauge the degree of uncertainty surrounding the predicted gradient, the gradient indicates the rate of change in the yield vector with respect to the information vector. The information and

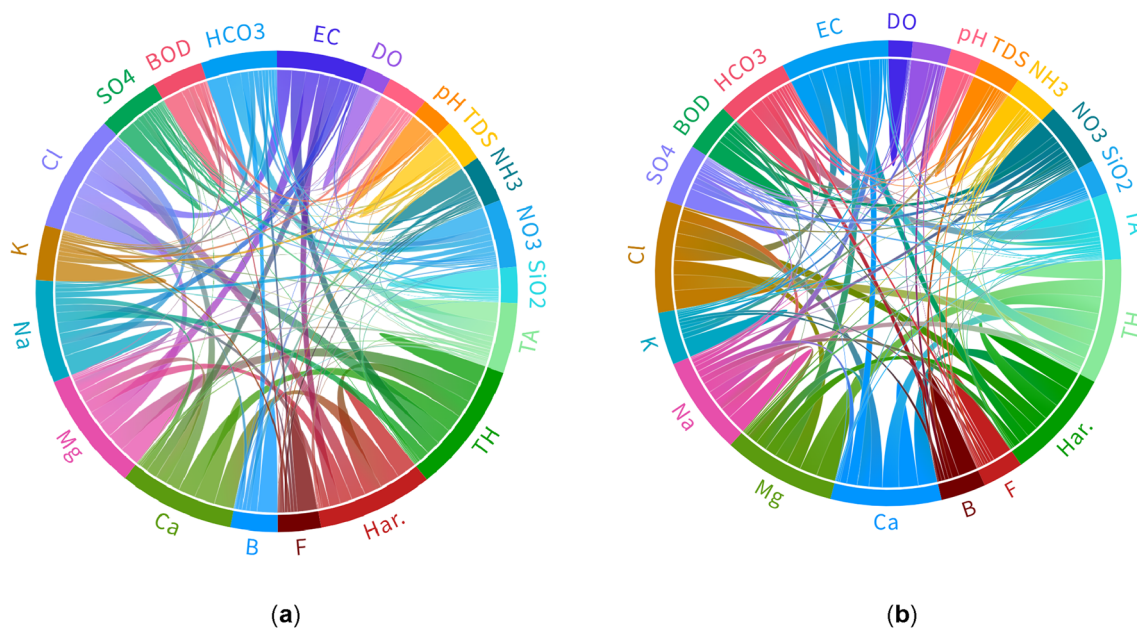


Fig. 3 Correlation chord diagram of parameters in Varanasi (a) and Prayagraj stations (b)

yield vectors have a strong link, which is indicated by the gamma value. Singh et al. (2020) investigated 22 models and chose the best inputs (sand, silt, clay, and OC) as the model's input variables. The best input variables were chosen based on the lowest gamma and SE values, which show a strong and consistent relationship between the chosen input factors and soil permeability. When we sorted the optimal input/output combination, we also discovered a similar outcome.

Comparison of different machine learning algorithms

The results of the developed models are presented in Tables 2 and 3. The MLP model was developed using the Tan hyperbolic (Tanh) transfer function and the Levenberg Marquard (LM) learning rule. The processing element

(neuron) and hidden layer were defined based on trial-and-error technique. The neuron was varied from 1 to 50 with a single hidden layer during the development of the model and also compared for each model based on error value during the training and testing periods. The best architectures of the MLP model for Varanasi and Prayagraj stations were obtained (3–29–1) and (3–11–1), respectively, among all other models. The values of PBIAS (0.11), RMSE (360.42), R (0.956), CE (0.932), PI (0.083), WI (0.908) and LMI (0.085) were found during the training period, and PBIAS (–0.46), RMSE (26.57), R (0.929), CE (0.915), PI (0.033), WI (0.965) and LMI (–0.028) during testing at Varanasi station. In addition, the performance index values of PBIAS (0.50), RMSE (368.3), R (0.909), CE (0.872), PI (0.015), WI (0.921) and LMI (0.083) at the Prayagraj site were estimated during training time. Similarly, values of PBIAS (–0.06),

Table 3 Performance of different machine learning algorithms at Prayagraj.

Output	Algorithms	Architecture	Training							Testing						
			PBIAS	RMSE	R	CE	PI	WI	LMI	PBIAS	RMSE	R	CE	PI	WI	LMI
EC ₀	MLP	3-11-1	1.16	395.8	0.865	0.785	0.099	0.910	0.082	-1.38	48.3	0.856	0.750	0.044	0.860	-0.093
EC ₀	CANFIS	MFs 2	0.13	389.8	0.877	0.866	0.089	0.882	0.096	1.30	52.3	0.826	0.759	0.103	0.882	-0.599
EC₀	RF	Tree=300, mtry = 6	0.50	368.3	0.909	0.872	0.015	0.921	0.083	-0.06	24.1	0.903	0.878	0.035	0.960	-0.038
EC ₀	DT	Size of tree = 5	-1.63	399.8	0.882	0.854	0.104	0.925	0.074	1.68	35.2	0.856	0.867	0.054	0.940	-0.175
EC ₀₊₁	MLP	3-11-1	-1.54	456.3	0.855	0.770	0.110	0.900	0.092	-4.76	47.2	0.753	0.818	0.106	0.722	0.139
EC ₀₊₁	CANFIS	MFs 2	-0.05	439.8	0.869	0.851	0.094	0.882	0.102	-4.50	50.1	0.725	0.739	0.092	0.701	0.096
EC₀₊₁	RF	Tree=300, mtry = 6	-0.22	398.3	0.902	0.892	0.088	0.907	0.085	-0.68	38.2	0.854	0.870	0.074	0.924	0.057
EC ₀₊₁	DT	Size of tree = 5	-1.36	414.8	0.857	0.821	0.104	0.906	0.075	-1.68	40.0	0.821	0.859	0.085	0.881	0.030
EC ₀₊₂	MLP	3-11-1	-1.55	456.3	0.828	0.748	0.144	0.888	0.092	-4.90	79.2	0.526	0.797	0.123	0.618	0.139
EC ₀₊₂	CANFIS	MFs 2	-0.08	409.7	0.838	0.831	0.094	0.882	0.102	-4.54	79.2	0.445	0.725	0.331	0.589	0.096
EC₀₊₂	RF	Tree=300, mtry = 6	-0.31	355.7	0.876	0.879	0.502	0.922	0.085	-0.78	61.2	0.721	0.859	0.487	0.682	0.057
EC ₀₊₂	DT	Size of tree = 5	-1.29	424.3	0.836	0.810	0.125	0.913	0.082	-1.83	63.0	0.575	0.849	0.097	0.635	0.034
EC ₀₊₃	MLP	3-11-1	0.18	487.2	0.809	0.736	0.175	0.888	0.102	-4.90	75.8	0.571	0.709	0.147	0.546	0.147
EC ₀₊₃	CANFIS	MFs 2	1.77	452.4	0.825	0.822	0.580	0.902	0.106	-3.86	73.2	0.575	0.720	0.359	0.550	0.100
EC₀₊₃	RF	Tree=300, mtry = 6	0.42	433.9	0.857	0.821	0.524	0.935	0.087	-0.92	52.3	0.701	0.848	0.512	0.628	0.058
EC ₀₊₃	DT	Size of tree = 5	0.31	476.6	0.839	0.769	0.064	0.931	-0.004	-1.72	57.0	0.641	0.826	0.064	0.594	-0.021

*Bold and blue color show best input output combination

*Bold and blue color show best input output combination

RMSE (24.1), R (0.903), CE (0.878), PI (0.035), WI (0.960) and LMI (-0.038) were found during the testing period. The CANFIS model was prepared by applying the Bell membership function (MFs) and Takagi–Sugeno–Kang (TSK) fuzzy model. In the CANFIS model, membership functions (MFs) per input were defined based on the trial-and-error technique and vary only from 2 to 7 because of the complexity of the model and available computer facilities. The membership function (4) per input was found in the best model among all developed models. The result of the CANFIS model originated in terms of PBIAS (0.24), RMSE (419.13), R (0.799), CE (0.638), PI (0.096), WI (0.865) and LMI (0.109) during the training period and PBIAS (-7.90), RMSE (54.19), R (0.708), CE (0.738), PI (0.082), WI (0.791) and LMI (0.060) during testing time at the Varanasi site.

Moreover, For MLP model, the values of PBIAS = 1.55, RMSE = 372.39, r = 0.932, CE = 0.867, PI = 0.102, WI = 0.964, LMI = 0.055 during training and PBIAS = -10.77 , RMSE = 48.02, r = 0.744, CE = 0.876, PI = 0.081, WI = 0.605, LMI = 0.305, respectively, during the testing period in Varanasi site. In the RF model, the number of trees was decided based on the error rate. The out-of-bag (OOB) was initially dropped, then became more and less constant at 300 trees. The number of m_{try} in the model was definite based on OOB error. The value of OOB was at its minimum (1.81) at 7 m_{try} in the developed model. After comparing the results of the developed RF model, the values of PBIAS = 0.11, RMSE = 360.42, r = 0.956, CE = 0.932, PI = 0.083, WI = 0.908, LMI = 0.083 were obtained during training and PBIAS = -0.46 , RMSE = 26.57, r = 0.929, CE = 0.915, PI = 0.033, WI = 0.965, LMI = -0.028 , respectively, during the testing period in Varanasi. Correspondingly, values of PBIAS = 0.50, RMSE = 368.3, R = 0.909, CE = 0.872, PI = 0.015, WI = 0.921, LMI = 0.083 during training and PBIAS = -0.06 , RMSE = 24.1, R = 0.903,

CE = 0.878, PI = 0.035, WI = 0.960, LMI = -0.038 , respectively, during the testing period in Prayagraj. In the DT model, the size of the tree was defined based on the cross validation and pruning of the model. The appropriate size of the tree was eight based on the cross validation and pruning of the tree. The size of the tree (6) in the developed DT model was obtained after cross validation and pruning of the tree. The results obtained from the DT model showed the values of PBIAS = 0.12, RMSE = 376.15, r = 0.842, CE = 0.709, PI = 0.086, WI = 0.896, LMI = 0.083 in the training period and PBIAS = -1.38 , RMSE = 35.23, r = 0.890, CE = 0.858, PI = 0.044, WI = 0.946, LMI = -0.093 , respectively, in the testing period at Varanasi. Likewise, the values of PBIAS = -1.63 , RMSE = 399.8, R = 0.882, CE = 0.854, PI = 0.104, WI = 0.925, LMI = 0.074 in the training period were PBIAS = 1.68, RMSE = 35.2, R = 0.856, CE = 0.867, PI = 0.054, WI = 0.940, LMI = -0.175 in the testing at Prayagraj.

