Performance Comparison of Machine Learning Approaches for Groundwater Quality Assessment

MINOR PROJECT

Bachelor of Technology
Civil Engineering

by: Jasika Kumari
Roll no.:BTECH/10892/22

Guide: Dr Jawed Iqbal
Department of Civil and Environmental Engineering
Birla Institute of Technology, Mesra (Jharkhand)

INTRODUCTION

1) Freshwater resources are rapidly degrading due to industrial discharge, urbanization, and seawater intrusion, posing a threat to global water availability.

2) Traditional Water Quality Index (WQI) assessment methods are time-consuming, data-intensive, and limited by local environmental variations.

3) Machine Learning models offer faster, more accurate, and scalable solutions for predicting groundwater quality compared to conventional techniques.

4) This study aims to evaluate groundwater properties, compute WQI, deploy ML models for prediction, and assess their performance using standard error metrics.

What is WQI?

Water Quality Index (WQI) is a numerical indicator that provides a single composite score representing the overall quality of water by combining multiple water quality parameters. The Brown et al. method (also called the Weighted Arithmetic Index Method) is one of the earliest and most widely accepted methods for calculating WQI. Lesser the value of WQI better will be the quality of water

STEP1 : UNIT WEIGHT

$$W_i = K/S_i$$

- W_i = unit weight
- S_i = standard permissible value (e.g., BIS, WHO)
- k = proportionality constant ensuring sum of weights = 1

$$k = 1/\Sigma(1/S_i)$$

STEP2 : QUALITY RATING

$$Q_i = (S_i - V_0)/(V_i - V_0) \times 100$$

- V_i = observed concentration
- V_0 = ideal value (usually 0 for most parameters; for pH, ideal = 7; for DO, ideal = 14.6 mg/L)
- S_i = standard permissible value

STEP3 : COMPUTE WQI

$$WQI = \Sigma(W_i \times Q_i)/\Sigma W_i$$

Since weights are normalized, $\Sigma W_i = 1$, so:

$$WQI = \Sigma(W_i \times Q_i)$$

PHYSICAL
Temperature
Turbidity
Total Dissolved Solids (TDS)
Electrical Conductivity (EC)

Biological
BOD
Coliform bacteria

CHEMICAL
pH
Dissolved Oxygen (DO)
Hardness
Alkalinity
Nitrates
Chlorides
Fluoride
Sulphates
Total Hardness (TH)

Machine Learning

Machine learning is a process through which a computer learns from a set of datasets with the help of algorithms and then makes predictions without being programmed again.

Supervised Machine Learning is a type of machine learning where the model is trained using labeled data—meaning each input has a known, correct output.The algorithm learns the relationship between inputs (features) and outputs (target) to make predictions on new, unseen data.

Unsupervised Machine Learning is a type of machine learning where the model is trained using unlabeled data, meaning there is no predefined output.The algorithm tries to find patterns, structure, or relationships within the data on its own.

Regression comes under supervised machine learning technique where a labeled data with output is used for training the model so that the machine could identify various patterns and relations in the dataset and predict the result for all the values continuously.

REGRESSION MODELS

1. Random Forest

Random Forest is an ensemble supervised learning algorithm that builds multiple decision trees on random subsets of data and features, then combines their predictions to improve accuracy and reduce overfitting. It effectively handles nonlinear relationships, high-dimensional data, and noisy or incomplete environmental datasets such as groundwater quality parameters.

2. Support Vector Regression (SVR)

Support Vector Regression extends Support Vector Machines to regression by fitting a function within an ε-insensitive margin, ensuring a balanced model that avoids overfitting. Using kernel functions like RBF or polynomial, SVR captures both linear and nonlinear relationships in groundwater quality datasets even with limited samples.

3. Gradient Boosting Regressor (GBR)

Gradient Boosting Regressor builds an ensemble of weak learners sequentially, where each new tree corrects the residual errors of the previous ones using gradient descent optimization. It captures intricate nonlinear dependencies in groundwater data and provides high predictive accuracy with strong control over overfitting.

4. Linear Regression

Linear Regression models the relationship between input variables and a continuous output by fitting the best-fit straight line or hyperplane using the least squares method.

