International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Review of Machine Learning and Artificial Neural Network Methods for Water Quality Prediction

Baliwada Madhavi[*]

Department of Computer Science and Engineering, GMR Institute of Technology, Rajam-532127, Andhra Pradesh, India

ABSTRACT :

The term "water quality" refers to the condition of water, including its chemical, physical, and biological characteristics. Domestic and industrial pollutions affected the water quality to a greater extent. Polluted water became a major reason behind several community diseases, mainly in undeveloped and developing countries. The public health condition is deteriorating and putting an extra burden of countermeasures to prevent such water borne diseases from spreading. The modelling of the water quality parameters is very essential in any analysis of any aquatic systems. The artificial neural network is a new technique with a flexible mathematical structure that can identify complex non-linear relationships between input and output data when compared to other classical modeling techniques. The ANN models are capable of dealing with different modeling problems in rivers, lakes, reservoirs, wastewater treatment plants (WWTPs), groundwater, ponds, and streams.

Keywords: Artificial Neural Networks (ANNs), Machine learning, Modelling Water Quality Parameters, Water Quality Prediction, Non-Linear Relationships.

Introduction :

Water plays a very huge role in maintaining human health and supporting agriculture besides those related to fishing and farming industries. Despite time, this world continues getting more industrialized and urbanized with its ever-widening city. Pollution from waters from rivers and lakes puts direct risks on the ecological set up but more significant threats come for human beings. Traditional methods of water quality monitoring, such as sampling for laboratory analysis, are cumbersome, expensive, and almost impossible to apply on a large scale. Recently, machine learning has been used as a more efficient technique for predicting water quality. Of all the techniques within this field, ANNs are one of the most effective. An advanced model, ANNs can analyze intricate patterns in data, making them useful for prediction in various water quality parameters like dissolved oxygen, which is fundamental to aquatic life. Unlike conventional approaches, ANNs present faster predictions and can even be applied when it becomes difficult to access water sources physically. This paper focuses on artificial neural networks to be applied in improving the quality of water prediction. The possibility with such models will offer timely and accurate assessments about conditions of water, allowin**g b**etter management of levels of pollution and ensuring safe and **s**ustainable use of water resources for the future

Literature Survey :

Wu & Wang (2022) propose a hybrid ANN model for predicting water quality and consumption, with an accuracy of 96% and R² of 0.997. It outperforms other ML models in terms of dual functionality and simplicity. Challenges include data imbalance and overfitting. Future work is suggested to be enhanced datasets, regularization, and real-time testing.

Chen et al. (2020) reviewed ANN models for predicting water quality parameters like dissolved oxygen and biochemical oxygen demand. Using FNN, RNN, RBFNN, and hybrid models, the CNN-LSTM achieved 95.1% accuracy. While offering high accuracy and non-linearity handling, challenges include data dependency and complexity. Improvements suggest cloud computing, cross-validation, and enhanced interpretability.

Shams et al. (2024) utilized ML models like RF, XGBoost, and MLP for water quality prediction, achieving an R² of 99.8% with MLP. Advantages include adaptability and non-linear handling, while challenges are interpretability and computational demands. Improvements suggest enhanced feature engineering, model diversity, cross-validation, and better handling of missing data Hussein et al. (2020) used ML models including SVR, RF, and MLP for the groundwater availability prediction with SVR having a MAE of 2.43 and RMSE of 5.413. Strengths include accuracy and scalability, but data imputation and computational costs are issues. Future enhancements include data augmentation, h ybrid models, and improved imputation techniques

Liu et al. (2024) applied the ML models ANN, RF, and SVM to handle water resource issues, resulting in up to 99% accuracy in flood prediction. The advantages are high accuracy and flexibility, whereas the challenges include data dependency and interpretability. Future improvements might include model integration, enhancing interpretability, and proper handling of non-stationary data.

