

Article

Wavelet-Enhanced Machine Learning for Seawater Alkalinity Prediction in the Arabian Gulf Using Monitored Water-Quality Variables

Saleh H. Alhathloul *  and Yazeed Algurainy 

Department of Civil Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

* Correspondence: saalhathloul@ksu.edu.sa

Abstract

Continuous monitoring of seawater alkalinity is essential for maintaining chemical stability in coastal environments and supporting efficient operation of desalination and water-treatment systems; however, direct alkalinity measurements are often limited in temporal resolution. This study develops and evaluates a machine learning framework for estimating seawater alkalinity using quality-controlled daily and sub-daily water-quality observations collected from a coastal monitoring station along the Arabian Gulf coast of eastern Saudi Arabia during 2017–2023. Five machine learning models, Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), and K-Nearest Neighbors (KNN), are assessed under two configurations: a baseline setup relying on the original predictor variables and an enhanced setup incorporating wavelet-decomposed features to represent multiscale temporal variability. Model performance is evaluated using five-fold cross-validation and quantified using R^2 , root mean square error (RMSE), and mean absolute error (MAE). Under the baseline configuration, ensemble-based models outperform single-estimator and distance-based approaches, with RF achieving the best performance ($R^2 = 0.77$, RMSE = 2.57 ppm, MAE = 1.71 ppm). The incorporation of wavelet-based feature enrichment leads to consistent performance improvements across all models, reflected by higher R^2 values and reduced RMSE and MAE. The wavelet-enhanced RF model exhibits the strongest overall performance, attaining a mean R^2 of approximately 0.91 together with an RMSE of about 1.6 ppm and an MAE of around 1.0 ppm, while also showing reduced variability across cross-validation folds. The XGB model shows notable improvement with wavelet enrichment, whereas SVR and KNN benefit mainly through moderate error reduction. Overall, the findings show that wavelet-based feature enrichment improves the accuracy and stability of ML models for seawater alkalinity estimation, with RF providing the most reliable performance for coastal monitoring applications.

Keywords: seawater alkalinity; machine learning; wavelet-based feature enrichment; coastal water quality; desalination monitoring; multiscale analysis



Academic Editor: Yunsong Mu

Received: 31 March 2026

Revised: 23 June 2026

Accepted: 25 June 2026

Published: 28 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Coastal and marine monitoring stations constitute a critical component of environmental management systems in arid and semi-arid regions, where coastal waters are exposed to increasing anthropogenic pressures and climatic variability [1,2]. In Saudi Arabia, continuous monitoring of seawater quality along the Arabian Gulf is essential for assessing

environmental conditions, supporting coastal management strategies, and ensuring the sustainability of marine and near-shore ecosystems [3,4]. Previous studies describe the Arabian Gulf as a shallow semi-enclosed sea, with a mean depth of about 35 m, strong evaporation, restricted exchange through the Strait of Hormuz, salinity that may exceed 43 practical salinity unit (psu) in some coastal areas, and a reported warming trend of approximately 0.2 °C per decade [5–7]. Maintaining stable seawater chemistry is particularly important in Gulf bay environments, where semi-enclosed circulation patterns and high evaporation rates can amplify temporal variability in water quality parameters [7–9].

Among the physicochemical parameters monitored in seawater environments, alkalinity plays a key role in regulating buffering capacity, pH stability, carbonate equilibrium, and biogeochemical processes. Alkalinity reflects the combined contribution of carbonate and bicarbonate species and is influenced by seawater composition, temperature, biological activity, and external inputs from coastal and offshore processes [10]. Variations in alkalinity can affect chemical equilibrium and may serve as an indicator of broader changes in marine water quality conditions [11–13].

Despite its environmental importance, alkalinity is not always measured continuously at high temporal resolution in coastal monitoring programs. Measurement constraints may arise from laboratory-based analytical requirements, sensor availability, maintenance limitations, or operational priorities at monitoring stations [14–16]. As a result, time-series datasets may include periods where alkalinity records are incomplete, while other water quality variables such as temperature, electrical conductivity, turbidity, and pH continue to be monitored. This creates a practical need for reliable predictive approaches capable of estimating alkalinity using routinely recorded parameters [17–19].

Machine learning (ML) techniques have been increasingly applied in water quality monitoring due to their ability to capture nonlinear relationships and complex interactions among environmental variables without relying on explicit physical or chemical formulations [20,21]. Ensemble learning models, including Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB), have demonstrated strong predictive skill and robustness in marine and coastal applications characterized by noisy, multivariate, and interdependent data [22–24]. These characteristics make such models particularly suitable for seawater monitoring and environmental assessment.

Prior studies in Saudi coastal waters and the Arabian Gulf region indicate that ML can effectively model routinely monitored seawater-quality indicators, including salinity (and its conductivity-related signal) and optical or nutrient-proxy variables such as water clarity/turbidity-related metrics and chlorophyll-a [25–27]. However, the systematic prediction of coastal seawater alkalinity remains comparatively limited, particularly in research that integrates (a) advanced ensemble-learning frameworks, (b) interpretable feature-importance assessment to clarify dominant controls, and (c) multiscale signal representations (e.g., wavelet-based decomposition) to capture time-varying variability in coastal carbonate chemistry [28–32].

Seawater quality records from coastal monitoring stations often exhibit variability across multiple temporal scales, even when observations are collected daily or at sub-daily intervals [33,34]. Short-term fluctuations driven by tidal dynamics and operational factors, together with longer-term environmental trends, can obscure relationships between observed variables [35–37]. Wavelet transform methods provide an effective framework for decomposing time-series data into components associated with different frequency bands, thereby enhancing representation of both high-frequency variability and low-frequency trends [38,39]. When integrated with ML models, wavelet-based feature enrichment has been shown to improve predictive accuracy and model stability in water quality and environmental monitoring studies [40,41].

Despite recent advances in ML applications for coastal water quality prediction, the systematic estimation of seawater alkalinity remains limited, particularly in studies that incorporate multiscale temporal variability. To the best of the author's knowledge, no previous study has integrated wavelet-based multiscale decomposition with ensemble ML for seawater alkalinity prediction in the Arabian Gulf. Therefore, the objective of this study is to develop a robust alkalinity prediction framework for a coastal seawater monitoring station located in the Arabian Gulf on the eastern coast of Saudi Arabia. Five ML models, RF, GB, XGB, Support Vector Regression, and K-Nearest Neighbors, are evaluated under two configurations: one using original water quality variables and another incorporating wavelet-enriched features. Feature importance analysis is employed to interpret the relative contribution of individual predictors, while all variables are retained during model development. Model performance is assessed using multiple statistical metrics to ensure reliable and generalizable outcomes.

2. Methodology

2.1. Data Acquisition and Quality Control

The data used in this study were collected from a seawater monitoring station located in the Arabian Gulf along the eastern coast of Saudi Arabia. Because the exact station location is of a confidential nature and is associated with operational coastal water-quality monitoring infrastructure, the precise coordinates are not disclosed. The measurements were obtained using an automated monitoring system installed at the station to record the analyzed water-quality variables. The original dataset contained 9644 monitoring records covering the period from January 2017 to December 2023 and included daily and sub-daily observations. Prior to analysis, the raw data underwent quality control procedures to identify and remove missing, incomplete, and inconsistent records. After data screening, a total of 5212 valid observations were retained and used in this study. The resulting quality-controlled dataset was not a fully continuous daily time series; rather, it represented an irregular monitoring record with temporal gaps caused by missing or incomplete measurements. The descriptive statistical properties of the analyzed variables are presented in Table 1.

Table 1. Descriptive statistical properties of the analyzed water quality variables.

Variable	Unit	Minimum	Maximum	Mean	Standard Deviation	Median
Temperature	°C	12.20	38.00	27.24	6.29	28.10
pH	-	7.95	8.28	8.04	0.04	8.02
Electrical Conductivity	µS/cm	59,000	66,100	62,162.70	1008.43	62,300
Turbidity	NTU	0.20	9.33	1.11	0.73	0.79
Chloride	ppm	24,184	27,233	25,642.32	449.83	25,725
Alkalinity	ppm as CaCO ₃	112	151	129.62	5.42	128
Residual Chlorine	ppm	0.00	2.24	0.29	0.08	0.28

2.2. Normalization of Input Variables

Normalization of predictor variables is performed to ensure compatibility with learning algorithms that are sensitive to feature scaling. Tree-based models, including Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB), are trained using the original variables since these models rely on threshold-based data partitioning and are insensitive to differences in variable magnitude [22–24]. In contrast, Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) are distance-based methods and are therefore affected by the relative scale of input variables. To avoid dominance of variables with large numerical ranges, predictor variables used in these models are standardized using z-score normalization. The standardized value of a predictor variable, denoted as x_i^*

is computed using Equation (1), where x_i is the original variable value, μ represents the mean of the variable, and σ denotes the standard deviation computed from the training dataset [42–44].

$$x_i^* = \frac{x_i - \mu}{\sigma} \quad (1)$$

A total of five ML models were evaluated in this study, including RF, GB, XGB, SVR, and KNN. Predictor variable normalization was applied selectively based on model requirements. The SVR and KNN models were trained using standardized input variables due to their sensitivity to feature scale, whereas RF, GB, and XGB models were trained using the original variables since they are insensitive to scaling. Wavelet decomposition was applied to the predictor variables and evaluated for all five models under two configurations, one using the original predictors and another using wavelet-decomposed predictors, enabling assessment of the impact of multiscale feature representation on alkalinity prediction.

2.3. Feature Importance Analysis

Feature importance is evaluated using an RF-based approach to quantify the contribution of each operational variable to alkalinity prediction. In RF regression, node impurity is commonly measured using variance. The decrease in variance resulting from a split on a given variable reflects its contribution to prediction accuracy. The importance of the j^{th} variable, denoted as I_j , is evaluated using Equation (2), where T represents the total number of trees in the forest, n is the node in tree t , and $\Delta\sigma^2(n, j)$ represents the variance reduction at node n resulting from a split on the variable j [22,45].

$$I_j = \sum_{t=1}^T \sum_{n \in t} \Delta\sigma^2(n, j) \quad (2)$$

Feature importance analysis was performed using the original, non-normalized predictor variables without wavelet decomposition. This choice was made because RF-based importance measures are insensitive to feature scaling and rely on variance reduction associated with tree-based splits, making normalization unnecessary [22,44]. In addition, wavelet-decomposed predictors were not used for feature importance analysis to ensure that the resulting importance scores remain physically interpretable in terms of the original operational variables [45].

This measure captures nonlinear effects and interactions among operational variables such as conductivity, chloride, pH, and free chlorine. The computed importance values are used only for interpretation and discussion, while all variables are retained in model training.

2.4. Wavelet-Based Feature Enrichment

To represent multiscale temporal variability in seawater quality observations from a coastal monitoring station, the discrete wavelet transform is applied to selected predictor variables. A discrete signal $x(t)$ can be decomposed into approximation and detail components as given in Equation (3), where $a_{J,k}$ are the approximation coefficients representing low-frequency trends, $\phi_{J,k}(t)$ represents the scaling function at level J , $d_{j,k}$ quantify high-frequency fluctuations associated with operational processes, and $\psi_{j,k}(t)$ is the wavelet function at level j [38,39].

$$x(t) = \sum_k a_{J,k} \phi_{J,k}(t) + \sum_{j=1}^J \sum_k d_{j,k} \psi_{j,k}(t) \quad (3)$$

Wavelet-based feature enrichment was applied only to the predictor variables, namely temperature, pH, electrical conductivity, turbidity, chloride, and residual chlorine. The target variable, alkalinity, was not decomposed. The Daubechies wavelet family was selected because it provides compact support and is widely used for representing local signal variations in wavelet analysis [38,46]. In this study, the db4 wavelet, referring to the fourth-order Daubechies wavelet, was used with a three-level discrete wavelet decomposition to capture both low-frequency variability and short-term fluctuations in the monitored seawater-quality variables.

For each predictor variable, the discrete wavelet transform decomposed the original signal into approximation and detail components. The approximation component represents the low-frequency structure of the signal, whereas the detail components represent higher-frequency fluctuations. The retained wavelet-derived components were reconstructed to the original observation length and aligned with the corresponding time-indexed observations. The reconstructed wavelet components were then concatenated with the original operational variables to form a wavelet-enriched feature set, which was used as input to the machine learning models.

To avoid information leakage, wavelet decomposition was not applied to the complete dataset before validation. Instead, wavelet-based feature construction was performed within the training workflow. For each cross-validation fold, the wavelet transformation was applied after data splitting, so validation observations were not used to construct training features. This ensured that the reported model performance was not inflated by information from validation observations entering the training process.

2.5. Machine Learning Models

The modeling workflow was implemented in Python version 3.11 using standard scientific and machine learning libraries, including pandas and NumPy for data handling, scikit-learn for RF, GB, SVR, KNN, preprocessing, and validation, XGBoost for the XGB model, PyWavelets for wavelet decomposition, SciPy for statistical analysis, and matplotlib for visualization. The selected models represent different learning structures: RF as a bagging-based ensemble, GB and XGB as boosting-based ensembles, SVR as a kernel-based regression model, and KNN as a distance-based regression model. Hyperparameters were selected within the training workflow, with model performance assessed using validation results based on R^2 , RMSE, and MAE. The independent test data were not used during hyperparameter selection.

The model hyperparameters used in this study were fixed before model evaluation and are reported to ensure reproducibility. No automated grid-search or Bayesian optimization procedure was applied. RF was trained using 500 trees with a fixed random seed. GB was implemented using the standard gradient-boosting regression configuration with a fixed random seed. XGB was implemented with 800 estimators, a learning rate of 0.05, a maximum tree depth of 4, a subsampling ratio of 0.90, a column-sampling ratio of 0.90, and an L_2 regularization parameter of 1.0. SVR used the radial basis function kernel with $C = 10$, $\epsilon = 0.1$, and $\gamma = \text{scale}$, while KNN used $k = 15$ neighbors with distance-based weighting. The same fixed machine learning hyperparameters were used for both the baseline and wavelet-enhanced configurations to isolate the effect of wavelet-based feature enrichment.

