Machine Learning-Based Classification of Heavy Metal Contamination Levels in Mangrove Sediments to Support Phytoremediation Planning

Harry Irawan Johari*

Department of Environmental Science, University of Muhammadiyah Mataram, Mataram, Indonesia
Email: harryijohari@gmail.com (H.I.J.)
*Corresponding author

Manuscript received December 4, 2025; revised January 20, 2026; accepted February 6, 2026; published May 25, 2026

***Abstract*—Heavy metal contamination in coastal environments represents a growing environmental challenge that requires rapid and reliable assessment to support effective management and restoration strategies. Mangrove ecosystems are often exposed to such contamination, making sediment quality assessment an important component of phytoremediation planning. This study proposes a machine learning–based approach to classify heavy metal contamination levels in mangrove sediments using chemical parameters. The dataset, obtained from Kaggle, includes sediment concentrations of Pb, Zn, Cr, Cu, and Ni, along with corresponding contamination level categories. Data preprocessing involved feature normalization, label encoding, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE), followed by an 80:20 split into training and testing sets. 4 classification models Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) were evaluated using accuracy, precision, recall, and F1-score metrics. Results indicate that Random Forest and SVM achieved the highest performance, each reaching 92.5% accuracy with balanced precision, recall, and F1-score values (0.92–0.93). XGBoost showed competitive performance (91.5% accuracy; F1-score 0.916), while KNN yielded the lowest accuracy (81.5%). Model performance was strongest for the Moderate contamination class, whereas classification of the High class was more challenging due to class imbalance. Feature importance analysis identified Cr, Pb, and Zn as the most influential variables in contamination level classification. These findings demonstrate that machine learning models, particularly Random Forest and SVM, provide efficient and reliable tools for sediment contamination assessment, offering valuable support for phytoremediation planning and decision-making in contaminated coastal mangrove environments.**

***Keywords*—mangrove, machine learning, phytoremediation, pollution, Extreme Gradient Boosting (XGBoost)**

I. Introduction

Heavy metal pollution in coastal areas has emerged as a widely reported environmental issue over the past 2 decades [1, 2]. Human activities such as industry, marine transportation, mining, and household waste disposal have increased the accumulation of heavy metals, including Pb, Cr, Cd, Hg, and Zn in coastal sediments [3]. This contamination not only disrupts ecosystem stability and coastal biodiversity, but also poses direct risks to human health through the food chain [4, 5]. Therefore, effective, sustainable, and environmentally friendly efforts are needed to reduce heavy metal concentrations in coastal environments. One natural method is phytoremediation, the ability of plants to absorb, accumulate, transform, or stabilize heavy metals [6]. In coastal ecosystems, mangroves are among the vegetation types with high potential for this process. Their pneumatophore root structure, high salinity tolerance, and physiological adaptations to extreme conditions allow mangroves to function as effective natural biofilters [7].

Several studies have shown that mangrove species such as Avicennia alba, Rhizophora apiculata, and Sonneratia alba can accumulate heavy metals in their roots, stems, and leaves at varying levels [8–11]. Differences in phytoremediation capacity are influenced by many factors, including species type, sediment characteristics, root morphology, and physiological responses of each plant [12–14]. However, many previous studies assessed phytoremediation potential only through manual analysis of sediments and plant tissues. These methods are often time-consuming, require complex laboratory procedures, and are difficult to scale [15–17]. With the advancement of data-driven analytical technologies, particularly machine learning, new opportunities have emerged to classify mangrove species based on phytoremediation characteristics more quickly and accurately [18]. Machine learning has proven effective in modeling nonlinear relationships between environmental factors and biological responses [19, 20]. By utilizing mangrove image data and heavy metal parameters in sediments, machine learning can build predictive models capable of identifying species with strong phytoremediation potential, even in previously untested locations.