Wu J (2024) reviewed ML and DL applications in water resource forecasting, highlighting RF's 99% accuracy in water-level prediction and ANN's 71–88% accuracy for water demand. Benefits include high predictive power and scalability, while challenges involve data dependency and complexity. Future work suggests hybrid models, data integration, and expanded sources Hussein et al. (2024) analyzed groundwater quality and predicted irrigation water quality indices using ML models like XGBoost, SVR, and KNN. SVR achieved the best performance (RMSE: 2.693, NSE: 0.961, R^2: 0.975). Advantages include predictive accuracy and statistical rigor. Improvements suggestdata augmentation, model optimization, and longitudinal studies.

Saleh & Rasel (2024) tested the ML models for rainfall forecasting, wherein Random Forest has achieved a highest correlation coefficient of 0.97, outperforming GBM with 0.41 and SVM with 0.20. Advantages are robustness with high accuracy; however, the challenges include data constraints and poor performance of SVM. Future enhancement would require data expansion, model optimization, and hybrid models.

Khan et al. (2023) evaluated streamflow forecasting models for the Hunza River Basin, where Adaptive Boosting showed the highest accuracy with R^2: 0.998. The benefits are high predictive accuracy and robustness, but the drawbacks include data dependency and limitations of KNN. Future directions include data augmentation, hybrid approaches, and model optimization.

Najwa Mohd Rizal et al. (2022) compares regression models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) for predicting river water quality. ANN achieved 85% accuracy. While offering adaptability and simplicity, challenges include model complexity and overfitting. Future improvements include expanding dataset diversity and using ensemble methods.

Yao et al. (2023) examined the effect of land use changes on water quality in urbanizing areas using Cellular Automata-Markov and Multiple Linear Regression models. The TN prediction presented a coefficient of determination of 0.691 and MRE of 12.14%. Future improvements include using more variables and advanced models.

Al-Adhaileh and Alsaade (2021) used KNN, FFNN, and ANFIS to classify water quality, aiming to meet drinking water standards. The ANFIS model achieved 96.17% accuracy in predicting WQI. While cost-effective and robust, it faces challenges like data dependency and complexity. Future improvements include data diversity and model optimization Othman et al. (2020) developed an ANN model that predicts the WQI using the least input variables, achieving a correlation of 98.78%. The model is efficient, even in the presence of missing data; however, it depends on data quality and requires expertise. Hybrid models and the extension of data willimprove it in the future for higher accuracy.

Methodology :

A hybrid model combining ANN, wavelet transform, and LSTM was developed for water quality prediction. Another study utilized machine learning models, including Random Forest and XGBoost, with grid search for hyperparameter tuning to predict water quality. Both approaches focus on data preprocessing and model evaluation for accurate predictions

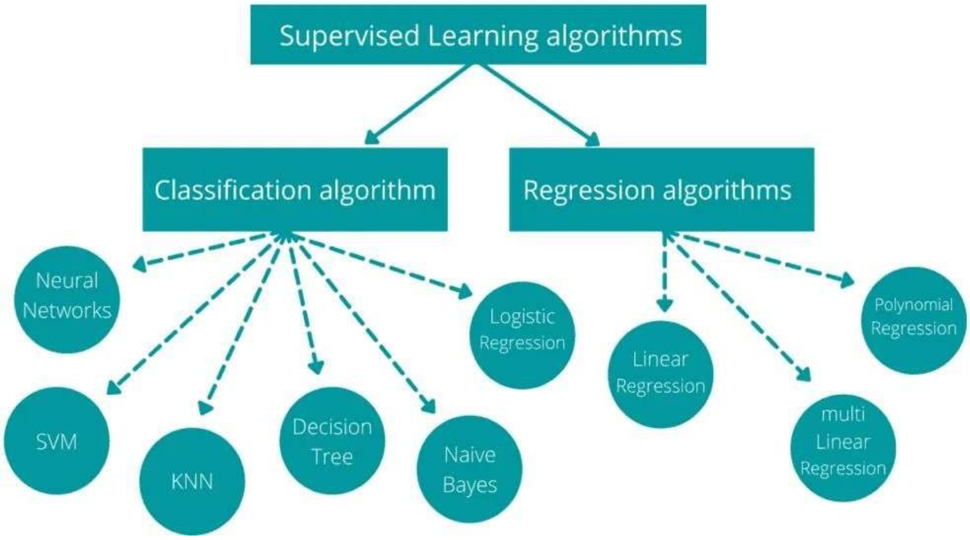
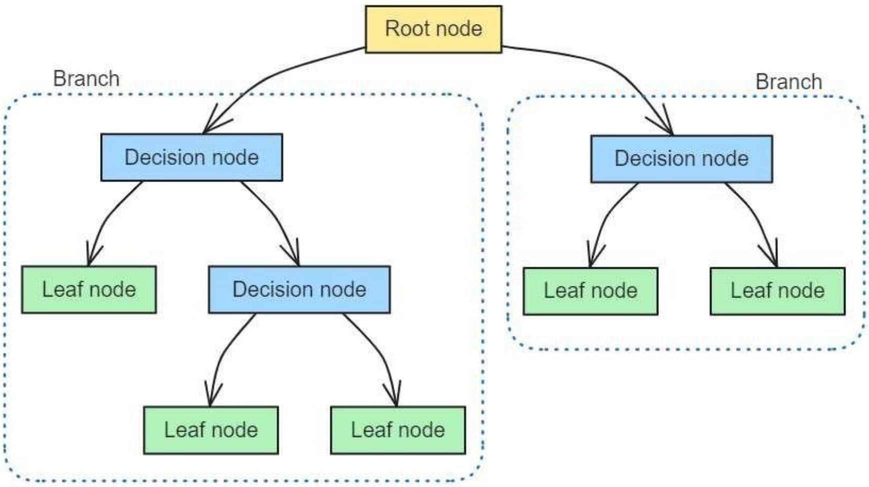