2.5.1. Random Forest Regression

The RF regression constructs an ensemble of decision trees using bootstrap samples of the training data and random subsets of input variables at each split [22]. For a given input vector, the predicted output is obtained by averaging the predictions of all trees, as shown

in Equation (4), where $\hat{y}(x)$ is the predicted alkalinity, $h_t(x)$ denotes the output of t^{th} tree, and T represents the total number of trees.

$$\hat{y}(x) = \frac{1}{T} \sum_{t=1}^T h_t(x) \quad (4)$$

2.5.2. Gradient Boosting Regression

The GB regression constructs an ensemble of decision trees in a sequential manner, where each new tree is fitted to the residual errors of the existing model [23]. For a given input vector x , the model prediction is updated iteratively by adding the contribution of the newly fitted tree, as shown in Equation (5), where $F_m(x)$ represents the updated ensemble model at iteration m , $F_{m-1}(x)$ denotes the model obtained from the previous iteration, $h_m(x)$ is the output of the m^{th} weak learner, and ν is the learning rate that governs the impact of each individual tree on the ensemble output.

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \quad (5)$$

2.5.3. Extreme Gradient Boosting

The XGB model extends the standard GB framework by incorporating explicit regularization terms to control model complexity and improve generalization performance [24]. During training, the model is optimized by minimizing an objective function that balances prediction accuracy with model simplicity, as expressed in Equation (6):

$$\ell = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (6)$$

This objective function (ℓ) consists of a loss component that measures the discrepancy between observed and predicted values and a regularization component that penalizes overly complex models, where $l(y_i, \hat{y}_i)$ denotes the loss function quantifying the prediction error for the i^{th} observation, f_k represents the k^{th} decision tree in the ensemble, $\Omega(f_k)$ is the regularization term applied to each tree to control its complexity, and n is the total number of training samples.

2.5.4. Support Vector Regression

The SVR model aims to determine a function that deviates from the observed values by no more than a predefined margin ε while maintaining minimal model complexity [42]. During training, the model is obtained by solving an optimization problem that minimizes a trade-off between the model flatness and the penalty for deviations exceeding the ε -insensitive zone, as expressed in Equation (7), where the prediction errors are constrained such that the deviation of the predicted value ($w^T \phi(x_i) + b$) from the observed value y_i does not exceed ε except for allowable slack variables. In this formulation, w denotes the weight vector, b is the bias term, C is a regularization parameter controlling the balance between model complexity and training error, and $\xi_i + \xi_i^*$ are slack variables that permit violations of the margin when necessary.

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + \xi_i^* \quad (7)$$

2.5.5. K-Nearest Neighbors

The KNN model predicts alkalinity by estimating the output value as the average of the k closest observations in the feature space, based on a chosen distance metric [43]. For

a given input vector x , the predicted value is computed as shown in Equation (8), where $\hat{y}(x)$ is the predicted alkalinity, y_i denotes the observed alkalinity of the i^{th} neighboring sample, and $N_k(x)$ indicates the set of the k nearest neighbors of x , such that the prediction is obtained by aggregating information from locally similar observations.

$$\hat{y}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i \quad (8)$$

2.6. Model Performance Evaluation

Model performance is evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), which collectively assess the accuracy and reliability of the predictive models. These metrics are defined in Equations (9)–(11), where R^2 measures the proportion of variance in the observed alkalinity explained by the model, RMSE quantifies the magnitude of prediction errors with greater sensitivity to large deviations, and MAE represents the average absolute difference between observed and predicted values. In these equations, y_i denotes the observed alkalinity, $\hat{y}_i$ represents the predicted alkalinity, $\bar{y}$ is the mean observed alkalinity, and n is the total number of observations [47].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (10)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (11)$$

To ensure a robust and unbiased assessment of model generalization capabilities, all predictive models were evaluated using a five-fold cross-validation strategy. The quality-controlled dataset was not a fully continuous or uniformly sampled time series because missing, incomplete, and inconsistent records were removed during data screening. In addition, date and time were not used as model predictors; therefore, the objective was to estimate alkalinity from simultaneously measured water-quality variables rather than to perform chronological forecasting. Accordingly, the retained irregular tabular observations were randomly partitioned into five approximately equal subsets ($k = 5$). During each iteration, four folds (80% of the data) were used for model training, while the remaining fold (20%) was reserved for validation. All preprocessing steps that required information from the data, including feature scaling for SVR and KNN and wavelet-based feature construction, were performed within the cross-validation workflow. Scaling parameters were estimated only from the training folds and then applied to the corresponding validation folds, thereby reducing the risk of information leakage during model evaluation. This procedure was repeated five times so that each fold served once as the validation set, and the reported performance metrics represent the average values across all folds.

The selection of five folds provides a practical balance between computational efficiency and statistical reliability, yielding stable performance estimates while avoiding excessive variance or training cost, particularly for ensemble and kernel-based learning algorithms [44,48]. Although this approach is suitable for evaluating predictive relationships in irregular tabular water-quality data, longer and more continuous records should consider chronological train–test splitting or independent temporal testing to further evaluate temporal transferability.

To further assess whether the observed model-performance differences were statistically supported, the Kruskal–Wallis test was applied to the fold-wise R^2 , RMSE, and MAE

values across the evaluated models. The test was conducted separately for the baseline and wavelet-enhanced configurations. The Kruskal–Wallis test was selected because it is a non-parametric test suitable for comparing multiple independent groups without requiring normally distributed data [49]. The test statistic (H) was computed using Equation (12), where k is the number of model groups, n_i is the number of fold-wise observations in group i , R_i is the sum of ranks for group i , and N is the total number of observations across all groups. The statistical significance of (H) was evaluated using the chi-square distribution with $(k - 1)$ degrees of freedom, and ($p < 0.05$) was used to indicate statistically significant differences among models.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \quad (12)$$

2.7. Sensitivity Analysis

To further evaluate the contribution of wavelet-derived features to the best-performing model, a wavelet-component sensitivity analysis was conducted. Four feature configurations were compared: original predictors only, original predictors plus approximation components (A), original predictors plus detail components (D), and the full wavelet-enriched feature set containing both approximation and detail components (A + D). The approximation components (A) represent smoother low-frequency variability, whereas the detail components (D) represent higher-frequency fluctuations. This analysis was designed to determine whether the improvement in alkalinity prediction was mainly associated with low-frequency approximation information, high-frequency detail information, or their combined contribution. Model sensitivity was assessed using cross-validated R^2 values.

In addition, a feature-selection sensitivity analysis was conducted for all evaluated models under both baseline and wavelet-enhanced configurations. The models were re-evaluated using different predictor subsets, including all predictors, top-ranked predictors based on feature-importance ranking, and reduced configurations excluding either electrical conductivity (EC) or chloride (Cl). This analysis was included to assess model stability under reduced predictor sets and to examine the influence of multicollinearity caused by the highly correlated electrical conductivity–chloride pair on predictive performance. Model sensitivity was evaluated using the cross-validated R^2 values.

3. Results

3.1. Exploratory Data Analysis of Water Quality Variables and Alkalinity

As shown in Figure 1, the monitored water quality variables display distinct distributional patterns that reflect the combined influence of environmental variability and local coastal processes. Temperature spans a relatively wide range, indicating pronounced temporal variability primarily associated with seasonal changes and natural marine dynamics. In contrast, pH values are tightly clustered within a narrow interval, suggesting a relatively stable carbonate buffering system in the seawater environment. Electrical conductivity and chloride concentrations exhibit unimodal distributions with moderate dispersion, indicating generally consistent salinity conditions throughout the study period. Turbidity and residual chlorine display right-skewed distributions, with the majority of observations concentrated at low values and occasional higher measurements, likely linked to short-term environmental disturbances or localized inputs. Alkalinity values are centered within a well-defined range, reflecting relatively stable carbonate chemistry with limited occurrence of extreme conditions.

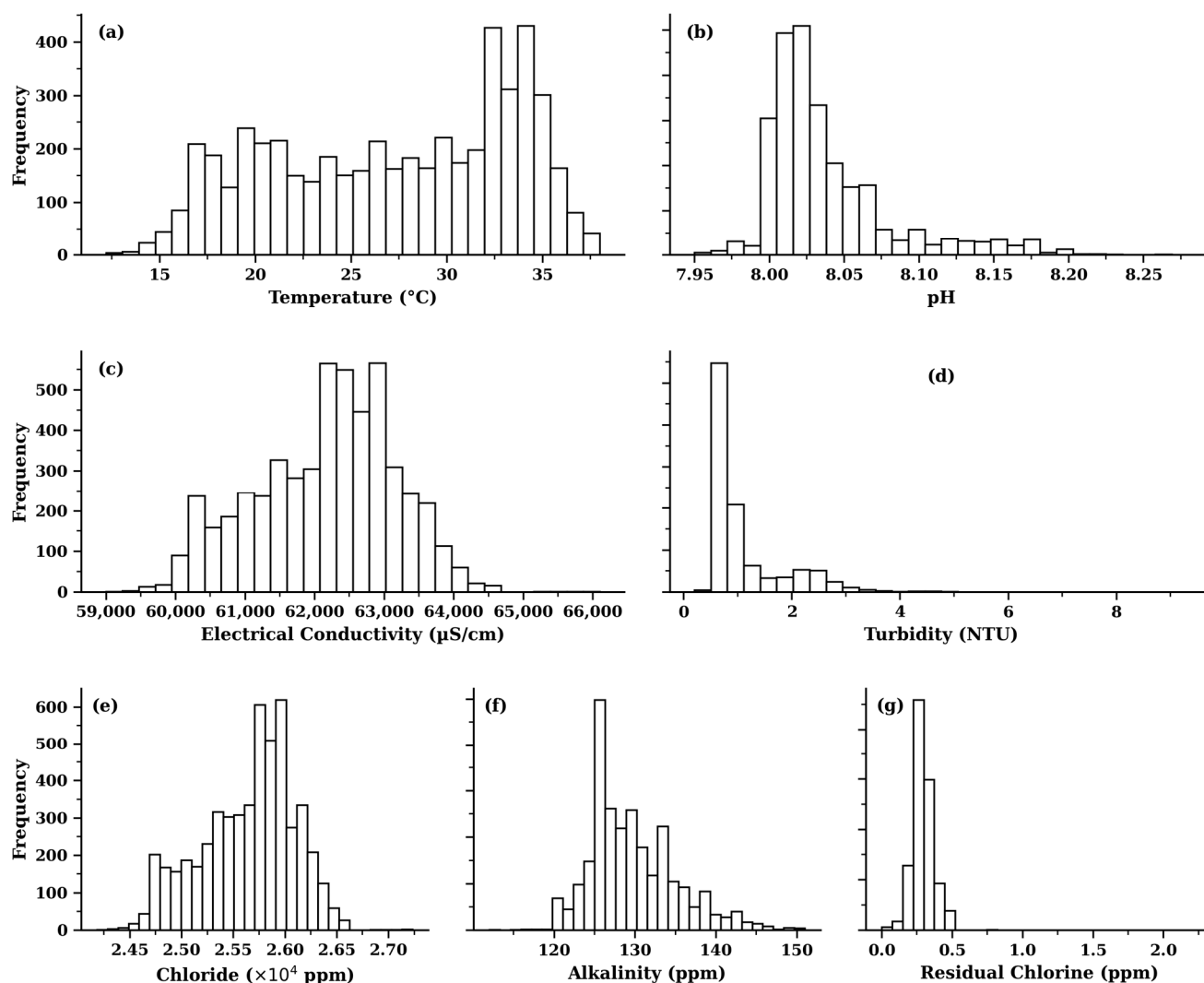


Figure 1. Statistical distributions of (a) temperature, (b) pH, (c) electrical conductivity, (d) turbidity, (e) chloride, (f) alkalinity, and (g) residual chlorine.

Clear patterns emerge in the relationships between alkalinity and selected water quality variables (Figure 2). Alkalinity shows a positive association with both electrical conductivity and chloride concentration, reflecting its close linkage with the ionic composition of seawater. In contrast, the relationship between alkalinity and turbidity is more dispersed, with alkalinity spanning a wide range even at low turbidity levels, indicating a weak direct dependence on particulate-related variability. The relationship between alkalinity and temperature also exhibits considerable scatter, suggesting that temperature influences alkalinity indirectly through environmental and physicochemical processes rather than acting as a primary controlling factor. Overall, these observed relationships highlight the multivariate and interconnected nature of alkalinity behavior in the coastal seawater environment. The horizontal striations observed in the alkalinity scatter plots are likely related to the measurement or reporting resolution of the monitoring system, including possible rounding during data logging. These discrete bands reflect the recorded structure of the original monitoring database and were not introduced during preprocessing or model development.

