Prediction of Physico-Chemical Characteristics of Groundwater Using Machine Learning Model

Dr L.Bhagya Lakshmi[1] , P Ramakoteswara rao [2], Ch.Chandra Mohan [3], Lella Kranthi Kumar[4], Dr.Kusuma Sundara Kumar[5*] Dr.Bandaru Venkata Shiva Kumar [6]

[1]Sr.Asst.Professor, Freshman Engineering Department, Lakireddy Bali Reddy College of Engineering, Mylavaram -521230, India
bhagyavvrk@gmail.com
[2]Asst. Professor, Freshman Engineering Department, PVP Siddhartha Institute of Technology, Kanuru, Vijayawada.A.P, India
ramakotesh26parasa@gmail.com
[3]Asst. Professor, CSIT Department, PVP Siddhartha Institute of Technology, Kanuru, Vijayawada, A.P, India
chetlachandramohan234@gmail.com
[4]Asst. Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India
kranthi1231@gmail.com
[5]Professor, Dept. of R &D, Bonam Venkata Chalamayya Engineering College-Odalarevu Konaseema, Andhra Pradesh, India.
*skkusuma123@gmail.com
[6]Professor, Department of Civil Engineering, WISTM Engineering College, Pendurthi, Visakhapatnam, Andhra Pradesh, India-531173.
shivakumar.bandaru@gmail.com

Abstract. — To maintain future supplies of clean drinking water, it is necessary to assess the state and degree of contamination in current groundwater. Predicting water quality properly is critical for reducing pollution and improving water management. This research offers a deep learning (DL)-based algorithm for predicting groundwater quality. To calculate the entropy weight-based groundwater quality index (EWQI), 200 groundwater samples are gathered from the research region, which is mostly utilized for agriculture in Krishna district, Andhra Pradesh, India. A variety of physicochemical characteristics are assessed in these samples. A set of five error metrics were created for assessing the performance of the model. The results show that the DL model, is working well with R2value of 0.996. It was shown that the most realistic and accurate method for predicting groundwater quality was the DL model.

Keywords: Ground water, Physico-chemical characteristics, Deep learning, EWQI, prediction

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_60

1 Introduction

Groundwater is not only an important supply of potable drinking water in many nations, but it also supplies a large amount of the water needed for domestic and agricultural use. Over the past few decades, there has been a massive decline in groundwater quality due to factors such as population development, urbanization, improper chemical fertilizer use, climate change, and inadequate management of water resources. Nevertheless, because there are no substitute sources of groundwater, Even if the quality of the groundwater is declining, people nevertheless utilize it for drinking in many parts of the world. Globally, there is a serious shortage of groundwater, a freshwater resource and essential component of the hydrological cycle [1].

Determining the quality of drinking water is so crucial in the modern world. Numerous studies have evaluated surface and groundwater quality for human use using the subjective technique known as the water quality index (WQI). The main drawback of this subjective evaluation method is that the experts who set the parameter weights for calculating the score of WQI have to account for ambiguity in the result.

However, assessing the quality of groundwater subjectively yields unrepresentative results. On the other hand, because they take into account local fluctuations in a dataset throughout the computing process, Techniques based on an objective weighing system are more reliable [2]. Several studies have also employed objective weighting systems based on statistical data analysis to avoid making errors in judgment as a result of an insufficient set of realistic weights in an assessment of the water quality. This has produced a more dependable result. By giving entropy-based weights to physicochemical factors,

A few researchers made an effort to lessen the usual WQI method's subjectivity. This shown that the subjective weighting approach is not as accurate or reliable as the objective-weighting system, or EWQI. As a result, it is imperative to improve the process of evaluating the quality of the water by employing an unbiased tool with a decision-making capacity that is both reliable and flexible [3].

However, there are a number of difficulties in assessing the quality of water, including large-scale sample collection, testing, and data handling. These jobs need a significant investment of time and money for labor, supplies, and equipment. Moreover, it requires a great deal of time and work to calculate quality indices. The financial value losses connected to using traditional techniques to assess water quality have an influence on decision-making on management plans for water quality. Consequently, a workable and affordable plan for the rapid and precise testing of quality of water needs to be implemented in order to solve these issues. In this case, use of Deep learning (DL) models gives us an efficient and trustworthy way to assess the ground water quality [4].

