Scientific Reports

https://doi.org/10.1038/s41598-026-48252-5

Article in Press

Forecasting toxic metal concentrations in an inland sea ecosystem with machine learning algorithms

Received: 1 March 2026

Accepted: 7 April 2026

Published online: 13 April 2026

Cite this article as: Ucan A., Tak N., Hocaoglu-Ozyigit A. *et al.* Forecasting toxic metal concentrations in an inland sea ecosystem with machine learning algorithms. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-48252-5

Aylin Ucan, Nihat Tak, Asli Hocaoglu-Ozyigit & Ibrahim Ilker Ozyigit

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Forecasting Toxic Metal Concentrations in an Inland Sea Ecosystem with Machine Learning Algorithms

Aylin Ucan[1], Nihat Tak[1,*], Asli Hocaoglu-Ozyigit[2], and Ibrahim Ilker Ozyigit[2]

[1]Marmara University, Department of Statistics, Istanbul, 34730, Turkiye
[2]Marmara University, Department of Biology, Istanbul, 34730, Turkiye
[*]nihat.tak@marmara.edu.tr, nihattak@gmail.com

ABSTRACT

In recent years, statistical and data-driven modeling approaches have been increasingly employed to predict element concentrations and to examine relationships among environmental features. In this context, the integration of feature selection techniques with machine learning models enhances model generalization and reduces model complexity by enabling the identification of key elements that are strongly associated with the target feature. This study applies machine learning models to investigate the relationships between Aluminum (Al) and other elements and to predict Al concentration levels in an inland marine ecosystem. Specifically, the study evaluates whether accurate predictions can be achieved using a reduced subset of informative elements rather than the full feature set. The findings demonstrate that machine learning methods, when combined with feature selection, can successfully predict Al concentrations while yielding more interpretable models based on a limited number of significant elements.

ARTICLE IN PRESS

1 Introduction

Nowadays, the environment is confronting many significant pollution challenges, particularly due to the escalating prevalence of human-induced activities. Rapid industrialization, population growth, urbanization, and irregular use of natural resources have contaminated essential environmental components, including water, air, and soil. This has led to the disruption of ecological balances[1,2]. One of the most notable forms of pollution is the accumulation of toxic metals/metalloids in the environment, which results in irreversible damage. Toxic metals are defined as elements that are toxic to living organisms at concentrations above a certain threshold, and which pose a permanent health risk due to their accumulation in biological systems. Toxic metals/metalloids that cause permanent effects, especially in aquatic environments, include arsenic (As), aluminum (Al), cadmium (Cd), chromium (Cr), lead (Pb), and mercury (Hg), The accumulation and distribution of toxic metals in saltwater ecosystems is a complex phenomenon[3]. Human activities, including industrial operations, agriculture, urbanization, and mining, contribute to the increased accumulation of heavy metals in these environments[4]. Through their participation in chemical reactions, these metals are activated by oxidation and dissolution, entering the food chain and causing irreversible damage to human health and water quality. A more profound comprehension of the accumulation and consequences of toxic metals in saltwater ecosystems is imperative for the sustainable management of marine environments[5,6].

Among these toxic metals, Al is the third most abundant element in the Earth's crust, after silicon and oxygen, and is transported to aquatic environments through both natural processes and anthropogenic activities[7,8]. Geological processes, especially weathering, and erosion of rocks, transport Al through soils and sediments into aquatic ecosystems such as oceans, seas, lakes, and rivers. In addition, volcanic activities and wind-borne soil particles also play important roles in the transfer of Al to water through natural sources. Anthropogenic sources refer to Al released into nature as a result of human activities. Al is transported to the marine ecosystem, especially as a result of mining and agricultural activities, industrial waste, burning of fossil fuels, and maritime transportation. Chemical fertilizers and pesticides from agricultural fields are also important sources of pollution. Industrial ports and discharges associated with alumina (Al_2O_3) and Al production can release high amounts of Al into the marine environment, posing serious ecological risks, especially to marine species[9–13].

Al can exist in many different forms in the aquatic environment, such as various inorganic monomeric species (Al^{3+}, $AlOH^{2+}$, $Al(OH)_2^+$, $Al(OH)_3^0$, $Al(OH)_4^-$) and amorphous amorphous $Al(OH)_3$[9]. The solubility and chemical form of Al in aqueous environments may vary depending on factors such as the pH level of the water, temperature, the amount of dissolved organic carbon (DOC), and the hardness of the water[14–16]. Therefore, the toxicity and bioavailability of Al produce varying effects on aquatic organisms as a result of environmental factors.

High concentrations of Al can cause serious toxic effects on marine organisms. In saline environments where acidification increases, Al solubility rises, and consequently, the toxic potential of Al increases[9,17]. These impacts disrupt the vital

ARTICLE IN PRESS

functions of organisms such as growth, reproduction and respiration, leading to negative impacts on biodiversity and ecosystem balance[3,12]. Studies on water and sediment samples have also shown that Al is generally found in higher concentrations than other toxic metals, and therefore there is increasing awareness of the toxic effects of Al. In addition, Al is present in high concentrations, especially in coastal waters; this is largely due to anthropogenic activities[18–21]. Like many other toxic metals, Al tends to bioaccumulate in both terrestrial and marine organisms. Bioaccumulation in marine organisms is usually more pronounced under low pH conditions, where water-soluble forms of Al become more toxic. Bioaccumulated Al can be carried up the food chain via marine organisms, posing greater risks to both predators that consume these organisms and humans[22–24]. The adverse effects of Al on human health include neurodegenerative diseases (especially Alzheimer's and epilepsy), nervous system toxicity, lung diseases (fibrosis and decreased lung function), and bone diseases. In addition, long-term exposure to Al can lead to serious health problems such as hormonal imbalances, kidney damage, and cancer[25–27].

Therefore, continuous monitoring of the values of Al and/or other toxic metals in water and sediment samples in marine ecosystems is crucial for both the protection of the ecosystem itself and the assurance of human health. Knowledge of these values is also critical for understanding bioaccumulation processes in marine ecosystems. In addition, information about the levels of toxic metals in different localities provides a significant contribution to understanding regional differences and developing pollution control strategies[28,29]. In this context, marine elemental monitoring studies are indispensable for sustainable marine management and environmental protection strategies.

Determining elemental concentrations in marine water and sediments is fundamental for environmental monitoring and pollution assessment; however, large-scale and long-term monitoring programs often generate complex datasets that are difficult to interpret through direct measurements alone. In this context, statistical and Machine Learning (ML) based approaches are increasingly used as complementary tools in environmental analysis to examine multivariate relationships and support environmental interpretation within monitored systems. Several recent studies have demonstrated the effectiveness of ML approaches in environmental prediction tasks. Yaseen et al. focused on predicting heavy metal concentrations in marine sediments using ML and metaheuristic optimization techniques and Deng et al. evaluated water and stream sediment quality in a coal mining area through chemical and statistical analyses[30,31]. Bhagat et al. and Mert & Kasapoğulları applied artificial intelligence (AI)-based models to predict heavy metal pollution in aquatic environments[32,33]. Yolcu et al. modeled the effects of essential heavy metals on environmental pollution using linear and nonlinear ANN-based approaches[34]. Chen et al. and Cui et al. applied ML approaches for analyzing and assessing environmental pollution and related risks[35,36]. Gryech et al. developed ML- and IoT-based frameworks for air pollution prediction and environmental health analysis, including the MOREAIR dataset[37,38]. Xu et al. proposed an ML-based approach for predicting water quality under incomplete datasets, achieving up to 75% accuracy despite 57% missing data[39]. Nguyen et al. applied multiple regression-based ML models to predict rare earth element concentrations, identifying Random Forest (RF) as the most effective model[40]. Han et al. compared LSTM, RF, Decision Tree, and SVM models for Water Quality Index prediction, demonstrating that LSTM achieved the highest accuracy due to its ability to capture temporal dependencies[41]. Chen et al. further proposed an interpretable ensemble ML framework for predicting dissolved oxygen levels in estuarine systems[42]. These studies collectively highlight the increasing importance and reliability of ML approaches in environmental prediction tasks.

In parallel, feature selection has become an essential component of machine learning applications, as it enhances model accuracy while reducing computational complexity in environmental prediction tasks. Dai et al. developed a hybrid model that uses XGBoost-based feature selection to eliminate redundant features, followed by MSCNN and GA-optimized LSTM for PM2.5 prediction[43]. Zhang et al. proposed a Bayesian optimization-based XGBoost-RFE feature selection method to determine the optimal feature subset, significantly improving ozone prediction performance[44]. In another study, Demir and Sahin applied GA-based feature selection and hyperparameter optimization to RF, SVM, and XGBoost models, demonstrating that XGBoost achieved the highest accuracy, especially when trained with balanced data[45]. Garabaghi et al. employed a wrapper-based feature selection approach and showed that dimensionality reduction improves prediction accuracy, with Random Forest emerging as the most robust model for dissolved oxygen prediction[46]. Ahajjam et al. used GA based feature selection to identify the most relevant features, then highlighted the key factors influencing wildfire occurrence, and observed significant improvements in prediction accuracy[47].

In this context, the study evaluates whether ML and feature selection methods can identify statistically informative element combinations to predict Al concentrations in water and sediment samples from the Marmara Sea. Unlike original dataset with all features, this study specifically investigates whether Al concentrations can be accurately predicted using a reduced subset of elements. The underlying assumption is that Al can be predicted more effectively with fewer but more informative element predictors, thus reducing the number of features while maintaining or even improving prediction performance. The goal is to evaluate the predictive utility of multivariate element relationships and demonstrate how feature selection can support more efficient, interpretable, and data-driven environmental monitoring strategies.