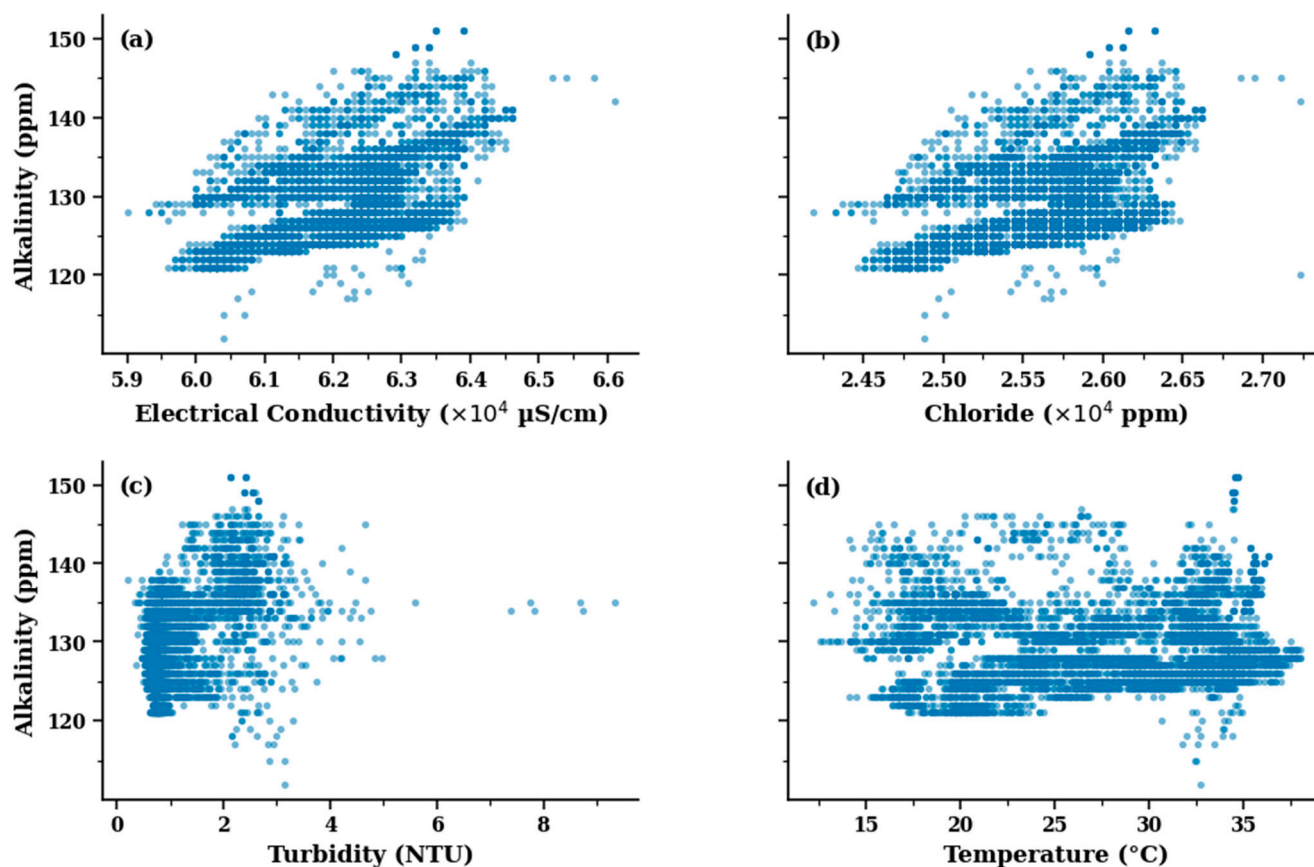


Figure 2. Relationships between alkalinity and (a) electrical conductivity, (b) chloride, (c) turbidity, and (d) temperature.

Monthly alkalinity exhibits clear seasonal variability when examined over the annual cycle (Figure 3). Median alkalinity values remain within a relatively narrow range of approximately 125–133 ppm, indicating a generally stable buffering capacity throughout the year. Slightly higher median levels are observed during late winter and spring, whereas marginally lower values tend to occur toward the end of the year. The degree of variability differs among months; wider interquartile ranges and extended whiskers in months such as March, June, and September indicate increased fluctuations and occasional elevated concentrations, while August displays a more compact distribution reflecting comparatively uniform conditions. Overall, the observed pattern suggests moderate seasonal variation accompanied by episodic variability rather than pronounced shifts, implying that alkalinity in the coastal seawater environment remains largely stable over time with limited short-term deviations.

3.2. Statistical Characteristics and Interrelationships of Water Quality Variables

The interrelationships among the monitored water quality variables provide important insight into the factors governing alkalinity variability in the coastal seawater environment. The Pearson correlation matrix (Figure 4) reveals a very strong positive association between electrical conductivity and chloride concentration, reflecting their shared dependence on dissolved ionic content. Alkalinity shows moderate positive correlations with conductivity, chloride, turbidity, and pH, indicating that its variability is influenced by a combination of salinity-related parameters and physicochemical conditions rather than by a single dominant factor. In contrast, residual chlorine exhibits weak to negative correlations with most variables, including alkalinity, suggesting that disinfection-related signals are largely decoupled from carbonate buffering processes. Temperature displays only a weak direct

relationship with alkalinity, implying that its influence is indirect and likely mediated through broader environmental and physicochemical mechanisms.

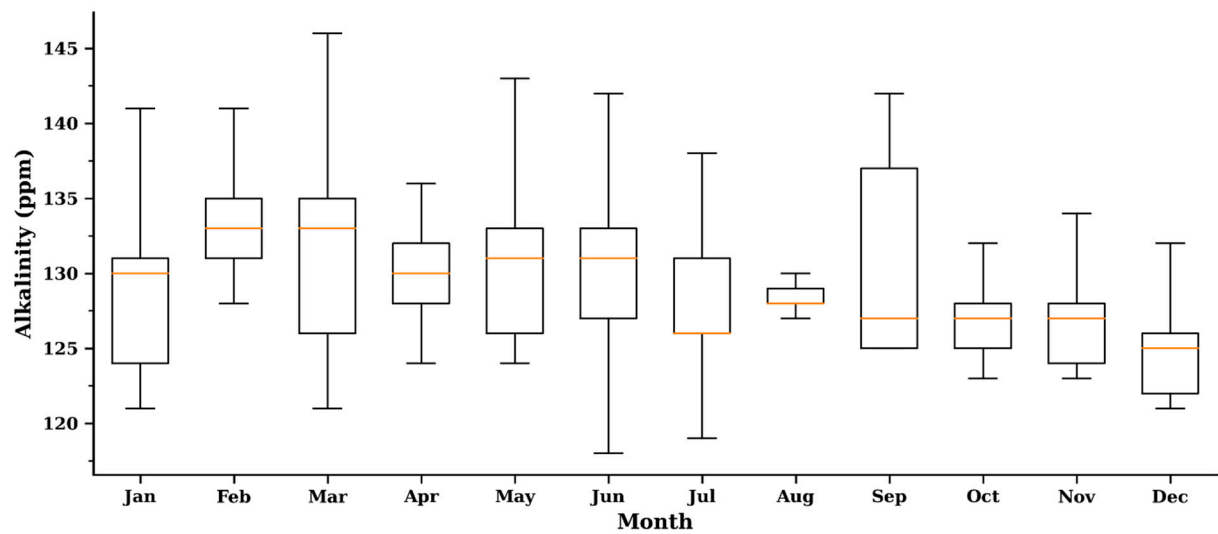


Figure 3. Monthly box-and-whisker distributions of alkalinity concentrations (ppm).

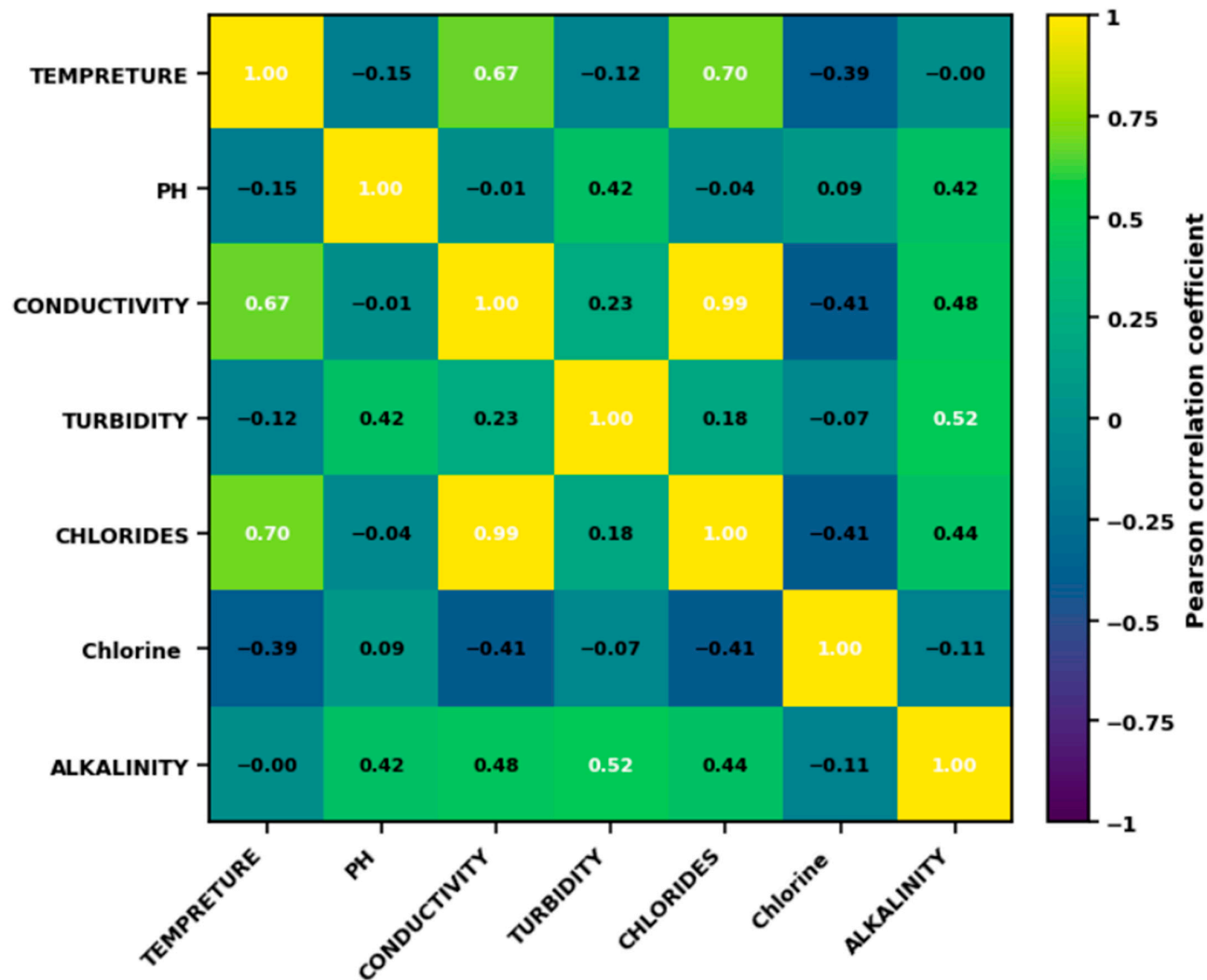


Figure 4. Pearson correlation matrix of the monitored water quality variables.

The relative importance of individual predictors in estimating alkalinity is further clarified through the RF feature importance analysis (Figure 5). Electrical conductivity emerges as the most influential variable, followed closely by chloride concentration, underscoring the dominant role of dissolved ionic content in governing alkalinity levels in coastal seawater environments. pH and temperature exhibit moderate contributions, reflecting their association with carbonate equilibrium and broader physicochemical conditions. In contrast, turbidity and residual chlorine show comparatively low importance, indicating a limited direct influence on alkalinity prediction. Collectively, these results emphasize the multivariate nature of alkalinity control, with salinity-related parameters providing the primary predictive information, supplemented by secondary chemical and environmental factors.

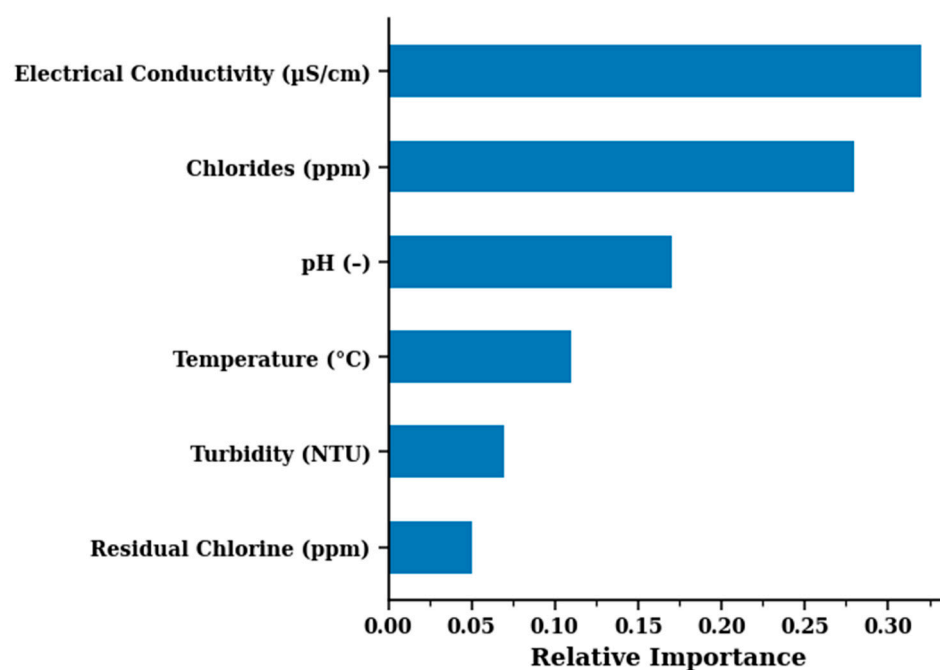


Figure 5. Relative importance of predictor variables in the RF alkalinity model.

