

# Improving water quality parameter prediction with multi-level linear regression model and hybrid feature selection

Aleefia Khurshid, Samruddhi Korke, Yudhir Kothari, Shruti Alone, Khushali Bais

Department of Electronics Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

## Article Info

### Article history:

Received May 29, 2024

Revised Sep 25, 2024

Accepted Dec 14, 2024

### Keywords:

Drinking water

Feature extraction

Machine learning

Multivariate linear regression

Water quality monitoring

## ABSTRACT

Predicting and modeling the quality of water is essential to guarantee that the water is safe to drink. The chlorine content in water needs to be monitored in real-time to provide a consistent supply of drinkable water. Additionally, potassium and chlorine have a major impact on how appealing the water is, as they are important components that influence taste and odor. Therefore, to evaluate the levels of chlorine and potassium, this work presents a multivariable linear regression approach backed by a hybrid feature extraction method. To bridge the gap between the filter and wrapper approaches, a hybrid approach is used to remove unnecessary information and reduce processing time and complexity. Here the quantitative parameters, in conjunction with categorical parameters, are instrumental in enabling accurate prediction of two water quality parameters. The two developed multi-level regression (MLR) models for the prediction of potassium and chloride are useful when factors affecting water parameters fluctuate at the site level as well as over larger spatial or temporal scales giving consumers a visual representation of how each parameter influences prediction. The converged model outperforms in comparison with other machine learning algorithms with a mean absolute error (MAE) of 7.42e-15 for potassium and 3.72e-14 for chloride.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Aleefia Khurshid

Department of Electronics Engineering, Shri Ramdeobaba College of Engineering and Management

Katol Road, Nagpur, Maharashtra, India

Email: khurshidaa@rknc.edu

## 1. INTRODUCTION

Water quality preservation is paramount as it directly impacts the health and well-being of organisms and ecosystems, highlighting the need to identify contaminants and pollutants that endanger human health and the environment. A critical aspect of water quality revolves around monitoring potassium (K) and chloride (Cl). For groundwater, a range of 35–125 mg/L is viewed as typical [3]. The water will taste salty when the amount of chloride is higher than 250–400 mg/L. Potassium is a direct indicator of contamination. For groundwater, a range of 35–125 mg/L is regarded as typical [1]. The water will taste salty if the chloride concentration is higher than 250–400 mg/L. In those who are vulnerable, it could have some negative health impacts [2]. Potassium also serves as a vital nutrient essential for the proper functioning of flora, fauna, and humans. However, elevated potassium levels in water may signify contamination from sources like agricultural runoff or industrial waste, necessitating monitoring to maintain water quality and safeguard aquatic life and potable water supplies. Similarly, chloride, often found in the form of sodium chloride (salt), can be harmless at low concentrations but poses risks at elevated levels originating from wastewater discharge or road salt runoff. Monitoring chloride levels is crucial to protect drinking water sources and mitigate adverse impacts on aquatic ecosystems. Ion-sensitive sensors commercially available

are costly. Developing a soft sensor for chloride or potassium in water is an alternative to commercially available sensors and therefore the significance of employing machine learning algorithms cannot be ruled out. Monitoring both potassium and chloride can help in understanding the efficiency and impacts of water treatment processes, such as softening, desalination, or reverse osmosis [3]. Commonly utilized machine learning algorithms for water quality parameter assessment include decision trees, support vector machines (SVM), random forests, and sophisticated deep learning methods like convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

River water chloride concentrations lead to rising salinity, which also poses a hazard to aquatic habitats. The significance of real-time river chloride prediction for managing and controlling chloride levels has drawn a lot of attention. However, increased values of potassium can noticeably affect the taste in water. The suggested deep learning model in [4] based on graphs acquired  $R^2$  and root mean square error (RMSE) values of 0.88 and 51.16 ppb, respectively. Conductivity, temperature, dissolved oxygen, PH, and turbidity are among the dependent parameters. Here the convolutional layers and a pooling layer make up the feed-forward structure, which results in a computationally demanding model for prediction of chloride.

Chinnappan *et al.* [5] present a fuzzy algorithm for determining chlorine levels in water, leveraging metrics such as recall, precision, and F-score for evaluation. With an F-score of 89%, a recall of 90%, and a precision of 92%, the suggested method performs better. The process uses chlorine levels and other variables, such as temperature (T), pH, and other chemicals, that may have an impact on chlorine levels as input. This study [6] uses four machine learning models to predict sodium adsorption ratio (SAR) and chloride concentration based on physical parameters such as EC, pH, temperature, and SAR: artificial neural network (ANN), k nearest neighbor (k-NN), and stochastic gradient descent (SGD). Trained on 176 samples and validated on 37 samples from Morocco's Chaouia coastal aquifer, the models demonstrated acceptable to good performances. The best chloride prediction models exhibit RMSE ranging from 1.74 to 2.67. The ANN and SGD models, offering the highest accuracy and stability, had 95% confidence bands of error at 1.39 for chloride.