Fig 3.1: Model classification

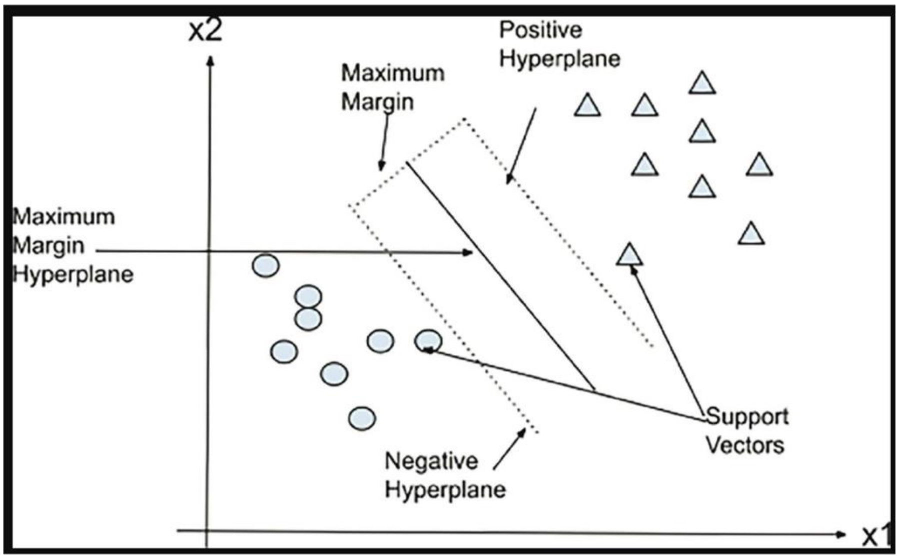
Machine learning models:

Decision Tree (DT): The Decision Tree (DT) is one of the supervised machine learning models used for classification or regression tasks. It does this by recursively splitting a dataset into subsets based on feature values, creating in the process a tree-like structure with nodes representing a decision or test on some features and branches representing their outcome. Each leaf represents a class label for a classification problem or a predicted value for a regression problem. DT is intuitive to interpret and deals with numeric as well as categorical variables. It does suffer from overfitting, but this problem has been shownto be much reduced through techniques like pruning or ensemble methods

