SciEn

Proceedings of 8th International Conference on Mechanical Engineering and Renewable Energy 2025
10-12 December, 2025, Chattogram, BANGLADESH

SciEn Conference Series: Engineering Vol. 4, 2026, pp 1-6 https://doi.org/10.38032/scse.2026.4.9

Comparative Analysis of Five Machine Learning Regression Models for Salinity Prediction in the California Current System

Mohammed Joobayear Hossain[1], Abbas Ali Khan[2], Nahid Hasan Munna[3], Aminur Rahman[4] and Forhad Mahmud[5]

[1]Department of Petroleum and Mining Engineering, SUST, Sylhet-3114, Bangladesh
[2]Department of Applied Mathematics, NSTU, Noakhali-3814, Bangladesh
[3]Department of Applied Mathematics, NSTU, Noakhali-3814, Bangladesh
[4]Department of Petroleum and Mining Engineering, SUST, Sylhet-3114, Bangladesh
[5]Department of Applied Mathematics, NSTU, Noakhali-3814, Bangladesh

ABSTRACT

Salinity significantly influences ocean water movement, climate change, marine ecosystems and heat transfer, impacting processes such as ocean circulation patterns, density-driven currents, sea ice production, nutrient distribution, and the global climate system. Various physics-based equations and oceanographic theories are utilized to measure saltwater salinity in specific locations. This approach fails to account for data fluctuation, harsh environmental conditions, and the detection of oceanographic data patterns. This type of salinity measurement is time-consuming and expensive. This study evaluates five superior regression models: multivariable linear regression (MVR), K nearest neighbor (KNN), random forest (RF), support vector regression (SVR), and extreme gradient boosting (XGBoost). After performing five regression models, the XGBoost model demonstrated superior performance compared to the other four models with the highest coefficient of determination (R^2). The score recorded was 0.9981, root mean square error (RMSE): 0.0232, mean square error (MSE): 0.0005, and mean absolute error (MAE): 0.0152. Cross-validation was also conducted for the best-performing model for generalization of model output. This study provides a robust and cost-efficient tool and proves the efficiency of the ML model for measuring seawater salinity.

Keywords: Machine Learning, Regression, Water Salinity, Environment, XGBoost

1. Introduction

Water salinity is essential for its direct implications in infrastructure, ecosystems, agriculture, and health [1-2]. Numerous physical or mathematical models have been created to predict and plan for the administration of water quality; however, these models are complex, data- intensive, and time-consuming. Therefore, these models present a challenge for consumers in developing countries due to the lack of data and other shortages. In order to resolve this issue, several statistical models have been developed that are for capturing both linear and non- linear relationships between input-output variables and enhancing predictive performance [3].

Several papers published in multiple journals for Automated salinity concentration measurement. Maier and Dandy [4], applied Artificial Neural Networks for short time salinity prediction in South Australia, with accuracy AAPE 5.3-7% even with limited input data. Eventually, Bowden et al. [5] and Melesse et al. [6] including the hybrid model like PMI-ANN and AR-M5P, performing very high coefficient of determination (R^2) 0.995, NSE 0.994 in Iranian river basins. In recent studies there are using a lot of diverse algorithms in various regions. Guillou et al. [7] applied MLR, SVR and RF models and forecast salinity in France R^2 is up to 0.51. Likewise, Mokhtar et al. [8] used the of SVM, XGBoost, RF for the causes of predicting of irrigation water quality in Egypt with achieving R^2 as high as 0.98 with minimal RMSE. Extreme Gradient Boosting has been identified an efficient solution. Sahour et al. [9] Applying XGBoost to the South Caspian coastal aquifer proves that XGBoost is more accurate and skilled than the Deep Neural Network (DNN) and Multiple Linear Regression (MLR) with achieving high accuracy R^2 1.3, NSE 1.5, NRMSD 1.5. Dang et al [10] in the context of Southeast Asia. In the Mekong Delta region of Vietnam, various machine learning models including RFR, SVR, LSTM and XGBoost showed extraordinary prediction capabilities, where the range of R^2 is between 1.5 to 1.5 and the maximum effectiveness is achieved by the RFR model.