To improve modeling accuracy, Zhang *et al.* [7] combined the perceptron model (MLP) and statistical inference model (SCA). An hourly river chloride prediction was conducted using the grand river in Canada as a case study, and the model performed well with RMSE of 11.58 mg/L, mean absolute percentage error (MAPE) of 27.55%, Nash–Sutcliffe efficiency (NSE) of 0.90, and  $R^2$  of 0.90. The provided data to the model include conductivity, water temperature, river flow rate, and rainfall. The study [8] develops an ANN model to predict increased chloride levels from road salt in a suburban watershed using measured rainfall volume and four other parameters (nitrate, suspended solids, turbidity, and dissolved organic carbon). Using three years of data at six sites, the ANN model, trained with backpropagation, shows a 91% fit between observed and predicted data. Spatial analysis reveals higher chloride clustering near impervious surfaces. The study suggests ANN modeling can be helpful for water quality prediction, particularly for chloride influenced by road salt.

Godo-Pla *et al.* [9] predicted the potassium permanganate demand for drinking water, using a multi-layer perceptron with four inputs resulting in an MAE of 0.128 mg·L<sup>-1</sup>. Artificial neural networks were explored in [10] for estimating chloride ion changes in urban ponds. When five water quality indices (COD-Cr, BOD5, DO, WS, and NO<sub>2</sub>) were used as inputs, the ANN model produced results with low error values and good predicted accuracy with MSE=4.94, RMSE=2.22, and MAPE of just 3.42%, despite a slightly higher  $R^2$  in the entrance zone. In this study [11], seven heavy metal parameters (Mg, SO<sub>4</sub>, K, Na, TH, Cl, and Ca) affecting water quality using deep learning techniques are predicted for measuring the water quality index. The input parameters temperature, EC, pH, and TDS were derived from 491 wells and the model performance indicates RMSE (train) of 8.12 and RMSE (test) of 11.36 for potassium (K). The chloride prediction using a deep neural network (DNN) resulted in RMSE (train) of 240.02 and RMSE (test) of 300.02.

Haghiabi *et al.* [12] developed the model using ANN and SVM, using distinct transfer and kernel functions, respectively. It was found that SVM had less data dispersion than the ANN. RMSE of 0.210 and  $R^2$  of 0.95 were obtained when tested SVM for prediction of chlorine. Using gradient boosting methods to build decision trees and produce predictions, the study [13] developed a machine learning model to forecast free chlorine residuals.

The possibilities for monitoring surface water quality using two machine learning algorithms: long short-term memory (LSTM) models and ANNs have been experimented in [14], [15]. However, they also come with specific challenges that need to be carefully managed when applied to surface water quality monitoring. These include data requirements, temporal correlations, model complexity, computational costs, and the ability to generalize across varying conditions.

Aldrees *et al.* [16] are of the opinion that the predictive models should be interpretable and have proposed a novel Shapley additive explanations (SHAP) technique for predicting water quality parameters. This technique is model-agnostic and has a high computational cost. While utilizing machine learning with boosted

trees, Schäfer *et al.* [17] were able to predict the changes in water quality with less than 1% error using two local and five global features including time stamp. These models require significant memory, with many trees. A feature importance study in [18] highlights the significant impact of specific variables and the effectiveness of machine learning models in differentiating between various parameters related to water quality.

In order to capture the complex correlations among water quality parameters, the study recommends using the highly accurate extreme gradient boosting (XGBoost), and random forest models-all of which are computationally expensive. The adaptive differential evolution algorithm proposed in study [19] uses the rank numbers to determine the positions of vectors in the mutation operation for solving various nonlinear regression problems. The method is self-adaptive but computation intensive. The Mamdani fuzzy technique excels at adapting to dynamic environmental shifts [20] in order to monitor critical parameters like pH, turbidity, temperature, and dissolved solids in shrimp cultivation. However, implementing complex membership functions can be difficult, particularly with limited hardware resources. Previous studies in the field of water quality research have explored the use of several machine-learning approaches to forecast the water quality index indicating that water quality measurements can be made with much greater precision due to machine learning and deep learning [21]–[24]. Future studies, according to the research team, should look into extending the applicability domain to enhancing predictivity. From the literature review, it can be concluded that the proposed models are either computationally expensive, depend on the appropriate choice of kernel, lack accuracy and interpretability, depend upon proper tuning of hyperparameters, or require many input parameters for the prediction of chloride and potassium. Few predictive frameworks target to predict chloride and potassium but require a large number of input features.

In order to close the research gap, the current work is focused on building a reliable and explainable predictive model with fewer input parameters that mitigates the aforementioned problems. Memory usage is managed with techniques like feature subsampling and using regression techniques. The robust regression technique employed for the presented work builds a model that explains to users how each parameter influences prediction. The model is constructed based on a substantial dataset from the river Ganga, sourced from the “Namami Ganga” project, where water contamination arises from effluents and various urban activities and is tested from data acquired from multiple sources. The combination of multilevel linear regression models and hybrid feature selection methods enhances the prediction of water quality parameters, emphasizing the novelty and effectiveness of the approach. It is anticipated that this work will advance the field by completing the following:

- a. Developing an explainable model for potassium and chloride concentration that will enable accurate outcomes by understanding the influence of each parameter on the predicted output.
- b. Employing a hybrid feature selection methodology to examine the significance of various input factors for prediction analysis.
- c. Establishing a framework for the creation of an effective hardware implementation with pre-knowledge of the regression coefficients.

The later sections also present a comparison between various machine-learning algorithms given their computational cost and efficiency.