Overall, the combined correlation and feature importance analyses indicate that alkalinity variability in the monitored seawater environment is primarily controlled by ionic composition, particularly electrical conductivity and chloride concentration, while pH and temperature exert a secondary but non-negligible influence. The limited contribution of turbidity and residual chlorine indicates that alkalinity is more closely linked to dissolved chemical processes than to particulate matter or disinfection practices. These findings provide a clear physical interpretation of the ML results and support the suitability of multivariate data-driven approaches for reliable alkalinity estimation using routinely monitored water quality parameters.

3.3. Predictive Performance of Machine Learning Models

3.3.1. Baseline Model Performance Using Original Predictors

The predictive performance of the ML models using the original water quality variables is summarized in Figure 6. Distinct differences are observed among the evaluated algorithms, with ensemble-based models consistently outperforming single-estimator approaches. The RF model demonstrates the strongest overall performance, achieving the highest mean coefficient of determination ($R^2 = 0.77$) along with the lowest RMSE (2.57 ppm) and MAE (1.71 ppm). In addition, RF exhibits the smallest standard deviations across all performance metrics, indicating high stability and robustness across the cross-validation

folds. The compact RF violin confirms low variability across validation folds, whereas the wider distributions of GB, SVR, and KNN indicate greater fold-to-fold sensitivity.

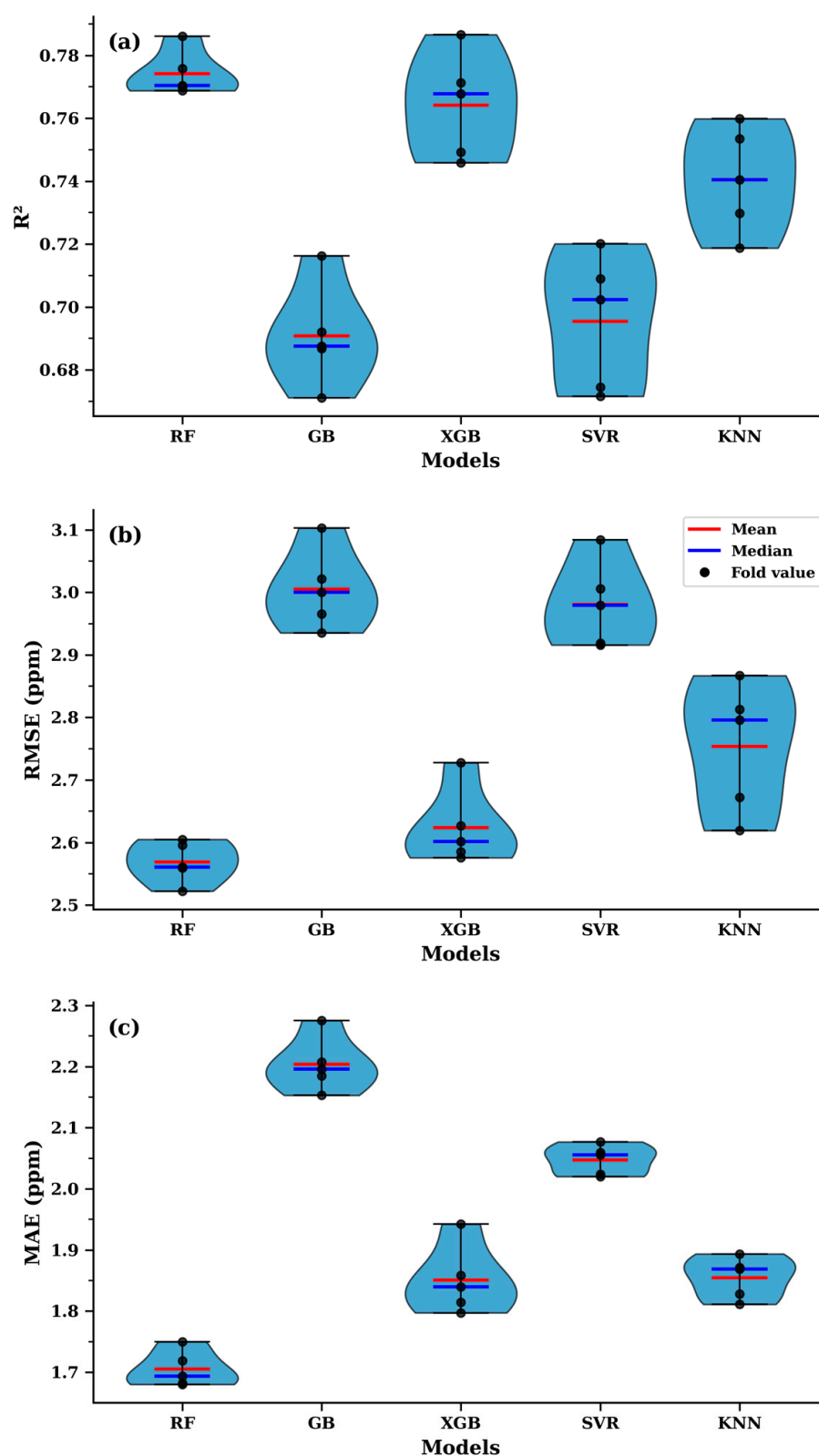


Figure 6. Cross-validated performance distributions of baseline machine learning models for alkalinity estimation using original predictor variables: (a) R^2 , (b) RMSE, and (c) MAE. Red horizontal lines indicate the mean, blue horizontal lines indicate the median, and black dots represent individual fold values from five-fold cross-validation.

XGB ranks second, exhibiting predictive accuracy comparable to that of RF ($R^2 = 0.76$) but with slightly higher error magnitudes and greater variability across cross-validation folds. The KNN model demonstrates moderate predictive capability ($R^2 = 0.74$), achieving intermediate error levels but showing reduced consistency relative to ensemble-based approaches. In contrast, SVR and GB yield lower predictive accuracy and higher error values, indicating a limited ability to fully capture the nonlinear relationships governing alkalinity when using only the original predictor set. Overall, these results underscore the superior performance and robustness of tree-based ensemble models, particularly RF, for baseline alkalinity prediction in coastal seawater monitoring applications.

Although the baseline results show that ensemble models, with RF exhibiting the strongest performance, achieve robust alkalinity predictions using the original water quality variables, residual uncertainty persists. Coastal seawater quality records exhibit variability across multiple temporal scales that may not be fully captured by the original predictors. To address this limitation, the following subsection investigates the effect of integrating wavelet-decomposed features into the modeling framework, with particular attention to gains in predictive accuracy, reductions in error metrics, and enhancements in model stability relative to the baseline performance.

3.3.2. Effect of Wavelet-Based Feature Enrichment on Model Performance

The influence of wavelet-based feature enrichment on model performance is evaluated using cross-validated distributions of R^2 , RMSE, and MAE, as shown in Figure 7. Relative to the baseline configuration based on the original predictor variables, the inclusion of wavelet-derived features leads to improved predictive performance across all considered models. This improvement is reflected by higher R^2 values and lower RMSE and MAE, indicating a better representation of alkalinity variability and reduced prediction errors when multiscale information is incorporated. The RF model achieves the highest mean R^2 (approximately 0.91), together with the lowest RMSE (about 1.6 ppm) and MAE (about 1.0 ppm), indicating the most favorable overall performance under the wavelet-enriched configuration. The compact RF and XGB distributions further indicate that the wavelet-enhanced ensemble models provide more consistent performance across validation folds. In contrast, the wider distributions observed for SVR and KNN suggest greater fold-to-fold sensitivity despite the overall improvement in their error metrics.

The performance improvements are more evident for the ensemble-based models, particularly RF and XGB, which show larger gains compared with the other approaches. In addition to improved central performance measures, these models exhibit reduced variability across cross-validation folds, suggesting enhanced robustness when wavelet-based features are included. SVR and KNN also show performance improvements with wavelet enrichment; however, the magnitude of these gains is smaller, indicating a more limited ability to exploit the additional multiscale information. Overall, the results indicate that wavelet-based feature enrichment enhances both predictive accuracy and model stability relative to the baseline configurations, supporting its use for improving alkalinity estimation under variable coastal seawater conditions.

3.3.3. Comparative Evaluation of Baseline and Wavelet-Enhanced Models

A direct comparison between the baseline models trained using the original predictor variables and the wavelet-enhanced models is presented in Figure 8, which summarizes the cross-validated mean $\pm$ standard deviation of R^2 , RMSE, and MAE for all evaluated algorithms. Overall, wavelet-based feature enrichment leads to improved predictive performance across all models, as indicated by increased R^2 values and reduced error metrics relative to the baseline configuration. In addition, a general reduction in variability across

cross-validation folds is observed, reflecting improved model stability and more consistent generalization when multiscale information is incorporated into the modeling framework.

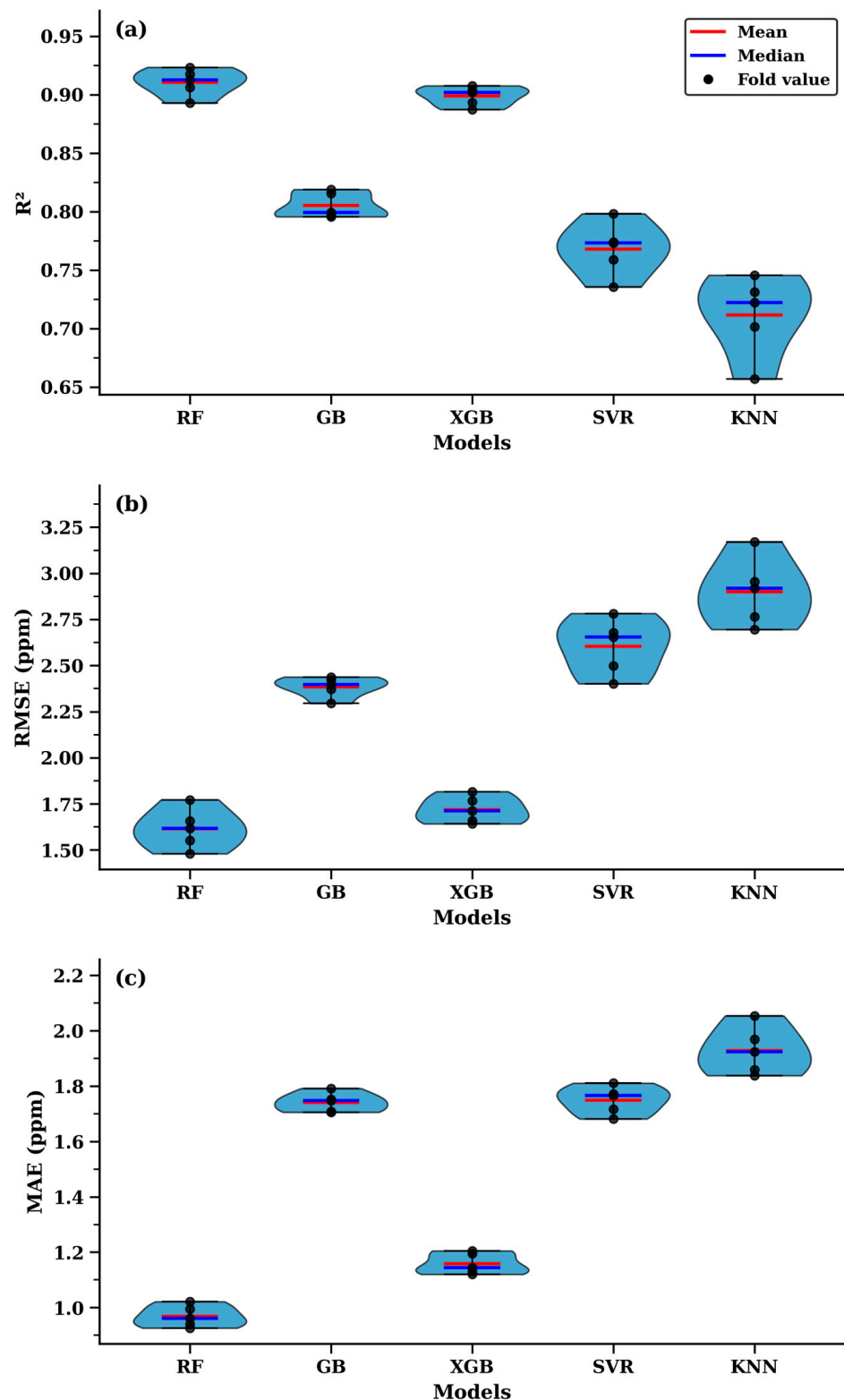


Figure 7. Cross-validated performance distributions of wavelet-enhanced machine learning models for alkalinity estimation: (a) R^2 , (b) RMSE, and (c) MAE. Red horizontal lines indicate the mean, blue horizontal lines indicate the median, and black dots represent individual fold values from five-fold cross-validation.