In accordance with these objectives, in this study, 6 water and sediment samples from 17 sites of the Sea of Marmara were collected and digested using a microwave oven, then the concentrations of Al (Al), Boron (B), Calcium (Ca), Cadmium (Cd),

ARTICLE IN PRESS

Chromium (Cr), Cobalt (Co), Copper (Cu), Iron (Fe), Potassium (K), Magnesium (Mg), Manganese (Mn), Sodium (Na), Nickel (Ni), Lead (Pb), and Zinc (Zn) were determined by ICP-OES (Inductively Coupled Plasma Optical Emission Spectroscopy) device. As part of this study, the collected data from water and sediment samples, observed at different times and locations, were combined to form two datasets with a total of 102 observations each. Feature selection methods were then used to select important features, and the prediction models were evaluated, with all results carefully documented for further analysis.

The rest of this article is organized as follows: Section 2 presents the data collection and preprocessing procedures. Section 3 focuses on data-driven Al concentration prediction. Section 4 presents the experimental evaluation of the study, while Section 5 provides a discussion of the results. Finally, Section 6 summarizes the conclusions.

2 Data collection and preprocessing

The study was conducted over a 10-day period, during which time six water and sediment samples were collected from 17 localities within the Sea of Marmara, as illustrated in Figure 1. The names and coordinates of the sampling stations are additionally presented in Table 1 (provided in Supplementary Information). In the course of the study, sea sediment and surface water samples were simultaneously collected from the same locality and placed in a glass jar. Sediment samples were collected using algarnas, while water samples were obtained directly from the sea surface. The sediment samples were then subjected to a drying process in an oven maintained at 120°C for approximately 6-7 hours until they reached a state of complete dryness. The dried sediment and water samples were stored in jars until they were analyzed in the laboratory.

Samples were provided to the laboratory for element analysis. Sediment samples of approximately 0.3 g were placed in Teflon vessels using 6 ml of 65% nitric acid (Merck), 2 ml of 48% hydrofluoric acid (Merck), and 3 ml of 37% hydrochloric acid (Merck). The digestion process was conducted using a Berghof MWS2 microwave device with a program set at 145°C for 10 minutes, 165°C for 15 minutes, and 175°C for 15 minutes. Subsequently, the samples were filtered using a 1-2 μm average diameter filter paper and transferred to sterile 50 ml Falcon tubes containing ultrapure water. The water samples were diluted with 1% nitric acid and filtered using blue band filter paper. All samples were prepared for analysis in an ICP-OES device, and the concentrations of 15 elements (Al, B, Ca, Cd, Cr, Co, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, and Zn) were determined with three replicates.

ARTICLE IN PRESS



Figure 1. Map of sampling stations in the Sea of Marmara

ARTICLE IN PRESS

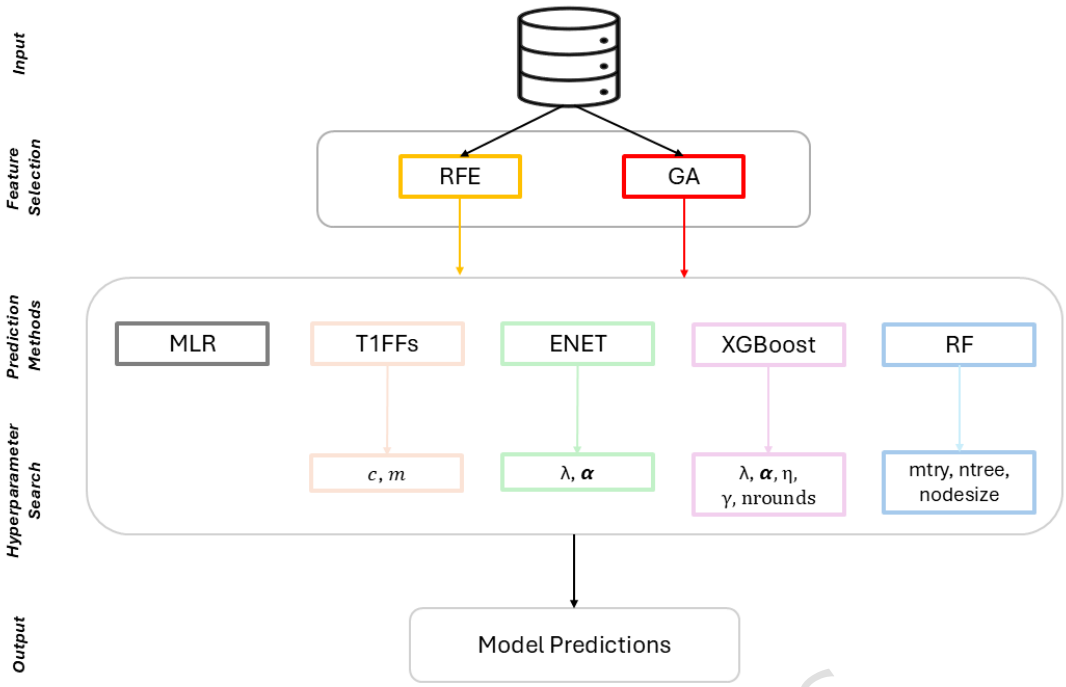


Figure 2. The workflow of the research

3 Data-Driven Aluminum concentration prediction

3.1 Overview

Figure 2 presents the workflow of this study, comprising several key steps. The process begins with the preprocessing of the input dataset, followed by the implementation of feature selection techniques, Recursive Feature Elimination (RFE) and Genetic Algorithm (GA), to identify the most relevant features. After selecting these key features, predictive modeling methods, including Multiple Linear Regression (MLR), Type 1 Fuzy Functions (T1FFs), Elastic Net (ENET), XGBoost, RF and Ensemble Learning Methods are applied to predict the target feature. During the prediction phase, hyperparameter optimization is conducted for models, focusing on parameters such as regularization coefficients (e.g., λ, α) and iteration counts, to enhance model performance. This workflow is designed to maximize predictive accuracy, with each step contributing to a more refined and robust analysis. In this section, detailed information regarding each phase is provided.

3.2 Feature Selection

Feature Selection (FS) plays a critical role in reducing model training time, enhancing the training process, and improving prediction accuracy. As noted by[48] FS effectively reduces the dimensionality of the feature space by eliminating irrelevant and redundant features, while maintaining the model's decision-making quality. The success of predictive analysis largely depends on the appropriate selection of features included in the model. In this study, the FS process is optimized using wrapper methods, which assess the relevance of features to select an optimal subset tailored to the specific mining algorithm[49]. This research employs two robust wrapper methods RFE and GA each using distinct strategies to identify the most effective subset of features for the same target features.

3.2.1 Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) systematically evaluates the importance of features by evaluating the performance of the model and gradually removing the least important ones[50]. Throughout this process, the model is retrained at each iteration to identify the most optimal subset of features. Each feature is evaluated on an individual basis.

A key parameter of RFE is the type of regression or classification model used in its configuration (e.g., linear regression, random forest). Additionally, parameters such as the number of features to be selected or the number of features to be eliminated at each iteration can also be adjusted to fine-tune the selection process. In this study, the implementation of the RFE algorithm is

ARTICLE IN PRESS

focused specifically on identifying the model type and the number of iterations. By allowing the model to leverage its inherent decision-making capabilities, the optimal features are identified without pre-specifying the number of features to be selected. Algorithm 1 (provided in Supplementary Information) presents pseudocode designed to fit this approach, utilizing the "rfe" function from the "caret" package for efficient and customizable feature elimination[51].

3.2.2 Genetic Algorithm (GA)

Genetic Algorithm (GA), introduced by John Holland in the 1970s[52], is an optimization technique inspired by the principles of biological evolution. Rather than assessing the importance of individual features, GA evaluates different combinations of features and employs genetic operations, such as crossover and mutation, to identify the optimal subset. As a population-based method, GA evaluates the fitness of entire populations and operates within a broad search space, striving to find the global optimum. This approach is particularly effective in uncovering nonlinear relationships within the data.

The GA is inherently more complex, and dynamic compared to RFE, necessitating the careful tuning of additional hyperparameters. Key parameters such as mutation rate, crossover, selection mechanisms, and population size are critical in identifying the optimal subset of features. The mutation probability introduces diversity within the population by randomly altering individuals, while the selection process determines which individuals are chosen for reproduction based on their fitness levels. The number of generations influences the duration of the evolutionary process, and elitism ensures that the best-performing individuals are preserved in subsequent generations. The fitness function is crucial as it evaluates the quality of solutions in relation to the optimization objective, guiding the algorithm toward the best results. Given the complexity and interdependence of these parameters, the design and tuning of the GA require careful attention to achieve optimal feature selection[53]

Due to its inherent nature, the GA requires more computational time compared to other methods. When dealing with smaller datasets, using wide ranges in the search for optimal hyperparameters may increase the risk of overfitting. Therefore, in this study, only specific hyperparameters such as the number of iterations, population size (popsize), mutation rate (pmutation), elitism (elite), and crossover (pcrossover) probability were considered to optimize the hyperparameter search, with the aim of speeding up the process and minimizing the risk of overfitting. Algorithm 2 (provided in Supplementary Information) presents the pseudocode for this approach, utilizing the "gafs" function from the "caret" package to achieve efficient feature selection by minimizing Mean Squared Residuals (MSR) with the fewest features[51].