## 2. METHODOLOGY

A multi-level regression model is used to understand factors affecting water parameters at site and spatial levels with a hybrid feature extraction approach to accurately predict water quality parameters like potassium and chloride. The step by approach is presented in Figure 1. The dataset used in this study is compiled by scrapping live data from the “Namami Ganga” project, consisting of 29,007 samples from 42 locations of river Ganga and 12 locations in river Yamuna on seasonal basis. The dataset comprises seventeen critical variables, including pH, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), temperature, color, total organic carbon (TOC), electrical conductivity (EC), and total dissolved solids (TDS), for six months. capturing variations resulting from weather changes. To ensure the enhanced quality of data, pre-processing techniques are implemented. The process involved data cleaning and transformation in Figure 1, which were aimed at improving the integrity of the dataset. Specifically, the linear scaling method was employed to normalize the data, resulting in a collection of approximately 23,000 clean and transformed samples. The sample data extracted from the database is presented in Table 1.

### 2.1. Correlation analysis and feature selection

In addition to water level, a correlation heat map was utilized to examine the relationship among the seventeen parameters that were taken into account. To simulate several learning models, variables with a correlation coefficient greater than 0.4 were used. Figure 2 displays the correlation heat map for the obtained parameters.

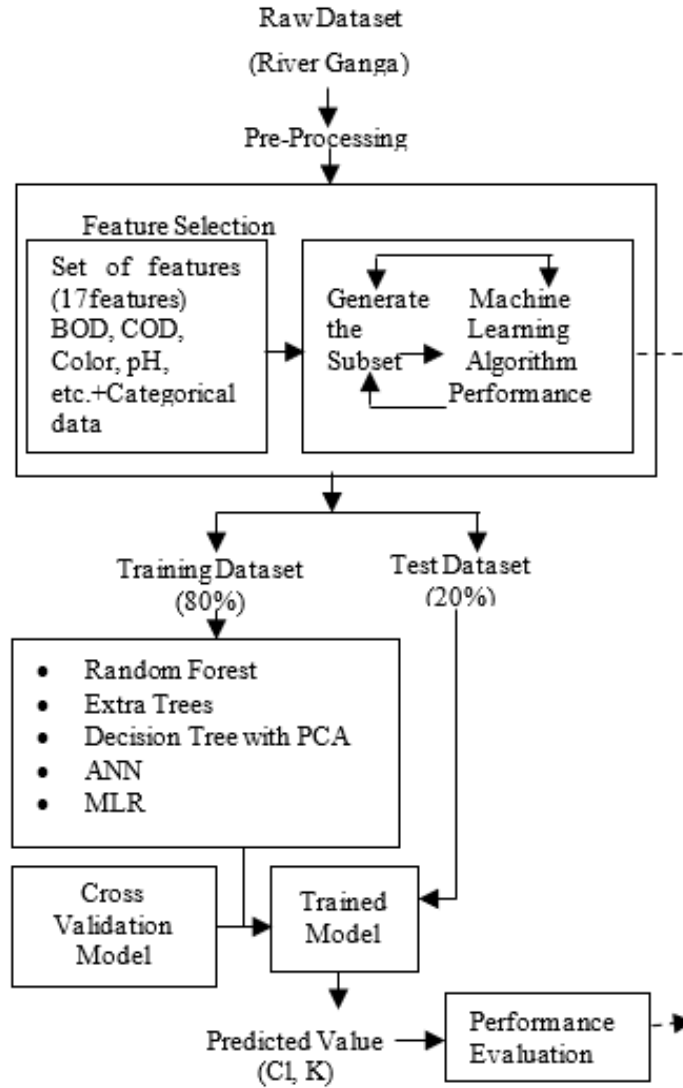


Figure 1. Methodology for feature selection and model design

Table 1. Sample data set

BOD (mg/l)	DO (mg/l)	Conductivity (µS/cm)	pH	Temperature (°C)	Potassium (mg/l)	Chloride (mg/l)	COD (mg/l)	TSS (mg/l)
3.42	8.61	1	0	26.4	10.42	10.7	17.35	13.77
1.58	5.48	160	8.74	26	10.42	0	14.97	120.73
1.84	6.99	288	7.65	30	10.42	0	13.15	38.12
1.99	6.83	190	8.53	26.7	10.42	0	17.36	143.34
1.18	8.07	234	7.21	26.9	10.42	0	13.32	128.95
1.16	1.96	365	8.7	31	9.54	19.9	12.86	22.34
4.64	9.06	727	8.3	31.3	5.64	18.1	25.2	129.33
3.34	7.43	182	7.84	28.5	1.8	4.2	18.14	220.43
2.14	6.76	198	7.6	29.6	3.64	9	14.2	158.23

To bridge the gap between the filter and wrapper approaches, a hybrid approach is used to remove unnecessary information and reduce processing time and complexity. The feature set is filtered using a correlation heatmap, and the ranking information that the filter method provides is then used to evaluate the features using particular machine-learning methods in Figure 1. Considering the correlation map and the interplay among different parameters, ten parameters which include BOD, DO, COD, pH, conductivity, total suspended solids (TSS), temperature, TOC, color and turbidity were finally utilized as inputs to the multi-level regression (MLR) model.