Similarly, MLP model Violin graph (Figs. 4 and 5) demonstrate the deviation between observed and modeled values of electrical conductivity (EC) using MLP, CANFIS, RF, and DT models. The MLP and CANFIS models had a high deviation between the observed and predicted model values of EC in Varanasi and Prayagraj, respectively. It was demonstrated from these graphs that the electrical conductivity of surface water has good prediction accuracy through the RF model, and other models have good arrangements between the observed and modeled EC of surface water. Similarly, the scatter diagram (Figs. 6 and 7) indicated that MLP and CANFIS models were over-predicted at Varanasi and under-predicted in Prayagraj. Also, RF model had good agreement with observed EC at both stations. The scatter diagram of the discrepancy ratio (DR) is illustrated in Figs. 8 and 9. The discrepancy ratio (DR) is defined as the logarithm of the ratio of modeled

Fig. 4 Violin plot of between observed and different predicted machine learning models at Varanasi site

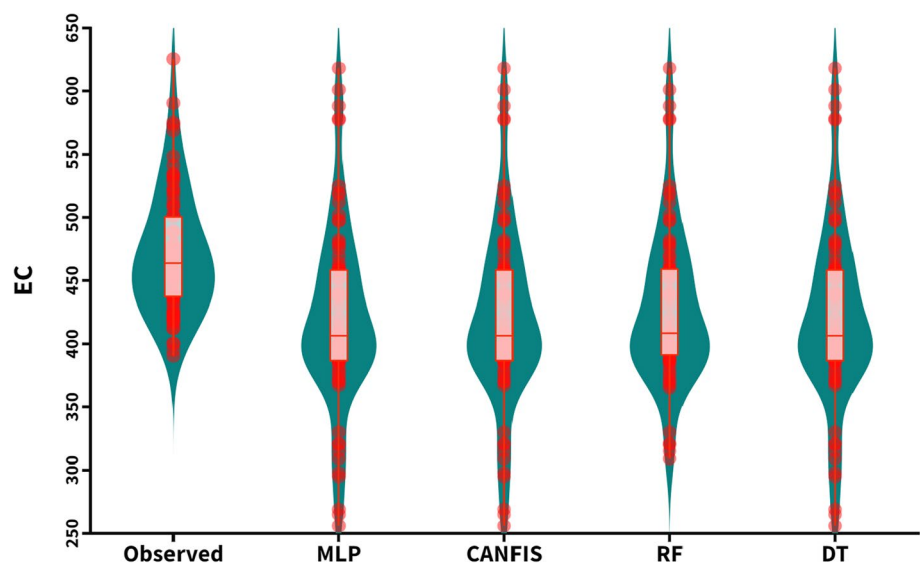
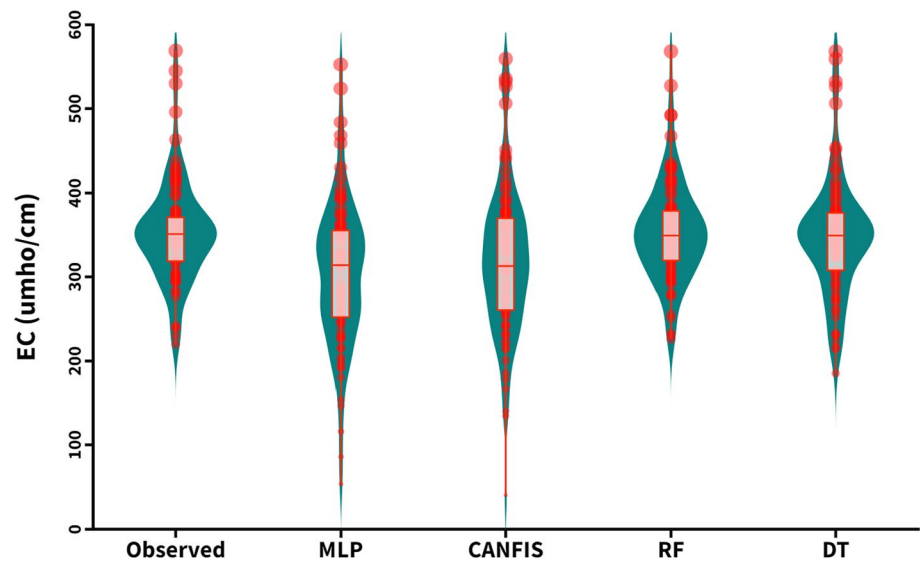


Fig. 5 Violin plot of between observed and different predicted machine learning models at Prayagraj site



data to observed data (Singh et al. 2018a). The DR values were near zero, which means the model was highly accurate. It was shown from the scatter diagram of DR values of different algorithms that the DR of the RF algorithm was plotted on about a zero line in both stations as compared to the MLP, CANFIS, and DT algorithms. After comparing all the results of four different algorithms, it was found that all models performed well for the prediction of the electrical conductivity of surface water. Also, the RF model has obtained the best results during training and testing periods in both sites as compared to the DT, CANFIS, and MLP models. The prediction capability of

the DT model was also good as compared to the CANFIS and MLP models. The prediction ability of the RF and DT models was found to be better than that of the CANFIS and MLP models, and the reason for this phenomenon may be overfitting and the complexity of the model. In addition, using RF and DT models can avoid the overfitting of large and complex data sets.

Multi-ahead electrical conductivity forecasting

In multi-ahead forecasting of electrical conductivity, current-month total hardness (TH), magnesium (Mg), and chloride