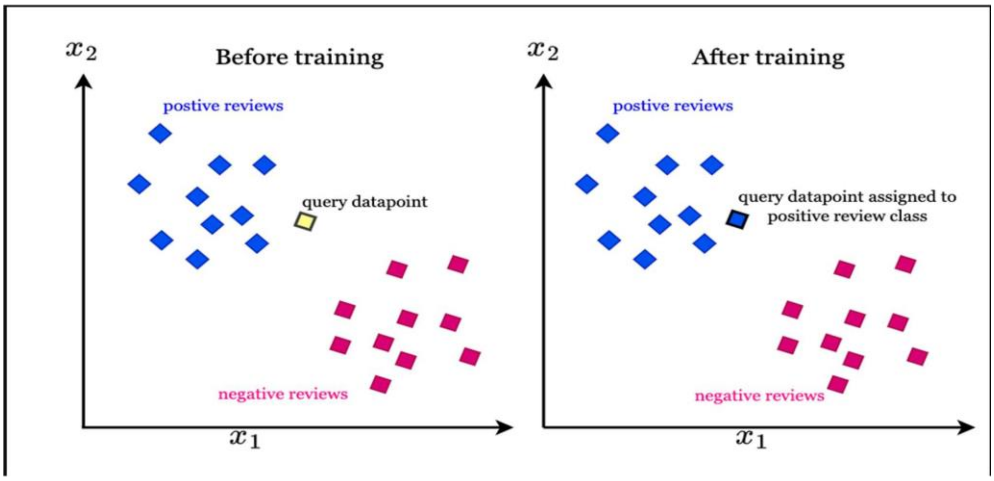


Decision Tree

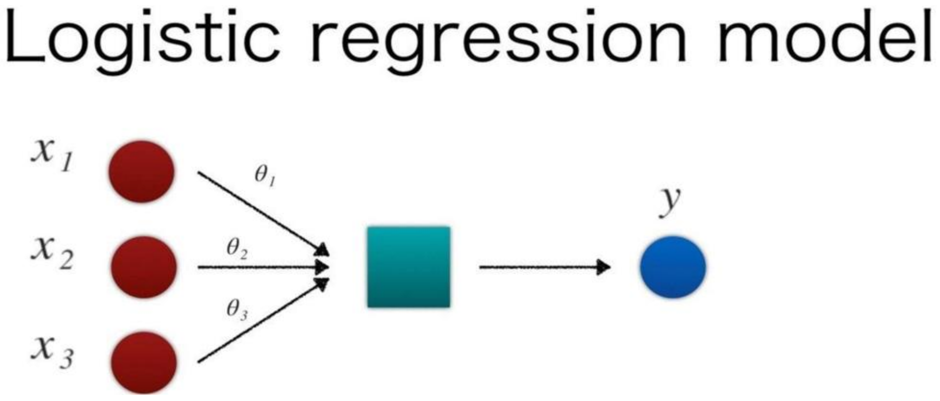
Support Vector Machine is a supervised machine learning model applied to classification and regression problems. It determines the optimal hyperplane that classifies different classes in the feature space with the maximum margin between the closest points (support vectors) of each class. SVM is used for both linear and nonlinear relationships using kernel functions like polynomial, radial basis function to transform the data into higher dimensions. It isefficient for high-dimensional spaces and also very robust against overfitting, especially when the margin is clearly separable. However, SVM can be computationally expensive and sensitive to the choice of kernel and hyperparameters.



KNN is one of the simplest, instance-based, supervised machine learning algorithms used in classification and regression tasks. It works on the 'k' nearest data points (neighbors) to the test point in the feature space, considering distance metrics like Euclidean or Manhattan distance. Classification assigns the test point with the majority class of the neighbors, whereas in regression, the prediction is taken as an average of the values of the neighbors. KNN is intuitive, non-parametric, and supports multi-class problems. Still, it is computationally expensive, especially when dealing with large datasets, and depends on the choice of 'k' and distance metric.