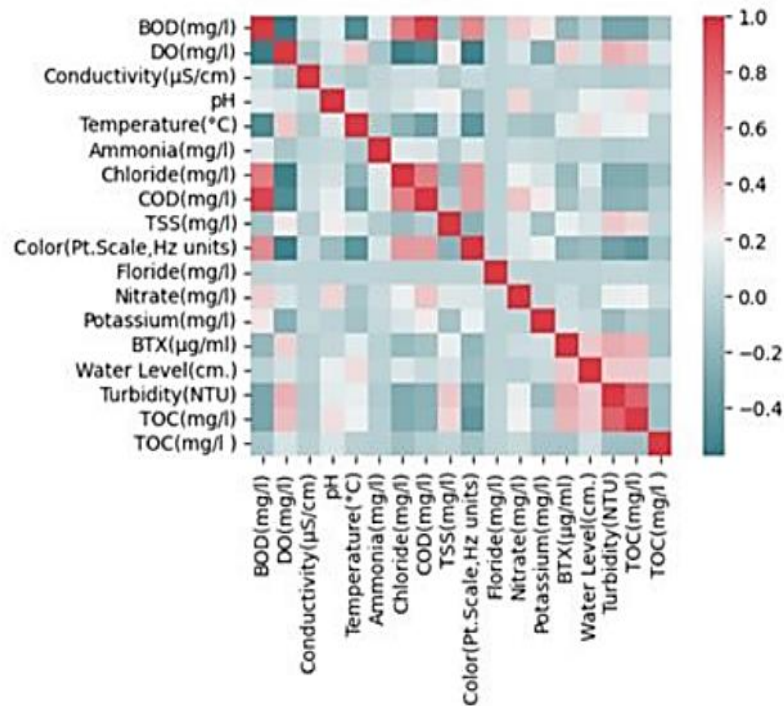


Figure 2. Correlation heatmap

## 2.2. Model selection and parameter tuning

From the literature review, it was derived that the researchers have utilized SVM, ANN, and deep learning techniques for predicting the water quality parameters. Theoretically, SVMs cannot converge to a solution, particularly when dealing with noisy data. Additionally, by employing a collaborative decision-making process that is aided by numerous trees offering their insights and producing accurate and consistent results, non-guaranteed convergence of neural networks can be avoided. Therefore, random forest, extra trees, k-means clustering, and decision tree are the algorithms that have been chosen in this context to provide reliable forecasts in various environments and extract additional performance from machine learning systems. These algorithms are chosen based on factors like the algorithm's interpretability, computational efficiency in resource-constrained environments, and ability to handle multivariate data, and the model is developed as shown in Figure 1. Two distinct models have been created to predict K and Cl. Metrics like F-score, accuracy, and recall are used to compare performance. These metrics such as accuracy, precision, recall, or mean squared error (MSE), provide a numerical assessment of the prediction accuracy and can be applied to evaluate how well the suggested model performs.

## 3. RESULTS AND DISCUSSION

This section presents and compares the results and analysis of the evaluations of the various prediction models with the multilinear regression model that has been suggested. The simulated machine learning models are optimized for increased prediction accuracy and employ either supervised or unsupervised learning for prediction using various transfer and kernel functions. Additionally provided are the findings from the prediction of the internal relationships between the components of water quality.

### 3.1. Performance analysis

In this study, decision tree, extra tree, random forest, ANN, MLR, models are used for the estimation of K and Cl in river water. The data set was partitioned for training and validation and verification purposes. Multilinear regression: Its basis is the linear relationships between inputs and outputs. This extracts the linear correlations between the dependent and independent variables using a constant regression in the formula [25]. The equation below is the basis of MLR work:

$$y = b_0 + b_1x_2 + \dots b_ix_i \quad (1)$$

where  $Y$  is the independent variable,  $B$  is the regression constant,  $X$  is the  $i^{\text{th}}$  predictor.

- a. Random forest (RF): The main method used in this approach is supervised learning. For each training set of data, this algorithm will create a decision tree. A tree does not need any features to be taken into account because it is a conceptual construct. As a result, there is less feature space. The final prediction for regression tasks is calculated by averaging the predictions made by each tree [26]–[28].
- b. Decision tree (DT): It consists of a single decision tree that can be trained. The prediction is fast as it adapts quickly to the dataset but is more susceptible to outliers.
- c. Extra tree: It generates a large number of decision trees, but there is no replacement in the random sample for each tree. A feature's uniqueness is determined by a random selection of its splitting value to calculate a locally optimal value. The trees become diverse and uncorrelated as a result.
- d. Artificial neural network (ANN): ANNs are weighted feedforward linked networks of neurons with weighted connections. The neurons are arranged in layers, where an input layer corresponds to a certain input data vector and an output layer yields the regression's result. The performance analysis for the evaluation metrics (MSE) of the different machine learning algorithms discussed above for potassium (K) and chloride (Cl) are stated in Tables 2 and 3.

From the Tables 2 and 3, it is evident that multilinear regression model outperforms the other stated models with ten input parameters both for K and CL. The output equation for MLR with ten input parameters is as stated in (2),

$$y = h_0 + h_1x_1 + h_2x_2 + h_3x_3 + h_4x_4 + h_5x_5 + h_6x_6 + h_7x_7 + h_8x_8 + h_9x_9 + h_{10}x_{10} \quad (2)$$

Table 2. Performance of machine learning algorithms for potassium level prediction