3.3 Prediction Models

In this study, we focus on these methods: Multiple Linear Regression, Elastic Net Regression, Type 1 Fuzzy Function, XGBoost, RF, Simple Average and Stacking[54] from Ensemble Learning Methods. Each method offers a unique approach to modeling and was chosen based on its ability to address different aspects of the prediction problem.

3.3.1 Multiple Linear Regression (MLP)

Multiple Linear Regression is one of the most used predictive models for examining the relationships among features[55]. This approach is particularly prevalent in determining the relationship between a dependent feature and multiple independent features. MLR, a statistical method, seeks to predict a dependent feature (or target feature) by utilizing more than one independent (or explanatory) feature. The goal of this model is to establish a linear relationship between the dependent features and multiple independent features. For the model to yield accurate results, certain assumptions must be satisfied, including the independence of independent features, normality of the error terms, and homoscedasticity. Failure to meet these assumptions can undermine the model's reliability and complicate the interpretation of its results. Nonetheless, multiple linear regression remains a fundamental tool for modeling and predicti relationships between features. The multiple linear regression model is given in Equation (1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{1}$$

where: x are the independent variables (features), β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, ε is the error term, accounting for the variation in y that is not explained by the independent variables.

3.3.2 Type-1 Fuzzy Functions (T1FFs)

Type-1 Fuzzy Functions (T1FFs) were developed by Turksen[56] as an alternative to rule-based fuzzy inference systems, eliminating the need for expert knowledge. In rule-based fuzzy inference systems, rules are required to establish a relationship between input and output. However, the proposed T1FFs approach represents the relationships between input and output through functions derived from clustering methods, without relying on such rules. This approach offers a significant advantage in modeling uncertainties and is commonly used in classification and regression problems[56–58].

The structure of T1FFs is based on clustering the input matrix using the FCM algorithm "ppclust" package[59].The algorithm results in grouping similar objects with higher membership degrees, where each cluster has a membership matrix containing the

ARTICLE IN PRESS

membership degrees of the observations. After this process, as many input matrices as the number of clusters are generated by adding the membership degrees to the input matrix separately. Finally, the predicts obtained from these input matrices are combined. The detailed steps of T1FFs are presented below:

- Step 1: Firstly, the number of fuzzy clusters (c), the fuzzy index value (m_i), and the centers of the clusters (v_i) are determined.

$$v_i = \frac{\sum_{k=1}^{n} \mu_{ik}^{m} x_k}{\sum_{k=1}^{n} \mu_{ik}^{m}}, i = 1,2,..,c \tag{2}$$

$$\mu_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d(x_k, v_i)}{d(x_k, v_j)} \right)^{\frac{2}{m-1}} \right]^{-1}, i = 1,2,..c; k = 1,2,\ldots,n \tag{3}$$

In Equation 3, x_k is a vector of elements that form the k^{th} row of X. μ_{ik} represents the degree to which the k^{th} observation belongs to the i^{th} cluster. Additionally, $d(x,v)$ the Euclidean distance, is calculated using Equation 4.

$$d(x_k, v_i) = \|x_k - v_i\| \tag{4}$$

- Step 2: The original inputs and the membership degrees μ_{ik} for each data point are combined for each cluster. Then, the multivariate regression model $Y^{(i)} = X^{(i)}\beta^{(i)} + \varepsilon^{(i)}$ is obtained for the i^{th} fuzzy function. Given that the number of independent features is p, the matrices $X^{(i)}$ and $Y^{(i)}$ are as follows:

$$\mathbf{X}^{(i)} = \begin{bmatrix} \mu_{i1} & x_{11} & \cdots & x_{p1} \\ \mu_{i2} & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{in} & x_{1n} & \cdots & x_{pn} \end{bmatrix}, \quad \mathbf{Y}^{(i)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{5}$$

- Step 3: The outputs that form the final predicted values of the system are obtained as the weighted averages of the results from the clusters, as shown in Equation 6.

$$\widehat{y}_k = \frac{\sum_{i=1}^{c} \widehat{y}_{ik} \mu_{ik}}{\sum_{i=1}^{c} \mu_{ik}} \tag{6}$$

ARTICLE IN PRESS

3.3.3 Elastic Net (ENET)

The Elastic Net (ENET) method is a type of regularized optimization applied to linear regression that compromises ridge regression and the lasso[60]. This approach is utilized to mitigate overfitting in regression models, while simultaneously ensuring both sparsity and the selection of groups of correlated features. ENET can handle cases where the linear regression model assumptions are unmet. For example, when there are cases of multiple collinearities within the dataset, it can handle this problem by co-selecting features with such dependencies.

The goal of ENET is to generate a sparse feature space by applying the l_1 penalty, while also enhancing stability and retaining correlated features through the l_2 penalty. Like the lasso, solving this optimization problem is computationally challenging; however, efficient algorithms have been developed to address it. Additionally, a free parameter, α, allows for the adjustment of the balance between the two penalties[61].

The ENET uses a mixture of the l_1 (lasso) and l_2 (ridge regression) penalties and can be formulated as in Equation (7):

$$\hat{\beta} = \arg\min_{\beta} \left(\|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2) \right) \tag{7}$$

As evident from the Equation (7), the performance of the ENET model is largely dependent on two critical hyperparameters: α and λ. Algorithm 3 (provided in Supplementary Information) presents the pseudocode showing how the ENET hyperparameters are tuned and implemented in this study, utilizing the "glmnet" function for this purpose[62].

ARTICLE IN PRESS

3.3.4 Extreme Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) is an advanced machine learning method based on the gradient boosting algorithm developed[63] in 2001. This method, known as MART (Multiple Additive Regression Trees), is a successful technique for making predictions using decision trees. Built on this foundation, XGBoost was introduced by[64] and quickly gained popularity. XGBoost focuses on minimizing error by generating sequential decision trees and is known for its exceptional speed, efficiency, and performance. It is widely applied in both classification and regression tasks. Unlike traditional models, XGBoost has less stringent assumptions, though having well-prepared and organized data is crucial for maximizing model accuracy. Since the algorithm is robust enough to handle missing data and multiple linear dependencies, it can be preferred over other methods in practice.

The objective function in XGBoost consists of two primary components: the loss function, which quantifies the discrepancy between the actual and predicted values, and a regularization term designed to penalize model complexity, thereby reducing the risk of overfitting. The model optimizes this objective function through the application of gradient descent.

If a dataset consists of n examples and p features and is defined as $D = \{(x_i, y_i)\}$ (where $|D| = n, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$), a tree ensemble model uses K additive functions to predict the output:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{8}$$

$$F = \{f(x) = w_{q(x)}\}, \quad q : \mathbb{R}^p \to T, \quad w \in \mathbb{R}^T \tag{9}$$

where F represents the space of regression trees, where q denotes the structure of each tree and maps a sample to its corresponding leaf index. T indicates the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and its associated leaf weights w. Unlike decision trees, each regression tree assigns a continuous score to each leaf, with the score at the i-th leaf represented as w_i[64].

To learn the set of functions used in the model, we minimize the following regularized objective in Equation (10):

$$L(\Phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{11}$$

In Equation (10), l is a differentiable convex loss function that quantifies the difference between the predicted value $\hat{y}_i$ and the actual target y_i. The second term, Ω, imposes a penalty on the model's complexity, specifically on the regression tree functions. This additional regularization term helps to refine the final learned weights, thereby reducing the risk of overfitting. Also, hyperparameter tuning is crucial in XGBoost to achieve optimal model performance. Key hyperparameters include maximum depth, which controls the maximum depth of each tree, learning rate (also known as eta (η)), iteration number (nrounds) and λ, which regulates the l_2 regularization term. Through techniques such as grid search and cross-validation, these parameters can be fine-tuned to improve model performance[64,65].

Algorithm 4 (provided in Supplementary Information) presents the pseudocode for the process of tuning these hyperparameters to achieve optimal model performance using the "xgboost" package. Also, XGBoost offers various booster types to enhance model performance. In this study, the booster = 'gbtree' was selected, as it iteratively builds improved models based on decision trees, making it widely applicable in regression problems[66].

3.3.5 Random Forest (RF)

The Random Forest algorithm (RF), introduced by Breiman (2001)[67], has become a widely used non-parametric method for both classification and regression. It builds predictive models using different types of predictor features without requiring any predefined assumptions about how these features relate to the response feature. The method is based on combining a large number of decision trees and producing predictions through aggregation. In regression problems, the final prediction is obtained by averaging the outputs of individual trees, whereas in classification problems, the final class is determined by majority voting.

Hyperparameter tuning plays an important role in improving the performance of the RF model. To achieve better prediction accuracy, hyperparameters can be optimized using a grid search method. Algorithm 5 presents the pseudocode that uses the hyperparameters *mtry*, *ntree*, and *nodesize* to determine the best-fitting model in this study (provided in Supplementary Information).

ARTICLE IN PRESS

4 Experimental Evaluation

In this part of the study, the prediction performance of the concentration values of the "Al" element in water and sediment samples obtained from the Sea of Marmara is evaluated with different individual and ensemble learning methods. In addition, the study aims to develop the most effective prediction model by applying feature selection techniques using fewer features to improve the simplicity and interpretability of the model while maintaining predictive performance.