While much of the existing research focused on forecasting salinity levels through time series models, the present study focused on estimating salinity at a specific point using a range of environmental and oceanographic factors and performing five optimal regression models: multivariable linear regression (MVR), K-nearest neighbor (KNN), random forest (RF), support vector regression (SVR), and extreme gradient boosting (XGBoost). These techniques can effectively capture the complex, nonlinear interactions among these variables without relying on temporal data. No prior studies have been undertaken in the California Current System off Southern and Central California, though this region contains significant ports (Los Angeles, Long Beach,

*Corresponding Author Email Address: mohammed05@student.sust.edu

and San Diego) essential to international commerce. Moreover, precise salinity data are crucial for enhancing the functionality of desalination facilities, such as the Carlsbad Desalination Plant, which rely on consistent salinity levels to maintain membrane efficacy and minimize energy expenditures [11].

Table 1 Statistical Summary of Target Variable

	Count	Mean	Standard Deviation	Minimum	Median (50%)	Maximum
Salinity	15897	33.74	0.53	32.06	33.86	35.28

2. Methodology

2.1 Study Area

The data set is sourced from the California Department of Fish and Wildlife (CalCOFI) and contains significant oceanographic features: Depth (m), Temperature (°C), Salinity, Dissolved O2 (ml/L), Potential Density (σθ), Phosphate (μmol/L), Silicate (μmol/L), Nitrite (μmol/L), Nitrate (μmol/L), and Dynamic Height (m).

2.2 Data Analysis and preprocessing

Table 1 provides a detailed description of the output variable, Salinity. Fig. 1 describes the correlation matrix, show the correlations between the variables and target in the dataset. Most variables exhibit weak correlations with Salinity, indicating limited direct interdependence. However, Salinity shows a strong correlation with potential density (0.83), highlighting the influence of density on Salinity concentration. A total of 15897 data points is utilized and to ensure robust model development, 95% of the dataset is allocated for training and 5% for testing. A 10-fold cross-validation strategy is employed for XGBoost model in Table 3 during training to minimize bias and achieve a comprehensive understanding of the relationships between input variables and output variable, Salinity. Fig. 2 depicts the Histogram plot that describe the distribution of data and demonstrate the skewness of each feature. A total of 15897 data points is utilized and to ensure robust model development, 95% of the dataset is allocated for training and 5% for testing.

2.3 Multivariant Linear Regression

The multiple variable regression (MVR) approach is an advanced form of regression analysis that integrates several predictor variables to assess the degree of linear correlation between independent and dependent variables [12].

2.4 Random Forest Regressor

Random Forest, a supervised learning technique that utilizes the ensemble learning method with several decision trees. Random Forest is a Bagging technique, wherein all computations are executed concurrently, and there is no interaction among the Decision Trees throughout their construction. Random Forest can be employed to address both classification and regression applications [13-14].

2.5 Support Vector Regression

Support Vector Regression (SVR) is a kind of Support Vector Machine (SVM) techniques, frequently employed for regression analysis. Support Vector Machines (SVM) seek to identify the ideal hyperplane that most effectively distinguishes between the two classes within the input data. In SVR, A margin is characterized the model's error tolerance, sometimes referred to as the ε-insensitive tube. This tube permits a certain departure of the data points from the hyperplane without being classified as mistakes. The hyperplane represents the optimal fit for the data contained within the Ɛ insensitive tube [15-16].