Sr. No	Input parameter	MSE value	Algorithm
1.	Bod, Cod, color	14.49	Random forest
2.	Bod, Cod, color	3.29	Extra trees
3.	Bod, Cod, color	51.84	Decision tree with PCA
4.	Bod, Cod, color, pH	99.91	Decision tree
5.	Bod, Cod, color, pH, conductivity	73.61	Decision tree
6.	Bod, Cod, color, conductivity	92.26	Decision tree
7.	Bod, Cod, pH, conductivity	82.97	Decision tree
8.	Bod, Cod, conductivity, temperature	56.97	Decision tree
9.	Bod, Cod, conductivity, temperature, pH	41.21	Decision tree
10.	Bod, Cod, color, pH	0.8603	ANN
11.	BOD, DO, Cod, pH, conductivity, TSS, temperature, TOC, color, turbidity	7.13e-29 MAE:7.42e-15	MLR

Table 3. Performance of machine learning algorithms for chloride level prediction

Sr. No	Input parameter	MSE value	Algorithm
1.	Bod, Cod, color, Do	30.30	Random Forest
2.	Bod, Cod, color, Do	18.94	Extra Trees
3.	Bod, Cod, color, Do	6.88	Decision Tree with PCA
4.	Bod, Cod, color, pH	1297.63	Decision Tree
5.	Bod, Cod, color, pH, conductivity	1199.88	Decision Tree
6.	Bod, Cod, color, conductivity	1131.91	Decision tree
7.	Bod, Cod, pH, conductivity	1182.97	Decision Tree
8.	Bod, Cod, conductivity, temperature	1109.27	Decision tree
9.	Bod, Cod, conductivity, pH	1.5008	ANN
10.	BOD, DO, COD, pH, conductivity, TSS, Temperature, TOC, color, turbidity	1.58e-27 MAE: 3.72e-14	MLR

The coefficients are stated in the Tables 4 and 5. It can be concluded that the conductivity has a statistically significant positive effect on potassium concentration in water. This aligns with the physical understanding that potassium ions contribute to the overall ionic strength and thus the conductivity of water. Also, the independent variables are not highly correlated with each other indicating a stable model. Also, the changes in pH are a significant predictor of changes in chloride concentration. This aligns with the chemical understanding that pH can affect the solubility and dissociation of chloride-containing compounds, thereby influencing chloride levels in water. Thus, by examining the magnitude and sign of the coefficients, users can determine which features are most important for the prediction. This helps in understanding which water quality parameters have the most significant impact.

Table 4. Converged MLR coefficients for potassium level prediction

Intercept	h0	5.853872596
BOD (mg/l)	h1	0.223606084
DO (mg/l)	h2	0.011443784
COD (mg/l)	h3	0.063117283
pH	h4	0.200946149
Conductivity (μS/cm)	h5	1.88847E-05
TSS (mg/l)	h6	-0.005238827
Temperature(°C)	h7	0.032758376
TOC (mg/l)	h8	-0.101306551
Color (Pt. Scale, Hz units)	h9	0.008637261
Turbidity (NTU)	h10	-0.006055858

Table 5 Converged MLR coefficients for chloride level prediction

Intercept	h0	-18.32195591
BOD (mg/l)	h1	1.338602419
DO (mg/l)	h2	-0.670665515
COD (mg/l)	h3	0.286509312
pH	h4	3.289346014
Conductivity (μS/cm)	h5	-3.91108E-06
TSS (mg/l)	h6	-0.015863574
Temperature(°C)	h7	0.17906353
TOC (mg/l)	h8	-0.508046421
Color (Pt. Scale, Hz units)	h9	0.037178484
Turbidity (NTU)	h10	0.012387753

The Figure 3 indicates the plot of predicted and actual values of potassium (K) for the multilinear regression model indicative of the good fit. The Figure 4 indicates the plot of predicted and actual values of chloride (Cl) for the multilinear regression model while testing. Both the models strike a good balance between bias and variance, avoiding underfitting and overfitting. Therefore, the independent variables chosen are relevant and have a strong relationship with the dependent variable indicating a robust feature selection. The relationships identified by the model are statistically significant with a p-value less than 0.05. The Table 6 presents a comparison of work published by other researchers for the prediction of heavy metals chloride and potassium. It can be concluded that the MLR is highly suitable for prediction of K and Cl utilizing ten input variables.

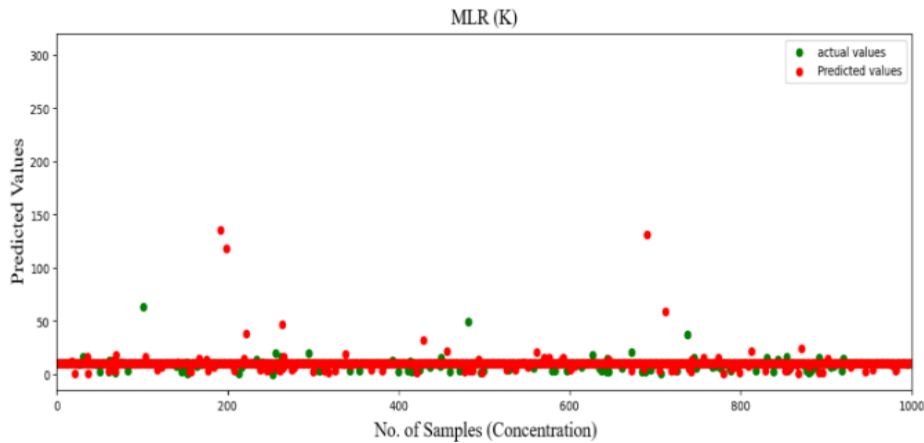


Figure 3. Actual vs predicted potassium concentration for different samples using MLR