Machine learning is a versatile, strong, and promising method in all fields of science. In several water research projects around the world, researchers have employed random forests (RF), eXtreme Gradient Boosting (XGBoost), and artificial neural networks (ANN) as machine learning (ML) techniques. Several researchers have estimated water quality, located pollution sources in the water supply network, forecast-

ed manganese (Mn) removal, and examined flood susceptibility using the RF model. Similarly, the XGBoost model was used to forecast lead (Pb) levels, biological water quality monitoring., and water quality parameters with different levels of accuracy.. Numerous studies, including those on water level forecasting, flood vulnerability, heavy metal pollution prediction, and wastewater heavy metal treatment, have made substantial use of ANN-based prediction models [5]. In addition to the research described above, a large number of studies have been undertaken over the last decade to forecast WQI by testing the performance of various ML models.

Researchers were able to estimate river WQI with 95% R2 accuracy by utilizing the ANN model. A support vector regression technique based on swarm optimization was used to predict WQI. Certain publications compared the standard back-propagation (BP) model to the BP model based on artificial bee colonies. The most reliable ANN model for WQI prediction was established utilising a cascade forward network. In a different research, machine learning methods were used to assess the same. The most effective models were the multilayer perception model for classification, gradient boosting combined with polynomial regression, and regression models like these. Additionally, WQI prediction models utilizing four traditional ML techniques and twelve hybrid machine learning approaches were provided. In every one of these studies, hybrid machine learning models outperformed traditional models in terms of prediction accuracy [6].

From the review of literature it is evident that most of the earlier studies concentrated on developing models for predicting water quality that took into account the conventional subjective weight-based WQI, which takes into account individual evaluations when deciding how much weight to assign the quality criterion. Therefore, a helpful computational technique for forecasting water quality in any place will be the objective based weighting model [7]. Traditional WQI loses important information about a place's water quality since it places greater weight on characteristics selected based on expert opinions or assessments. EWQI, on the other hand, is advancement above conventional WQI.

Additionally, performing normal WQI calculations takes a lot of effort and time. The current study applied the ground water quality prediction modeling. In this model EWQI is used DL technique also adopted [8]. The DL technique has garnered a lot of attention lately because to its capacity to assess nonlinear correlations and dataset complexity. This machine learning technique's disregard for specific properties that are more representative than those required by traditional machine learning methods is the main factor in its performance.

The model's success is mostly dependent on the proper feature selection; therefore it might be difficult and time-consuming to remove attributes that are not useful for the algorithm. The DL technique, on the other hand, uses feature learning as a self-deterministic method to identify the necessary representation for a specific job. It is well-suited for processing multi-dimensional inputs due to its additional complex topologies, robust learning capacity, significant flexibility in model configuration, and higher generalization capabilities.[9]

A few recent research have demonstrated the efficacy of deep learning algorithms in forecasting soil moisture content, flash flood vulnerability, heavy metal contamination of groundwater [10]. Above all, deep learning models have shown effective in a wide range of research domains. Still, there are very few applications of the DL model in hydrology, much less for groundwater quality forecasting. To the best of the authors' knowledge, no prior study has examined and confirmed the efficacy of the DL model in groundwater quality prediction using the objective-weighting method. The current study proposes to develop a model in the widespread agricultural area of Krishna district of Andhra Pradesh, India, in an effort to close this gap [11].

Because human pollution severely damages this critical supply of drinking water in the research study area, a detailed evaluation of groundwater quality is necessary. Furthermore, no such comprehensive analysis has been finished at the current study site. It is more accurate to forecast groundwater quality using multiple machine learning techniques than it is to evaluate it using only one. The performance of each prediction model is compared with three popular machine learning methods (ANN, XG Boost, and RF) in order to decide which is best for the field being studied [12].