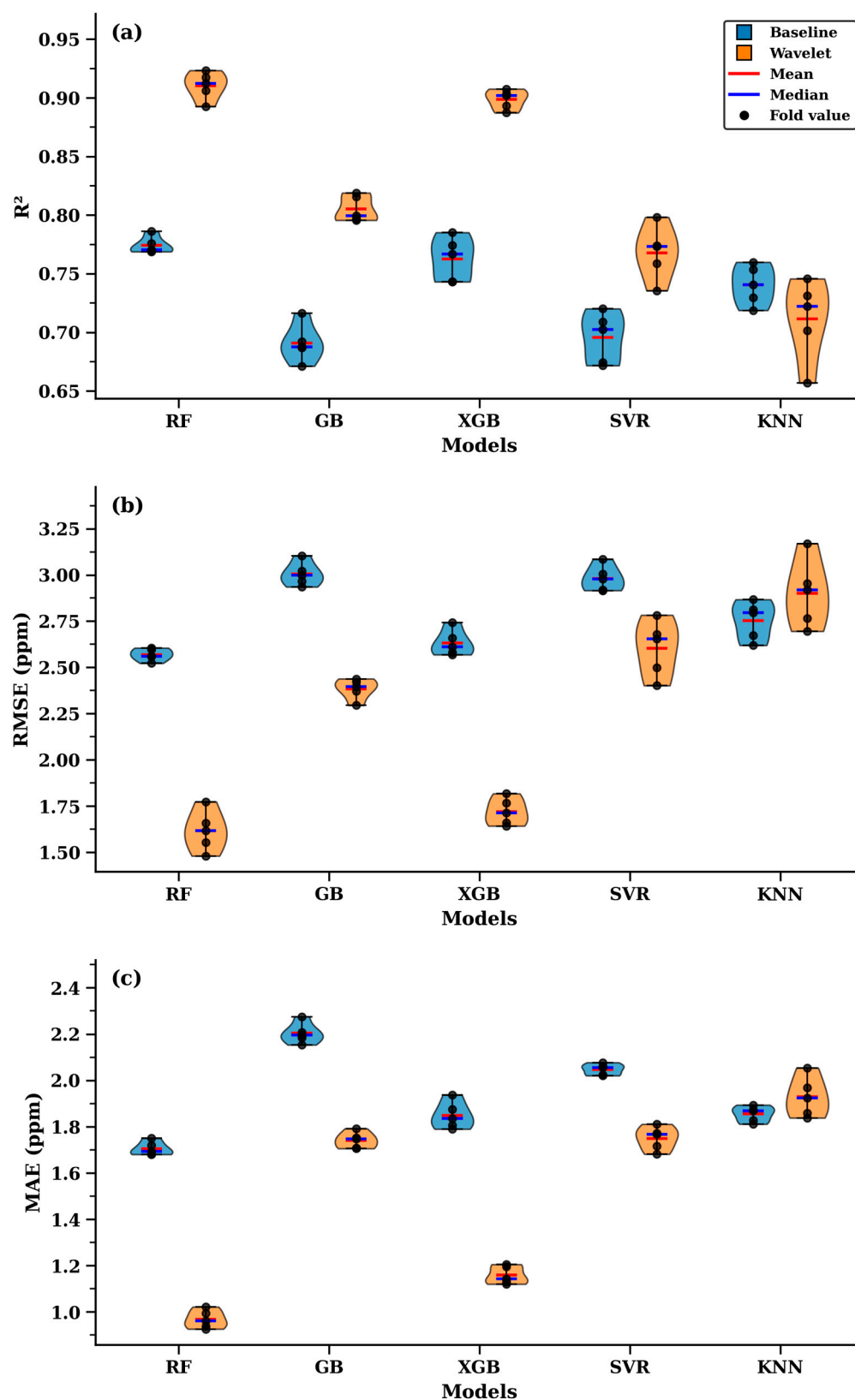


Figure 8. Comparative robustness analysis of baseline and wavelet-enhanced machine learning models for alkalinity estimation. Cross-validated performance is shown as fold-wise distributions for (a) R^2 , (b) RMSE, and (c) MAE.

Among the evaluated algorithms, RF consistently exhibits the strongest performance in both the baseline and wavelet-enhanced configurations, achieving the highest R^2 values together with the lowest RMSE and MAE. Under the wavelet-enriched setting, RF attains a mean R^2 of approximately 0.91, accompanied by an RMSE of about 1.6 ppm and an MAE of around 1.0 ppm, indicating a clear improvement over the baseline case. The integration of wavelet-decomposed features further enhances RF performance and reduces cross-validation variability, reinforcing its robustness relative to the other models. XGB also shows notable performance gains with wavelet enrichment, reflected by increased R^2 and reduced error metrics, although its performance remains slightly below that of RF. The SVR and KNN models benefit primarily through moderate reductions in RMSE and MAE, with smaller increases in R^2 . Overall, the results presented in Figure 8 confirm that wavelet-based feature enrichment improves predictive accuracy and robustness across all models, with RF emerging as the most reliable approach for alkalinity estimation under the studied coastal seawater conditions. The paired violin distributions also show that the improvement is most pronounced for RF and XGB, where the wavelet-enhanced configurations shift toward higher R^2 and lower error values while maintaining relatively compact fold-wise spreads. This indicates that wavelet enrichment improves not only average predictive accuracy but also the consistency of model performance across validation folds.

To further support the robustness of the model comparison, the Kruskal–Wallis test was applied to the fold-wise R^2 , RMSE, and MAE values across the five evaluated models: RF, GB, XGB, SVR, and KNN. This analysis provided an additional statistical basis for evaluating whether the apparent differences in model performance across validation folds were meaningful. As shown in Table 2, statistically significant differences were found among the models for all metrics under both the baseline and wavelet-enhanced configurations ($p < 0.05$). This indicates that the observed performance differences were statistically supported rather than purely descriptive.

Table 2. Kruskal–Wallis test results for fold-wise model-performance differences under baseline and wavelet-enhanced configurations.

Configuration	Metric	Kruskal–Wallis (H)	p -Value	Significant at ($p < 0.05$)
Baseline	R^2	19.311	0.000513	Yes
Baseline	RMSE	20.817	0.000344	Yes
Baseline	MAE	21.397	0.000206	Yes
Wavelet-enhanced	R^2	21.807	0.000190	Yes
Wavelet-enhanced	RMSE	21.630	0.000237	Yes
Wavelet-enhanced	MAE	21.692	0.000204	Yes

3.4. Hyperparameter Configuration for Baseline and Wavelet-Enriched Models

The hyperparameter settings reported in Table 3 were selected to balance robust generalization and model stability under cross-validation, while avoiding excessive tuning that may artificially inflate performance estimates. For ensemble tree-based methods, employing a relatively large number of trees is a common practice, as it reduces variance and stabilizes predictions, which is particularly important for noisy and heterogeneous water-quality datasets. Previous studies in water-quality and coastal environmental monitoring have shown that RF models with several hundred trees provide reliable ensemble averaging and improved robustness when applied to complex hydro-environmental data [22,50].

Table 3. Hyperparameter configuration of ML models used for baseline and wavelet-based alkalinity prediction.

Model	Configuration	Key Hyperparameters
RF	Baseline & Wavelet	Number of trees = 500; Maximum depth = None; Minimum samples per split = 2; Feature selection = maximum features (1.0); Bootstrap sampling = True
GB	Baseline & Wavelet	Number of boosting stages = 100; Learning rate = 0.1; Maximum tree depth = 3; Loss function = squared error
XGB	Baseline & Wavelet	Number of trees = 800; Learning rate = 0.05; Maximum depth = 4; Subsample ratio = 0.9; Column subsample ratio = 0.9; I_2 regularization (λ) = 1.0
SVR	Baseline & Wavelet	Kernel = radial basis function (RBF); Regularization parameter (C) = 10.0; ϵ -insensitive loss = 0.1; Kernel coefficient γ = scale
KNN	Baseline & Wavelet	Number of neighbors (k) = 15; Distance weighting = inverse distance; Distance metric = Euclidean
Wavelet Decomposition	Wavelet only	Wavelet type = Daubechies (db4); Decomposition level = 3; Feature construction = approximation and detail coefficients concatenated with original predictors

For boosted tree models, the XGBoost configuration summarized in Table 3 follows a conservative and widely adopted boosting strategy, combining moderate tree depth with a small learning rate and subsampling. This configuration limits overfitting while preserving the ability to capture nonlinear relationships. Water-quality modeling studies commonly employ learning rates in the range of 0.05–0.10 together with row and column subsampling ratios close to 0.8–0.9 to enhance generalization performance under variable environmental conditions [24,51].

For kernel-based and distance-based models, the SVR and KNN hyperparameters in Table 3 are consistent with values frequently reported in water-quality monitoring and environmental prediction studies, where smooth nonlinear behavior and noise tolerance are critical. The SVR model is commonly implemented using ϵ -insensitive loss values on the order of 0.1 in water-quality and coastal monitoring applications, providing a balance between model flexibility and robustness to measurement uncertainty [52,53]. Similarly, KNN models are often configured with distance-based weighting to emphasize nearby observations and with moderate neighborhood sizes to avoid excessive smoothing. Such settings have been widely adopted in water-quality and water-quality-index prediction frameworks, including applications related to drinking water and coastal environmental assessment [51,54].

3.5. Error Diagnostics and Model Uncertainty Analysis

3.5.1. Diagnostic Evaluation of Baseline Model Performance

A comprehensive diagnostic evaluation of the baseline RF model used for alkalinity prediction based on the original water quality variables is illustrated in Figure 9. The observed versus predicted scatter plot (Figure 9a) shows a strong linear relationship, indicating that the model successfully captures the overall variability and central tendency of alkalinity across the observed range. Most predictions cluster close to the 1:1 line, demonstrating good agreement between measured and estimated values. However, some dispersion is evident, particularly at higher alkalinity levels, suggesting localized prediction uncertainty under certain conditions.

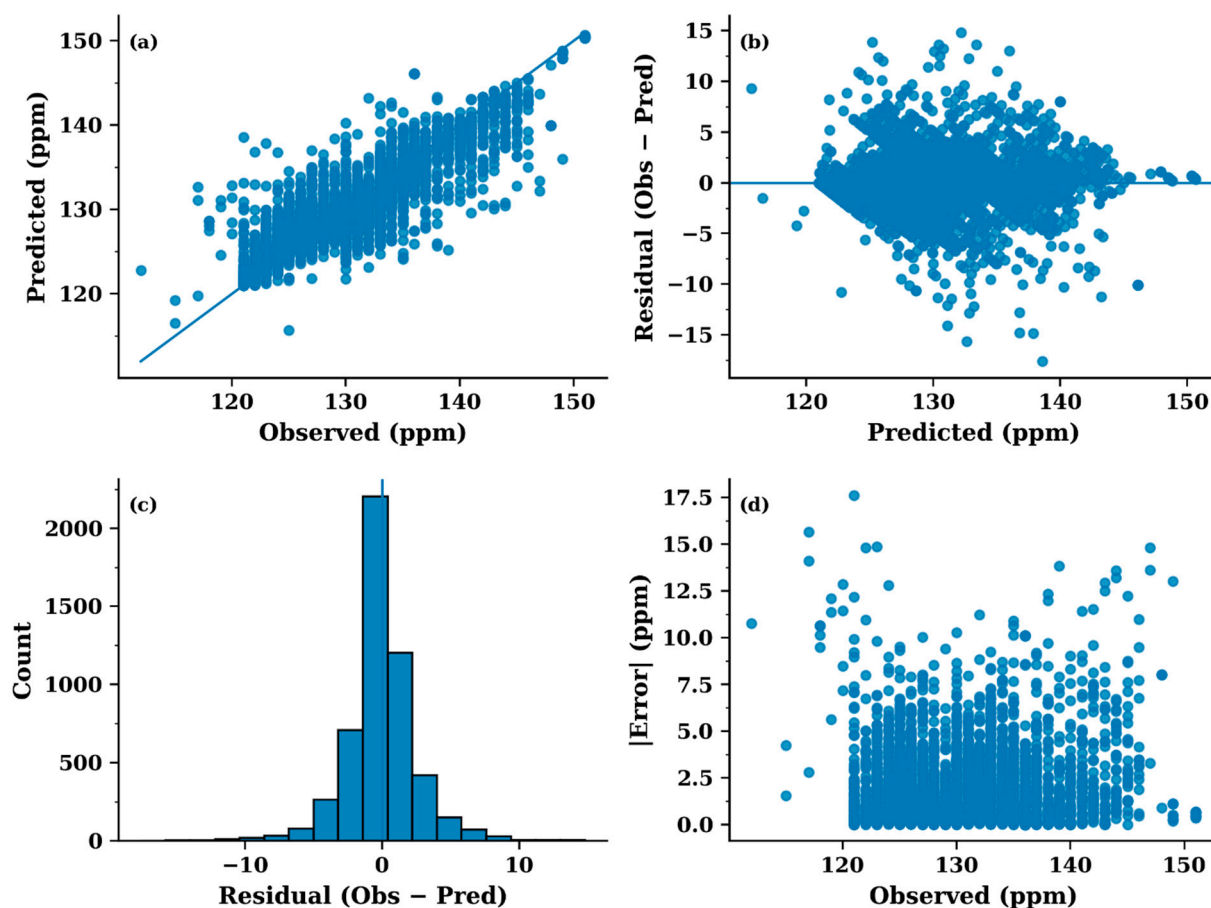


Figure 9. Diagnostic evaluation of the baseline RF model: (a) observed versus predicted alkalinity, (b) residuals versus predicted values, (c) residual distribution, and (d) absolute error versus observed alkalinity.

The residual diagnostics further clarify model behavior. As shown in Figure 9b, residuals are generally centered around zero across the range of predicted alkalinity values, indicating the absence of systematic bias. Nevertheless, the spread of residuals increases slightly at intermediate and higher prediction levels, pointing to mild heteroscedasticity. The residual distribution shown in Figure 9c is approximately symmetric with a clear central peak near zero, confirming that prediction errors are largely unbiased, although the presence of extended tails reflects occasional larger deviations from observations.

Error variability across the alkalinity range is illustrated in Figure 9d. Absolute prediction errors remain relatively small for most observations but increase sporadically at higher alkalinity values, indicating that model uncertainty is not uniform across the range of conditions. This behavior suggests that while the baseline RF model provides reliable overall performance, certain regimes exhibit increased complexity that is not fully captured by the original predictors alone. Collectively, these diagnostics highlight the strengths of the baseline model while also motivating the incorporation of multiscale features to further reduce residual variability and improve predictive robustness.