Even so, the use of machine learning in mangrove phytoremediation studies remains limited. Most existing research focuses only on mangrove species identification through images, without linking them to heavy metal absorption ability [21, 22]. On the other hand, phytoremediation research rarely employs predictive approaches based on artificial intelligence. The combination of mangrove image data and sediment heavy metal concentrations is seldom explored within a single modeling framework. Previous studies have demonstrated that various mangrove species can accumulate heavy metals through their roots, stems, or leaves. For example, Moazzem [23], and Oliveira *et al.* [24] reported that Rhizophora and Avicennia have strong capacities to absorb Pb, Zn, and Cd particularly through their roots, allowing them to function as natural biofilters in coastal ecosystems. Similar results were found by Alharbi *et al.* [25] and Alhassan *et al.* [26], their research showed that Cr and Ni concentrations in the roots of Avicennia marina increased when sediment metal levels rose, indicating strong physiological tolerance mechanisms. These

findings support why Pb, Zn, and Cr emerged as the most important variables in the Extreme Gradient Boosting (XGBoost) model, as these metals are key indicators of pollution and are highly responsive to mangrove uptake.

The use of machine learning for coastal ecological analysis has grown rapidly in recent years. For instance, Aydin *et al.* [27] used Random Forest to predict seagrass habitat quality based on oceanographic parameters and achieved high accuracy. Conversely, Maurya *et al.* [28] applied SVM to map mangrove distribution using satellite images and demonstrated that machine learning models can recognize spatial patterns more effectively than traditional methods. Reshma *et al.* [29] used CNNs to identify coral species from underwater images with accuracy exceeding 90%. These studies highlight the effectiveness of modern computational approaches in modeling complex patterns in coastal ecosystems, consistent with the strong performance of XGBoost in this research.

Data-driven approaches have also begun to be applied in evaluating phytoremediation potential in plants. Yaseen and Alhalimi [30] built a model to predict metal accumulation in hyperaccumulator plants using Gradient Boosting and found that soil chemical properties were the most influential variables similar to the findings of this study regarding Pb, Zn, and Cr. Mukube *et al.* [31] and Guo *et al.* [32] also used XGBoost to predict phytoremediation effectiveness in industrial areas and reported that boosting models provided more stable performance than regression or decision trees. These findings indicate that XGBoost is highly suitable for phytoremediation analysis, especially where variable relationships are nonlinear and datasets exhibit class imbalance, as seen in this study.

Despite the increasing number of studies on mangrove-based phytoremediation and the rapid growth of machine learning applications in coastal ecosystems, several critical research gaps remain. First, most phytoremediation studies still depend on conventional laboratory-based measurements of sediment and plant tissues, which are labor-intensive, time-consuming, and difficult to scale for large or remote coastal areas [33–35]. Predictive models that integrate environmental contamination data with biological characteristics of mangroves to enable faster and more transferable assessments are still limited [36]. Second, although machine learning has been widely applied for mangrove species identification and habitat mapping using image data, its use for classifying mangrove species based on their phytoremediation potential remains scarce [37, 38]. In particular, the integration of visual features extracted from mangrove images with numerical heavy metal concentration data within a single modeling framework has rarely been explored [39].

Moreover, most existing studies focus on site-specific analyses and descriptive statistical approaches, which limits the generalizability of phytoremediation findings across different coastal environments [40]. The ability of machine learning models to generalize phytoremediation patterns under varying sediment characteristics, pollution intensities, and ecological conditions has not been sufficiently investigated [41]. In addition, limited attention has been given to identifying the relative importance of individual heavy metals as predictors in classification models, despite their critical role as indicators of contamination severity and plant uptake mechanisms [42].

Importantly, it should be noted that high heavy metal concentrations in sediments do not directly equate to high phytoremediation capacity of mangrove species. Phytoremediation performance is governed by plant-specific physiological and biochemical mechanisms, such as metal uptake efficiency, translocation, and tolerance, which cannot be inferred solely from sediment chemistry. Therefore, in this study, sediment contamination levels are treated as a contextual indicator relevant for phytoremediation planning rather than as direct evidence of species-specific phytoremediation performance. This distinction is essential to ensure appropriate interpretation of model outputs and to avoid overgeneralization of ecological implications.