2 Materials And Method

Study Area:

The Krishna district of Andhra Pradesh is home to the research study area, which is located in its eastern and central parts. The investigation's scope is approximately 3430 km^2. Geographically, the study area lies between 15°42’ 05’’ and 16°45’ 06’’ N latitude and between 80° 48’ 06’’ and 81° 35’ 11’’ E longitude, as Figure 1 illustrates, almost third of study area is made up of agricultural land.

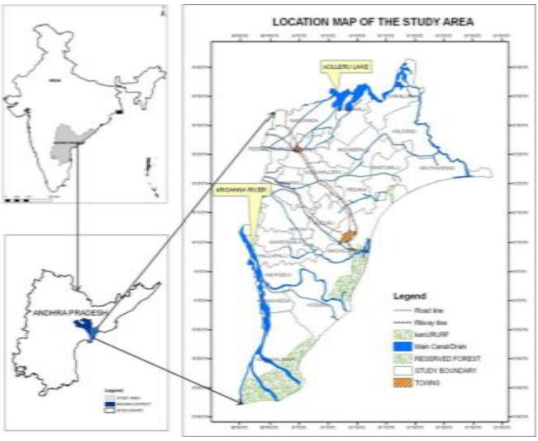

Fig.1: Location map of the study area

The primary crop is paddy, as per reports. Its production accounts for 61% of total crop area during the Kharif season, and it has intensity of cropping around 163%. Because of human activity, the current study region has the potential to become contaminated groundwater. Irrigation techniques are widespread throughout the research region, with the exception of the monsoon season. The District Census Handbook, 2011 states that surface and groundwater sources irrigate about 76% of the entire cultivable area.

The suggested approach of the current study is divided into four main phases: gathering of groundwater samples, study of physicochemical parameters, computation of EWQI. The data includes the creation of ANN, DL, XGBoost, and RF model techniques as well as the splitting of the dataset for training and testing. Selecting the optimal prediction model and calculating error metrics are two aspects of model validation. Throughout the model calibration process, training data is used to optimize machine learning models. The best machine learning models are validated with testing datasets.

Groundwater sampling and analysis: 200 locations were sampled in April–June 2023 from a variety of sources (dug and bore wells), including the whole research area. Afterwards, samples were gathered, and several physicochemical parameters related to water quality were examined. The following were listed: pH, total dissolved solids (TDS), hardness, SO_4, NO_3, F, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO3, Cl, and PO4. Groundwater samples were collected at the field, delivered to the lab, and refrigerated at 40 degrees Celsius in order to be examined further. The samples were stored in a cold icebox with an EPS thermo box.

Groundwater quality index computation based on entropy:

To prevent subjectivity in the weighting approach, objective weights are calculated based on the entropy strategy. The entropy actually generates an information network based on the origin of a collection of weights, which may be used to assess the implicit links between parameters or within the dataset. The main method used to assign weights to the designated criteria is the community difference between the values of each alternate criterion and the other criteria. Greater unpredictability and less information are implied by higher entropy values in the data.

Using prediction models based on machine learning:

RF model: Reinforcement learning, has gained favor recently as a tool for water resource applications. It might be consistent with the concepts of classification trees, regression, and bagging with extra randomization. The average forecast of n-tree trained models is the bagger's prediction. By decreasing the correlation between the trees, randomization also contributes to a reduction in the variation of the predictions. During the randomization process, the responsible variables are selected at random to be candidates for data partitioning.

XGBoost model: Lately, a number of data mining applications have made extensive use of the XGBoost tool. It generates a lot of shallow decision trees, and the total of all the trees yields an extremely accurate prediction. The XGBoost technique handles the loss function to produce decision trees that regularize the tree to avoid over fitting and minimize an objective function. Moreover, the XGBoost model's adapta-

bility in modifying its hyper parameter adds to its extensive application in various study fields.

ANN model: An artificial neural network (ANN) is a type of parallel information system that simulates brain activity by using layers of neurons to replicate the structure and functionality of real neurons. This study uses multilayer feed-forward perception neural network with a nonlinear function.

DL model: Several connected neurons make up the input, hidden, and output layers. The inputs for the model's hidden layers come from the outputs. Finally, the output layer determines the outcome by using the last hidden layer's primary abstract data.