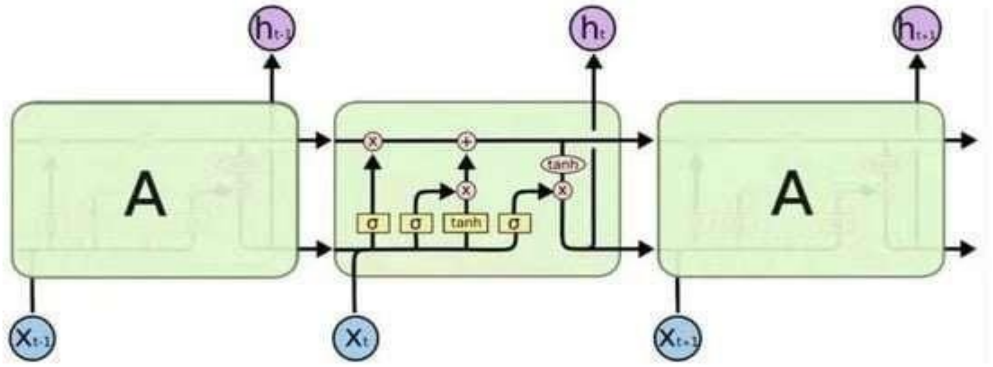


Logistic Regression is a type of supervised machine learning algorithm used for binary classification problems. It predicts the probability of a binary outcome by using a logistic (sigmoid) function to a linear combination of the input features. The output from the logistic function is between 0 and 1, which indicates the probability of the positive class. The model estimates the parameters by minimizing the log-loss function, which compares the predicted probabilities with the actual outcomes. It's simple, interpretable, and efficient for small- and medium-sized datasets but is assuming a linear relationship between the features and the log odds of the target. It fails if data is non-linear.

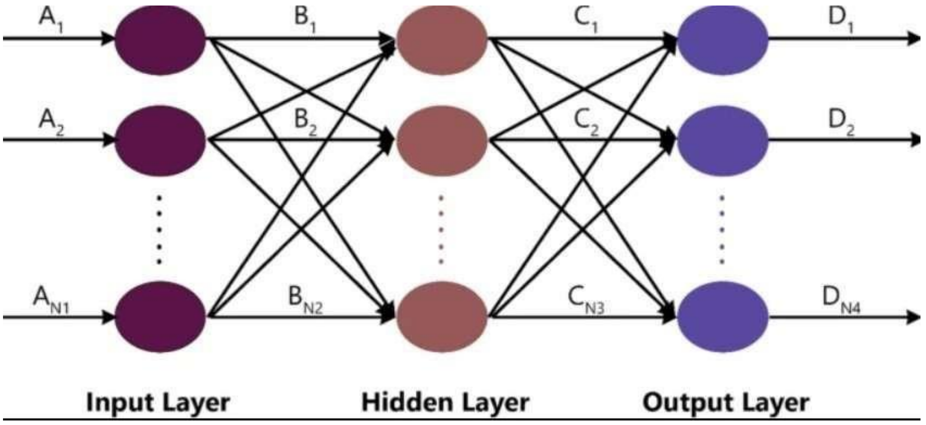


Deep Learning models:

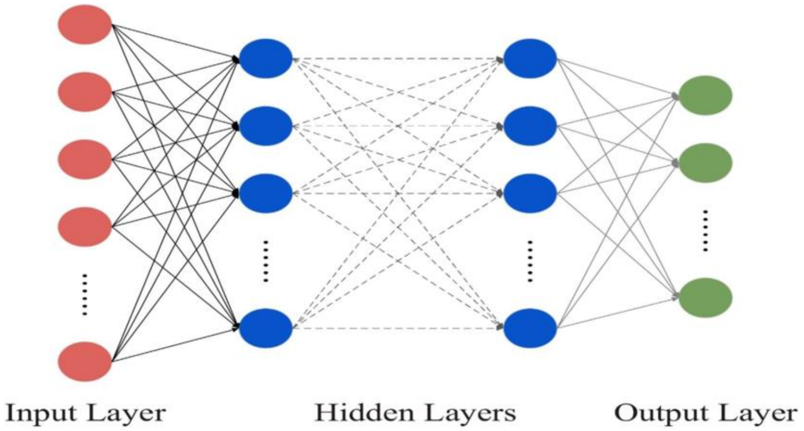
Long and Short-Term Memory (LSTM): It is an extension of RNN where a memory unit is introduced to remember information for an extended length of time. Each memory cell has three components: three different types of gates: input, forget, and output. Information flows into and out of the memory cell through these gates. The input gate controls what comes into the memory cell. The output gate controls what goes out of the memory cell. The forget gate controls which information might be removed from the memory cell. Long sequences of data are stored, updated, and retrieved by LSTM with the help of these gates.