3.5.2. Diagnostic Evaluation of Wavelet-Enriched Model Performance

The diagnostic plots for the wavelet-enhanced RF model, presented in Figure 10, provide further insight into model behavior and error characteristics beyond aggregate performance metrics. The observed-predicted scatter plot (Figure 10a) shows a strong linear relationship with most points closely aligned along the 1:1 reference line, indicating that the model effectively captures the magnitude and variability of alkalinity across the

observed range. Minor dispersion is evident at higher concentration levels, suggesting localized uncertainty under specific conditions; however, no systematic bias is apparent. The residuals plotted against predicted values (Figure 10b) are symmetrically distributed around zero, confirming the absence of consistent over- or under-prediction across the prediction domain.

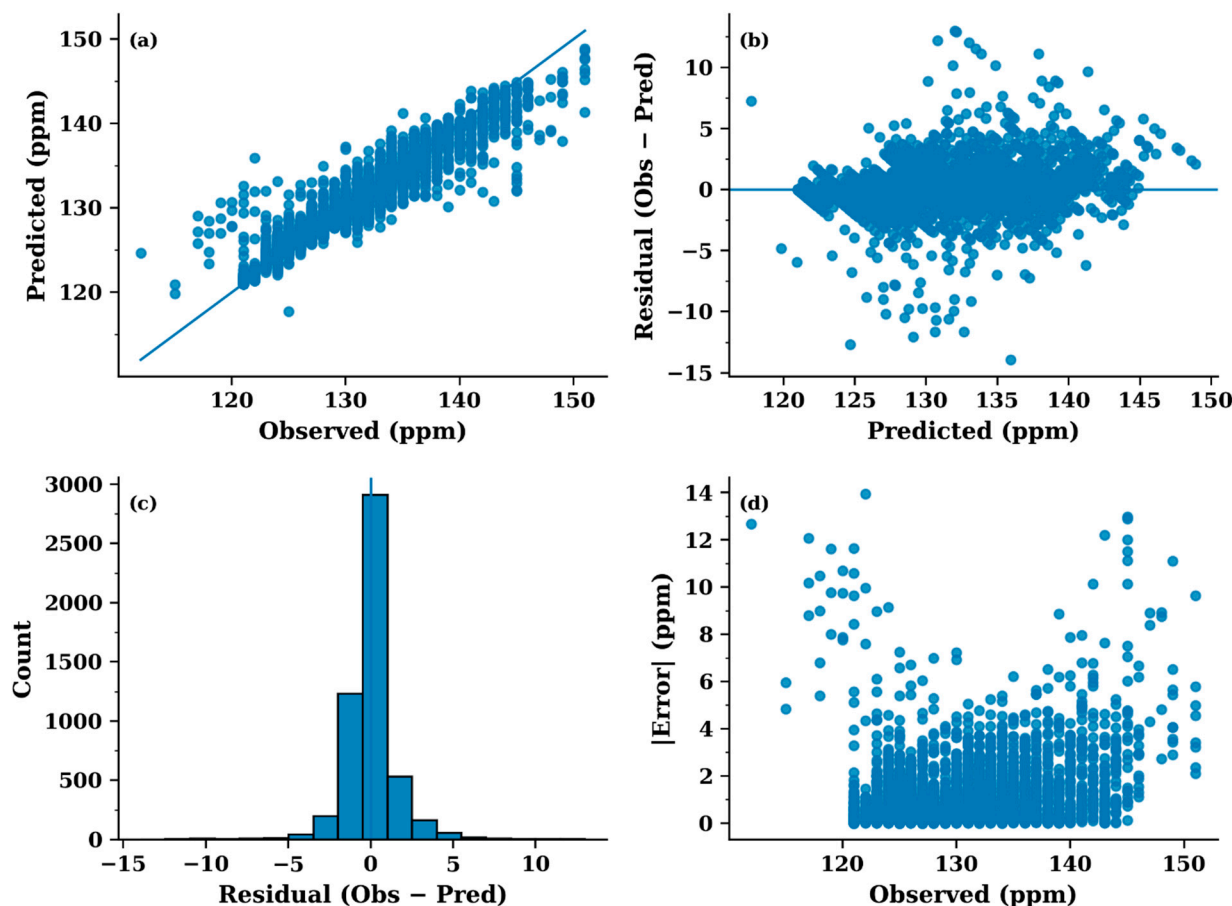


Figure 10. Diagnostic evaluation of the wavelet-enhanced RF model: (a) observed versus predicted alkalinity, (b) residuals versus predicted values, (c) residual distribution, and (d) absolute error versus observed alkalinity.

The residual distribution shown in Figure 10c is approximately centered at zero with a narrow spread, indicating that prediction errors are largely unbiased and dominated by small residuals. Although a limited number of larger deviations are present in the distribution tails, their frequency is low relative to the overall sample size. The absolute error versus observed alkalinity plot (Figure 10d) further indicates that most prediction errors remain small across the full concentration range, with slightly increased dispersion at higher alkalinity values. Collectively, these diagnostics confirm that the wavelet-enhanced RF model exhibits stable and reliable predictive behavior, with residual patterns consistent with a well-calibrated model and no evidence of structural deficiencies or systematic error.

3.6. Sensitivity Analysis Results

The wavelet-component sensitivity analysis was conducted for the RF model, which was the best-performing model and represents a bagging-based ensemble learning approach, as summarized in Table 3. The wavelet-enhanced configuration reported in Table 3 corresponds to the Original + A + D setting, in which the original predictors were concatenated with both approximation components (A) and detail components (D). This analysis

aimed to clarify the relative contribution of these components to alkalinity prediction. As shown in Figure 11, the original predictor configuration provided the baseline level of performance, while the Original + A configuration produced a clear increase in R^2 . The full wavelet-enriched configuration, referred to as Original + A + D, also maintained strong predictive performance. In contrast, the Original + D configuration resulted in lower performance, indicating that high-frequency fluctuations alone contributed less effectively to alkalinity prediction. Overall, the results suggest that the performance improvement from wavelet enrichment was mainly driven by approximation components, which represent smoother low-frequency variability in the monitored water-quality variables. This component-level behavior suggests that the detail components may primarily capture short-term fluctuations or noise that are less informative for RF-based alkalinity prediction when used without the approximation components. The similar performance of the Original + A and Original + A + D configurations further indicates that adding detail components provides only limited marginal benefit once the smoother low-frequency structure has already been incorporated.

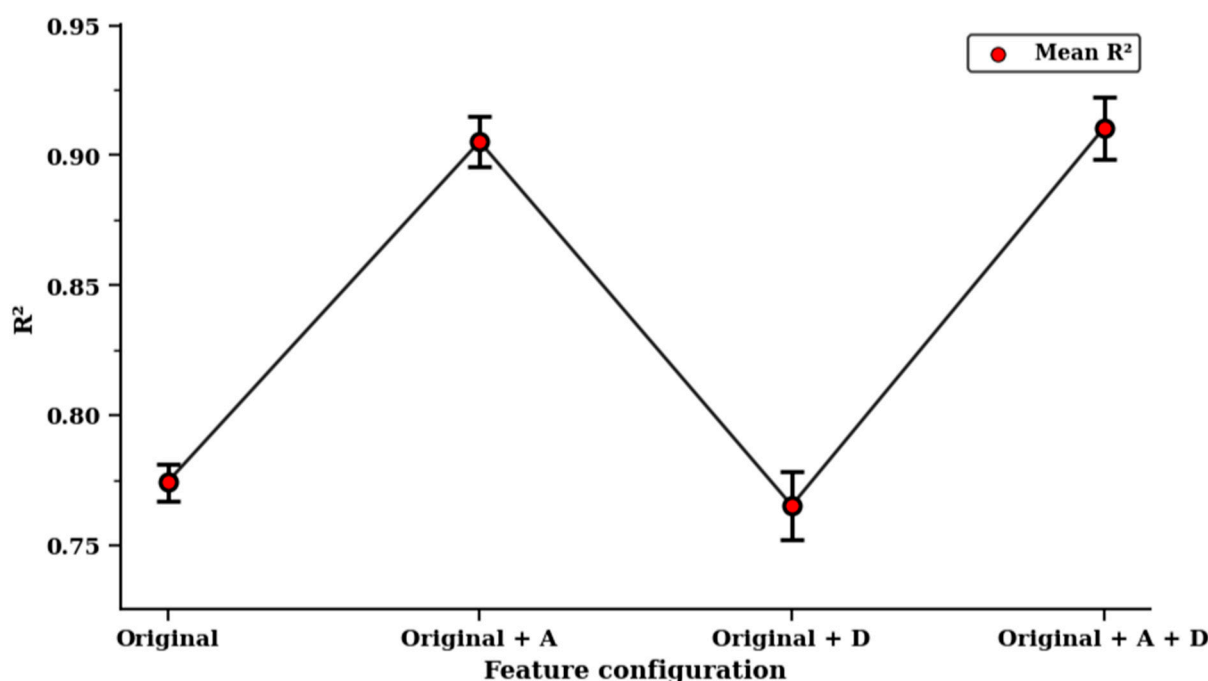


Figure 11. Wavelet-component sensitivity analysis for the best-performing model based on cross-validated R^2 .

Figure 11 indicates that the main wavelet-related improvement was associated with the approximation components. The Original + A configuration improved R^2 relative to the baseline, whereas Original + D performed below the baseline. This suggests that approximation components captured useful low-frequency variability related to alkalinity behavior, while detail components alone mainly represented high-frequency fluctuations or noise. As a result, adding detail components without the approximation components likely increased feature noise and model complexity without adding useful low-frequency structure, causing the model to generalize less effectively than the baseline using only original predictors. The comparable performance of Original + A and Original + A + D indicates that detail components added limited additional benefit once the approximation components were included.

The feature-selection sensitivity analysis was conducted to evaluate model robustness under reduced predictor sets and to examine the effect of multicollinearity between electri-

cal conductivity (EC) and chloride (Cl). As shown in Figure 12, the highest performance was achieved by RF under the wavelet-enhanced configuration using all original predictors, together with their wavelet-derived approximation and detail components. Across the tested scenarios, model performance generally declined when only the top two predictors were retained, indicating that alkalinity prediction benefits from multivariate information rather than a very limited predictor subset. The “No EC” and “No Cl” scenarios maintained relatively high R^2 values, especially for RF and XGB, suggesting that the high EC-Cl correlation did not solely control model performance because either salinity-related variable could partly compensate for the removal of the other. Overall, the wavelet-enhanced configuration showed stronger and more stable performance across most feature-selection scenarios, confirming RF as the most reliable model for alkalinity prediction.

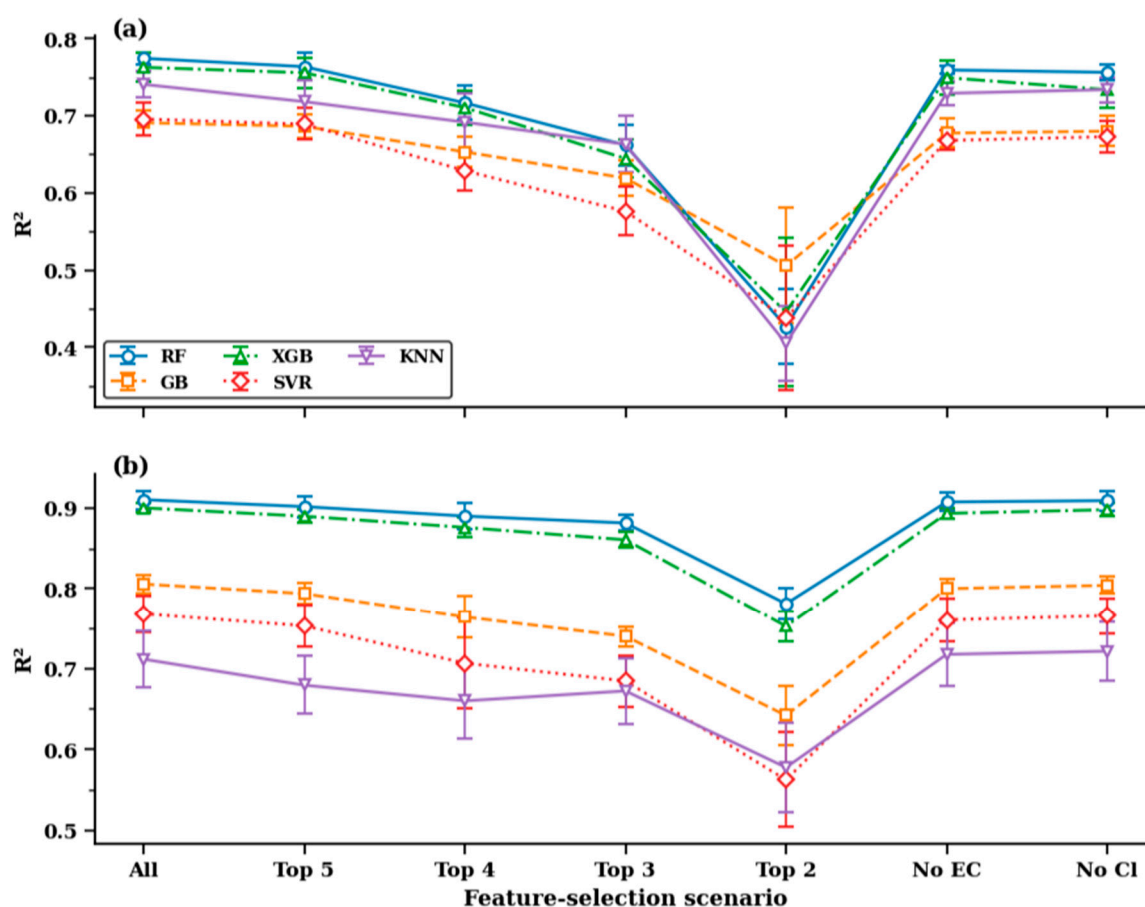


Figure 12. Feature-selection sensitivity analysis for all evaluated models under (a) baseline and (b) wavelet-enhanced configurations. Symbols indicate mean cross-validated R^2 values, and error bars represent standard deviations (SD) across the five validation folds.