The values of all features in the dataset are continuous. The concentration values of these elements, measured as mg/L for water and mg/kg for sediment, are used as input features in the analysis. All analyses were performed using the R statistical programming language. In addition, the datasets were divided into 50% for training, 20% for validation, and 30% for testing. Table 1 presents the number of samples in each subset for the water and sediment datasets. In this study, to ensure a fair comparison, the training, evaluation, and test sets were kept the same across all methods, and the performance comparisons of the models were made using the evaluation values of the test set.

Dataset Split	Split Ratio	Water	Sediment
Train	50%	51	51
Validation	20%	20	20
Test	30%	31	31
Total	100%	102	102

Table 1. Number of samples subsets for water and sediment samples

Within the scope of this study, feature selection was performed as a preprocessing step using RFE and GA-based methods with "function = random forest". Consequently, in addition to the original dataset, two reduced datasets, including selected features for both water and sediment datasets, were also included in the analysis.

The search space used to optimize model performance are presented in Table 2. The remaining parameters were retained at their default settings to ensure a consistent evaluation framework and to focus on the influence of the selected hyperparameters. To reduce the sensitivity of hyperparameters optimization, data partitioning (training and validation sets) was applied 10 times using random sampling (except MLR). For each iteration, models were trained on the training set and hyperparameters were selected based on validation set performance in terms of RMSE. Using the optimal hyperparameters, predictions were obtained on the test set. Finally, the test predictions obtained from the 10 iterations were averaged and compared with the actual values. Since the optimal hyperparameters varied across iterations, optimal hyperparameter value could not be reported.

The performance of the evaluated methods is compared using root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). These metrics assess the accuracy of prediction models by measuring the discrepancy between predicted and actual values, which is crucial in evaluating how well the model predicts continuous outcomes. In regression problems, they provide valuable insights into both the magnitude and distribution of residuals, helping to evaluate model performance[68]

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{13}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of observations.

4.1 Al concentration prediction in water samples

The water dataset, collected from various time periods and locations, provides valuable insights into environmental fluctuations and serves as the foundation for developing predictive models to predict Al concentration using feature selection techniques.

Figure 3 presents a scatter plot matrix, which visually summarizes the relationships between Al concentration and other elements in the dataset. In the upper triangle, Pearson correlation coefficients reveal strong positive correlations between Al and several elements, including Cd (0.783), Co (0.808), and Fe (0.844), suggesting the presence of multicollinearity. Notably, Al

ARTICLE IN PRESS

Parameters	TIFFs	ENET	XGBoost	Random Forest	GA-based FS
λ	-	$10^{\text{seq}(-3,5,\text{length.out}=100)}$	$10^{\text{seq}(-1,1,\text{by}=0.2)}$	-	-
α	-	$\text{seq}(0,1,\text{by}=0.1)$	$0,1,\text{by}=0.1$	-	-
η	-	-	$0,0.30,\text{by}=0.05$	-	-
γ	-	-	$0,1,2,3,4,5$	-	-
nrounds	-	-	25, 50, 100, 250	-	-
c	2, 3, 4, 5	-	-	-	-
m	1.5, 2, 2.5	-	-	-	-
mtry	-	-	-	$\sqrt{p}$, $p/3$	-
ntree	-	-	-	100, 200, 300, 400, 500	-
nodesize	-	-	-	1, 2, 3, 4, 5	-
popSize	-	-	-	-	(30, 40, 50, 60)
pcrossover	-	-	-	-	(0.70, 0.80, 0.90)
pmutation	-	-	-	-	(0.05, 0.10, 0.15)
elite	-	-	-	-	(0, 1)

Table 2. Hyperparameter search space for the water and sediment datasets

also shows moderate negative correlations with elements such as K (-0.240) and Mn (-0.293). The scatter plots in the lower triangle display both linear and non-linear relationships between pairs of features, while the diagonal contains density plots, offering a clear view of each feature's distribution.This graphical analysis helps identify potential predictors for Al concentration while simultaneously revealing the presence of non-linear interactions and outliers in the data. These challenges, along with the high degree of multicollinearity, indicate that traditional linear methods may struggle to provide accurate predictions. Consequently, advanced machine learning models such as T1FFs, XGBoost, and RF which are capable of handling complex interactions and multicollinearity, will be explored as alternatives to linear regression. Tree-based models (e.g., XGBoost and RF) are generally scale-invariant; however, models such as MLR and ENET are highly sensitive to the magnitude of input features. In the water dataset, substantial differences exist in concentration ranges between major elements (e.g., Fe, Na, K) and trace elements (e.g., Cd, Pb). Therefore, prior to model training, all features were standardized using Z-score.

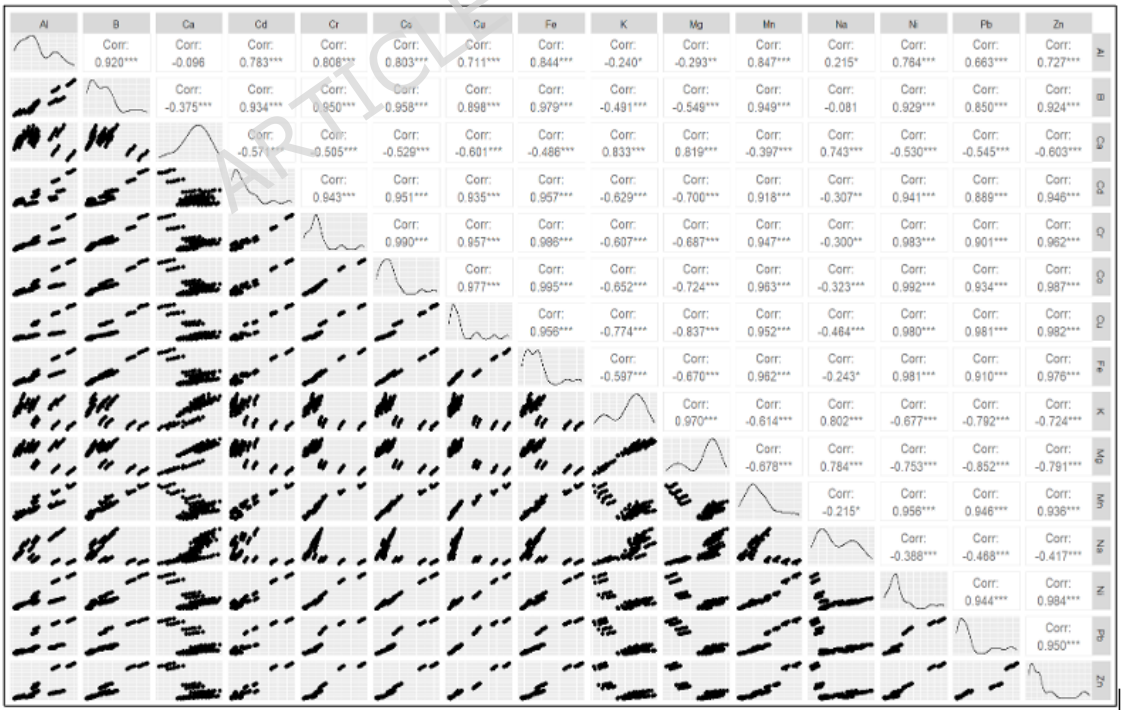


Figure 3. Correlation and scatter plot matrix of element concentrations in water dataset

Table 3 presents the number and names of the selected features due to the feature selection methods. It was observed

ARTICLE IN PRESS

Methods	Number of features	Feature Names
RFE	6	B, Cr, Cu, Fe, Ni, Mn
GA	6	Cr, Cu, Mn, Na, Ni, Pb

Table 3. Selected features for water dataset

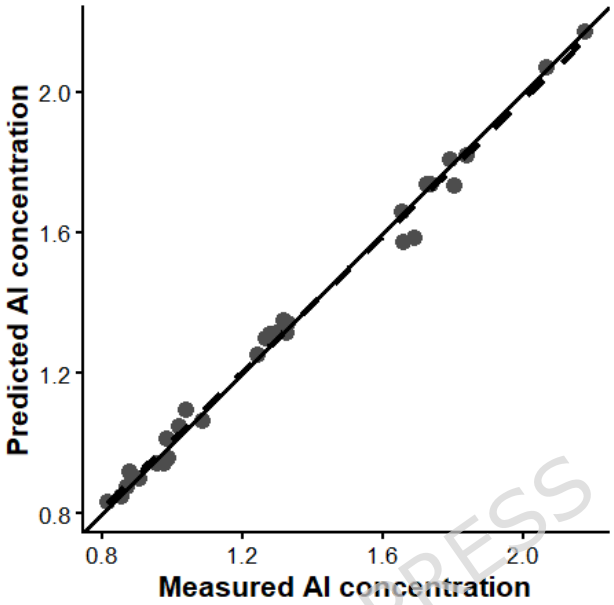


Figure 4. Scatter plot of measured and predicted values of the best individual model for the water test set.

that both the methods selected 6 features as more important to predict the amount of 'Al' as the dependent feature (best hyperparameter for GA: $popSize = 60$, $pcrossover = 0.80$, $pmutation = 0.1$, $elite = 0$).

Table 4 presents the RMSE, MAE, and MAPE on the test set for the original dataset (No FS) and the reduced datasets (respectively RFE, GA). When looking the original dataset and with the dataset reduced using the RFE method, T1FFs achieved the best performance in terms of RMSE. In contrast, for the dataset reduced using GA selected features, XGBoost was observed as the best-performing method with in terms of RMSE. Overall, in terms of RMSE, the best performance among the individual methods was performed using the XGBoost model on the dataset with features selected using GA. The MAE and MAPE further support this finding. The stacking method demonstrated the best performance in terms of RMSE, while also exhibiting highly competitive results with respect to MAE and MAPE.