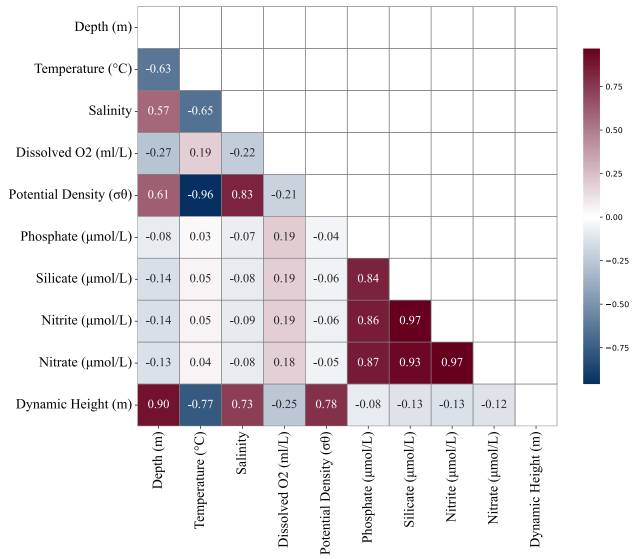

Fig.1 Correlation Matrix of Input and target Features

2.6 Support Vector Regression

Support Vector Regression (SVR) is a kind of Support Vector Machine (SVM) techniques, frequently employed for regression analysis. Support Vector Machines (SVM) seek to identify the ideal hyperplane that most effectively distinguishes between the two classes within the input data. In SVR, A margin is characterized the model's error tolerance, sometimes referred to as the ε-insensitive tube. This tube permits a certain departure of the data points from the hyperplane without being classified as mistakes. The hyperplane represents the optimal fit for the data contained within the Ɛ insensitive tube [15-16].

2.7 K-Nearest Neighbors (KNN)

The K-nearest neighbor (KNN) method is employed for classification and regression tasks. The model works by using k nearest neighbor of given data set and make prediction based on most of the class or averaging continuous value [17].

2.8 Extreme Gradient Boosting (XGBoost)

XGBoost is an ensemble method utilizing decision trees, where the final prediction for a given sample is derived from the aggregate of the outputs of each tree. This will enhance performance by leveraging the strengths of each tree while mitigating their own model deficiencies. XGBoost employs two significant strategies to address overfitting in machine learning: the utilization of second-order derivatives and the incorporation of a regularization framework. The integration of second-order derivatives in XGBoost improves the optimization process during training [18-19].

2.9 Cross Validation

Cross-validation is a repeated sampling technique employed to assess the efficacy of a machine learning model. The dataset is being divided into two halves. One is the training set, while the other is the testing set. It generally provides a reliable assessment of various model performances and aids in identifying overfitting. To ensure accuracy, it is repeated several times. Among the several types of Cross-Validation, we are employing K- Fold Cross-Validation in this context. It functions incrementally. Initially, it divides the entire dataset into two subsets. Divide the entire dataset into k equal segments. The model is trained on k-1 folds, with the remaining fold designated for testing. This procedure is executed k times. [20]