Linear SVR enhances this by applying margin maximization and regularization, making it suitable for moderately linear groundwater quality patterns with good interpretability.

5.LightGBM Regressor

LightGBM is a fast, efficient gradient boosting algorithm that grows trees leaf-wise and uses techniques like histogram-based learning and GOSS to optimize speed and accuracy. It excels in modeling complex nonlinear groundwater quality interactions, especially in large or high-dimensional datasets.

What is Comparative Analysis

Comparative Analysis is the process through which different ML algorithms are examined in terms of their performance they have for that specified set of data provided.
Certain metrics such as R^2 score, MAE error, RMSE error, Cross-Validation, Test Set Evaluation R and Qualitative Remarks are used for comparing these models.

MAE (Mean Absolute Error)

MAE represents the average absolute difference between predicted and actual values, showing how far predictions deviate on average.

Test Set Evaluation (Test R)

Test Set Evaluation measures how well the model generalizes to unseen data using the R^2 score on the test split.

R^2 Score (Coefficient of Determination)

R^2 measures how much variance in the target variable is explained by the model, indicating the overall goodness of fit.

RMSE (Root Mean Squared Error)

RMSE is a standard accuracy metric in regression that calculates the square root of the average squared prediction error.

LITERATURE REVIEW

Groundwater Quality Concerns Rapid industrialization, agricultural runoff, and urban expansion have led to severe groundwater contamination.Traditional hydrochemical testing and WQI calculations are accurate but time-consuming, expensive, and limited to small areas.Increasing complexity of groundwater pollution (heavy metals, nitrates, microbial load) requires smarter, data-driven tools.

Importance of Water Quality Index (WQI) WQI is widely used to evaluate drinking water suitability by combining multiple parameters (pH, EC, TDS, nitrates, hardness etc.).The conventional Weighted Arithmetic Index method (Brown et al., 1970) is simple but cannot capture nonlinear interactions among variables.Researchers highlight a need for automated, scalable, and accurate prediction approaches for WQI estimation.

Emergence of Machine Learning in Water Quality Studies ML models can process large, noisy, heterogeneous datasets and uncover complex patterns that traditional statistics fail to detect. Increasingly used in environmental studies to classify contamination levels, map pollution hotspots, and predict WQI in real time.

LITERATURE REVIEW

ML Models Used in Past Research Classical models: Linear Regression, Support Vector Regression (SVR).Ensemble models: Random Forest (RF), Gradient Boosting (GBR), XGBoost, LightGBM
Deep models: ANN, MLP, ANFIS (in some studies)

Key Research Findings Ensemble-based models consistently deliver the highest accuracy due to their ability to capture nonlinear, multi-parameter interactions.Krishnamoorthy & Lakshmanan (2024): Random Forest achieved 97% accuracy for classification and >0.90 R^2 for WQI prediction.
Ibrahim et al. (2023): Integrating GIS, IWQI and ML enables spatial mapping of salinity, alkalinity, and contaminant hotspots. Studies show ML models outperform traditional methods in speed, scalability, and predictive reliability.

Research Gap & Relevance to Present Study Very few studies directly compare multiple ML algorithms on the same groundwater dataset.Limited research focusing on Indian groundwater conditions using multi-parameter datasets. This project addresses the gap by evaluating five ML regressors (LR, SVR, RF, GBR, LightGBM) for WQI prediction and identifying the most effective model.

METHODOLOGY

Data Loading & Preprocessing

- Imported dataset (Results_MADE.csv) containing 10 physicochemical parameters and WQI.
- Checked dataset structure, handled missing values using mean imputation.
- Applied log1p transformation to skewed features (BOD, FS, FC, TC, Conductivity) to reduce skewness.
- Renamed columns for consistency and improved readability.

Feature Preparation & Splitting

- Separated input features (X) and target variable (WQI).
- Performed train–test split (80/20) to evaluate generalization.
- Ensured reproducibility using a fixed random state (42).