Nevertheless, a single hidden layer neural network is unable to adequately represent a wide range of measurable functions. However, having multiple hidden layers gives you more flexibility and makes it possible to use fewer neurons to approximate complex tasks. A deep learning model is better suited to modelling extremely challenging perceptions and providing estimates of highly nonlinear functions. Optimization of neural network architecture is usually a situation-specific process where the common approach to the desired result is trial and error.

Weight selection, learning rate, epoch termination, goal error setting, and network error derivative correction are some of the variables that affect model calibration, often known as neural network training. Several researchers have successfully trained a multilayer feed-forward network utilizing back-propagated (BP) techniques.

Model creation:

The suggested DL model for the current investigation is organized by ReLU activation, regularization, and strategy. The inquiry makes use of R version 4.0.2 and installs the packages keras, randomForest, xgboost, caret, and neuralnet. Applied machine learning models may be impacted by varying input variable ranges. Stated otherwise, the model's computation may be skewed by differing input variable ranges preferring those with a higher range, even when short range input variables have a greater influence on the target variable's prediction. Therefore, data normalization must be completed before developing these prediction models.

The minimum and maximum normalization processes are used to normalize the datasets used in ANN, DL, RF, and XGBoost models so that they range from zero to one. By reducing computing time and error during model execution, this enhances model performance. The normalization of input data is established by

$$Xn = (X \_ Xmin)/ (Xmax \_ Xmin)$$

where Xn is the normalized value and Xmin and Xmax are the minimum and maximum values, respectively. Training and validation datasets are separated from the normalized input dataset. The validation dataset's usual value should fall between 10% and 40% of the whole dataset. However, after some trial and error and a modification of the splitting ratios from 75:25 to 85:15, respectively, the data partitioning for the relevant models is finally established. The training dataset is used to calculate the model's weights.

At this stage, the model may exhibit overfitting, a common defect that impedes training and generates noise. Since the hyper-parameters cannot obtain information directly from the training dataset, they must become more complex in order to achieve the ideal model design. The hyper parameters of each model have therefore been changed during the calibration process.

Validation of the model:
During training and validation, the model's accuracy is evaluated using the following metrics: mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE)

3 Results And Discussion

Table 1 displays the statistical findings for the chosen physicochemical characteristics. The average TDS, Ca^{2+}, Mg^{2+}, NO3 , and PO4, in the groundwater in the study area are all greater than the BIS (2012) recommended level for drinking reasons. With the exception of TDS, Table 1 demonstrates that there is a slight variation between the mean and median values of every parameter. The pH of is generally alkaline, ranging from 6.90 to 8.90. The current research region's groundwater has increased amounts of TDS, carbonates, magnesium, by-carbonates, nitrates, phosphates.

Table 1. Statistical Results

Parameter	Median	Mean	Max.	Skewness	BIS Standards	No. of samples exceeding standard	Entropy weights
pH	7.80	7.82	8.90	1.29	8.5	05	0.109
TDS	619.9	695.74	2561.00	2.01	500	169	0.047
TH	137.0	145.18	515.00	1.66	500	01	0.058
Ca2+	86.11	90.87	298.00	2.00	75	159	0.056
Mg2+	31.82	33.77	121.50	1.65	30	121	0.058
Na+	50.17	57.95	248.40	1.64	200	01	0.056
K+	5.14	7.67	47.30	3.23	12	23	0.099
$HCO3^-$	307.80	298.31	405.90	—0.77	300	126	0.082
Cl^-	113.89	147.37	918.00	2.50	250	29	0.068

SO2-	29.86	33.97	136.00	1.59	200	00	0.051
NO^-	39.17	48.4	247.00	1.94	45	99	0.076
F^-	0.47	0.48	2.03	2.54	1.0	06	0.080
PO3-	0.21	0.29	1.17	1.40	0.1	137	0.160
EWQI	83.36	92.17	231.91	1.21	-	-	-

Spatial Distribution of EWQI:

Table 1 displays the EWQI value range for the research region, which is 14.33 to 231.91. The skewness, mean, and median values are 1.21, 83.36, and 92.17, in that order. PO4 and TDS produced the largest and lowest entropy weights respectively, indicating their highest and lowest importance. The EWQIs are then used for interpolation in ArcGIS 10.3 utilizing the inverse distance weighted (IDW) and conventional kriging approaches. Based on the estimated quality indices of the groundwater samples, five subcategories have been identified: 51-100 (Good), 101-150 (Moderate), 151-200 (Poor), and >200 (Very Poor). Based on the results, groundwater categories of "Excellent," "Good," "Moderate," "Poor," and "Very poor" have been assigned to twenty-five, fourteen, forty-two, five, and five samples. Figure 3 shows that while most of the investigated area (about 87.10% of the total area) is classified as being of good to moderate quality, the majority of the center, southeast, and northwest sections are classified as having moderate to severely low quality.

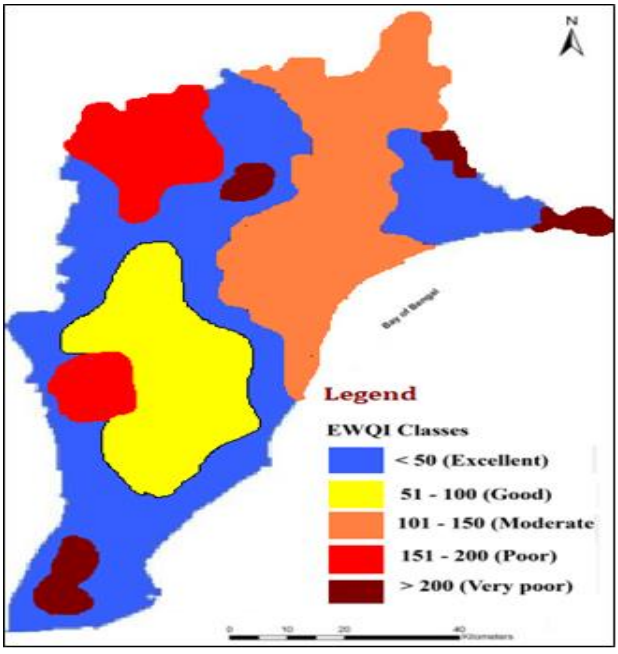

Fig.2. Spatial distribution of groundwater quality

The presence of shallow groundwater levels or the region's ongoing intensive agricultural activities could be the cause of the relatively low groundwater quality in certain places. The efficiency of models for machine learning :All the models are run in advance before the selected machine learning methods are decided upon in order to ascertain their optimal architectures. This study uses the available dataset to test split ratios between 75:25 and 85:15. Using the trial-and-error procedure, the lowest RMSE value is used to determine the final dataset partitioning. According to the findings, testing (calibration) and training (validation) account for 18% (41 numbers) and 82% (185 numbers) of the whole dataset, respectively. Plotting the expected data against the test data was shown in Figure3.

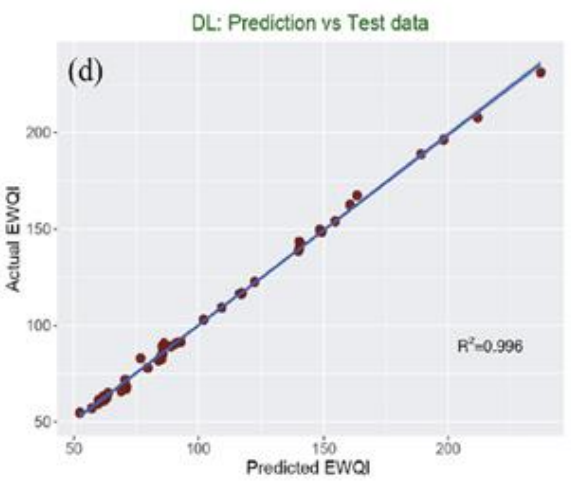

Fig.3. Plot comparing test and prediction data

Salt leaching via fertile soils and home sewage seeping into groundwater are the causes of the high TDS concentration in the study area's groundwater. The primary cause of the elevated levels of Ca2Í and Mg2á is geogenic sources, specifically carbonate-derived materials including calcite, dolomite, and aragonite. Dissolution of calcite and dolomite are responsible for these elevated levels. Furthermore, the ion exchange mechanism-may be the source of the Ca2+ and Mg2+ enrichment. Farmers' extensive use of chemical fertilizers based on nitrogen, phosphate, and potassium (N-P-K), such as urea, is the primary source of the high levels of NO3 and PO4 in the groundwater in the studied region.