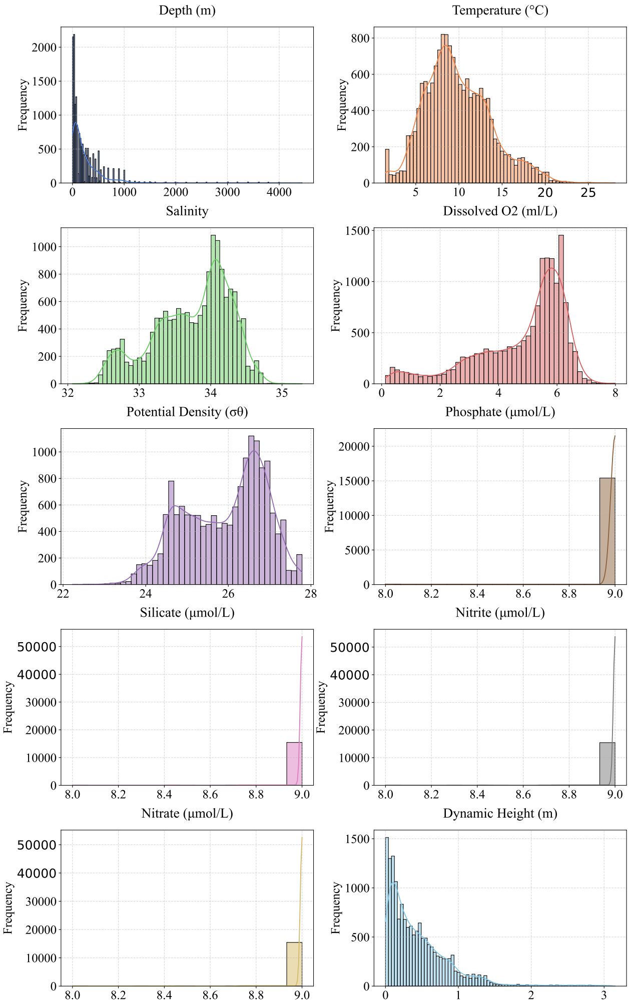

Fig. 2: Histogram plot of Oceanographic Parameters

3. Result and Discussion:

The models are evaluated based on four key metrics: R^2, MAE, MSE and RMSE, across both training and testing datasets Table 2. The Random Forest (RF) model showed very high performance on the training data, with an R^2 of 0.9996, but its testing R^2 dropped to 0.9881, indicating slight overfitting. Its testing MAE was 0.0089, and RMSE was 0.0233, reflecting fewer amounts of errors. The support vector regression (SVR) model had a training R^2 of 0.78136 and a testing R^2 of 0.7586, showing a notable decline in performance on the test set and the potential for overfitting. Its testing MAE 0.1740 and RMSE 0.2476 were the highest among the three models, suggesting poorer performance in both accuracy and error. On the other hand, the Extreme Gradient Boosting (XGBoost) demonstrated the best balance between training and testing performance, with a training R^2 of 0.9992 and a testing R^2 of 0.9981, indicating minimal overfitting. It also had the lowest MAE for both training 0.0105 and testing 0.0152 and RMSE 0.0232, showcasing superior predictive accuracy and precision Table 2.

The performance of XGBoost model was assessed using the K-fold cross-validation technique, employing various numbers of folds (K) including 3-fold, 5-fold, 7- fold, 9-fold, and 10-fold in Table 3.

Table 2 Model Evaluation Metrics

Model	Train R^2	Test R^2	Train RMSE
MVR	0.983	0.979	0.076
RF	0.9996	0.988	0.011
SVR	0.781	0.758	0.266
XGBoost	0.9992	0.9981	0.0148
KNN	0.9682	0.9524	0.0956

Table 2 Continued

Model	Test RMSE	Train MSE	Test MSE	Train MAE	Test MAE
MVR	0.067	0.005	0.004	0.052	0.050
RF	0.023	0.0001	0.0005	0.003	0.008
SVR	0.2476	0.0695	0.0613	0.1856	0.1740
XGBoost	0.0232	0.0002	0.0005	0.0105	0.0152
KNN	0.1155	0.0091	0.0133	0.0570	0.0717

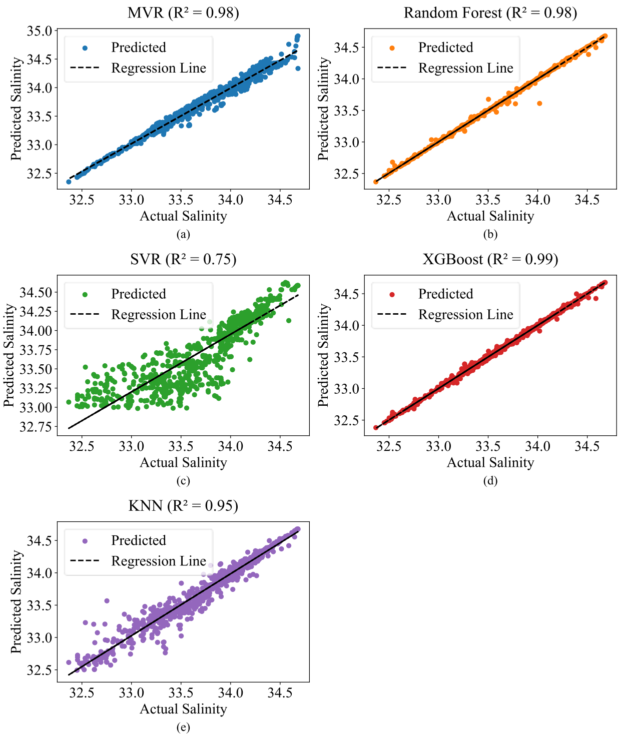

Fig.3 Graphical Presentation of testing Data (a: MVR, b: RF, c: SVR, d: XGBoost, e: KNN)