Fig. 6 Scatter diagram of different machine learning models at Varanasi Site

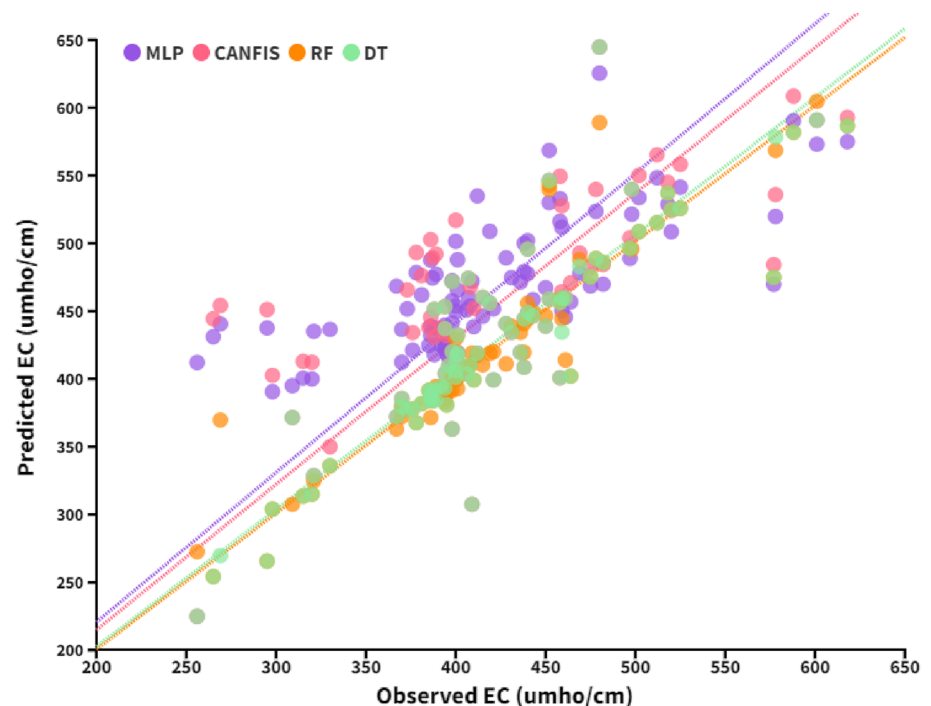


Fig. 7 Scatter diagram of different machine learning models at Prayagraj Site

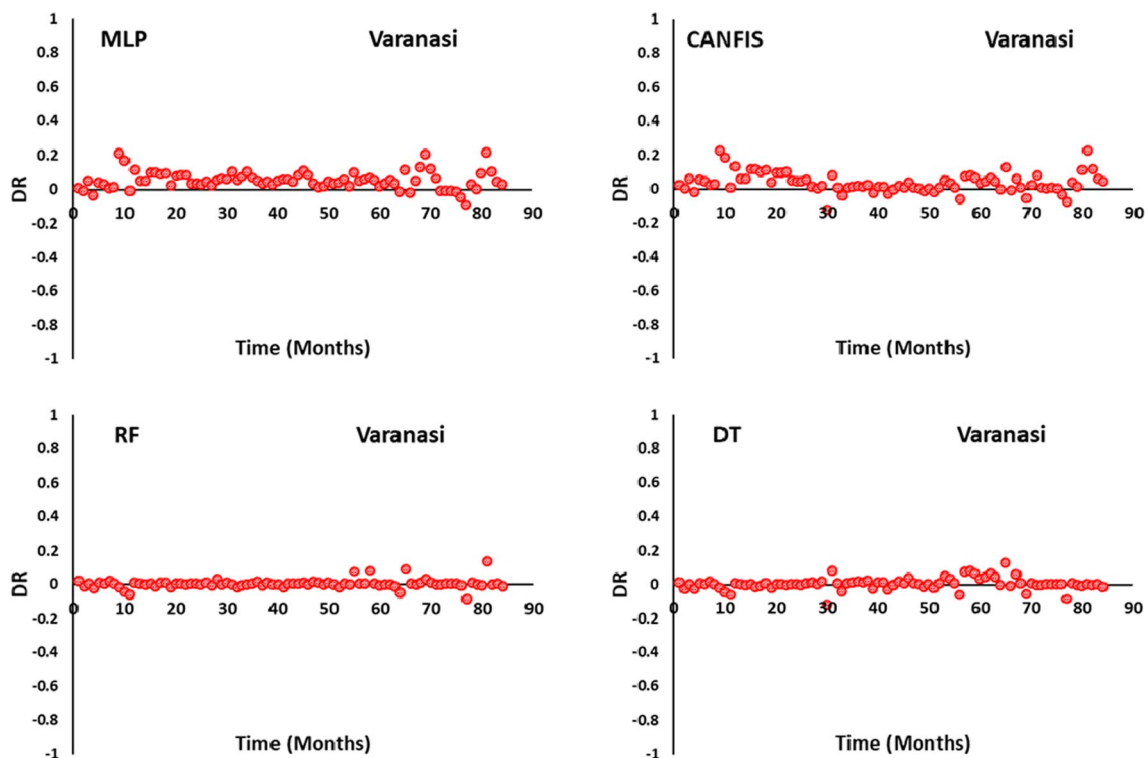
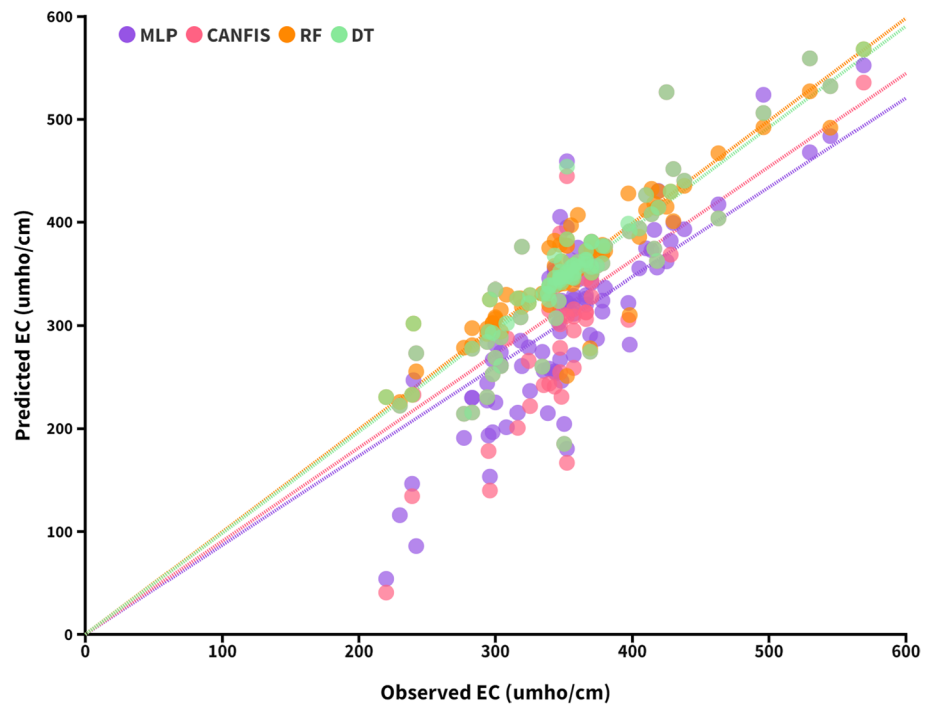


Fig. 8 Discrepancy ratio (DR) values of MPL, CANFIS, RF and DT model in Varanasi

(CI) parameters were used as input variables for forecasting multi-ahead electrical conductivity (EC_{t+1} , EC_{t+2} , EC_{t+3}). The same architecture of MLP, CANFIS, RF, and DT algorithms was used for the forecasting of multi-ahead electrical

conductivity, which was developed during the prediction of current electrical conductivity (EC_t). The results of multi-ahead models are presented in Tables 2 and 3, respectively, for Varanasi and Prayagraj stations. The performance of

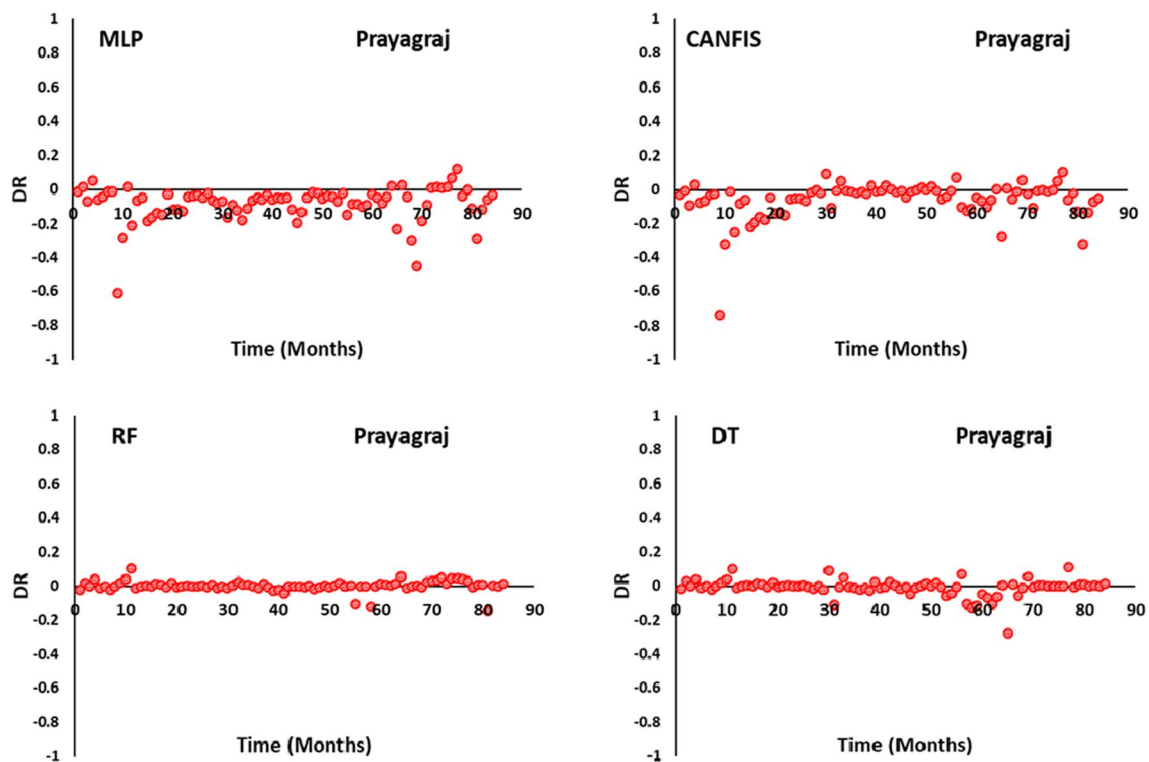


Fig. 9 Discrepancy ratio (DR) values of MPL, CANFIS, RF and DT model in Prayagraj

multi-ahead electrical conductivity forecasting models was assessed using statistical indices PBIAS, RMSE, R, CE, PI, WI, and LMI, while the comparative assessment was done based on the Violin plot (Figs. 4 and 5) and the Taylor diagram (Figs. 12 and 13).

Radar plot is platform that show multi-indices in single diagram. Radar plot (Figs. 10 and 11) illustrates the comparative assessment of MLP, CANFIS, RF, and DT models. The radar plot of RF and DT models represented a good aggregate with observed EC of surface water as compared to CANFIS and MLP models and had low error as compared to CANFIS and MLP models. Also, RF and DT models had high correlation and efficiency, as compared to CANFIS and MLP models. It also observed that RF model had good aggregate, low error, high correlation and high potential for forecasting multi-ahead electrical conductivity (EC_{t+1} , EC_{t+2} , EC_{t+3}) at both stations. The Taylor diagram is a single stand that provides three different statistical indices platforms to compare multiple models in a single diagram. RMSE, correlation coefficient, and standard deviation indices were used in the Taylor diagram. It was observed from the results of multi-ahead forecasting models, which are presented in Tables 2 and 3, that the potential of developed models was reduced as compared to current-month electrical conductivity forecasting models. It was concluded that the applicability of MLP, CANFIS RF, and DT algorithms

was satisfactory for multi-ahead electrical forecasting at both Varanasi and Prayagraj stations of the Upper Ganga basin.