4 Conclusion

This research established a deep learning approach to anticipate groundwater quality, or EWQI, and verified all ML models (RF, XGBoost, and ANN). Goodness of fit measures, also known as performance metrics, was used to compare each model's

prediction capabilities. However, in terms of prediction performance, the XGBoost model performed better than the RF and ANN models. The models function as follows, according to the R2 values of 0.996, 0.927, 0.917, and 0.886 that were acquired during the validation step: DL > ANN > XGBoost > RF. Therefore, it is possible to assess groundwater quality in the current research region using the recommended DL model. This study's prediction models are restricted to utilizing information from a single monsoon dataset. Using data from numerous seasons, more information about the present study's groundwater quality may be revealed. These kinds of reasoning are most likely responsible for the high level of ML model adoption in water resources studies. The current work may be improved by comparing the DL model's forecast capacity to that of other machine learning models and taking into account a wide range of different hydro-geo-meteorological inputs.

5 References

1. Abbasnia, A., Radfard, M., Yousefi, M., Alimohammadi, M., Mahvi, A.H., Nabizadeh, R., and Yousefi, N., 2019. An analysis of Sistan and Baluchistan province in Iran's groundwater quality using the water quality index and its applicability for determining water quality for irrigation and drinking purposes is presented in this case study. Risk Assess. Hum. Ecol. 25 (4), 988-1005.
2. García-Nieto, J., Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., and Irfan, R., 2019.effective machine learning supervision for the prediction of water quality. 11 (11), 2210. Water.
3. Birikundavyi, S., Rousselle, J., Trung, H.T., and Labib, R., 2002. Neural network performance in anticipating daily stream flow. Journal of Hydrol. Eng. 7 (5), 392–398.
4. Indian Standard Drinking Water Specification, Second Revision, Bureau of Indian Standards, ISO: 10500:2012. Sectional Committee on Drinking Water, New Delhi, India, FAD 25.
5. Fagbote, E.O., Olanipekun, E.O., Uyi, H.S., 2014. Water quality index of the ground water of bitumen deposit impacted farm settlements using entropy weighted method. Int. J. Environ. Sci. Technol. 11, 127-138.
6. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: a review. Neurocomputing 187, 27-48.
7. Gupta, R., Singh, A.N., Singhal, A., 2019. Application of ANN for water quality index.International Journal of Machine Learning and Computing 9 (5), 688-693.
8. Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks areuniversal approximators. Neural Network. 2 (5), 359-366.
9. Kazakis, N., Mattas, C., Pavlou, A., Patrikaki, O., Voudouris, K., 2017. Multivariate statistical analysis for the assessment of groundwater quality under different hydrogeological regimes. Environmental Earth Sciences 76 (9), 349.
10. K. Sundara Kumar, P. Sundara Kumar, Dr. M. J. Ratnakanth Babu & Dr. Ch. Hanumantha Rao, Assessment and Mapping Of Ground Water Quality Using Geographical Information Systems, International Journal of Engineering Science and Technology Vol. 2(11), 2010, 6035-6046.
11. K. Sundarakumar, P. Udayabhaskar, K. Padma Kumari, and Ch. Kannam Naidu., 2010. Ground water quality assessment of Srikakulam district of Andhra pradesh, India, using GIS. Int. J. of Applied Env. Sciences, 5(4), 495–504.

12. K. Sundara Kumar, Gurjeet Singh, Dr. G.V. Rao, and S. Chandra Mouli , 2011. Spatial distribution and multiple linear regressions modeling of ground water quality with geostatistics, , International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 6, Number 24 , pp. 2719-2730.