	No FS			RFE			GA		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
MLR	0.0455	0.0349	0.0271	0.0807	0.0698	0.0549	0.0928	0.0742	0.0639
T1FFs	0.0358	0.0299	0.0246	0.0688	0.0603	0.0468	0.0754	0.0559	0.0493
ENET	0.0511	0.0449	0.0359	0.0857	0.0746	0.0575	0.0991	0.0814	0.0662
XGBoost	0.0637	0.0451	0.0330	0.0727	0.0471	0.0342	**0.0354**	**0.0255**	**0.0201**
RF	0.0800	0.0621	0.0470	0.0781	0.0620	0.0470	0.0660	0.0465	0.0366
Simple Average	0.0388	0.0313	0.0237	0.0678	0.0545	0.0415	0.0648	0.0509	0.0429
Stacking	0.0437	0.0314	0.0234	0.0640	0.0497	0.0376	0.0353	0.0284	0.0215

Table 4. Test set performance metrics for the water dataset

Figure 4 illustrates the relationship between measured and predicted Al concentrations obtained from best individual model. The figure indicates that the model predictions are highly consistent with the measured values and exhibit a linear relationship.

ARTICLE IN PRESS

4.2 Al concentration prediction in sediment samples

The sediment samples dataset reveals the stable and accumulated nature of the elements over time. Collecting data from various locations and time periods provides insight into the long-term environmental variability and changes in the elemental composition of the Sea of Marmara. This dataset forms the basis for applying feature selection techniques and developing predictive models to accurately predict the concentration of Al based on the levels of other elements present in the sediment.

Figure 5 provides a visual summary of the relationships between Al concentration and other elements in the sediment dataset. Like the water dataset, the upper triangle displays Pearson correlation coefficients. Strong positive correlations are observed between Al and elements such as Cd (0.872), Co (0.874), and Fe (0.895). Notably, there are also significant correlations between Al and Ni (0.833) as well as Zn (0.862). Negative correlations are seen with K (-0.268) and Mg (-0.285). The scatter plots in the lower triangle depict both linear and non-linear relationships, while the diagonal shows density plots illustrating the distribution of each feature. This graphical analysis highlights the relationships between elements in the sediment dataset, in a manner similar to the water dataset. Multicollinearity and non-linear relationships are also prevalent in the sediment data, presenting similar challenges for prediction models. Therefore, features were standardized prior to analysis, and the analysis process was conducted similarly to the water dataset.

Table 5 summarizes the number and names of the features selected by the feature selection methods. The RFE method selected 9 features, whereas the GA method selected 6 features (with optimal hyperparameters of $popSize = 60$, $pcrossover = 0.80$, $pmutation = 0.15$, and $elite = 1$).

Table 6 reports the RMSE, MAE, and MAPE obtained by comparing the predicted and actual values on the test set for the sediment dataset. Overall, XGBoost consistently demonstrated strong performance across all datasets (No FS, RFE, and GA), achieving the best results among the individual models in terms of RMSE. Looking at the MAE and MAPE, it is clear that the XGBoost model performs best for the dataset reduced using features selected with RFE. For ensemble learning methods, the stacking method demonstrated highly competitive performance in terms of RMSE on the GA-based dataset, and in terms of MAE and MAPE on the RFE-selected feature dataset.

Figure 6 presents the relationship between the measured and predicted Al concentrations obtained by the best individual model for the sediment dataset. The figure shows a high level of agreement between predicted and measured values.

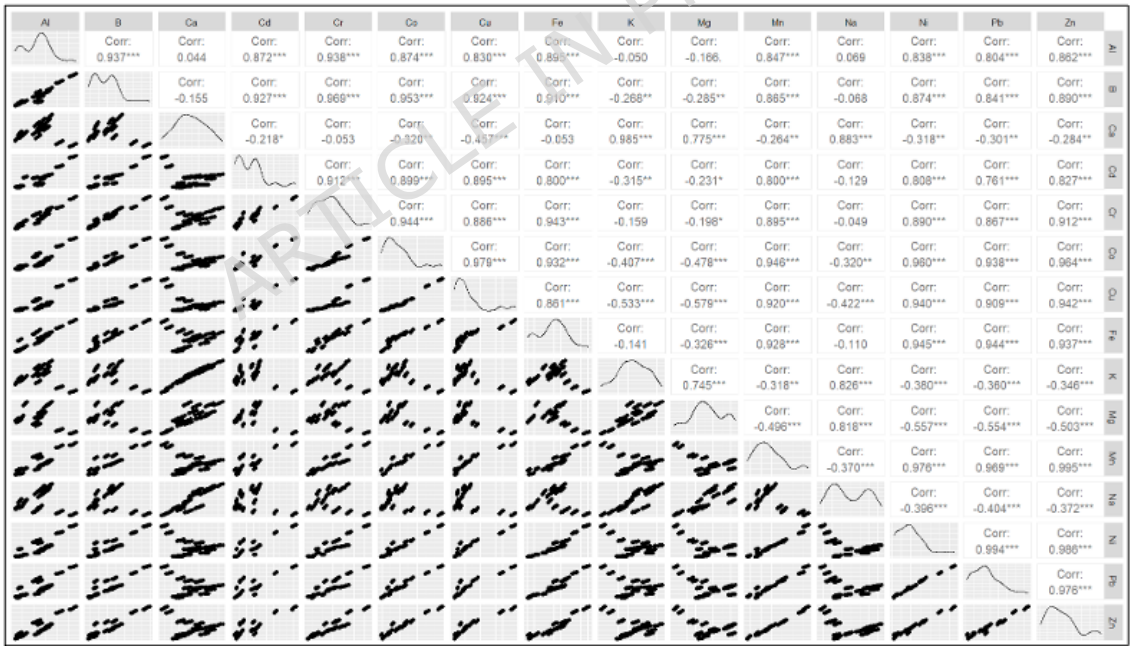


Figure 5. Correlation and scatter plot matrix of element concentrations in sediment dataset

ARTICLE IN PRESS

Methods	Number of features	Feature Names
RFE	9	B, Ca, Cd, Co, Cr, Cu, K, Na, Mn
GA	6	B, Cd, K, Ni, Pb, Zn

Table 5. Selected features for the sediment dataset

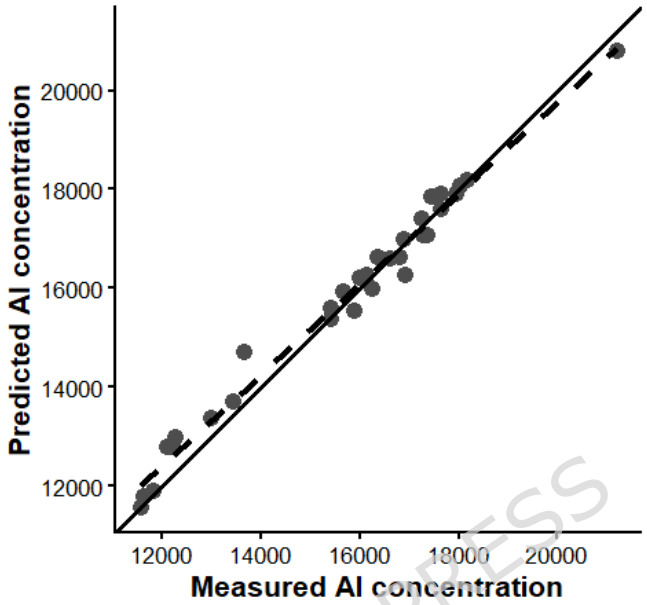


Figure 6. Scatter plot of measured and predicted values of the best individual model for the sediment test set.

	No FS			RFE			GA		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
MLR	655.5621	505.6465	0.0337	662.2309	574.0016	0.0376	702.7171	554.7387	0.0368
T1FFs	549.4160	407.0236	0.0257	597.2927	510.1607	0.0339	575.8882	429.6723	0.0283
ENET	603.6719	471.8115	0.0310	721.2473	623.3698	0.0411	699.7990	556.8776	0.0366
XGBoost	447.0389	284.9075	0.0196	469.0149	**253.2664**	**0.0170**	**356.5694**	267.1739	0.0181
RF	591.2105	417.2046	0.0293	546.5986	384.1385	0.0262	522.5457	367.2470	0.0255
Simple Average	462.7250	363.5111	0.0246	493.1931	384.0733	0.0259	514.7512	396.4661	0.0263
Stacking	481.3372	336.3688	0.0222	427.2833	311.5760	0.0208	400.7323	320.6840	0.0214

Table 6. Test set performance metrics for the sediment dataset

5 Discussion

Experimental evidence from water and sediment data shows that feature selection approaches applied to selected datasets can provide superior prediction performance compared to the original dataset. Several observations further clarify the method's behavior.

The results obtained from the water data show that the best prediction performance in terms of RMSE is achieved by modeling the selected features by GA using XGBoost. In the original dataset, the XGBoost model was trained using 14 features and obtained RMSE value of 0.0637 on the test set. After feature selection, the number of features was reduced to 6, obtained lower RMSE of 0.0354. The feature selection process improved the RMSE of the XGBoost model by 44.4% compared to the original dataset. Furthermore, when compared to the MLR (RMSE = 0.0455), the best individual method demonstrates an improvement of approximately 22.2%. As illustrated in Figure 7, the Shapley Additive Explanations (SHAP) based feature importance analysis of the best-performing individual model reveals that Cu and Cr are the most influential features. Mn and Na exhibit moderate contributions, whereas Ni and Pb have a lower impact on the model output. In addition, the ensemble

ARTICLE IN PRESS

learning results demonstrate that stacking outperforms simple averaging. This can be attributed to the ability of stacking to learn an optimal combination of base learners through a meta-model. Moreover, as presented in Table 4, the RF model showed a performance increase of 2% with RFE and 17% with GA in terms of RMSE after feature selection.