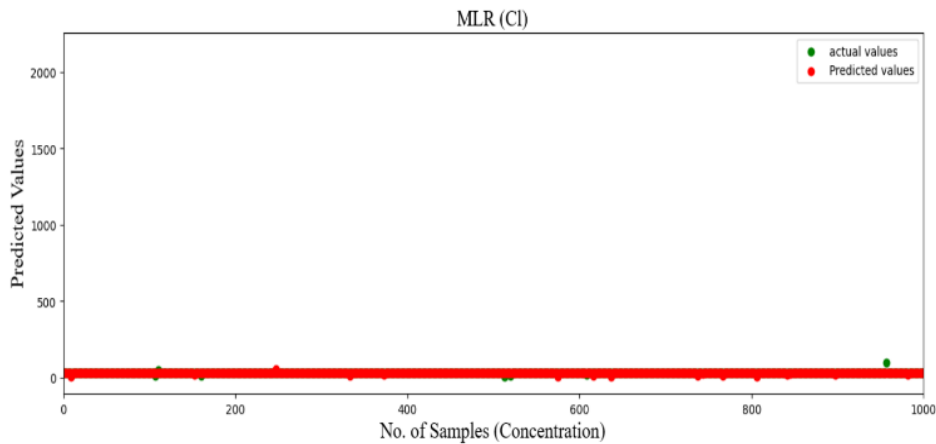


Figure 4. Actual vs predicted chloride concentration for different samples using MLR

The experimental findings indicate that since multilevel regression models are made to take into consideration nested data structures, they allow for random effects at many levels. This is because water quality data frequently have a hierarchical structure, such as measurements made from multiple locations over different time periods. When factors affecting water parameters fluctuate at the site level as well as over larger spatial or temporal scales, this can be helpful in studies on water quality. Simpler explanations of the correlations between the predictors and the response variable are offered by multilevel regression models substantiating the objective of the proposed work. In contrast to the intricate, non-linear interactions found in decision trees or random forests, the coefficients in a regression model show the direct influence of each predictor, which is frequently easier to understand. Figures 5 and 6 presents a plot of actual vs predicted potassium and chloride concentration when the proposed model is tested for out of bag samples from different locations.

Table 6. Performance comparison with reported work

Year of publication/ Reference No.	Input parameters	Performance metric	Algorithm
2023 [5]	T, pH, water flow	Cl level prediction: recall: 90%, precision: 92%, F-score: 89%	Decision tree
2020 [8]	Rainfall volume, turbidity, total suspended solids (TSS), dissolved organic carbon (DOC), sodium, chloride, and total nitrate concentrations	Cl level; prediction accuracy-91% RMSE: 3.09(averaged over six different sites data)	ANN
2021 [6]	EC, T, pH	Cl level prediction: RMSE: 0.01 (ANN), 0.13 (SGD) 7.17 (KNN), 10.53 (SVM)	SGD, ANN, k-NN, and SVM
2012 [7]	EC, T, river flow rate, and rainfall	Cl level prediction: R <sub>2</sub> : 0.9 RMSE: 11.78	Stepwise cluster analysis with MLP
2019 [9]	UV254, turbidity, T, inflow	For K level prediction: MAE: 0.128	ANN
2023 [12]	T, pH, EC, and TDS	Cl level prediction: RMSE: 300.42 R <sub>2</sub> : 0.98 K level prediction: RMSE: 11.36 R <sub>2</sub> : 0.92	Deep neural network
Proposed MLR model	BOD, DO, Cod, pH, EC, TSS, T, TOC, color, turbidity	K level prediction: MSE: $7.13e^{-29}$ MAE: $7.42e^{-15}$ Cl level prediction: MSE: $1.58e^{-27}$ MAE: $3.72e^{-14}$	MLR