Discussion

The focus of the present study is to compare the performances of MLP, CANFIS RF, and DT models by using the observed and simulated electrical conductivity plots based on the training and testing periods data sets. The forecasting accuracy of MLP, CANFIS RF, and DT algorithms deteriorated when the lead time of electrical conductivity was increased. This finding is in agreement with Nanda et al. (2019). The capability of MLP and CANFIS algorithms for electrical conductivity prediction was lower as compared to RF and DT models. Similar results were found in the prediction of soil permeability by Singh et al. (2020). It was shown from the Taylor diagram in Figs. 12 and 13 that the RMSE of MLP, CANFIS RF, and DT algorithms increased with increasing electrical conductivity, whereas the correlation coefficient of the RMSE of MLP, CANFIS RF, and DT algorithms decreased with increasing electrical conductivity at Varanasi and Prayagraj stations. It was also concluded based on statistical indices and the Taylor diagram that the regression-based random forest model was the best performing model for prediction of

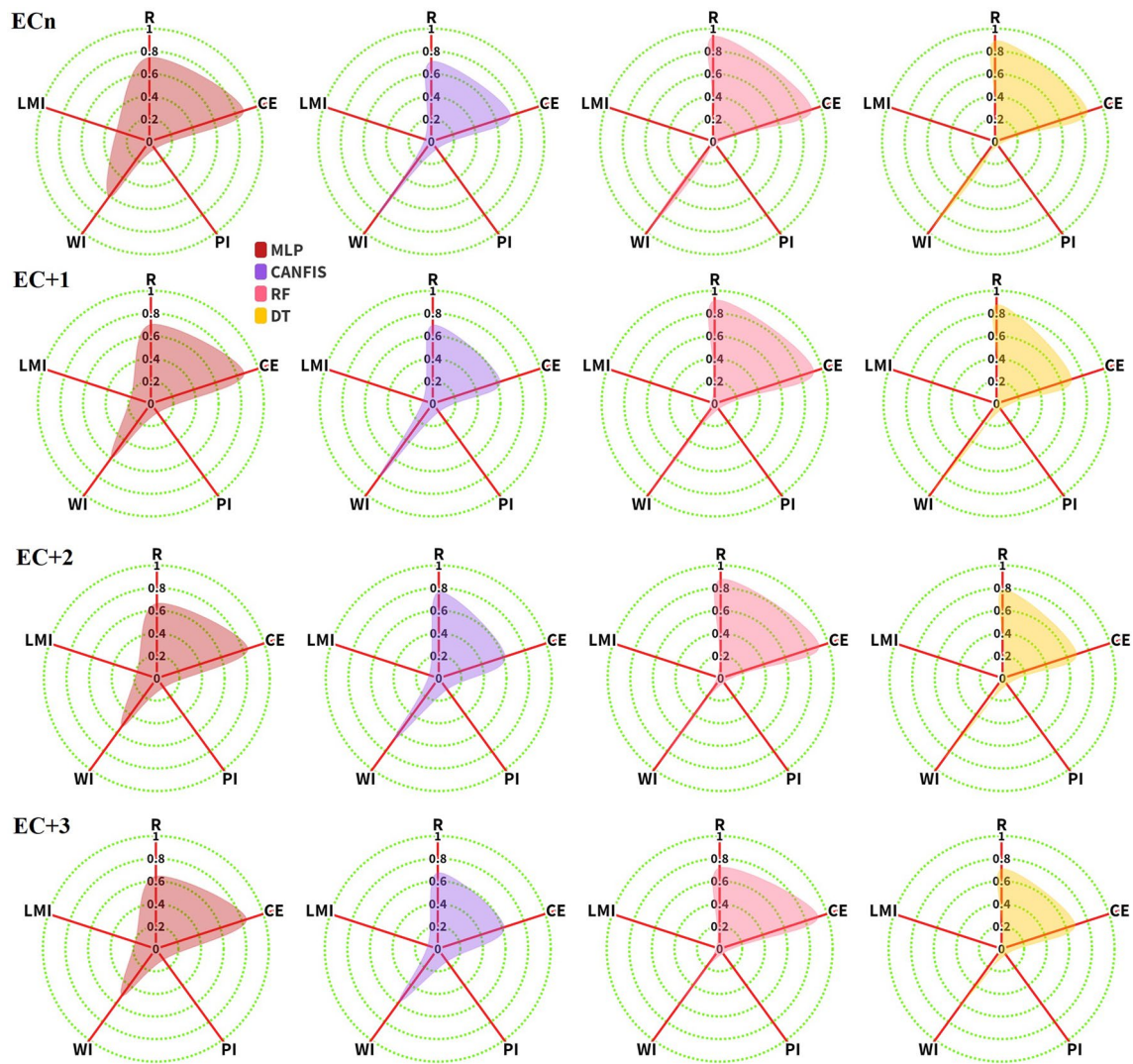


Fig. 10 Radar Plot of different indices for current and multi-ahead electrical conductivity at Varanasi site

multi-ahead electrical conductivity at both Varanasi and Prayagraj stations of the Upper Ganga basin. It was determined that current moth total hardness (TH), magnesium (Mg), and chloride (Cl) parameters can be used as input variables for forecasting multi-ahead electrical conductivity (EC_{t+1} , EC_{t+2} , and EC_{t+3}) in the future study in the Upper Ganga basin. It seemed that RF and DT models had better performance as compared to MLP and CANFIS models. RF and DT models can be used for multi-ahead forecasting of monthly electrical conductivity at both Varanasi and Prayagraj stations of the Upper Ganga basin.

Similar results also found by researchers, Ravansalar and Rajaei (2015) studied the effectiveness of the hybrid WANN model for predicting monthly EC. The integration of wavelet analysis and neural networks improves the prediction accuracy compared to using ANN alone, making it a valuable

tool for assessing water quality in similar environmental systems. Ahmadianfar et al. (2020) compared various standalone approaches, including locally weighted linear regression (LWLR), support vector regression (SVR), wavelet support vector regression (W-SVR), autoregressive integrated moving average (ARIMA), wavelet ARIMA (W-ARIMA), multivariate linear regression (MLR), and wavelet MLR (W-MLR), for predicting the monthly electrical conductivity (EC) values. The W-LWLR method demonstrated the highest correlation coefficient among the compared methods. A higher correlation coefficient indicates a stronger linear relationship between the predicted and observed EC values. Therefore, the W-LWLR method exhibited superior performance in capturing the EC variations. Ahmadianfar et al. (2022) compared the performance of several models: standalone ANFIS, least-square support vector machine

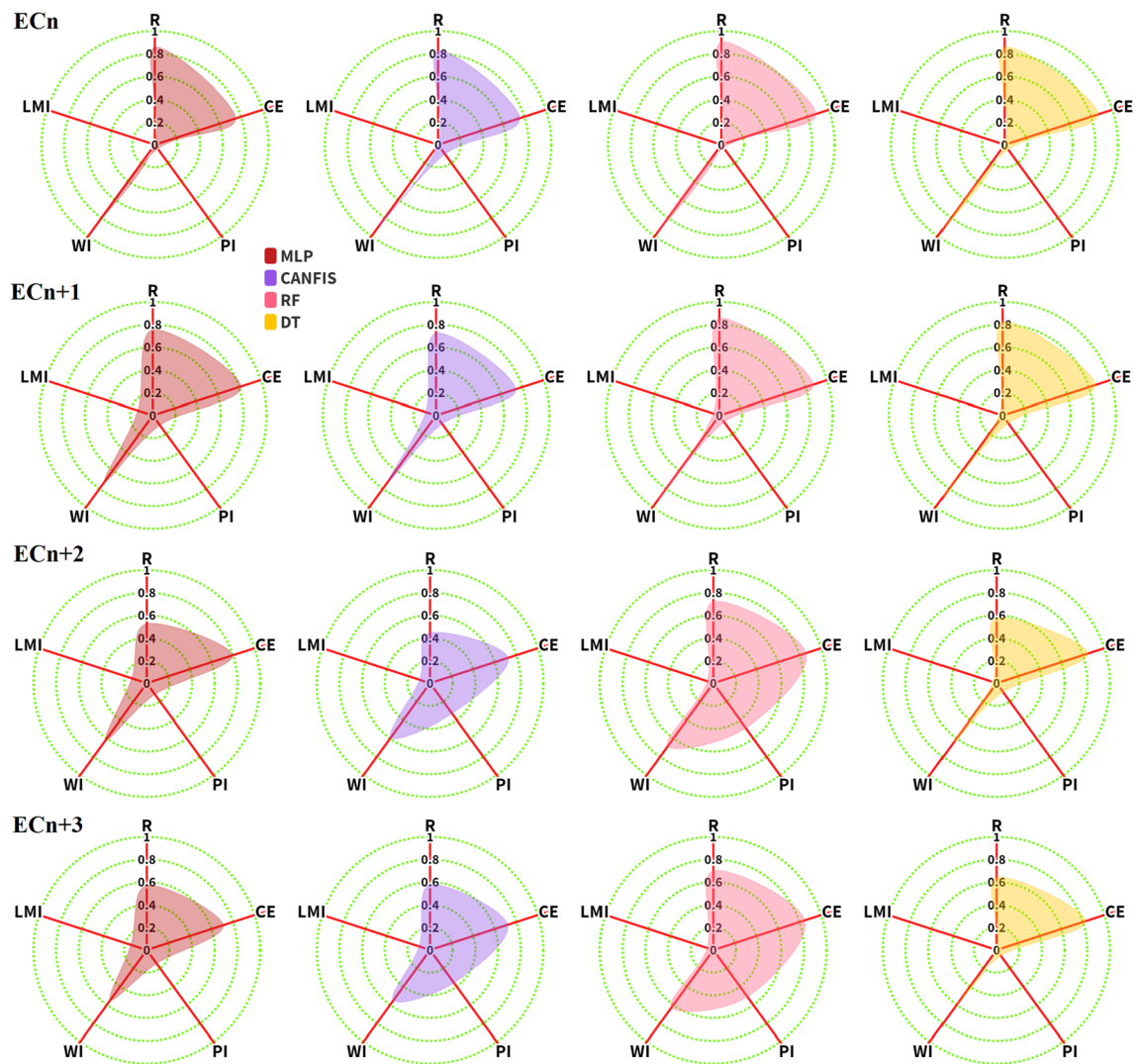


Fig. 11 Radar plot of different indices for current (ECn) and multi-ahead electrical conductivity at Prayagraj site

(LSSVM), multivariate adaptive regression spline (MARS), generalized regression neural network (GRNN), wavelet-LSSVM (WLSSVM), wavelet-MARS (W-MARS), wavelet-ANFIS (W-ANFIS), and wavelet-GRNN (W-GRNN). The results indicated that the W-ANFIS model significantly improves the prediction certainty of EC, outperforming the other models with the Dmey mother wavelet in terms of EC prediction. Melesse et al. (2020) also compared the conventional random forest and hybrid random forest. It was found that hybridization of machine learning methods significantly improves the model performance in capturing maximum EC values, which is crucial for effective water resource management. Abba et al. (2023) studied the direct modeling of electrical conductivity (EC) in a sandstone groundwater (GW) aquifer in Southern Saudi Arabia. The research explores the application of hydrogeochemical and physical variables to predict EC using an adaptive neuro-fuzzy inference system

(ANFIS)-based method. ANFIS method was outperforming model for predicting EC of ground water.