Model Selection: e.g. Random Forest Regressor

- Chosen due to strong performance on nonlinear environmental datasets.
- Research-optimized hyperparameters:
- n_estimators = 200, max_depth = 10, min_samples_split = 3
- min_samples_leaf = 2, bootstrap enabled
- Parallel computation enabled with n_jobs = -1.

Model Training & Cross-Validation

- Trained Random Forest on the training dataset.
- Performed 10-fold cross-validation using KFold:
- Evaluated using R^2, MAE, RMSE.
- Ensured stability by shuffling and using random_state = 42.

METHODOLOGY

Model Evaluation

- Generated predictions on the test set.
- Computed evaluation metrics:
- R^2 Score, Mean Absolute Error (MAE),
- Root Mean Squared Error (RMSE).
- Compared CV performance with test performance for reliability.

Visualization & Interpretation

- Actual vs Predicted Plot to assess prediction alignment.
- Residual Distribution Plot to examine error spread and normality.
- Feature Importance Plot to identify key contributors (pH, EC, Nitrate, BOD).

Saving Results

- Exported model metrics to CSV format (results_random_forest_research.csv) for documentation.
- Ensures reproducibility and comparison with other ML models.

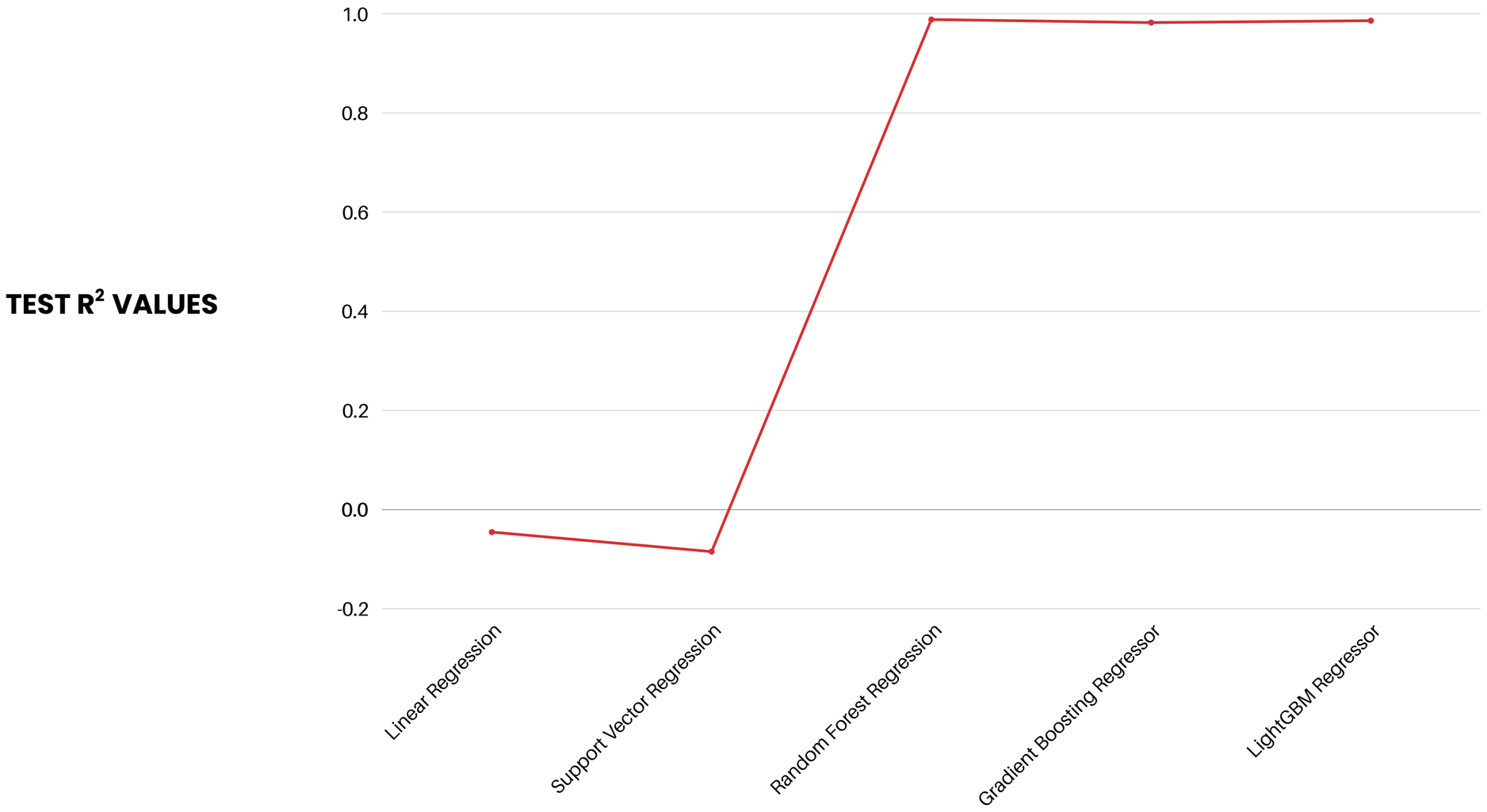
```
--- 10-Fold Cross-Validation ---
CV Mean R²: 0.9511 ± 0.1239
CV Mean MAE: 8.7089 ± 16.8835
CV Mean RMSE: 31.7103 ± 70.1074
```

```
============================
Random Forest Regression (Research-Optimized)
============================
Test R² Score: 0.9885
Test MAE: 5.5204
Test RMSE: 15.7573
```