The findings from the sediment dataset indicate that the best prediction, in terms of RMSE, is achieved when features selected by GA are using XGBoost. In the original dataset, the XGBoost was trained with 14 features and produced an RMSE of 447.0389 on the test set. Following feature selection, the feature count was reduced to 6, resulting in a lower RMSE of 356.5694. This reduction corresponds to a 20% improvement in RMSE compared to the No FS. Additionally, when compared with the MLR (RMSE = 655.5621), the best model showed an approximate improvement of 45%. As shown in Figure 8, SHAP-based feature importance analysis for the model highlights B and Cd as the most significant faetures. Furthermore, as presented in Table 6, the RF model demonstrated a 7% and 11% improvement in RMSE with RFE and GA, respectively.

Feature selection improved the prediction performance of XGBoost and RF models, while no improvement was observed in MLR, T1FF, and ENET methods. This difference may stem from the structural characteristics of the methods. XGBoost and RF models are tree-based models. They have flexible structures based on a splitting mechanism, which can better model nonlinear relationships and intervariable interactions than classical methods. Therefore, feature selection and elimination of noisy features can improve model performance by providing more optimal splits. However, other methods, such as MLR, rely on the contribution of all features under a linear assumption. Therefore, removing some features from the model as a result of feature selection analysis can lead to a decrease in explained variance and negatively impact prediction performance.

While the predictive performance of the models provides quantitative evidence of the effectiveness of feature selection, it is also important to interpret the selected predictors within a geochemical context in order to better understand the underlying environmental processes that may be represented by the models. In the water dataset, the GA based selected features (Cr, Cu, Mn, Na, Ni, and Pb) may represent overlapping environmental signals related to anthropogenic inputs, particulate transport, and adsorption–scavenging processes. Transition metals such as Cr, Cu, Ni, and Pb are commonly reported to originate from similar sources and may exhibit co-variation due to shared transport pathways. Mn may influence trace metal behavior through redox-driven scavenging processes, while Na can reflect salinity-driven variability and hydrographic mixing. These factors may collectively suggest that Al distribution in the water dataset is governed by broader physicochemical conditions rather than a single controlling feature.

In the sediment dataset, the selected features (B, Cd, K, Ni, Pb, and Zn) may indicate that Al prediction is associated with a combination of lithogenic background and anthropogenic influences. K is generally linked to the mineral structure of sediments, whereas Cd, Ni, Pb, and Zn are often associated with anthropogenic inputs and their accumulation in reactive sediment phases. B may reflect both natural processes and marine geochemical conditions. These features can be interpreted as indicators of sediment composition and binding processes that may influence elemental distributions. Overall, the selected features may be interpreted as co-varying indicators of the same geochemical environment, rather than direct causal determinants of Al concentration.

To further contextualize these relationships, it is important to consider the environmental behavior of Al. In marine systems, Al distribution is primarily governed by physicochemical processes such as particle interactions, hydrographic variability, and source inputs, rather than a single controlling factor. In sediments, Al is typically associated with lithogenic background and mineral composition, while anthropogenic influences may contribute to localized variability. In this regard, factors such as sediment texture, organic matter, and reactive mineral phases play a key role in regulating metal binding and distribution. Accordingly, Al concentrations should not be interpreted as an isolated indicator of contamination, but rather as the result of the combined influence of natural geochemical processes and human activities.

Within this framework, the Al concentrations observed in the present dataset gain environmental context from both recent marine guideline values and regional sediment studies. For marine waters, the 2025 ANZG guideline defines aluminium thresholds of 9, 37, 72, and 160 µg/L corresponding to 99%, 95%, 90%, and 80% species protection levels, respectively, and identifies 37 µg/L as indicative of slightly to moderately disturbed ecosystems. The guideline further specifies that dissolved aluminium measurements, operationally defined as 0.45 µm-filtered samples, are appropriate for such comparisons (ANZG, 2025). In sediments, however, Al is more meaningfully interpreted within a regional geochemical context rather than against a universal threshold. Evidence from the Sea of Marmara indicates that Al and Fe are the most abundant sedimentary elements, with localized enrichment patterns linked to industrialization, maritime activity, and urbanization[69]. Source-apportionment studies further classify Al and Fe within lithogenic components, whereas Cd, Pb, and Zn are more closely associated with anthropogenic inputs, with sediment texture and organic matter influencing metal binding[70]. Similarly, studies from the East China Sea highlight grain size and organic matter as primary controls on heavy metal distribution, with fine-grained terrestrial inputs playing a significant role in metal transport and accumulation[71]. Additionally, manganese oxides act as efficient natural scavengers through adsorption and oxidation processes, supporting the role of Mn-bearing phases in trace metal retention[72]. Taken together, these findings indicate that the observed Al distribution reflects natural mineralogical background conditions together with superimposed anthropogenic pressures, rather than an isolated contamination signal.

ARTICLE IN PRESS

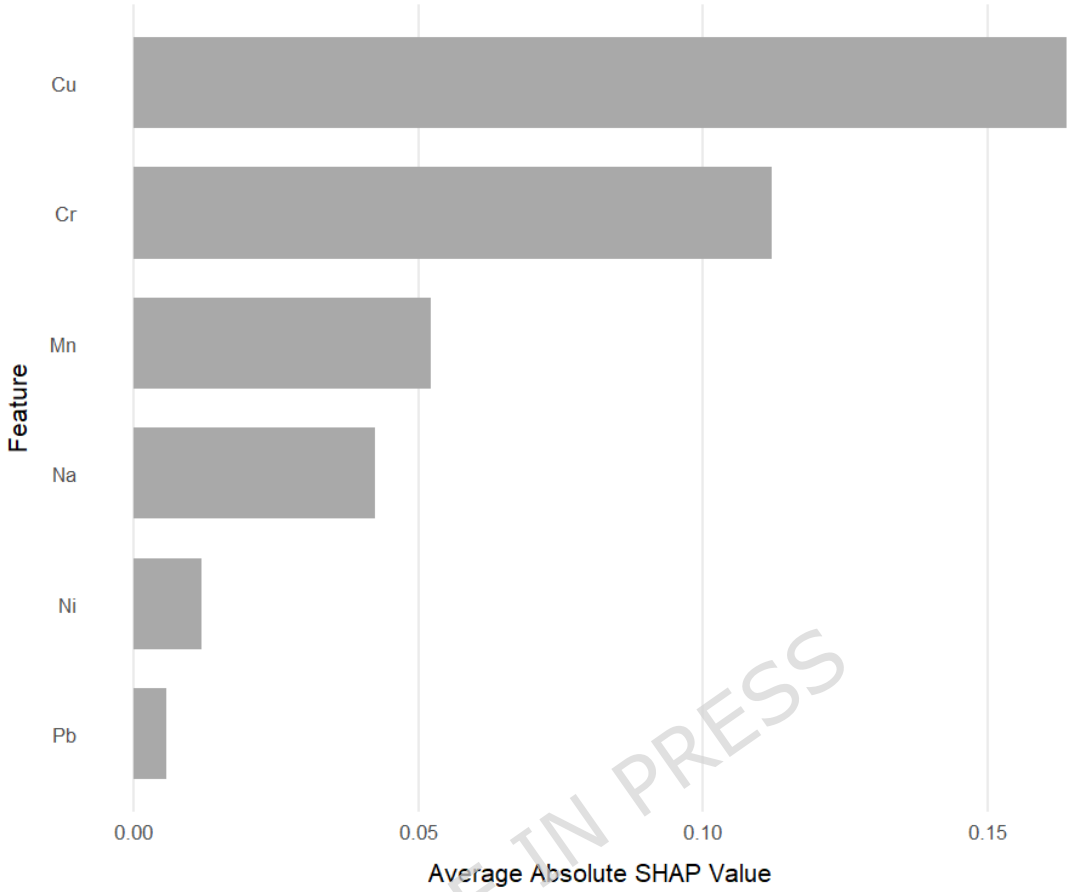


Figure 7. Feature importance analysis of the best-performing individual model on the water dataset.