Conclusions

In this work, the electrical conductivity of surface water was predicted in the upper Ganga basin of India based on different machine learning algorithms, including multi-layer perceptron (MLP), co-adaptive neuro-fuzzy inference system (CANFIS), random forest (RF), and decision tree (DT), and also examined the capabilities of the models. It was suggested based on the gamma test that current-month total hardness (TH), magnesium (Mg), and chloride (Cl) parameters were established as the best suitable independent vector combination for forecasting current-month electrical conductivity (EC_t) as well as multi-ahead electrical

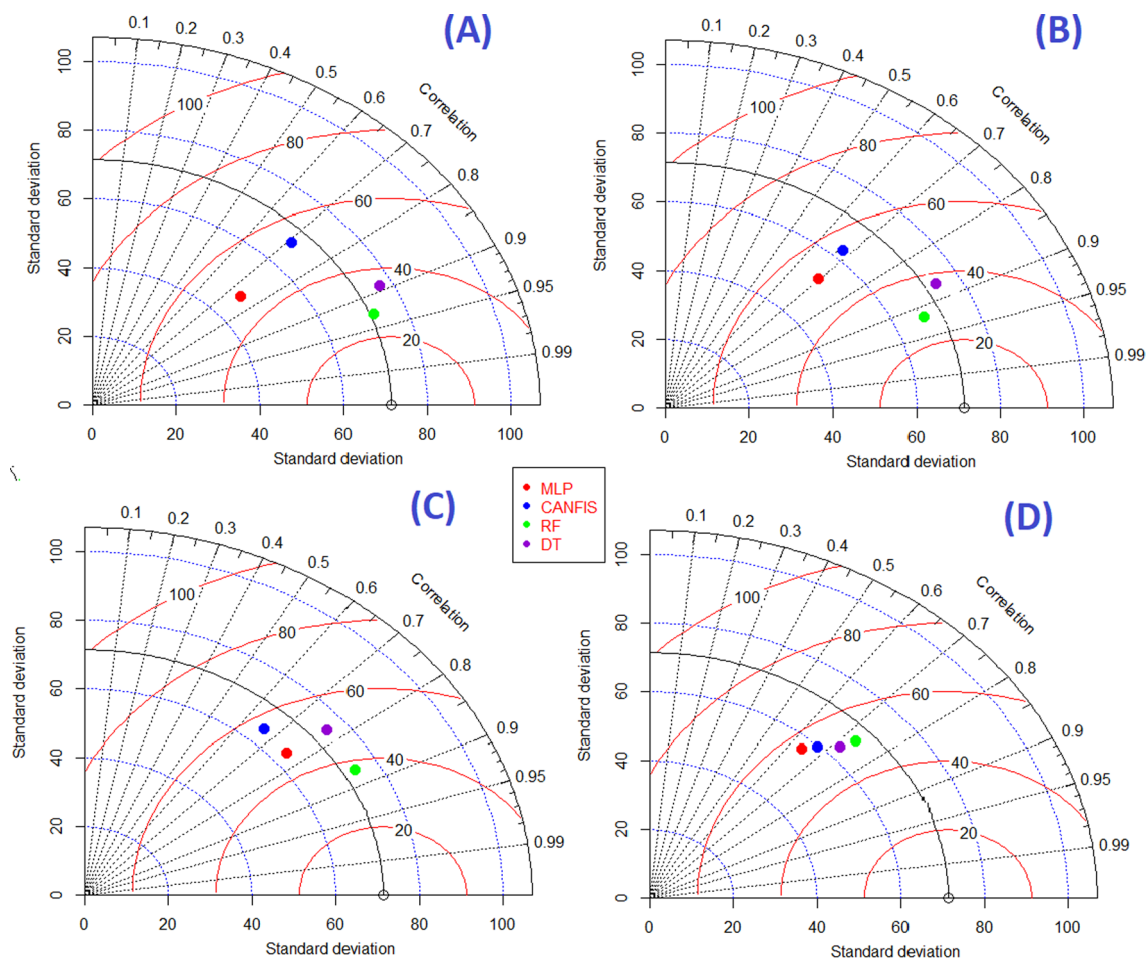


Fig. 12 Taylor diagram of present month EC_t (A), one month ahead EC_{t+1} (B), two month ahead EC_{t+2} (C) and ahead three-month EC_{t+3} (D) during testing period at Varanasi station

conductivity (EC_{t+1} , EC_{t+2} , and EC_{t+3}). The results indicate, based on statistical indices and visual interpretation, that the random forest algorithm demonstrated the best performance for forecasting the current month's electrical conductivity of surface water among MLP, CANFIS and decision tree algorithms while decision tree algorithms perform better than the multilayer perceptron, co-adaptive neuro-fuzzy inference system. The obtainable results of multilayer perceptron, co-adaptive neuro-fuzzy inference system, random forest, and decision tree algorithms for prediction of the current month electrical conductivity (EC_t) as well as multi-ahead electrical conductivity (EC_{t+1} , EC_{t+2} , and EC_{t+3}) were satisfactory at both Varanasi and Prayagraj stations of Upper Ganga basin. Moreover, the different machine learning algorithms and the results presented now provide an orientation line on the upcoming investigation of the upper Ganga basin, and it was also suggested that policies of irrigation and drinking water quality deterrence should be applied continuously for proper supervision along the upper Ganga basin.

There is still work to be done to improve the situation, though, due to various restrictions. The following issues should be researched more in the future.

1. Temporal prediction is the subject of the current prediction outcomes. Future work will examine the performance of our model in multi-head predictions of electrical conductivity using both spatial and temporal prediction.
2. Only the indications of water quality itself are taken into account in the current projection; however, additional variables like location, precipitation, and temperature may also have an impact. We shall keep these things in mind going forward to better enhance the forecasting outcomes.
3. In the future, we want to experiment with the multi objective optimization approach to create hybrid algorithms for machine learning (HMLA) and evaluate their performance against traditional models.

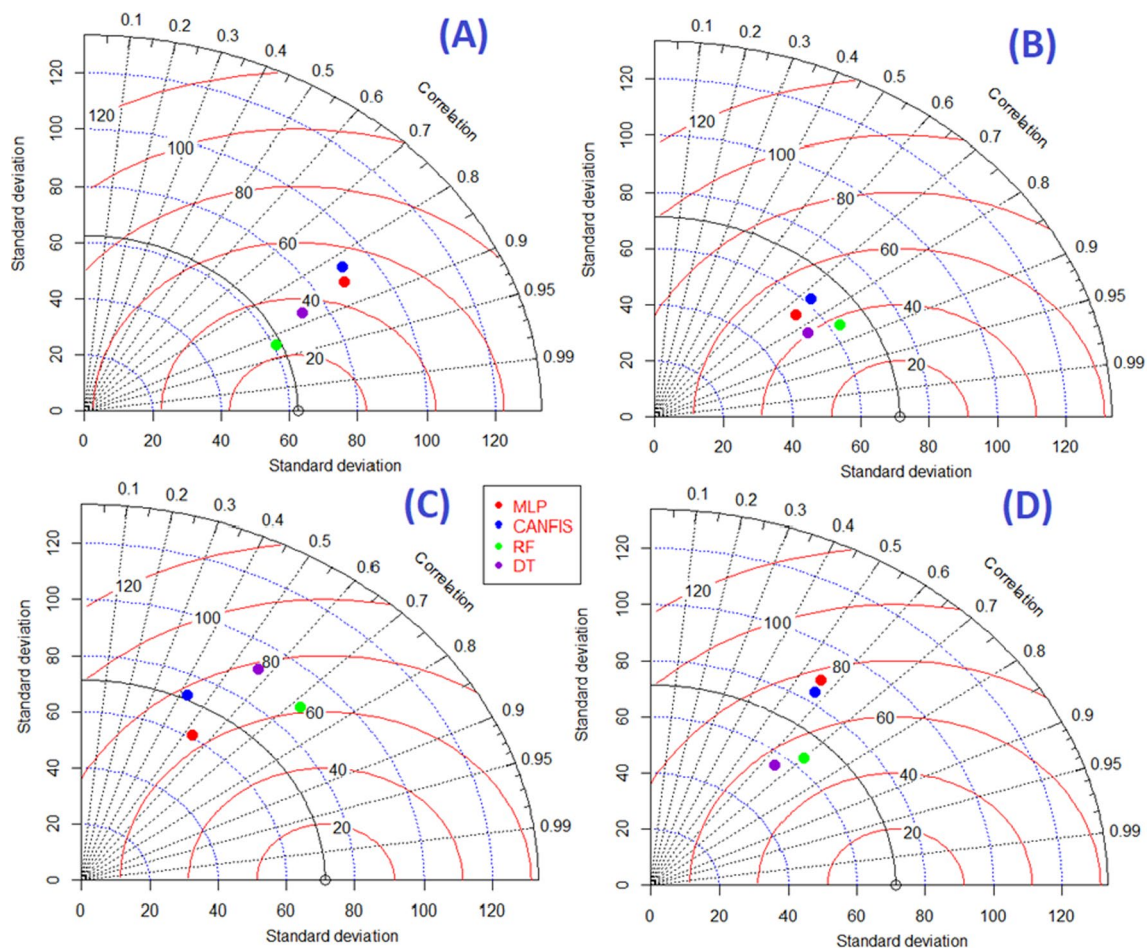


Fig. 13 Taylor diagram of present month EC_t (A), one month ahead EC_{t+1} (B), two month ahead EC_{t+2} (C) and ahead three—month EC_{t+3} (D) during testing period at Prayagraj station

- Other climatic factors also impact water quality, however in this particular study, only the quality of the water data is considered. We will use these variables as inputs in the future for long-term forecasting.

Acknowledgements The authors extend their appreciation to the Deanship of Scientific Research, King Saud University for funding through the Vice Deanship of Scientific Research Chairs; Research Chair of Prince Sultan Bin Abdulaziz International Prize for Water.