4. Discussion

4.1. Model Performance and Algorithmic Behavior

The results demonstrate that ensemble-based machine learning models outperform single-estimator and distance-based approaches in predicting seawater alkalinity. This superior performance arises from the ability of ensemble methods to combine multiple decision trees, thereby reducing variance and improving generalization in systems characterized by nonlinear behavior and interacting predictors [22,23]. In coastal seawater environments, alkalinity is governed by coupled physicochemical processes, including carbonate equilibria, salinity effects, and operational controls, which introduce nonlinear dependencies that are difficult to capture using linear or distance-based formulations.

Ensemble models are particularly well suited to such conditions because they can represent complex response surfaces without requiring explicit assumptions regarding the functional form of predictor-response relationships [44].

Among the evaluated algorithms, RF consistently emerges as the best-performing model, achieving the highest coefficient of determination together with the lowest RMSE and MAE across both baseline and wavelet-enhanced configurations. The strong performance of RF can be attributed to its use of bootstrap aggregation and randomized feature selection, which enhance robustness to noise, multicollinearity, and overfitting [22,55]. Furthermore, RF is well-suited to modeling the nonlinear interactions that govern alkalinity dynamics, as it can naturally partition the predictor space to capture combined effects of pH, dissolved ions, temperature, and treatment-related variables. This ability to account for interaction effects and heterogeneous responses explains the consistent superiority of RF in modeling alkalinity under complex and variable coastal seawater conditions [44,45].

Although Long Short-Term Memory (LSTM) models are widely used for sequential time-series prediction because they can learn temporal dependencies [56], their conventional application is more suitable for continuous or regularly structured time-series inputs. In this study, the quality-controlled record contained temporal gaps after the removal of missing and incomplete observations. Therefore, applying LSTM would require additional missing-data treatment or irregular-time modeling assumptions [57]. For this reason, the present study focused on robust and interpretable machine learning models suitable for irregular tabular water-quality data, while LSTM-based approaches are recommended for future work when longer and more continuous records become available.

The strong influence of electrical conductivity and chloride on alkalinity prediction reflects the dominant role of salinity under arid coastal conditions. In the Arabian Gulf, high evaporation rates relative to precipitation lead to progressive concentration of dissolved ions, resulting in elevated salinity and alkalinity. This evaporation-driven enrichment enhances both conservative ions, such as chloride, and carbonate species, explaining the close association between alkalinity and conductivity-related variables. This relationship is consistent with observations from coastal and semi-enclosed marine systems, where salinity strongly influences seawater chemistry and buffering capacity [1,9], as well as findings from coastal monitoring and environmental assessment studies.

From a geochemical perspective, alkalinity is governed primarily by the carbonate system, including bicarbonate (HCO_3^-) and carbonate (CO_3^{2-}) ions, with additional contributions from minor species such as borate. The moderate importance of pH observed in this study aligns with its role in regulating carbonate speciation and buffering capacity, while temperature influences equilibrium constants and reaction kinetics. Although pH contributed to alkalinity prediction, its interpretation should be treated cautiously because the observed pH range was narrow. Therefore, the pH contribution likely reflects its role within the multivariate carbonate-chemistry context rather than large independent pH variability. These physicochemical controls are well established in studies of natural water chemistry and marine biogeochemical processes [10,11,13]. The combined effects of salinity concentration and carbonate equilibrium result in nonlinear relationships between alkalinity and environmental variables, which are effectively captured by ensemble-based ML models.

4.2. Effect of Wavelet-Based Feature Enrichment

The integration of wavelet-based feature enrichment leads to consistent improvements in predictive performance across all evaluated machine learning models. By decomposing the original water-quality predictors into components that represent different temporal scales, wavelet analysis enables the models to capture both short-term fluctuations and

longer-term trends embedded in the alkalinity signal. These multiscale representations provide additional explanatory information that is not fully preserved in the raw variables, resulting in increased coefficients of determination and reduced error metrics. The observed reduction in variability across cross-validation folds further indicates that wavelet-enriched models generalize more reliably under different training-testing partitions. The benefit of wavelet-based preprocessing is consistent with previous hydrological and environmental modeling studies, where multiscale signal decomposition improved the representation of complex and nonstationary processes [20,58].

The magnitude of improvement associated with wavelet-based feature enrichment varies among algorithms, reflecting differences in their capacity to exploit high-dimensional and nonlinear feature spaces. Ensemble-based models, particularly RF and XGB, benefit most from the inclusion of wavelet-derived predictors, as their tree-based structures can implicitly select relevant features and model nonlinear interactions across multiple scales. In contrast, SVR and KNN exhibit more modest gains, suggesting limitations in fully leveraging the expanded feature set despite improvements in error reduction. These findings indicate that wavelet decomposition is most effective when combined with flexible learning algorithms capable of handling multiscale, nonlinear relationships, reinforcing the suitability of wavelet-enhanced ensemble models for alkalinity estimation in complex coastal seawater systems [24,44].

The component-level pattern observed in Figure 11 provides additional insight into the source of the wavelet-related improvement. The stronger performance of the Original + A configuration indicates that the approximation components captured smoother low-frequency variability that is more consistent with gradual alkalinity dynamics in the monitored seawater system. In contrast, the lower performance of the Original + D configuration suggests that the detail components mainly represented high-frequency fluctuations or short-term noise that were less informative for RF-based alkalinity prediction when used without the approximation components. Because these detail components were added to the original predictors without the stabilizing low-frequency information provided by the approximation components, they likely increased feature noise and model complexity, leading to poorer generalization than the baseline model. The comparable performance of the Original + A and Original + A + D configurations further indicates that adding detail components provided limited marginal benefit once the smoother low-frequency structure had already been incorporated.

4.3. Diagnostic Evaluation and Model Reliability

The diagnostic analyses provide additional insight into the reliability and generalization capability of the baseline and wavelet-enhanced RF models. The observed-predicted relationships demonstrate strong agreement across the full alkalinity range, with predictions closely aligned along the 1:1 reference line, indicating effective capture of the underlying response behavior. Residuals are symmetrically distributed around zero with no systematic trends relative to predicted values, suggesting the absence of structural bias or persistent over- or underestimation. Such residual behavior is commonly interpreted as evidence of a well-calibrated model in data-driven hydrological and environmental applications [44,59].

The residual distributions and absolute error patterns further confirm model robustness. Errors remain generally small and approximately symmetric, with only modest increases in dispersion at higher alkalinity values. This localized behavior is likely associated with increased process variability or operational influences rather than deficiencies in the modeling framework itself. This residual pattern is consistent with previous machine learning-based water-quality prediction studies, where limited heteroscedasticity at

extreme values is commonly observed in complex environmental systems [20,60]. Overall, these diagnostic results support the stability and reliability of the wavelet-enhanced RF model for alkalinity estimation under variable coastal seawater conditions.

4.4. Practical Implications for Coastal Seawater Monitoring

The strong predictive performance of the RF model, particularly when enhanced with wavelet-based features, highlights the practical value of the proposed framework for coastal seawater monitoring applications. Reliable estimation of alkalinity is critical for managing chemical stability, optimizing treatment processes, and supporting downstream operations such as desalination and corrosion control. Traditional alkalinity measurements are often laboratory-based and subject to sampling delays, whereas the presented data-driven approach enables continuous estimation using routinely monitored water-quality variables [16,19]. This capability can enhance operational responsiveness and support real-time decision-making in coastal monitoring systems [20,21].

The robustness of the wavelet-enhanced RF model across cross-validation folds further suggests that the framework can maintain stable predictive performance under variable monitoring conditions. This is important for coastal seawater systems, where alkalinity may be influenced by seasonal forcing, operational changes, and short-term environmental fluctuations. From an operational perspective, improved model stability can reduce prediction uncertainty and support more consistent decision-making in monitoring and treatment applications [44,60].

Beyond the specific study site, the proposed methodology demonstrates potential for application to other coastal or marine environments with similar monitoring infrastructures. The reliance on commonly measured water-quality parameters, combined with a flexible ensemble-learning framework, makes the approach adaptable to different coastal monitoring settings, provided that appropriate site-specific validation and recalibration are performed when needed. As coastal monitoring programs increasingly adopt automated sensing and data-driven analytics, integrating multiscale signal processing with ensemble models offers a scalable and reliable pathway for improving water-quality assessment and supporting sustainable coastal water management strategies [20,24].

Some limitations should be considered when interpreting the results. The analysis was based on data from a single coastal monitoring station; therefore, future studies should evaluate the proposed framework using additional monitoring sites. The quality-controlled dataset also contained temporal gaps, which should be considered when interpreting the wavelet-enhanced results and when comparing the present approach with sequence-based models. Future work could evaluate gap-aware wavelet approaches, longer continuous monitoring records, and separate hyperparameter optimization for each model-feature configuration to further assess potential performance improvements.

5. Conclusions

In this study, we evaluated the performance of several machine learning models for estimating seawater alkalinity using routinely monitored water-quality variables from a coastal monitoring station. The results indicated that ensemble-based algorithms generally outperformed single-estimator and distance-based approaches, reflecting their stronger capability to capture nonlinear relationships and interaction effects commonly present in environmental systems. Among the evaluated models, RF consistently achieved the best performance, characterized by higher coefficients of determination, lower error magnitudes, and greater fold-to-fold stability under both baseline and wavelet-enhanced configurations. This consistent behavior highlighted the suitability of RF for modeling complex alkalinity dynamics governed by coupled physical and chemical processes.

The incorporation of wavelet-decomposed features led to systematic performance improvements across all evaluated models by enhancing the representation of both short-term variability and longer-term trends in alkalinity behavior. By decomposing the original predictors into multiscale components, the modeling framework was better able to capture nonstationary patterns that were difficult to represent using raw variables alone. These improvements were most pronounced for tree-based ensemble models, which were well suited to handling high-dimensional feature spaces and nonlinear interactions. Diagnostic analyses further supported the reliability of the wavelet-enhanced RF model, showing a strong agreement between observed and predicted values and largely unbiased residuals across the alkalinity range.

From an applied perspective, the proposed framework provides a practical and scalable approach for continuous alkalinity estimation in coastal seawater monitoring systems. Its reliance on commonly measured water-quality parameters facilitates integration into automated monitoring and decision-support platforms, particularly in desalination and coastal water-treatment facilities where timely chemical control is critical. The demonstrated robustness of the wavelet-enhanced RF model suggests strong potential for operational deployment and transferability to similar coastal environments. Future work may focus on extending the framework to additional regions and incorporating real-time sensor data to support proactive coastal water-quality management. Future studies should also consider longer and more continuous monitoring records, independent temporal testing, and configuration-specific hyperparameter optimization to further evaluate model transferability and robustness.

Author Contributions: S.H.A.: formal analysis, investigation, writing—original draft. Y.A.: writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ongoing Research Funding program (ORF-2026-1807), King Saud University, Riyadh, Saudi Arabia.

Data Availability Statement: The data are available from the authors upon reasonable request due to governmental privacy and data-sharing regulations imposed by the original data-providing agencies.

Conflicts of Interest: The author declares no competing interests.

References

1. Hosseini, H.; Saadaoui, I.; Moheimani, N.; Al Saidi, M.; Al Jamali, F.; Al Jabri, H.; Hamadou, R.B. Marine Health of the Arabian Gulf: Drivers of Pollution and Assessment Approaches Focusing on Desalination Activities. *Mar. Pollut. Bull.* **2021**, *164*, 112085. [[CrossRef](#)] [[PubMed](#)]
2. Tremblay, L.A.; Chariton, A.A.; Li, M.-S.; Zhang, Y.; Horiguchi, T.; Ellis, J.I. Monitoring the Health of Coastal Environments in the Pacific Region-A Review. *Toxics* **2023**, *11*, 277. [[CrossRef](#)] [[PubMed](#)]
3. Lattemann, S.; Höpner, T. Environmental Impact and Impact Assessment of Seawater Desalination. *Desalination* **2008**, *220*, 1–15. [[CrossRef](#)]
4. Ghaffour, N.; Missimer, T.M.; Amy, G.L. Technical Review and Evaluation of the Economics of Water Desalination: Current and Future Challenges for Better Water Supply Sustainability. *Desalination* **2013**, *309*, 197–207. [[CrossRef](#)]
5. Vasou, P.; Krokos, G.; Langodan, S.; Sofianos, S.; Hoteit, I. Contribution of Surface and Lateral Forcing to the Arabian Gulf Warming Trend. *Front. Mar. Sci.* **2024**, *10*, 1260058. [[CrossRef](#)]
6. Ibrahim, H.D.; Xue, P.; Eltahir, E.A.B. Multiple Salinity Equilibria and Resilience of Persian/Arabian Gulf Basin Salinity to Brine Discharge. *Front. Mar. Sci.* **2020**, *7*, 550181. [[CrossRef](#)]
7. Campos, E.J.D.; Gordon, A.L.; Kjerfve, B.; Vieira, F.; Cavalcante, G. Freshwater Budget in the Persian (Arabian) Gulf and Exchanges at the Strait of Hormuz. *PLoS ONE* **2020**, *15*, e0233090. [[CrossRef](#)] [[PubMed](#)]
8. Lachkar, Z.; Mehari, M.; Lévy, M.; Paparella, F.; Burt, J.A. Recent Expansion and Intensification of Hypoxia in the Arabian Gulf and Its Drivers. *Front. Mar. Sci.* **2022**, *9*, 891378. [[CrossRef](#)]
9. Paparella, F.; D'Agostino, D.; Burt, J.A. Long-Term, Basin-Scale Salinity Impacts from Desalination in the Arabian/Persian Gulf. *Sci. Rep.* **2022**, *12*, 20549. [[CrossRef](#)] [[PubMed](#)]