Multilayer Perceptron (MLP) is a type of feedforward artificial neural network which is used for supervised learning, including classification and regression. It consists of one or more layers of neurons; input layer, one or more hidden layers, and an output layer. Every neuron in a layer is connected to neurons in the next layer, with weights that are adjusted during training. MLP applies activation functions, such as ReLU or sigmoid, to introduce non- linearity and allow the model to learn complex patterns. The model is trained using backpropagation and optimization algorithms such as gradient descent. MLP is powerful for complex tasks but can suffer from overfitting and computational cost.



Artificial Neural Networks, or ANN, are a class of machine learning models inspired by the human brain. They are used in classification, regression, and pattern recognition tasks. An ANN consists of layers of interconnected neurons: an input layer, one or more hidden layers, and an output layer. Each connection has an associated weight that adjusts during training using backpropagation and optimization techniques like gradient descent. Activation functions such as ReLU or sigmoid introduce nonlinearity, allowing the network to learn complex patterns. Such networks are powerful for anything like image recognition or natural language processing but require large data sets and a lot of computational resources to train suitably



Results :

For water quality prediction, each model's accuracy score on the water quality dataset is displayed, providing insights into which methods perform best. Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) each produce an accuracy score based on 5-fold cross- validation, while the Artificial Neural Network (ANN) is trained and tested directly on the test set after training on the training set. Among the traditional machine learning models, the Decision Tree and SVM models generally perform well, with slight variations depending on the dataset's structure and complexity. However, these models are often limited by simpler decision boundaries, which may hinder their ability to generalize on more complex data. The

ANN model, in contrast, achieves the highest accuracy among all models due to its ability to learn non-linear relationships and complex patterns present in the data. By including multiple hidden layers with ReLU activation functions, the ANN captures subtle interactions between features like pH, turbidity, and temperature, which are crucial for determining water potability. This performance is further boosted by the use of categorical cross-entropy loss and the Adam optimizer, which enables efficient training on multiclass classification. As a result, the ANN is especially well-suited for this dataset, where predicting water quality requires sophisticated, multi-layer processing.

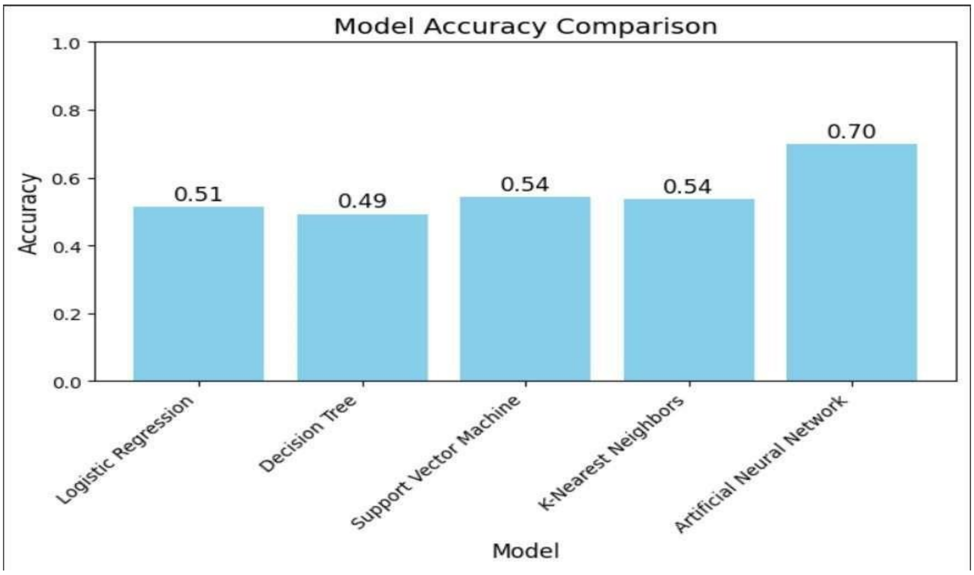


Fig 4: ANN gives the best accuracy

This reveals that the ANN model outperforms traditional machine learning algorithms in predicting water quality levels. This finding is consistent with the ANN's strength in handling complex, non-linear datasets like environmental data. Therefore, deep learning is likely the best approach for this type of classification problem. However, tuning parameters for each model and using additional feature engineering or data augmentation techniques could improve performance across all models. In practice, this analysis highlights the importance of testing various algorithms for any classification task, as the optimal model depends on both the dataset's characteristics and the specific application.