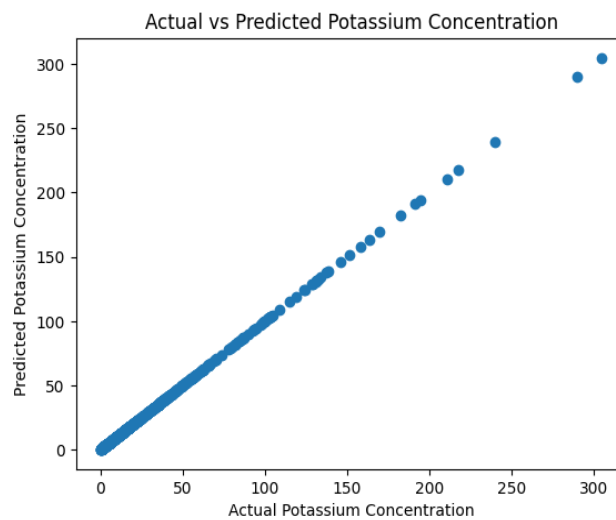


Figure 5. Scatter plot with the forecasted and measured potassium concentrations



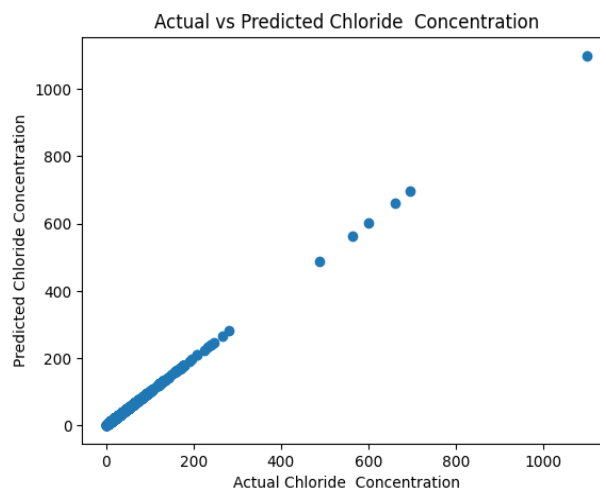


Figure 6. Scatter plot with the forecasted and measured chloride concentrations

#### 4. CONCLUSION

The analysis of water quality encompasses the examination of potassium (K) and chloride (Cl), both of which play crucial roles in environmental and human health. In the realm of predictive analysis, various intelligent algorithms are employed to forecast water quality parameters. However, upon comparing the accuracy of MLR from the literature with the results obtained from our experimental implementation, it is evident that MLR exhibited superior accuracy and yielded the best-fit results. This underscores the robustness and efficacy of the MLR model in predicting water quality parameters giving consumers a visual representation of how each parameter influences prediction.

By combining the broad efficiency of filter methods with the specific accuracy of wrapper methods, this hybrid approach enhances computational feasibility, model performance, and robustness, making it a valuable strategy in machine learning. While the literature review did not thoroughly address the implementation of MLR, the potential for integrating this software methodology onto low-cost, low-powered hardware for real-time water quality monitoring presents an intriguing prospect. This avenue suggests the possibility of applying advanced analytical techniques to practical, real-world scenarios, paving the way for cost-effective, real-time monitoring solutions. In future residual analysis can be carried out to suggest other factors influencing the dependent variable, if any. Also, the work can be directed for real-time implementation of the converged model. The provided conclusion integrates the significance of potassium and chloride ions in water quality, the prowess of the MLR model, and the potential for real-time monitoring solutions to support public health by ensuring safe drinking water aiding environmental management





#### REFERENCES

- [1] CPCB, "CPCB's technical guidelines," *Central Pollution Control Board*. <https://cpcb.nic.in/cpcb-technical-guidelines-sops/> (accessed May 29, 2024).
- [2] BIS, "Drinking water-specification," Bureau of Indian Standards, New Delhi, 2012.
- [3] M. Kumar and A. Puri, "A review of permissible limits of drinking water," *Indian Journal of Occupational and Environmental Medicine*, vol. 16, no. 1, p. 40, 2012, doi: 10.4103/0019-5278.99696.
- [4] V. Oliveira Santos, P. A. Costa Rocha, J. V. G. Thé, and B. Gharabaghi, "Graph-based deep learning model for forecasting chloride concentration in urban streams to protect salt-vulnerable area," *Environments - MDPI*, vol. 10, no. 9, 2023, doi: 10.3390/environments10090157.
- [5] C. V. Chinnappan *et al.*, "IoT-enabled chlorine level assessment and prediction in water monitoring system using machine learning," *Electronics (Switzerland)*, vol. 12, no. 6, 2023, doi: 10.3390/electronics12061458.
- [6] A. El Bilali, A. Taleb, A. Nafii, B. Alabjah, and N. Mazigh, "Prediction of sodium adsorption ratio and chloride concentration in a coastal aquifer under seawater intrusion using machine learning models," *Environmental Technology and Innovation*, vol. 23, 2021, doi: 10.1016/j.eti.2021.101641.
- [7] Q. Zhang *et al.*, "Real-time prediction of river chloride concentration using ensemble learning," *Environmental Pollution*, vol. 291, 2021, doi: 10.1016/j.envpol.2021.118116.
- [8] K. Jahan and S. M. Pradhanang, "Predicting Runoff chloride concentrations in suburban watersheds using an artificial neural network (ANN)," *Hydrology*, vol. 7, no. 4, pp. 1–17, 2020, doi: 10.3390/hydrology7040080.
- [9] L. Godo-Pla, P. Emiliano, F. Valero, M. Poch, G. Sin, and H. Monclús, "Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural network: Uncertainty and sensitivity analysis," *Process Safety and Environmental Protection*, vol. 125, pp. 317–327, 2019, doi: 10.1016/j.psep.2019.03.017.
- [10] T. Miller and G. Poleszczuk, "Prediction of the seasonal changes of the chloride concentrations in urban water reservoir," *Ecological Chemistry and Engineering S*, vol. 24, no. 4, pp. 595–611, 2017, doi: 10.1515/eces-2017-0039.