10. Florence, T.M.; Batley, G.E.; Benes, P. Chemical Speciation in Natural Waters. *C R C Crit. Rev. Anal. Chem.* **1980**, *9*, 219–296. [[CrossRef](#)]
11. Middelburg, J.J.; Soetaert, K.; Hagens, M. Ocean Alkalinity, Buffering and Biogeochemical Processes. *Rev. Geophys.* **2020**, *58*, e2019RG000681. [[CrossRef](#)] [[PubMed](#)]
12. Rheuban, J.E.; Gassett, P.R.; McCorkle, D.C.; Hunt, C.W.; Liebman, M.; Bastidas, C.; O'Brien-Clayton, K.; Pimenta, A.R.; Silva, E.; Vlahos, P.; et al. Synoptic Assessment of Coastal Total Alkalinity through Community Science. *Environ. Res. Lett.* **2021**, *16*, 024009. [[CrossRef](#)] [[PubMed](#)]
13. Egleston, E.S.; Sabine, C.L.; Morel, F.M.M. Revelle Revisited: Buffer Factors That Quantify the Response of Ocean Chemistry to Changes in DIC and Alkalinity. *Glob. Biogeochem. Cycles* **2010**, *24*, GB1002. [[CrossRef](#)]
14. Schaap, A.; Papadimitriou, S.; Mawji, E.; Walk, J.; Hammermeister, E.; Mowlem, M.; Loucaides, S. Autonomous Sensor for In Situ Measurements of Total Alkalinity in the Ocean. *ACS Sens.* **2025**, *10*, 795–803. [[CrossRef](#)] [[PubMed](#)]
15. Sonnichsen, C.; Atamanchuk, D.; Hendricks, A.; Morgan, S.; Smith, J.; Grundke, I.; Luy, E.; Sieben, V.J. An Automated Microfluidic Analyzer for In Situ Monitoring of Total Alkalinity. *ACS Sens.* **2023**, *8*, 344–352. [[CrossRef](#)] [[PubMed](#)]
16. Seelmann, K.; Aßmann, S.; Körtzinger, A. Characterization of a Novel Autonomous Analyzer for Seawater Total Alkalinity: Results from Laboratory and Field Tests. *Limnol. Oceanogr. Methods* **2019**, *17*, 515–532. [[CrossRef](#)]
17. Rosenau, N.A.; Galavotti, H.; Yates, K.K.; Bohlen, C.C.; Hunt, C.W.; Liebman, M.; Brown, C.A.; Pacella, S.R.; Largier, J.L.; Nielsen, K.J.; et al. Integrating High-Resolution Coastal Acidification Monitoring Data Across Seven United States Estuaries. *Front. Mar. Sci.* **2021**, *8*, 679913. [[CrossRef](#)] [[PubMed](#)]
18. Shyu, H.-Y.; Castro, C.J.; Bair, R.A.; Lu, Q.; Yeh, D.H. Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System. *ACS Environ. Au* **2023**, *3*, 308–318. [[CrossRef](#)] [[PubMed](#)]
19. Qiu, L.; Jiang, K.; Li, Q.; Yuan, D.; Chen, J.; Yang, B.; Achterberg, E.P. Variability of Total Alkalinity in Coastal Surface Waters Determined Using an in-Situ Analyzer in Conjunction with the Application of a Neural Network-Based Prediction Model. *Sci. Total Environ.* **2024**, *908*, 168271. [[CrossRef](#)] [[PubMed](#)]
20. Nourani, V.; Hosseini Baghanam, A.; Adamowski, J.; Kisi, O. Applications of Hybrid Wavelet–Artificial Intelligence Models in Hydrology: A Review. *J. Hydrol.* **2014**, *514*, 358–377. [[CrossRef](#)]
21. Mosavi, A.; Hosseini, F.S.; Choubin, B.; Abdolshahnejad, M.; Gharechae, H.; Lahijanzadeh, A.; Dineva, A.A.; Mosavi, A.; Hosseini, F.S.; Choubin, B.; et al. Susceptibility Prediction of Groundwater Hardness Using Ensemble Machine Learning Models. *Water* **2020**, *12*, 2770. [[CrossRef](#)]
22. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
23. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
24. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
25. Goma, M.N.; Mulla, D.J.; Galzki, J.C.; Sheikho, K.M.; Alhazmi, N.M.; Mohamed, H.E.; Hannachi, I.; Abouwarda, A.M.; Hassan, E.A.; Carmichael, W.W.; et al. Red Sea MODIS Estimates of Chlorophyll a and Phytoplankton Biomass Risks to Saudi Arabian Coastal Desalination Plants. *J. Mar. Sci. Eng.* **2020**, *9*, 11. [[CrossRef](#)]
26. Rajabi-Kiasari, S.; Hasanlou, M. An Efficient Model for the Prediction of SMAP Sea Surface Salinity Using Machine Learning Approaches in the Persian Gulf. *Int. J. Remote Sens.* **2020**, *41*, 3221–3242. [[CrossRef](#)]
27. Abedini, M.; Esmailpour, Y.; Gholami, H.; Bazrafshan, O.; Nafarzadegan, A.R. Change Analysis of Surface Water Clarity in the Persian Gulf and the Oman Sea by Remote Sensing Data and an Interpretable Deep Learning Model. *Environ. Sci. Pollut. Res. Int.* **2025**, *32*, 5987–6004. [[CrossRef](#)] [[PubMed](#)]
28. García-Ibáñez, M.I.; Gualart, E.F.; Lucas, A.; Pascual, J.; Gasol, J.M.; Marrasé, C.; Calvo, E.; Pelejero, C. Two New Coastal Time-Series of Seawater Carbonate System Variables in the NW Mediterranean Sea: Rates and Mechanisms Controlling pH Changes. *Front. Mar. Sci.* **2024**, *11*, 1348133. [[CrossRef](#)]
29. Qiu, L.; Esposito, M.; Martínez-Cabanas, M.; Achterberg, E.P.; Li, Q. Autonomous High-Frequency Time-Series Observations of Total Alkalinity in Dynamic Estuarine Waters. *Mar. Chem.* **2023**, *257*, 104332. [[CrossRef](#)]
30. Broullón, D.; Pérez, F.F.; Velo, A.; Hoppema, M.; Olsen, A.; Takahashi, T.; Key, R.M.; Tanhua, T.; González-Dávila, M.; Jeansson, E.; et al. A Global Monthly Climatology of Total Alkalinity: A Neural Network Approach. *Earth Syst. Sci. Data* **2019**, *11*, 1109–1127. [[CrossRef](#)]
31. Grbčić, L.; Družeta, S.; Mauša, G.; Lipić, T.; Lušić, D.V.; Alvir, M.; Lučin, I.; Sikirica, A.; Davidović, D.; Travaš, V.; et al. Coastal Water Quality Prediction Based on Machine Learning with Feature Interpretation and Spatio-Temporal Analysis. *Environ. Model. Softw.* **2022**, *155*, 105458. [[CrossRef](#)]
32. Mohammed, M.A.A.; Miklós, R.; Darabos, E.; Szabó, N.P.; Szűcs, P. Chemometrics of Karst Systems: Monitoring Climate Impacts on Groundwater Quality in Garadna Spring, Northern Hungary, Using Self-Organizing Maps and Wavelet Transform Analysis. *Results Eng.* **2025**, *28*, 108298. [[CrossRef](#)]

33. Takeshita, Y.; Frieder, C.A.; Martz, T.R.; Ballard, J.R.; Feely, R.A.; Kram, S.; Nam, S.; Navarro, M.O.; Price, N.N.; Smith, J.E. Including High Frequency Variability in Coastal Ocean Acidification Projections. *Biogeosciences* **2015**, *12*, 5853–5870. [\[CrossRef\]](#)
34. Fettweis, M.; Riethmüller, R.; Van der Zande, D.; Desmit, X. Sample Based Water Quality Monitoring of Coastal Seas: How Significant Is the Information Loss in Patchy Time Series Compared to Continuous Ones? *Sci. Total Environ.* **2023**, *873*, 162273. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Nascimento, Â.; Biguino, B.; Borges, C.; Cereja, R.; Cruz, J.P.C.; Sousa, F.; Dias, J.; Brotas, V.; Palma, C.; Brito, A.C. Tidal Variability of Water Quality Parameters in a Mesotidal Estuary (Sado Estuary, Portugal). *Sci. Rep.* **2021**, *11*, 23112. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Al-Kaabi, A.; Al-Sulaiti, H.; Al-Ansari, T.; Mackey, H.R. Assessment of Water Quality Variations on Pretreatment and Environmental Impacts of SWRO Desalination. *Desalination* **2021**, *500*, 114831. [\[CrossRef\]](#)
37. Nelson, N.G.; Muñoz-Carpena, R.; Neale, P.J.; Tzortziou, M.; Megonigal, J.P. Temporal Variability in the Importance of Hydrologic, Biotic, and Climatic Descriptors of Dissolved Oxygen Dynamics in a Shallow Tidal-Marsh Creek. *Water Resour. Res.* **2017**, *53*, 7103–7120. [\[CrossRef\]](#)
38. Mallat, S. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way | Guide Books | ACM Digital Library. Available online: <https://dl.acm.org/doi/book/10.5555/1525499> (accessed on 15 January 2026).
39. Torrence, C.; Compo, G.P. A Practical Guide to Wavelet Analysis. *Bull. Am. Meteorol. Soc.* **1998**, *79*, 61–78. [\[CrossRef\]](#)
40. Nourani, V.; Kisi, Ö.; Komasi, M. Two Hybrid Artificial Intelligence Approaches for Modeling Rainfall–Runoff Process. *J. Hydrol.* **2011**, *402*, 41–59. [\[CrossRef\]](#)
41. Wang, Y.; Yuan, Y.; Pan, Y.; Fan, Z.; Wang, Y.; Yuan, Y.; Pan, Y.; Fan, Z. Modeling Daily and Monthly Water Quality Indicators in a Canal Using a Hybrid Wavelet-Based Support Vector Regression Structure. *Water* **2020**, *12*, 1476. [\[CrossRef\]](#)
42. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2000; ISBN 978-1-4419-3160-3.
43. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [\[CrossRef\]](#)
44. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009; ISBN 978-0-387-84857-0.
45. Louppe, G.; Wehenkel, L.; Suter, A.; Geurts, P. Understanding Variable Importances in Forests of Randomized Trees. In *Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 1, 5 December 2013*; Curran Associates Inc.: Red Hook, NY, USA, 2013; Volume 1, pp. 431–439.
46. Daubechies, I. *Ten Lectures on Wavelets*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1992; ISBN 978-0-89871-274-2.
47. Willmott, C.J.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.* **2005**, *30*, 79–82. [\[CrossRef\]](#)
48. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2, 20 August 1995*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp. 1137–1143.
49. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [\[CrossRef\]](#)
50. Xiong, Y.; Zhang, T.; Sun, X.; Yuan, W.; Gao, M.; Wu, J.; Han, Z.; Xiong, Y.; Zhang, T.; Sun, X.; et al. Groundwater Quality Assessment Based on the Random Forest Water Quality Index—Taking Karamay City as an Example. *Sustainability* **2023**, *15*, 14477. [\[CrossRef\]](#)
51. Masood, A.; Niazkar, M.; Zakwan, M.; Piraei, R.; Masood, A.; Niazkar, M.; Zakwan, M.; Piraei, R. A Machine Learning-Based Framework for Water Quality Index Estimation in the Southern Bug River. *Water* **2023**, *15*, 3543. [\[CrossRef\]](#)
52. Reza Nikoo, M.; Bahrami, N.; Madani, K.; Al-Rawas, G.; Vanda, S.; Nazari, R. A Robust Decision-Making Framework to Improve Reservoir Water Quality Using Optimized Selective Withdrawal Strategies. *J. Hydrol.* **2024**, *635*, 131153. [\[CrossRef\]](#)
53. Arias-Rodriguez, L.F.; Tüzün, U.F.; Duan, Z.; Huang, J.; Tuo, Y.; Disse, M.; Arias-Rodriguez, L.F.; Tüzün, U.F.; Duan, Z.; Huang, J.; et al. Global Water Quality of Inland Waters with Harmonized Landsat-8 and Sentinel-2 Using Cloud-Computed Machine Learning. *Remote Sens.* **2023**, *15*, 1390. [\[CrossRef\]](#)
54. Carbureanu, M.; Gheorghe, C.G.; Carbureanu, M.; Gheorghe, C.G. A Machine Learning-Based Data-Driven Model for Predicting Wastewater Quality Parameters in the Industrial Domain. *Appl. Sci.* **2026**, *16*, 694. [\[CrossRef\]](#)
55. Cutler, D.R.; Edwards, T.C., Jr.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random Forests for Classification in Ecology. *Ecology* **2007**, *88*, 2783–2792. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* **2018**, *8*, 6085. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Wang, D.; Ding, H.; Singh, V.P.; Shang, X.; Liu, D.; Wang, Y.; Zeng, X.; Wu, J.; Wang, L.; Zou, X. A Hybrid Wavelet Analysis–Cloud Model Data-Extending Approach for Meteorologic and Hydrologic Time Series. *J. Geophys. Res. Atmos.* **2015**, *120*, 4057–4071. [\[CrossRef\]](#)

59. Cook, R.D. Influential Observations in Linear Regression. *J. Am. Stat. Assoc.* **1979**, *74*, 169–174. [[CrossRef](#)]
60. Wilks, D. *Statistical Methods in the Atmospheric Sciences*; Academic Press: London, UK, 2011; ISBN 978-0-12-385022-5.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.