RESULTS

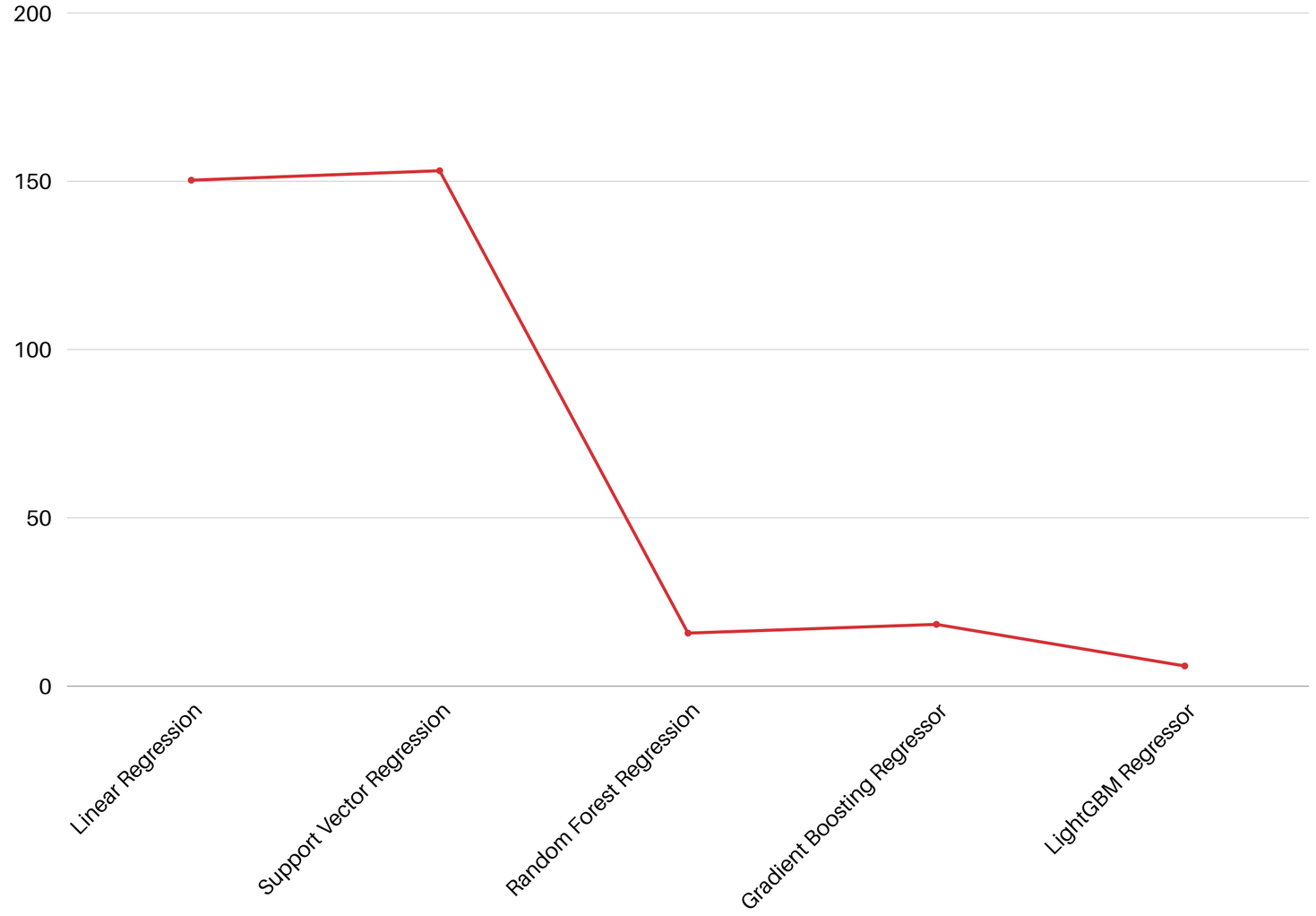
Algorithm	R^2 (Mean ± SD)	MAE (Mean ± SD)	RMSE (Mean ± SD)	Test R^2	Test MAE	Test RMSE
Linear Regression	0.0089 ± –	75.8671 ± –	150.3254 ± –	-0.0452	75.8671	150.3254
Support Vector Regression	0.0122 ± 0.1378	53.4255 ± –	130.6976 ± –	-0.0846	61.9205	153.1343
Random Forest Regression	0.9511 ± 0.1239	8.7089 ± 16.8835	31.7103 ± 70.1074	0.9885	5.5204	15.7573
Gradient Boosting Regressor	0.9123 ± 0.0912	10.5421 ± 14.2321	28.6709 ± 52.8702	0.9824	6.9342	18.3529
LightGBM Regressor	0.9348 ± 0.1015	9.3281 ± 15.1024	26.1458 ± 49.7611	0.9862	5.9927	5.9927