Funding Open access funding provided by Lulea University of Technology. This research was funded by the Deanship of Scientific Research, King Saud University through the Vice Deanship of Scientific Research Chairs; Research Chair of Prince Sultan Bin Abdulaziz International Prize for Water.

Declarations

Conflict of interest No conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abba SI, Benaafi M, Usman AG, Aljundi IH (2023) Sandstone groundwater salinization modelling using physicochemical variables in Southern Saudi Arabia: application of novel data intelligent algorithms. *Ain Shams Eng J* 14:101894. <https://doi.org/10.1016/j.asej.2022.101894>

- Achite M, Elshaboury N, Jehanzaib M et al (2023) Performance of machine learning techniques for meteorological drought forecasting in the Wadi mina basin. *Algeria Water* 15:765. <https://doi.org/10.3390/w15040765>
- Adams A, Sterling L (1992) Learning with continuous classes. In: Adams A, Sterling L (eds) *AI '92, Proceedings of the 5th Australian joint conference on artificial intelligence*. World scientific, pp 343–348
- Ağbulut Ü, Gürel AE, Biçen Y (2021) Prediction of daily global solar radiation using different machine learning algorithms: evaluation and comparison. *Renew Sustain Energy Rev* 135:110114. <https://doi.org/10.1016/j.rser.2020.110114>
- Ahmadianfar I, Shirvani-Hosseini S, He J et al (2022) An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction. *Sci Rep* 12:4934. <https://doi.org/10.1038/s41598-022-08875-w>
- Ahmadianfar I, Jamei M, Chu X (2020) A novel hybrid wavelet-locally weighted linear regression (W-LWLR) model for electrical conductivity (EC) prediction in surface water. *J Contam Hydrol* 232:103641–103716. <https://doi.org/10.1016/j.jconhyd.2020.103641>
- Al-Ghobari HM, El-Marazky MS, Dewidar AZ, Mattar MA (2018) Prediction of wind drift and evaporation losses from sprinkler irrigation using neural network and multiple regression techniques. *Agric Water Manag* 195:211–221. <https://doi.org/10.1016/j.agwat.2017.10.005>
- Atieh M, Mehlretter SL, Gharabaghi B, Rudra R (2015) Integrative neural networks model for prediction of sediment rating curve parameters for ungauged basins. *J Hydrol* 531:1095–1107. <https://doi.org/10.1016/j.jhydrol.2015.11.008>
- Aytek A (2009) Co-active neurofuzzy inference system for evapotranspiration modeling. *Soft Comput* 13:691–700. <https://doi.org/10.1007/s00500-008-0342-8>
- Bajirao TS, Kumar P, Kumar M et al (2021a) Superiority of hybrid soft computing models in daily suspended sediment estimation in highly dynamic rivers. *Sustain* 13:1–29. <https://doi.org/10.3390/su13020542>
- Bajirao TS, Kumar P, Kumar M et al (2021b) Potential of hybrid wavelet-coupled data-driven-based algorithms for daily runoff prediction in complex river basins. *Theor Appl Climatol* 145:1207–1231. <https://doi.org/10.1007/s00704-021-03681-2>
- Barzegar R, Asghari Moghaddam A, Adamowski J, Ozga-Zielinski B (2018) Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. *Stoch Environ Res Risk Assess* 32:799–813. <https://doi.org/10.1007/s00477-017-1394-z>
- Bhat SU, Nisa AU, Sabha I, Mondal NC (2022) Spring water quality assessment of Anantnag district of Kashmir Himalaya: towards understanding the looming threats to spring ecosystem services. *Appl Water Sci* 12:180. <https://doi.org/10.1007/s13201-022-01691-7>
- Bouaziz M, Medhioub E, Csaplovics E (2021) A machine learning model for drought tracking and forecasting using remote precipitation data and a standardized precipitation index from arid regions. *J Arid Environ* 189:104478–104516. <https://doi.org/10.1016/j.jaridenv.2021.104478>
- Boubakri A, Elgharbi S, Dhaouadi I et al (2023) Optimization and prediction of lead removal from aqueous solution using FO–MD hybrid process: statistical and artificial intelligence analysis. *J Environ Manage* 337:117731. <https://doi.org/10.1016/j.jenvman.2023.117731>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman L (1999) Random forests—random features. Technical Report 567, Statistics Department
- Chapman DV (1996) Water quality assessments: a guide to the use of biota, sediments and water in environmental monitoring, 2nd edn. CRC Press
- Chen H, Lin Z, Wu H et al (2015) Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. *Spectrochim Acta Part A Mol Biomol Spectrosc* 135:185–191. <https://doi.org/10.1016/j.saa.2014.07.005>
- Chen G, Long T, Xiong J, Bai Y (2017) Multiple random forests modelling for urban water consumption forecasting. *Water Resour Manag* 31:4715–4729. <https://doi.org/10.1007/s11269-017-1774-7>
- Cornejo-Bueno L, Casanova-Mateo C, Sanz-Justo J, Salcedo-Sanz S (2019) Machine learning regressors for solar radiation estimation from satellite data. *Sol Energy* 183:768–775. <https://doi.org/10.1016/j.solener.2019.03.079>
- Czajkowski M, Kretowski M (2016) The role of decision tree representation in regression problems—an evolutionary perspective. *Appl Soft Comput* 48:458–475. <https://doi.org/10.1016/j.asoc.2016.07.007>
- Deng T, Chau K-W, Duan H-F (2021) Machine learning based marine water quality prediction for coastal hydro-environment management. *J Environ Manage* 284:112051–112116. <https://doi.org/10.1016/j.jenvman.2021.112051>
- Dunlop J, McGregor G, Horrigan N (2005) Potential impacts of salinity and turbidity in riverine ecosystems: characterisation of impacts and a discussion of regional target setting for riverine ecosystems in Queensland
- Ekemen Keskin T, Özler E, Şander E et al (2020) Prediction of electrical conductivity using ANN and MLR: a case study from Turkey. *Acta Geophys* 68:811–820. <https://doi.org/10.1007/s11600-020-00424-1>
- Elbeltagi A, Kumar M, Kushwaha NL et al (2022) Drought indicator analysis and forecasting using data driven models: case study in Jaisalmer. *Stoch Environ Res Risk Assess*, India. <https://doi.org/10.1007/s00477-022-02277-0>
- Elbeltagi A, Al-Mukhtar M, Kushwaha NL et al (2023a) Forecasting monthly pan evaporation using hybrid additive regression and data-driven models in a semi-arid environment. *Appl Water Sci* 13:42. <https://doi.org/10.1007/s13201-022-01846-6>
- Elbeltagi A, Pande CB, Kumar M et al (2023b) Prediction of meteorological drought and standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. *Environ Sci Pollut Res* 30:43183–43202. <https://doi.org/10.1007/s11356-023-25221-3>
- Elbeltagi A, Seifi A, Ehteram M et al (2023c) GLUE analysis of meteorological-based crop coefficient predictions to derive the explicit equation. *Neural Comput Appl*. <https://doi.org/10.1007/s00521-023-08466-4>
- Esfahani FS, Keshavarzi A (2020) Circulation cells topology and their effect on migration pattern of different multi-bend meandering rivers. *Int J Sediment Res* 35:636–650. <https://doi.org/10.1016/j.ijsrc.2020.04.004>
- Ezea VC, Ihedioha JN, Abugu HO, Ekere NR (2022) A multi-criteria approach to drinking and irrigation water assessment of spring water in Igbo-Etiti. *Nigeria Appl Water Sci* 12:220. <https://doi.org/10.1007/s13201-022-01747-8>
- Fakhari A, Moghadam AME (2013) Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Appl Soft Comput* 13:1292–1302. <https://doi.org/10.1016/j.asoc.2012.10.019>
- Feng Y, Gong D, Zhang Q et al (2019) Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. *Energy Convers Manag* 198:111780. <https://doi.org/10.1016/j.enconman.2019.111780>
- Fu L (1994) Neural networks in computer intelligence. McGraw-Hill, New York
- Gholami V, Khaleghi MR, Pirasteh S, Booij MJ (2021) Comparison of self-organizing map, artificial neural network, and co-active neuro-fuzzy inference system methods in simulating groundwater quality: geospatial artificial intelligence. *Water Resour Manag*. <https://doi.org/10.1007/s11269-021-02969-2>