Conclusion :

In conclusion, this review focuses on the strength of machine learning and deep learning, particularly ANN, in predicting water quality. Comparative analysis of models such as Logistic Regression, Decision Trees, Support Vector Machines, and K-Nearest Neighbors reveals that traditional machine learning techniques offer reliable predictions for water quality classification. However, ANN goes beyond those models in accuracy and robustness because of its multilayer architecture, allowing it to capture complex, non-linear relationships in the data. This is why ANN is a better predictive tool forwater quality compared to traditional machine learning models.

REFERENCES :

1. Asefa, T., & Kumar, S. (2022). Machine learning methods for river flow forecasting: A review. Water Resources Research, 58(2), e2021WR031263.DOI: 10.1029/2021WR031263.
2. Bhat, S., & Mishra, V. (2021). Forecasting groundwater levels using deep learning algorithms. Journal of Hydrology, 603, 127095. DOI: 10.1016/j.jhydrol.2021.127095.
3. . Chen, J., & Li, Y. (2020). Application of support vector machines in water quality prediction: A review. Environmental Modelling & Software, 125, 104639. DOI: 10.1016/j.envsoft.2019.104639.
4. Cohen, S., & Elimelech, M. (2019). Machine learning for water resource management: Challenges and opportunities. Journal of Water Resources Planning and Management, 145(9), 04019047. DOI: 10.1061/(ASCE)WR.1943-5452.0001101.
5. Deng, Z., & Zhuang, Q. (2021). A comparative study of machine learning algorithms for streamflow prediction. Hydrological Sciences Journal, 66(11), 1725-1742. DOI: 10.1080/02626667.2021.1902158
6. Gao, H., & Li, X. (2022). Ensemble machine learning models for reservoir inflow prediction. Water, 14(8), 1275. DOI: 10.3390/w14081275
7. Goswami, S., & Sharma, S. (2020). Predicting groundwater level fluctuations using machine learning: An empirical study. Water Resources Management, 34(10), 3317-3332. DOI: 10.1007/s11269-020-02623-w
8. Khan, S., & Wu, S. (2021). Leveraging deep learning for water demand forecasting: An urban case study. Computers, Environment and Urban Systems, 87, 101645. DOI: 10.1016/j.compenvurbsys.2021.101645.5.01645
9. Kumar, R., & Raj, P. (2023). Hybrid machine learning models for improved water quality prediction. Environmental Science & Technology, 57(3), 1542-1553. DOI: 10.1021/acs.est.2c06163
10. Liu, L., & Zhao, Y. (2020). Application of artificial neural networks in water quality modeling: A review. Water, 12(11), 3037. DOI: 10.3390/w12113037
11. Mao, X., & Zhang, Y. (2021). Predicting reservoir storage using machine learning algorithms: A case study. Journal of Hydrology, 597, 126081. DOI: 10.1016/j.jhydrol.2021.126081
12. Mishra, V., & Singh, S. (2022). Performance of machine learning models in predicting river discharge: A systematic review. Hydrology and Earth System Sciences, 26(7), 2323-2345. DOI: 10.5194/hess-26-2323-2022
13. Ochoa, G., & Bender, T. (2021). Machine learning approaches groundwater level forecasting. Water Resources Management, 35(8), 2745-2760. DOI: 10.1007/s11269-021-02823-5
14. Park, J., & Kim, S. (2020). Comparison of machine learning techniques for water consumption prediction. Journal of Water Resources Planning and Management, 146(7), 04020045. DOI: 10.1061/(ASCE)WR.1943-5452.0001140
15. Shao, X., & Zhang, X. (2022). Application of random forest for predicting water quality in lakes: A case study. Environmental Monitoring and Assessment, 194(1), 1-16. DOI: 10.1007/s10661-021-09170-2.