The R^2 testing value indicates that the all of values are approximately 0.997. The most effective configuration is the nine-fold cross-validation, with a remarkable lowest RMSE of 0.027, whereas 3-fold witnessed the highest maxE 0.599 and RMSE 0.029 value among all folds in Table 3. The maximum and minimum absolute errors verify the model's consistency across multiple folds. The findings reveal that the model has good predictive power and ability to be generalized across diverse subsets of data. The order of accuracy higher to lower of the different k-fold based on the R2, adjusted R^2, MSE, MAE, MAPE, maxE scores: 9-fold, 10-fold, 7-fold, 5-fold, and 3 folds. on the R2, adjusted R2, MSE, MAE, MAPE, maxE scores: 9-fold, 10-fold, 7-fold, 5-fold, and 3 folds. The residual plots Fig. 5 shows how five models perform in predicting salinity, with XGBoost and Random Forest exhibiting the tightest clustering around zero, indicating high accuracy. SVR has the widest spread and largest residuals, suggesting poorer predictions. KNN and MVR perform moderately, but still show more variability compared to tree-based models.

Table 3 Cross validation Score (XGBoost)

K	R^2	Adjusted R^2	MAE	MSE
03	0.997	0.997	0.016	0.001
05	0.997	0.997	0.016	0.001
07	0.997	0.997	0.016	0.001
09	0.997	0.997	0.015	0.001
10	0.997	0.997	0.015	0.001

Table 3 Continued

K	MAPE	RMSE	maxE
03	0.0	0.029	0.599
05	0.0	0.028	0.519
07	0.0	0.029	0.422
09	0.0	0.027	0.392
10	0.0	0.028	0.419

4. Input feature Ranking

From Table 2, it is evident that XGBoost performs better than all other proposed models. To investigate how individual inputs influence the prediction of water salinity, inputs are ranked by systematically excluding each parameter and evaluating the model's performance. Specifically, one parameter is excluded at a time, leaving 10 inputs, and the model's performance metrics R^2 is analyzed in Table 4.

Table 4 Feature importance analysis based on model performance

Dropped Feature	MVR	KNN	SVR	RF	XGBoost
Depth (m)	0.971	0.951	0.751	0.987	0.995
Dissolved O2 (ml/L)	0.970	0.950	0.755	0.988	0.996
Dynamic Height (m)	0.973	0.951	0.753	0.987	0.995
Nitrate (µmol/L)	0.978	0.952	0.751	0.983	0.995
Nitrite (µmol/L)	0.976	0.951	0.753	0.981	0.995
Phosphate (µmol/L)	0.977	0.950	0.755	0.983	0.995
Potential Density(σθ)	0.886	0.887	0.641	0.880	0.855
Silicate (µmol/L)	0.972	0.952	0.755	0.986	0.995
Temperature (°C)	0.835	0.897	0.698	0.868	0.856

The ranking is based on the impact of each input on these metrics, with higher ranks indicating greater influence on Salinity prediction. Fig. 5 presents the input rankings, showing that depth has the significant impact on water Salinity. As depth increases water salinity also increases but after certain depth around 2000 m salinity becomes constant. When potential density and temperature is dropped all of the model performance dropped significantly. Best performer XGBoost model performance dropped 0.9981 to 0.855 and 0.856 for both potential density and temperature mean this two feature is the highly correlated with target variable. Feature importance methods, SHAP capture nonlinear interaction and accounts for feature dependencies ensuring robust and reliable insights into model behavior. Fig 6 also confirms that potential density and temperature are the most critical input.