ARTICLE IN PRESS

ARTICLE IN PRESS

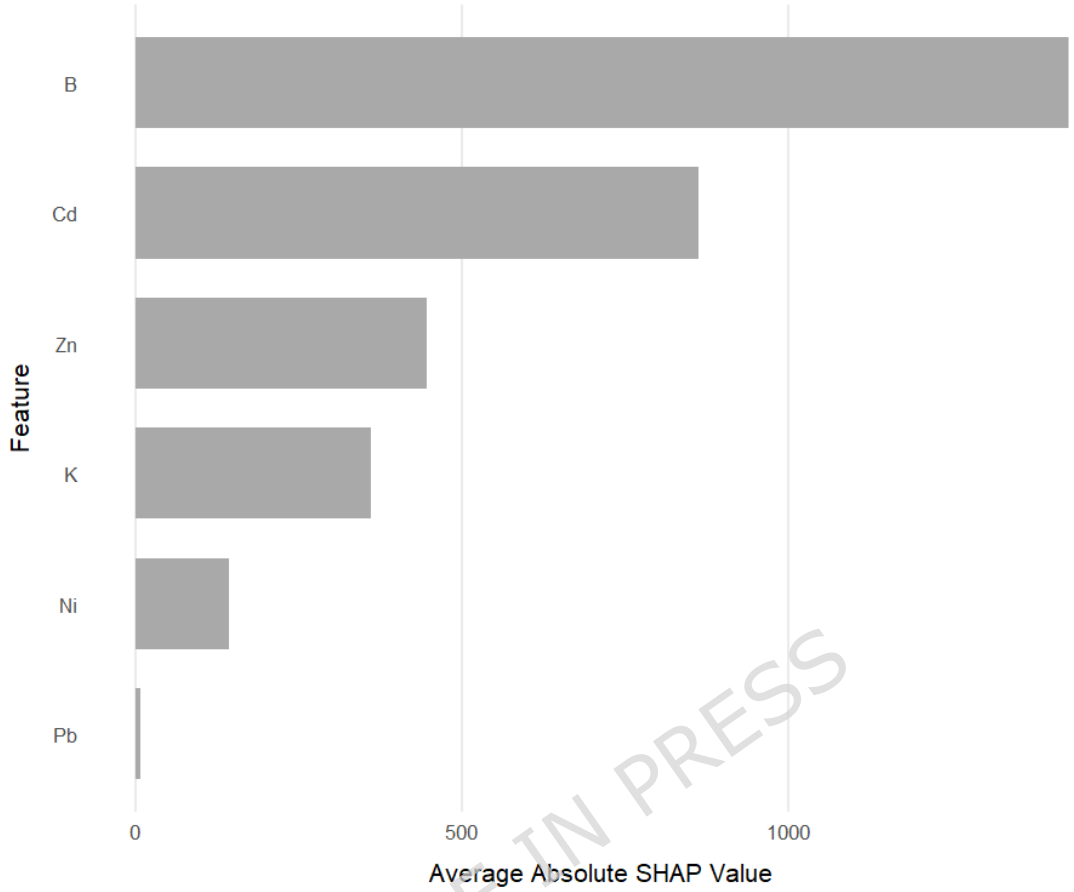


Figure 8. Feature importance analysis of the best-performing individual model on the sediment dataset.

6 Conclusion

In this study, water and sediment samples collected from 17 sites in the Sea of Marmara were analyzed, and the concentrations of 15 elements were determined using ICP-OES. The resulting datasets enabled the investigation of relationships among elements for predicting Al concentrations. In this context, the study evaluates whether machine learning and feature selection methods can identify statistically informative combinations of elements for Al prediction. Unlike the use of the full feature set, this study specifically examines whether Al concentrations can be accurately predicted using a reduced subset of elements.

In this context, MLR, T1FFs, ENET, XGBoost, RF, and Ensemble Learning methods were applied to predict the concentration levels of the Al element. To enhance model performance, achieve similar or better predictive performance with fewer features, and gain a better understanding of the relationships between features, feature selection methods RFE and GA were employed in addition to the prediction techniques. During the model evaluation phase, the datasets were divided into training, validation and testing , and hyperparameter optimization was carried out. The findings indicated that the features selected through GA (Cr, Cu, Mn, Na, Ni, Pb) achieved the highest performance for the XGBoost model in the water dataset. In the sediment dataset, the features determined by GA (B, Cd, K, Ni, Pb, Zn) yielded the best results for the XGBoost model on the test dataset.

At the conclusion of the study, it is observed that the number of important features for predicting Al concentration can be reduced from 14 to 6 when comparing the original dataset to the dataset created using the selected features for the both water and sediment dataset. This demonstrates that similar or even superior results can be obtained using only 6 features instead of 14, and that analysis processes can be effectively conducted with data optimized through feature selection. Furthermore, the findings highlight the importance of element interactions in predicting Al concentrations and demonstrate that feature selection improves model performance while maintaining prediction accuracy. Overall, the study shows that integrating feature selection with machine learning offers an effective approach to more accurately predicting Al concentrations using a reduced set of informative elements, while also contributing to a clearer and more interpretable presentation of relationships between

ARTICLE IN PRESS

elements.

In future studies, the application of additional machine learning models, such as deep learning algorithms, may further enhance the prediction accuracy of element concentrations in environmental datasets. In this context, deep learning approaches based on neural networks, which have been reported to demonstrate superior performance compared to conventional machine learning models in similar prediction tasks, may also be considered[73]. Furthermore, expanding the dataset by incorporating seasonal and spatial variations across a broader sampling area in the Sea of Marmara and other similar inland seas would provide more comprehensive insights.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. Arslan, H., Agir, A. & Demir, G. Impacts of PM10 exposure on hospitalization for acute bronchitis in Ankara, Türkiye. *Front. Life Sci. Relat. Technol.* **5**, 1–5 (2024).
2. Gjergjizi Nallbani, B. *et al.* Utilizing a medicinal plant and soil assays for heavy metal pollution assessment and its impacts in a Turkish Industrial Zone. *Spectrosc. Lett.* 1–18 (2024).
3. Banaee, M., Zeidi, A., Mikušková, N. & Faggio, C. Assessing metal toxicity on crustaceans in aquatic ecosystems: a comprehensive review. *Biol. Trace Elem. Res.* 1–19 (2024).
4. Fulke, A. B., Ratanpal, S. & Sonker, S. Understanding heavy metal toxicity: Implications on human health, marine ecosystems and bioremediation strategies. *Mar. Pollut. Bull.* **206**, 116707 (2024).
5. Jean-Lavenir, N. M. *et al.* Potentially toxic metals contamination in stream sediments of mbal area (Pan-African fold belt, Cameroon). *Discov. Geosci.* **2**, 15 (2024).
6. Prabakaran, K., Sompongchaiyakul, P., Bureekul, S., Wang, X. & Charoenpong, C. Heavy metal bioaccumulation and risk assessment in fishery resources from the Gulf of Thailand. *Mar. Pollut. Bull.* **198**, 115864 (2024).
7. Borrell, A. *et al.* High aluminum content in bone of marine mammals and its relation with source levels and origin. *Environ. Pollut.* **331**, 121936 (2023).
8. Yazicioglu, H. *et al.* Physiological alterations and genotoxic damage under combined aluminum and cadmium treatments in Bryophyllum daigremontianum clones. *Mol. biology reports* **51**, 1–21 (2024).
9. Botté, A., Zaidi, M., Guery, J., Fichet, D. & Leignel, V. Aluminium in aquatic environments: abundance and ecotoxicological impacts. *Aquatic Ecol.* **56**, 751–773 (2022).
10. Tiwari, A. K., Pal, D. B. & Prasad, N. Agricultural waste biomass utilization in waste water treatment. In *Utilization of Waste Biomass in Energy, Environment and Catalysis*, 19–41 (CRC Press, 2022).
11. Islam, M. S. *et al.* Contamination and ecological risk assessment of Cr, As, Cd and Pb in water and sediment of the southeastern Bay of Bengal coast in a developing country. *Mar. Pollut. Bull.* **197**, 115720 (2023).
12. Lao, Y., Ma, J., Pan, K., Chen, F. & Zhang, Z. A Brief Review of Effects of Aluminum on Marine Diatoms. *Bull. Environ. Contamination Toxicol.* **113**, 31 (2024).
13. Sáez-Guinoa, J., García-Franco, E., Llera-Sastresa, E. & Romeo, L. M. The effects of energy consumption of alumina production in the environmental impacts using life cycle assessment. *The Int. J. Life Cycle Assess.* **29**, 380–393 (2024).
14. Santore, R. C. *et al.* Development and application of a biotic ligand model for predicting the chronic toxicity of dissolved and precipitated aluminum to aquatic organisms. *Environ. toxicology chemistry* **37**, 70–79 (2018).
15. Feng, J., Li, X., Yang, Y. & Zhou, Z. Preparation of low-cost aluminum-loaded longan shell adsorbent for fluoride removal: Experimental and modeling studies. *J. Environ. Chem. Eng.* **10**, 108917 (2022).
16. Wang, J., Liu, Q., Xu, L., Siddique, M. S. & Yu, W. Impacts of water hardness on coagulation-UF-NF process using aluminum salts. *Sep. Purif. Technol.* **314**, 123611 (2023).
17. Gantayat, R. R. & Elumalai, V. Salinity-Induced Changes in Heavy Metal Behavior and Mobility in Semi-Arid Coastal Aquifers: A Comprehensive Review. *Water* **16**, 1052 (2024).
18. Ghani, S. A. A. Trace metals in seawater, sediments and some fish species from Marsa Matrouh Beaches in north-western Mediterranean coast, Egypt. *The Egypt. J. Aquatic Res.* **41**, 145–154 (2015).

ARTICLE IN PRESS

ARTICLE IN PRESS

19. Angel, B. M., Apte, S. C., Batley, G. E. & Golding, L. A. Geochemical controls on aluminium concentrations in coastal waters. *Environ. Chem.* **13**, 111–118 (2015).

20. Abdelaal, A., Abdelkader, A. I., Alshehri, F., Elatiar, A. & Almadani, S. A. Assessment and spatiotemporal variability of heavy metals pollution in water and sediments of a coastal landscape at the Nile Delta. *Water* **14**, 3981 (2022).

21. Mbandzi-Phorego, N., Puccinelli, E., Pieterse, P. P., Ndaba, J. & Porri, F. Metal bioaccumulation in marine invertebrates and risk assessment in sediments from South African coastal harbours and natural rocky shores. *Environ. Pollut.* 124230 (2024).