- Ghorbani MA, Aalami MT, Naghipour L (2017) Use of artificial neural networks for electrical conductivity modeling in Asi river. *Appl Water Sci* 7:1761–1772. <https://doi.org/10.1007/s13201-015-0349-0>
- Giri A, Bharti VK, Kalia S et al (2022) Hydrochemical and quality assessment of irrigation water at the trans-himalayan high-altitude regions of Leh, Ladakh. *India Appl Water Sci* 12:197. <https://doi.org/10.1007/s13201-022-01716-1>
- Gomaa MM (2020) Salinity and water effect on electrical properties of fragile clayey sandstone. *Appl Water Sci* 10:116. <https://doi.org/10.1007/s13201-020-01189-0>
- Gu Z, Zhu T, Jiao X et al (2021) Neural network soil moisture model for irrigation scheduling. *Comput Electron Agric* 180:105801. <https://doi.org/10.1016/j.compag.2020.105801>
- Harb R, Yan X, Radwan E, Su X (2009) Exploring precrash maneuvers using classification trees and random forests. *Accid Anal Prev* 41:98–107. <https://doi.org/10.1016/j.aap.2008.09.009>
- Hassan MA, Khalil A, Kaseb S, Kassem MA (2017) Potential of four different machine-learning algorithms in modeling daily global solar radiation. *Renew Energy* 111:52–62. <https://doi.org/10.1016/j.renene.2017.03.083>
- Hastie T, Friedman J, Tibshirani R (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, New York, NY
- Igboama WN, Hammed OS, Fatoba JO et al (2022) Review article on impact of groundwater contamination due to dumpsites using geophysical and physiochemical methods. *Appl Water Sci* 12:130. <https://doi.org/10.1007/s13201-022-01653-z>
- Kakaei Lafdani E, Moghaddam Nia A, Ahmadi A (2013) Daily suspended sediment load prediction using artificial neural networks and support vector machines. *J Hydrol* 478:50–62. <https://doi.org/10.1016/j.jhydrol.2012.11.048>
- Karimi S, Sadraddini AA, Nazemi AH et al (2018) Generalizability of gene expression programming and random forest methodologies in estimating cropland and grassland leaf area index. *Comput Electron Agric* 144:232–240. <https://doi.org/10.1016/j.compag.2017.12.007>
- Khadr M, Elshemy M (2017) Data-driven modeling for water quality prediction case study: the drains system associated with Manzala Lake. *Egypt Ain Shams Eng J* 8:549–557. <https://doi.org/10.1016/j.asej.2016.08.004>
- Khatti J, Grover KS (2023) Prediction of UCS of fine-grained soil based on machine learning part 1: multivariable regression analysis, gaussian process regression, and gene expression programming. *Multiscale Multidiscip Model Exp Des* 6:199–222. <https://doi.org/10.1007/s41939-022-00137-6>
- Kisi O, Ay M (2014) Comparison of Mann-Kendall and innovative trend method for water quality parameters of the Kizilirmak river, Turkey. *J Hydrol* 513:362–375. <https://doi.org/10.1016/j.jhydrol.2014.03.005>
- Kisi O, Shiri J (2012) River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. *Comput Geosci* 43:73–82. <https://doi.org/10.1016/j.cageo.2012.02.007>
- Kumar A, Kumar P, Singh VK (2019) Evaluating different machine learning models for runoff and suspended sediment simulation. *Water Resour Manag* 33:1217–1231. <https://doi.org/10.1007/s11269-018-2178-z>
- Kumar A, Singh VK, Saran B et al (2022) Development of novel hybrid models for prediction of drought- and stress-tolerance indices in teosinte introgressed maize lines using artificial intelligence techniques. *Sustainability* 14:2287. <https://doi.org/10.3390/su14042287>
- Kushwaha NL, Rajput J, Elbeltagi A et al (2021) Data intelligence model and meta-heuristic algorithms-based pan evaporation modelling in two different agro-climatic zones: a case study from northern India. *Atmosphere (basel)* 12:1654. <https://doi.org/10.3390/atmos12121654>
- Lakrou A, Meddah B, Ali Rahmani SE (2022) Geochemical assessment of thermal water in Bouhanifia aquifer system: suitability for touristic therapeutic treatment. *Appl Water Sci* 12:204. <https://doi.org/10.1007/s13201-022-01727-y>
- Lee D-L, Deng LY, Lin K-H, et al (2013) Using decision tree analysis for personality to decisions of the national skills competition participants. In: Park JJ (Jong H, Barolli L, Khafa F, Jeong H-Y (eds) Information technology convergence. Lecture notes in electrical engineering, 253rd edn. Springer Netherlands, Dordrecht, pp 683–691
- Liu Z, Feng S, Zhang D et al (2023) Effects of precipitation, irrigation, and exploitation on groundwater geochemical evolution in the people's victory canal irrigation area. *China Appl Water Sci* 13:1. <https://doi.org/10.1007/s13201-022-01808-y>
- Loh W-Y (2014) Fifty years of classification and regression trees. *Int Stat Rev* 82:329–348. <https://doi.org/10.1111/insr.12016>
- Maansi JR, Wats M (2022) Evaluation of surface water quality using water quality indices (WQIs) in Lake Sukhna, Chandigarh. *India Appl Water Sci* 12:2. <https://doi.org/10.1007/s13201-021-01534-x>
- Mahanta AR, Rawat KS, Singh SK et al (2022) Evaluation of long-term nitrate and electrical conductivity in groundwater system of Peninsula. *India Appl Water Sci* 12:17. <https://doi.org/10.1007/s13201-021-01568-1>
- Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environ Model Softw* 25:891–909. <https://doi.org/10.1016/j.envsoft.2010.02.003>
- Malik A, Kumar A (2015) Pan evaporation simulation based on daily meteorological data using soft computing techniques and multiple linear regression. *Water Resour Manag* 29:1859–1872. <https://doi.org/10.1007/s11269-015-0915-0>
- Markuna S, Kumar P, Ali R et al (2023) Application of innovative machine learning techniques for long-term rainfall prediction. *Pure Appl Geophys* 180:335–363. <https://doi.org/10.1007/s00024-022-03189-4>
- Masoudi R, Mousavi SR, Rahimabadi PD et al (2023) Assessing data mining algorithms to predict the quality of groundwater resources for determining irrigation hazard. *Environ Monit Assess* 195:319. <https://doi.org/10.1007/s10661-022-10909-9>
- Mattar MA (2018) Using gene expression programming in monthly reference evapotranspiration modeling: a case study in Egypt. *Agric Water Manag* 198:28–38. <https://doi.org/10.1016/j.agwat.2017.12.017>
- Mattar MA-A, El-Marazky MS, Ahmed KA (2017) Modeling sprinkler irrigation infiltration based on a fuzzy-logic approach. *Spanish J Agric Res* 15:e1201. <https://doi.org/10.5424/sjar/2017151-9179>
- Mattar MA, Roy DK, Al-Ghobari HM, Dewidar AZ (2022) Machine learning and regression-based techniques for predicting sprinkler irrigation's wind drift and evaporation losses. *Agric Water Manag* 265:107529. <https://doi.org/10.1016/j.agwat.2022.107529>
- Melesse AM, Khosravi K, Tiefenbacher JP et al (2020) River water salinity prediction using hybrid machine learning models. *Water* 12:2951. <https://doi.org/10.3390/w12102951>
- Mensah-Akutteh H, Buamah R, Wiawe S, Nyarko KB (2022) Optimizing coagulation–flocculation processes with aluminium coagulation using response surface methods. *Appl Water Sci* 12:188. <https://doi.org/10.1007/s13201-022-01708-1>
- Mirbagheri SA, Nourani V, Rajaei T, Alikhani A (2010) Neuro-fuzzy models employing wavelet analysis for suspended sediment concentration prediction in rivers. *Hydrol Sci J* 55:1175–1189. <https://doi.org/10.1080/02626667.2010.508871>
- Mirzania E, Vishwakarma DK, Bui Q-AT et al (2023) A novel hybrid AIG-SVR model for estimating daily reference evapotranspiration. *Arab J Geosci* 16:301. <https://doi.org/10.1007/s12517-023-11387-0>