RESULTS


TEST R^2 VALUES
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
Linear Regression
Support Vector Regression
Random Forest Regression
Gradient Boosting Regressor
LightGBM Regressor
ML MODELS

RESULTS

TEST RMSE VALUES



ML MODELS

Conclusions

Ensemble models like Random Forest (Test R^2 = 0.9885) and LightGBM (Test R^2 = 0.9862) showed outstanding performance by effectively capturing the non-linear and high-dimensional relationships among groundwater parameters.Their ability to model complex interactions made them the most reliable predictors of Water Quality Index (WQI).

Linear Regression and SVR struggled due to the non-linearity present in groundwater datasets, producing poor R^2 scores and higher errors. This demonstrates that simple linear models are unsuitable for interpreting groundwater quality patterns involving chemical, biological, and physical interactions..

Models such as Gradient Boosting (Test R^2 = 0.9824) also performed strongly, confirming that boosting and bagging methods handle noise, missing values, and nonlinear variations more effectively. Their low MAE and RMSE values validate that ensemble approaches significantly outperform traditional regression methods in groundwater quality prediction.

References

Brown, R. M., McClelland, N. I., Deininger, R. A. (1970). A water quality index: Do we care? Water & sewage works, 117, 339-343.

L. Krishnamoorthy and V. R. Lakshmanan, “Groundwater quality assessment using machine learning models: a comprehensive study on the industrial corridor of a semi-arid region,” Environmental Science and Pollution Research, vol. 31, no. 28, pp. 83041–83060, Jul. 2024, doi: 10.1007/s11356-024-34119-7.

H. Ibrahim, Z. M. Yaseen, M. Scholz, M. Ali, M. Gad, S. Elsayed, M. Khadr, H. Hussein, H. H. Ibrahim, M. H. Eid, A. Kovács, S. Péter, and M. M. Khalifa, “Evaluation and Prediction of Groundwater Quality for Irrigation Using an Integrated Water Quality Indices, Machine Learning Models and GIS Approaches.

C. Tebbutt, “Sustainable Water Development: Opportunities and Constraints,” Water International, vol. 13, pp. 189, 1992.

Thank You.