22. Walton, R. C., McCrohan, C. R., Livens, F. & White, K. N. Trophic transfer of aluminium through an aquatic grazer–omnivore food chain. *Aquatic toxicology* **99**, 93–99 (2010).

23. Hardisson, A., Revert, C., Gonzales-Weler, D. & Rubio, C. Aluminium exposure through the diet. *Food Sci. Nutr* **3**, 19 (2017).

24. Barnhart, B., Flinders, C., Ragsdale, R., Johnson, G. & Wiegand, P. Deriving human health and aquatic life water quality criteria in the United States for bioaccumulative substances: a historical review and future perspective. *Environ. Toxicol. Chem.* **40**, 2394–2405 (2021).

25. Alasfar, R. H. & Isaifan, R. J. Aluminum environmental pollution: the silent killer. *Environ. Sci. Pollut. Res.* **28**, 44587–44597 (2021).

26. Bonfiglio, R., Scimeca, M. & Mauriello, A. The impact of aluminum exposure on human health. *Arch. Toxicol.* **97**, 2997–2998 (2023).

27. Ergashov, A. THE GENERAL EFFECT OF ALUMINUM ON THE BODY. *Int. J. Med. Sci. And Clin. Res.* **3**, 90–95 (2023).

28. Yilmaz, N., Ozyigit, I. I., Dogan, I., Demir, G. & Yalcin, I. E. A case study performed in Küçükçekmece Lagoon channel/Istanbul, Turkey: how the heavy metal contamination and the seasonal variations on phytoplankton composition influence water quality. *Desalination Water Treat.* **239**, 126–136 (2021).

29. Trevizani, T. H., Domit, C., Santos, M. C. d. O. & Figueira, R. C. L. Bioaccumulation of heavy metals in estuaries in the southwest Atlantic Ocean. *Environ. Sci. Pollut. Res.* **30**, 26703–26717 (2023).

30. Yaseen, Z. M. *et al.* Heavy metals prediction in coastal marine sediments using hybridized machine learning models with metaheuristic optimization algorithm. *Chemosphere* **352**, 141329 (2024).

31. Deng, D., Wu, Y., Ren, B. & Yin, H. Applying Chemical and Statistical Analysis Methods to Evaluate Water and Stream Sediments around the Coal Mine Area in Dazhu, China. *Water* **15**, 1421 (2023).

32. Bhagat, S. K. *et al.* Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models. *Environ. Pollut.* **268**, 115663 (2021).

33. Mert, B. K. & Kasapogullari, D. A case study of using artificial neural networks to predict heavy metal pollution in Lake Iznik. *Environ. Monit. Assess.* **196**, 1–26 (2024).

34. Yolcu, U., Yalcin, I. E., Uras, M. E. & Ozyigit, I. I. Modeling the effects of essential heavy metals on environmental pollution: A linear and nonlinear prediction model via cascade forward-neural network. *Math. Methods Appl. Sci.* **47**, 4306–4318 (2024).

35. Chen, D. *et al.* Delineating and identifying risk zones of soil heavy metal pollution in an industrialized region using machine learning. *Environ. Pollut.* **318**, 120932 (2023).

36. Cui, S. *et al.* Advances and applications of machine learning and deep learning in environmental ecology and health. *Environ. Pollut.* 122358 (2023).

37. Gryech, I., Ghogho, M., Mahraoui, C. & Kobbane, A. An exploration of features impacting respiratory diseases in urban areas. *Int. J. Environ. Res. Public Heal.* **19**, 3095 (2022).

38. Gryech, I., Asaad, C., Ghogho, M. & Kobbane, A. Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: A systematic literature review. *Eng. Appl. Artif. Intell.* **137**, 109182 (2024).

39. Xu, X., Lai, T., Jahan, S. & Farid, F. Water and Sediment Analyse Using Predictive Models. *arXiv preprint arXiv:2203.03422* (2022).

40. Nguyen, H. N. *et al.* Predicting element concentrations by machine learning models in neutron activation analysis. *J. Radioanal. Nucl. Chem.* **333**, 1759–1768 (2024).

ARTICLE IN PRESS

ARTICLE IN PRESS

41. Han, Z., Zhang, S., He, L. *et al.* Predicting and investigating water quality index by robust machine learning methods. *J. Environ. Manag.* **381**, 125156 (2025).

42. Chen, X. *et al.* Water quality parameters-based prediction of dissolved oxygen in estuaries using advanced explainable ensemble machine learning. *J. Environ. Manag.* **380**, 125146 (2025).

43. Dai, H., Huang, G., Zeng, H. & Yang, F. PM2.5 concentration prediction based on spatiotemporal feature selection using XGBoost-MSCNN-GA-LSTM. *Sustainability* **13**, 12071 (2021).

44. Zhang, B., Zhang, Y. & Jiang, X. Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm. *Sci. Reports* **12**, 9244 (2022).

45. Demir, S. & Sahin, E. K. Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing. *Environ. Earth Sci.* **81**, 459 (2022).

46. Garabaghi, F. H., Benzer, S. & Benzer, R. Modeling dissolved oxygen concentration using machine learning techniques with dimensionality reduction approach. *Environ. Monit. Assess.* **195**, 879 (2023).

47. Ahajjam, A. *et al.* Enhancing prediction of wildfire occurrence and behavior in Alaska using spatio-temporal clustering and ensemble machine learning. *Ecol. Informatics* **85**, 102963 (2025).

48. Theng, D. & Bhoyar, K. K. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowl. Inf. Syst.* **66**, 1575–1637 (2024).

49. Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artif. intelligence* **97**, 273–324 (1997).

50. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. learning* **46**, 389–422 (2002).

51. Kuhn, M. *et al.* Package 'caret'. *The R J.* **223**, 48 (2020).

52. Holland, J. Adaptation in Natural and Artificial Systems, lstedn (1975).

53. Eiben, A. E. & Smith, J. E. *Introduction to evolutionary computing* (Springer, 2015).

54. Wolpert, D. H. Stacked generalization. *Neural networks* **5**, 241–259 (1992).

55. Seber, G. A. & Lee, A. J. *Linear regression analysis* (John Wiley & Sons, 2012).

56. Turksen, I. B. Fuzzy functions with LSE. *Appl. Soft Comput.* **8**, 1178–1188 (2008).

57. Aladag, C. H., Turksen, I. B., Dalar, A. Z., Egrioglu, E. & Yolcu, U. Application of type-1 fuzzy functions approach for time series forecasting. *Turkish J. Fuzzy Syst.* **5**, 1–9 (2014).

58. Tak, N. & Inan, D. Type-1 fuzzy forecasting functions with elastic net regularization. *Expert. Syst. with Appl.* **199**, 116916 (2022).

59. Cebeci, Z. *et al.* Package 'ppclust' (2019).

60. Hans, C. Elastic net regression modeling with the orthant normal prior. *J. Am. Stat. Assoc.* **106**, 1383–1393 (2011).

61. Kelly, J. W., Degenhart, A. D., Siewiorek, D. P., Smailagic, A. & Wang, W. Sparse linear regression with elastic net regularization for brain-computer interfaces. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4275–4278 (IEEE, 2012).

62. Friedman, J. *et al.* Package 'glmnet'. *CRAN R Repositary* **595** (2021).

63. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals statistics* 1189–1232 (2001).

64. 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*, 785–794 (2016).

65. Qiu, Y. *et al.* Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Eng. with Comput.* **38**, 4145–4162 (2022).

66. Chen, T., He, T., Benesty, M. & Khotilovich, V. Package 'xgboost'. *R version* **90**, 40 (2019).

67. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

68. González-Sopeña, J., Pakrashi, V. & Ghosh, B. An overview of performance evaluation metrics for short-term statistical wind power forecasting. *Renew. Sustain. Energy Rev.* **138**, 110515 (2021).

69. Bat, L., Şahin, F., Öztekin, A., Özsandıkçı, U. & Özkan, E. Y. Trace elements pollution in surface sediment of the sea of marmara coastal and transition water. *Mar. Pollut. Bull.* **218**, 118067 (2025).

ARTICLE IN PRESS

70. Pourang, N. *et al.* Spatiotemporal patterns, source apportionment, and ecological risk of major and trace elements in sediment cores from anzali international wetland. *Sci. Reports* **15**, 41965 (2025).

71. Li, J. Natural and anthropogenic controls on heavy metal distribution in east china sea sediments. *Front. Mar. Sci.* **12**, 1689901 (2025).

72. Li, F., Yin, H., Zhu, T. & Zhuang, W. Understanding the role of manganese oxides in retaining harmful metals: Insights into oxidation and adsorption mechanisms at microstructure level. *Eco-Environment & Heal.* **3**, 89–106 (2024).

73. Zeng, P. *et al.* RWKV-SKF: A recurrent architecture with state-space and frequency-domain filtering for dissolved oxygen predicting and revealing influencing mechanisms. *Inf. Sci.* 123018 (2025).

Funding

This study was partially supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Grant Number 123F266. The authors thank TÜBİTAK for their support. (Funding: Not received for publication fee)

Author contributions statement

AU: Formal analysis, software, visualization, writing-original draft, writing-review; NT: Supervision, conceptualization, formal analysis, writing-original draft, writing-review; AHO: Financial support, material preparation, data curation, writing-review; IIO: Financial support, data curation, writing-review, and editing.

Additional information

Competing interests. On behalf of all authors of the paper, I declare that there is no competing interests.

ARTICLE IN PRESS