- Mizutani E, Jang J-SR (1995) Coactive neural fuzzy modeling. In: Proceedings of ICNN'95 - International Conference on Neural Networks. IEEE, pp 760–765
- Mokhtar A, Elbeltagi A, Gyasi-Agyei Y et al (2022) Prediction of irrigation water quality indices based on machine learning and regression models. *Appl Water Sci* 12:76. <https://doi.org/10.1007/s13201-022-01590-x>
- Nagalla R, Pothuganti P, Pawar DS (2017) Analyzing gap acceptance behavior at unsignalized intersections using support vector machines, decision tree and random forests. *Procedia Comput Sci* 109:474–481. <https://doi.org/10.1016/j.procs.2017.05.312>
- Naiel MAE, Abdelghany MF, Khames DK et al (2022) Administration of some probiotic strains in the rearing water enhances the water quality, performance, body chemical analysis, antioxidant and immune responses of Nile tilapia. *Oreochromis Niloticus Appl Water Sci* 12:209. <https://doi.org/10.1007/s13201-022-01733-0>
- Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H et al (2019) Machine learning methods for better water quality prediction. *J Hydrol* 578:124084. <https://doi.org/10.1016/j.jhydrol.2019.124084>
- Nanda T, Sahoo B, Chatterjee C (2019) Enhancing real-time stream-flow forecasts with wavelet-neural network based error-updating schemes and ECMWF meteorological predictions in Variable Infiltration Capacity model. *J Hydrol* 575:890–910. <https://doi.org/10.1016/j.jhydrol.2019.05.051>
- Napagoda N, Tilakaratne CD (2012) Artificial neural network approach for modeling of soil temperature: a case study for Bathalagoda area. *Sri Lankan J Appl Stat* 13:39–59
- Nasridinov A, Ihm S-Y, Park Y-H (2013) A Decision Tree-Based Classification Model for Crime Prediction. In: Park JJ (Jong H, Barolli L, Xhafa F, Jeong H-Y (eds) Information Technology Convergence. Lecture Notes in Electrical Engineering, 253rd edn. Springer Netherlands, Dordrecht, pp 531–538
- Nemati S, Fazelifard MH, Terzi Ö, Ghorbani MA (2015) Estimation of dissolved oxygen using data-driven techniques in the Tai Po River, Hong Kong. *Environ Earth Sci* 74:4065–4073. <https://doi.org/10.1007/s12665-015-4450-3>
- Nemčić-Jurec J, Ruk D, Orešćanin V et al (2022) Groundwater contamination in public water supply wells: risk assessment, evaluation of trends and impact of rainfall on groundwater quality. *Appl Water Sci* 12:172. <https://doi.org/10.1007/s13201-022-01697-1>
- Noori R, Karbassi AR, Moghaddamnia A et al (2011) Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J Hydrol* 401:177–189. <https://doi.org/10.1016/j.jhydrol.2011.02.021>
- Pal M, Mather PM (2003) An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens Environ* 86:554–565. [https://doi.org/10.1016/S0034-4257\(03\)00132-9](https://doi.org/10.1016/S0034-4257(03)00132-9)
- Podolsky MD, Barchuk AA, Kuznetsov VI et al (2016) Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific J Cancer Prev* 17:835–838. <https://doi.org/10.7314/APJCP.2016.17.2.835>
- Quinlan JR (1992) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. World Scientific, pp 343–348
- Ravansalar M, Rajaei T (2015) Evaluation of wavelet performance via an ANN-based electrical conductivity prediction model. *Environ Monit Assess* 187:366. <https://doi.org/10.1007/s10661-015-4590-7>
- Rezaeianzadeh M, Tabari H, Arabi Yazdi A et al (2014) Flood flow forecasting using ANN, ANFIS and regression models. *Neural Comput Appl* 25:25–37. <https://doi.org/10.1007/s00521-013-1443-6>
- Roy DK, Sarkar TK, Kamar SSA et al (2022) Daily prediction and multi-step forward forecasting of reference evapotranspiration using LSTM and Bi-LSTM models. *Agronomy* 12:594. <https://doi.org/10.3390/agronomy12030594>
- Samantaray S, Sahoo A, Satapathy DP (2022) Prediction of ground-water-level using novel SVM-ALO, SVM-FOA, and SVM-FFA algorithms at Purba-Medinipur. *India Arab J Geosci* 15:723. <https://doi.org/10.1007/s12517-022-09900-y>
- Saroughi M, Mirzania E, Vishwakarma DK et al (2023) A novel hybrid algorithms for groundwater level prediction. *Iran J Sci Technol Trans Civ Eng*. <https://doi.org/10.1007/s40996-023-01068-z>
- Shukla R, Kumar P, Vishwakarma DK et al (2021) Modeling of stage-discharge using back propagation ANN-, ANFIS-, and WANN-based computing techniques. *Theor Appl Climatol*. <https://doi.org/10.1007/s00704-021-03863-y>
- Sihag P, Esmaeilbeiki F, Singh B et al (2019a) Modeling unsaturated hydraulic conductivity by hybrid soft computing techniques. *Soft Comput* 23:12897–12910. <https://doi.org/10.1007/s00500-019-03847-1>
- Sihag P, Tiwari NK, Ranjan S (2019b) Prediction of unsaturated hydraulic conductivity using adaptive neuro-fuzzy inference system (ANFIS). *ISH J Hydraul Eng* 25:132–142. <https://doi.org/10.1080/09715010.2017.1381861>
- Singh VK, Kumar P, Singh BP (2016a) Rainfall-runoff modeling using artificial neural networks (ANNs) and multiple linear regression (MLR) techniques. *Indian J Ecol* 43:436–442
- Singh VK, Singh BP (2016b) Basin suspended sediment prediction using soft computing and conventional approaches in India. *Environ Sci* 7:459–468
- Singh B, Sihag P, Singh K (2017) Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Model Earth Syst Environ* 3:999–1004. <https://doi.org/10.1007/s40808-017-0347-3>
- Singh VK, Kumar D, Kashyap PS, Kisi O (2018a) Simulation of suspended sediment based on gamma test, heuristic, and regression-based techniques. *Environ Earth Sci* 77:708. <https://doi.org/10.1007/s12665-018-7892-6>
- Singh VK, Singh BP, Kisi O, Kushwaha DP (2018b) Spatial and multi-depth temporal soil temperature assessment by assimilating satellite imagery, artificial intelligence and regression based models in arid area. *Comput Electron Agric* 150:205–219. <https://doi.org/10.1016/j.compag.2018.04.019>
- Singh AK, Kumar P, Ali R et al (2022a) An integrated statistical-machine learning approach for runoff prediction. *Sustainability* 14:8209. <https://doi.org/10.3390/su14138209>
- Singh VK, Panda KC, Sagar A et al (2022b) Novel genetic algorithm (GA) based hybrid machine learning-pedotransfer Function (ML-PTF) for prediction of spatial pattern of saturated hydraulic conductivity. *Eng Appl Comput Fluid Mech* 16:1082–1099. <https://doi.org/10.1080/19942060.2022.2071994>
- Singh VK, Kumar D, Kashyap PS et al (2020) Modelling of soil permeability using different data driven algorithms based on physical properties of soil. *J Hydrol* 580:124223. <https://doi.org/10.1016/j.jhydrol.2019.124223>
- Singha S, Pasupuleti S, Singha SS et al (2021) Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* 276:130265. <https://doi.org/10.1016/j.chemosphere.2021.130265>
- Srinivasulu S, Jain A (2006) A comparative analysis of training methods for artificial neural network rainfall-runoff models. *Appl Soft Comput* 6:295–306. <https://doi.org/10.1016/j.asoc.2005.02.002>
- Stefánsson A, Končar N, Jones AJ (1997) A note on the Gamma test. *Neural Comput Appl* 5:131–133. <https://doi.org/10.1007/BF01413858>
- Szabolcs I (1986) Agronomical and ecological impact of irrigation on soil and water salinity. In: Stewart BA (ed) *Advances in*

- soil science, 4th edn. Springer, New York, New York, NY, pp 189–218
- Tabari H, Sabziparvar A-A, Ahmadi M (2011) Comparison of artificial neural network and multivariate linear regression methods for estimation of daily soil temperature in an arid region. *Meteorol Atmos Phys* 110:135–142. <https://doi.org/10.1007/s00703-010-0110-z>
- Tahmasebi P, Hezarkhani A (2011) Application of a modular feed-forward neural network for grade estimation. *Nat Resour Res* 20:25–32. <https://doi.org/10.1007/s11053-011-9135-3>
- Taşan S (2023) Estimation of groundwater quality using an integration of water quality index, artificial intelligence methods and GIS: case study, central mediterranean region of Turkey. *Appl Water Sci* 13:15. <https://doi.org/10.1007/s13201-022-01810-4>
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Atmos* 106:7183–7192. <https://doi.org/10.1029/2000JD900719>
- Tutmez B, Hatipoglu Z, Kaymak U (2006) Modelling electrical conductivity of groundwater using an adaptive neuro-fuzzy inference system. *Comput Geosci* 32:421–433. <https://doi.org/10.1016/j.cageo.2005.07.003>
- Venkatesan E, Velmurugan T (2015) Performance analysis of decision tree algorithms for breast cancer classification. *Indian J Sci Technol* 8:1–8. <https://doi.org/10.17485/ijst/2015/v8i29/84646>
- Vishwakarma DK, Ali R, Bhat SA et al (2022a) Pre- and post-dam river water temperature alteration prediction using advanced machine learning models. *Environ Sci Pollut Res*. <https://doi.org/10.1007/s11356-022-21596-x>
- Vishwakarma DK, Kumar R, Pandey K et al (2018) Modeling of rainfall and ground water fluctuation of gonda district Uttar Pradesh, India. *Int J Curr Microbiol Appl Sci* 7:2613–2618. <https://doi.org/10.20546/ijcmas.2018.705.302>
- Vishwakarma DK, Pandey K, Kaur A et al (2022b) Methods to estimate evapotranspiration in humid and subtropical climate conditions. *Agric Water Manag* 261:107378–107416. <https://doi.org/10.1016/j.agwat.2021.107378>
- Vishwakarma DK, Kuriqi A, Abed SA et al (2023) Forecasting of stage-discharge in a non-perennial river using machine learning with gamma test. *Heliyon* 9:e16290. <https://doi.org/10.1016/j.heliyon.2023.e16290>
- Voyant C, Notton G, Kalogirou S et al (2017) Machine learning methods for solar radiation forecasting: a review. *Renew Energy* 105:569–582. <https://doi.org/10.1016/j.renene.2016.12.095>
- Wen X, Fang J, Diao M, Zhang C (2013) Artificial neural network modeling of dissolved oxygen in the Heihe river, Northwestern China. *Environ Monit Assess* 185:4361–4371. <https://doi.org/10.1007/s10661-012-2874-8>
- Willmott CJ (1981) On the validation of models. *Phys Geogr* 2:184–194. <https://doi.org/10.1080/02723646.1981.10642213>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Deepak Kumar¹ · Vijay Kumar Singh² · Salwan Ali Abed³  · Vinod Kumar Tripathi¹ · Shivam Gupta⁴ · Nadhir Al-Ansari⁵  · Dinesh Kumar Vishwakarma⁶  · Ahmed Z. Dewidar^{7,8} · Ahmed A. Al-Othman⁸ · Mohamed A. Mattar^{7,8,9} 

✉ Vijay Kumar Singh
vijaysinghswce@gmail.com

✉ Nadhir Al-Ansari
nadhir.alansari@ltu.se

✉ Mohamed A. Mattar
mmattar@ksu.edu.sa

Deepak Kumar
dpk94@bhu.ac.in

Salwan Ali Abed
salwan.abed@qu.edu.iq

Vinod Kumar Tripathi
vktripathi@bhu.ac.in

Shivam Gupta
shivamiitg@gmail.com

Dinesh Kumar Vishwakarma
dinesh.vishwakarma4820@gmail.com

Ahmed Z. Dewidar
adewidar@ksu.edu.sa

Ahmed A. Al-Othman
othmana@ksu.edu.sa

¹ Department of Agricultural Engineering, Institute of Agricultural Sciences, Banaras Hindu University, [BHU], Varanasi, Uttar Pradesh 221005, India

² Department of Soil and Water Conservation Engineering, Acharya Narendra Deva University of Agriculture & Technology, Kumarganj, Ayodhya, Uttar Pradesh 224229, India

³ College of Science, University of Al-Qadisiyah, P.O. Box.1895, Diwaniya 58001, Iraq

⁴ Department of Irrigation and Drainage Engineering, Acharya Narendra Deva University of Agriculture & Technology, Kumarganj, Ayodhya, Uttar Pradesh 224229, India

⁵ Department of Civil, Environmental, and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden

⁶ Department of Irrigation and Drainage Engineering, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand 263145, India

⁷ Prince Sultan Bin Abdulaziz International Prize for Water Chair, Prince Sultan Institute for Environmental, Water and Desert Research, King Saud University, Riyadh 11451, Saudi Arabia

⁸ Department of Agricultural Engineering, College of Food and Agriculture Sciences, King Saud University, Riyadh 11451, Saudi Arabia

⁹ Agricultural Engineering Research Institute (AENRI), Agricultural Research Centre, Giza 12618, Egypt