- [11] H. Moeinzadeh, P. Jegakumaran, K. T. Yong, and A. Withana, "Efficient water quality prediction by synthesizing seven heavy metal parameters using deep neural network," *Journal of Water Process Engineering*, vol. 56, 2023, doi: 10.1016/j.jwpe.2023.104349.
- [12] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," *Water Quality Research Journal*, vol. 53, no. 1, pp. 3–13, 2018, doi: 10.2166/wqrj.2018.025.
- [13] NSF, "Inexpensive monitoring process powered by machine learning could aid in water treatment," *U.S. National Science Foundation*. <https://new.nsf.gov/news/inexpensive-monitoring-process-powered-machine> (accessed May 29, 2024).
- [14] R. Rana *et al.*, "Artificial intelligence for surface water quality evaluation, monitoring and assessment," *Water (Switzerland)*, vol. 15, no. 22, 2023, doi: 10.3390/w15223919.
- [15] T. Y. Wei, E. S. Tindik, C. F. Fui, Haviluddin, and M. H. A. Hijazi, "Automated water quality monitoring and regression-based forecasting system for aquaculture," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 12, no. 1, pp. 570–579, 2023, doi: 10.11591/eei.v12i1.4464.
- [16] A. Aldrees, M. Khan, A. T. B. Taha, and M. Ali, "Evaluation of water quality indexes with novel machine learning and shapley additive explanation (SHAP) approaches," *Journal of Water Process Engineering*, vol. 58, 2024, doi: 10.1016/j.jwpe.2024.104789.
- [17] B. Schäfer *et al.*, "Machine learning approach towards explaining water quality dynamics in an urbanised river," *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-16342-9.
- [18] F. Abbas *et al.*, "Machine learning models for water quality prediction: a comprehensive analysis and uncertainty assessment in Mirpurkhas, Sindh, Pakistan," *Water (Switzerland)*, vol. 16, no. 7, 2024, doi: 10.3390/w16070941.
- [19] W. Wongsu, P. Puphasuk, and J. Wetweeraopong, "Differential evolution with adaptive mutation and crossover strategies for nonlinear regression problems," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 13, no. 5, pp. 3503–3514, 2024, doi: 10.11591/eei.v13i5.6417.
- [20] M. Qomaruddin, A. Riansyah, and H. M. Hermawan, "Mamdani fuzzy-based water quality monitoring and control system in vannamei shrimp farming using the internet of things," *International Journal of Advances in Applied Sciences*, vol. 13, no. 1, pp. 180–187, 2024, doi: 10.11591/ijaas.v13.i1.pp180-187.
- [21] Z. Wang, Q. Wang, and T. Wu, "A novel hybrid model for water quality prediction based on VMD and IGOA optimized for LSTM," *Frontiers of Environmental Science and Engineering*, vol. 17, no. 7, 2023, doi: 10.1007/s11783-023-1688-y.
- [22] M. G. Uddin *et al.*, "Marine waters assessment using improved water quality model incorporating machine learning approaches," *Journal of Environmental Management*, vol. 344, 2023, doi: 10.1016/j.jenvman.2023.118368.
- [23] P. L. Georgescu *et al.*, "Assessing and forecasting water quality in the Danube River by using neural network approaches," *Science of the Total Environment*, vol. 879, 2023, doi: 10.1016/j.scitotenv.2023.162998.
- [24] A. Y. Sun and B. R. Scanlon, "How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions," *Environmental Research Letters*, vol. 14, no. 7, 2019, doi: 10.1088/1748-9326/ab1b7d.
- [25] W. Liu, T. Liu, Z. Liu, H. Luo, and H. Pei, "A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction," *Environmental Research*, vol. 224, 2023, doi: 10.1016/j.envres.2023.115560.
- [26] P. Sihag, A. Angelaki, and B. Chaplot, "Estimation of the recharging rate of groundwater using random forest technique," *Applied Water Science*, vol. 10, no. 7, p. 182, Jul. 2020, doi: 10.1007/s13201-020-01267-3.
- [27] N. Bournas, A. Galdeano, M. Hamoudi, and H. Baker, "Interpretation of the aeromagnetic map of Eastern Hoggar (Algeria) using the Euler deconvolution, analytic signal and local wavenumber methods," *Journal of African Earth Sciences*, vol. 37, no. 3–4, pp. 191–205, Oct. 2003, doi: 10.1016/j.jafrearsci.2002.12.001.
- [28] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *Journal of Biomedical Informatics*, vol. 69, pp. 218–229, May 2017, doi: 10.1016/j.jbi.2017.04.001.

## BIOGRAPHIES OF AUTHORS






**Aleefia Khurshid**     received the Ph.D. degree in electronics engineering from Visvesvaraya National Institute of Technology, Nagpur, India in the year 2011. She is a life member of IETE, ISTE, and IE. Her current research interests include data-driven system identification, modeling, and image processing. She is currently serving as professor in the Department of Electronics Engineering at Shri Ramdeobaba College of Engineering and Management, Nagpur, India. She can be contacted at email: khurshidaa@rknc.edu.






**Samruddhi Korke**     is a recent graduate with a B.Tech. degree in electronics engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur. Samruddhi is curious about the emerging technology in the field of VLSI and AIML. Her dedication to interdisciplinary innovation underscores the significance of her contributions to this research, promising to advance the field of water quality monitoring. She can be contacted at email: korkesd@rknc.edu.






**Yudhir Kothari**    recently completed his graduation with a degree in B.Tech. Electronics Engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur. His research interest includes data analysis and management. Yudhir is currently working in sales and business development in Mumbai. He can be contacted at email: kotharira@rknec.edu.



**Khushali Bais**    is a recent electronics engineering graduate from Shri Ramdeobaba College of Engineering and Management, Nagpur. Khushali has a strong passion for AI generative technologies and web development and is actively pursuing advancements in the field through active learning. She can be contacted at email: baik@rknec.edu.



**Shruti Nitin Alone**    is a recent graduate with a degree in electronics engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur. She is currently working as an assistant system engineer at Tata Consultancy Services, Nagpur. With a background in AI and machine learning, she is committed to enhancing her expertise in the field and making significant contributions. She can be contacted at email: alones@rknec.edu.