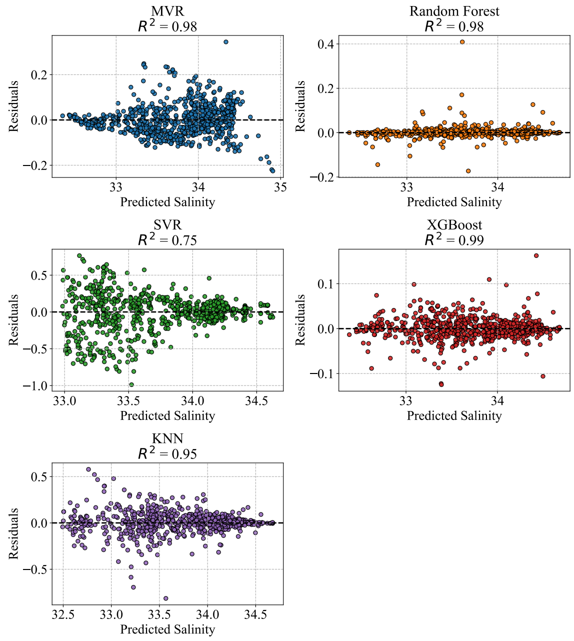

Fig.4 Evaluation of Prediction Errors

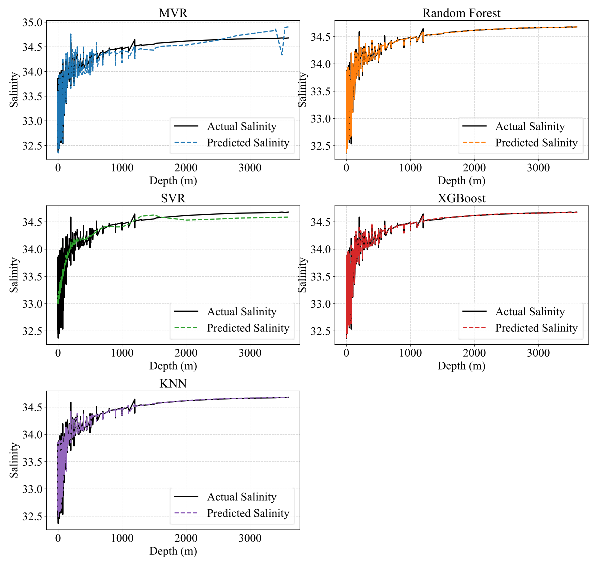

Fig.5 Salinity Profile Comparison Across Depth: Observed vs. Predicted

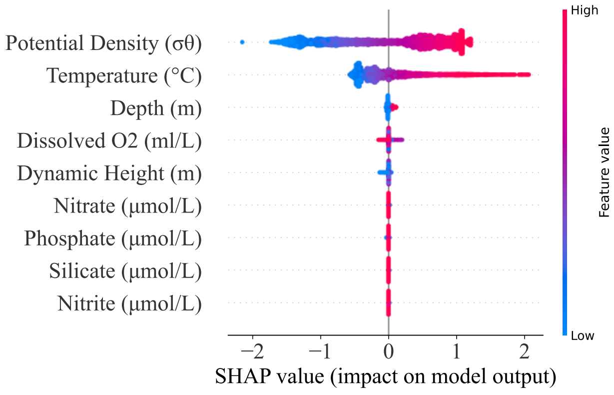

Fig.6 SHAP Feature Importance XGBoost

5. Conclusions

The XGBoost model, constructed using several Trees and Random Forest (RF) without grid search or hyperparameter optimization, surpassed individual models in forecasting Salinity. The predictions of XGBoost and RF achieved accurate Salinity predictions in 2.56 and 3.67 seconds, ensuring efficiency and speed. XGBoost performing best than other model achieving coefficient of determination R^2 0.9992 for training and 0.9981 for testing meaning that no overfitting occurs. Error analysis also conducted RMSE 0.0232, MSE 0.0005, MAE 0.0152 indicates lower error compared to others model. Model generalization proved that XGBoost is generalized for real world complex application. This quick and reliable performance makes it ideal for real-time monitoring and decision-making. The model's simplicity and computational efficiency, without the need for extensive tuning, make it suitable for deployment water Salinity prediction.

6. References

[1] M. A. Nahian, A. Ahmed, A. N. Lázár, C. W. Hutton, M. Salehin, and P. K. Streatfield, "*Drinking water salinity associated health crisis in coastal Bangladesh*," Elementa: Science of the Anthropocene, vol. 6, no. 2, 2018.

[2] T. Khanom, "Effect of salinity on food security in the context of interior coast of Bangladesh," *Ocean & Coastal Management*, vol. 130, pp. 205–212, 2016,

[3] X. Yan, T. Zhang, W. Du, Q. Meng, X. Xu, and X. Zhao, "*A comprehensive review of machine learning for water quality prediction over the past five years*," Journal of Marine Science and Engineering, vol. 12, no. 1, p. 159, 2024.

[4] H. R. Maier and G. C. Dandy, "*Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications*," Environmental Modelling & Software, vol. 15, no. 1, pp. 101–124, 2000.

[5] G. J. Bowden, H. R. Maier, and G. C. Dandy, "*Input determination for neural network models in water resources applications. Part 1—Background and methodology*," Journal of Hydrology, vol. 301, pp. 75–92, 2004.

[6] M. A. Melesse et al., "*Comparison of M5P and regression-based models in salinity prediction," Hydrological Processes*, vol. 34, no. 4, pp. 453–464,2020.

[7] N. Guillou, G. Chapalain, and S. Petton, "*Short-term forecast of sea surface salinity using machine learning," Journal of Marine Systems*, vol. 225, 2022.

[8] A. Mokhtar et al., "*Machine learning approaches for evaluating irrigation water quality in arid regions," Agricultural Water Management*, vol. 244, 2022.

[9] H. Sahour, V. Gholami, and M. Vazifedan, "*A comparative* analysis *of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer*," Journal of Hydrology, vol. 591, p. 125321, Aug. 2020.

[10] L. T. T. Dang, H. Ishidaira, K. P. Nguyen, K. Souma, and J. Magome, "*Short-term salinity prediction for coastal areas of the Vietnamese Mekong Delta using various machine learning algorithms: a case study in Soc Trang Province*," *Applied Water Science*, vol. 15, no. 4, 2025.

[11] "Carlsbad Desalination Project Overview," Poseidon Water.

[12] A. Hussen, T. A. Munshi, L. N. Jahan, and M. Hashan, "*Advanced machine learning approaches for predicting permeability in reservoir pay zones based on core analyses*," Heliyon, vol. 12, p. e32666, Jun. 2024.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001,

[14] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Analysis and Recognition (ICDAR '95)*, Montreal, QC, Canada, Aug. 14–18, 1995, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994

[15] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural*

[16] N. R. Fathi, "Support Vector Regressor – Theory and Coding Exercise in Python," *Medium*, Apr. 24 2021

[17] J. W. Sun, W. Du, and N. Shi, "A survey of kNN algorithm," *Information Engineering and Applied Computing*, vol. 1, no. 1, Art. no. 770, 2019,

[18] A. B. Siddique, T. A. Munshi, N. I. Rakin, M. Hashan, S. S. Chnapa, and L. N. Jahan, "Application of supervised machine learning and unsupervised data compression models for pore pressure prediction employing drilling, petrophysical, and well log data," *Scientific Reports*, vol. 15, no. 1, p. 24706, 2025

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794

[20] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of machine learning algorithms with different k values in k-fold cross-validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021

Nomenclature:

Symbol	Meaning
RF	Random Forest
MVR	Multivariable Linear Regression
SVR	Support Vector Regression
KNN	K Nearest Neighbor

XGBoost	Extreme Gradient Boosting
RMSE	Root Means Square Error
MSE	Mean Square Error
MAE	Mean Absolute Error
maxE	Maximum Error
minE	Minimum Error

R2	Coefficient of determination
m	Meter
°C	Celsius
O_2	Oxygen
μmol/L	Micromole/Liter