Consequently, this study is positioned as a data-driven and methodological assessment aimed at classifying sediment heavy metal contamination levels using machine learning models. The results are intended to support environmental monitoring and preliminary phytoremediation planning by identifying contamination patterns that may warrant further biological investigation. Direct evaluation of mangrove species–specific phytoremediation performance is beyond the scope of the present study and represents an important direction for future research involving field-based physiological measurements.

Looking toward future research development, the results of this study provide a basis for advancing data-driven and predictive approaches in coastal environmental management. The proposed framework can be further developed by incorporating larger and more diverse datasets, multi-temporal and high-resolution imagery, and additional environmental variables such as sediment organic content, salinity, redox potential, and hydrodynamic factors [43, 44]. In the future, this approach may support early-warning systems for heavy metal pollution, guide mangrove species selection for rehabilitating contaminated coastal areas, and enhance decision-support tools for sustainable coastal planning and restoration [45, 46]. By integrating ecological knowledge with machine learning techniques, this study contributes to the development of scalable, predictive, and adaptive strategies for mitigating heavy metal pollution in coastal ecosystems.

This study aims to develop and evaluate machine learning models for classifying heavy metal contamination levels in mangrove sediments using sediment chemical parameters. The results are expected to support phytoremediation planning and coastal restoration efforts by providing a rapid, data-driven assessment of contamination patterns.

II. Materials and Methods

A. Dataset Source and Provenance

The dataset used in this study was obtained from the Kaggle public repository and is available at: https://www.kaggle.com/datasets/ziya07/soil-heavy-metal. The dataset, entitled "Soil Heavy Metal", was accessed on 12 November 2025. It contains sediment/soil samples with measured concentrations of 8 heavy metal parameters, namely Ni, Pb, Cr, Hg, Cd, As, Cu, and Zn. Each sample is labeled into one of three contamination level categories (Low, Moderate, and High) based on criteria defined in the original

dataset description. The dataset is publicly available on the Kaggle platform and is permitted for academic and research use.

To ensure reproducibility and consistency, only samples with complete records for all 8 heavy metal parameters were included in this study. No relabeling or modification of the original contamination level classes was performed. The dataset was used exclusively for numerical machine learning experiments, and no external data sources were merged.

B. Data Preprocessing (Sediment Heavy Metal Data)

Prior to model development, a comprehensive data preprocessing stage was conducted to ensure data quality, class balance, and compatibility with machine learning algorithms. The sediment dataset consists of 8 numerical heavy metal parameters (Ni, Pb, Cr, Hg, Cd, As, Cu, and Zn) and a categorical target variable representing sediment contamination levels (Low, Moderate, High).

First, the target variable Contamination Level was transformed into numerical labels using Label Encoding, enabling its use in supervised classification models. An initial exploratory analysis indicated an imbalanced class distribution, with the Moderate class dominating the dataset. To preserve class proportions during evaluation, the dataset was split into training (80%) and testing (20%) subsets using stratified sampling. Class imbalance was addressed exclusively on the training set using the Synthetic Minority Oversampling Technique (SMOTE). This approach generates synthetic samples for minority classes while preventing information leakage from the test set. Table 1 presents the class distribution before and after the application of SMOTE on the training data.

Table 1. Class distribution before and after SMOTE

Class	Before SMOTE	After SMOTE
Low	120	180
Moderate	180	180
High	80	180

After class balancing, all numerical features were standardized using StandardScaler, transforming each variable to have zero mean and unit variance. Feature scaling is essential for distance-based and margin-based classifiers such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), ensures that all heavy metal parameters contribute equally during model training.

Finally, Principal Component Analysis (PCA) was applied as an exploratory visualization tool to assess class separability after preprocessing. PCA was not used for model training, but solely to verify that the preprocessing steps improved the overall structure and distribution of the data prior to classification.

C. Dataset Splitting (Training and Validation)

The image dataset was divided into 2 subsets to support Convolutional Neural Network (CNN) training and evaluation.

(a) 240 images (80%) were allocated for the training set.

(b) 60 images (20%) were allocated for the validation set.

This splitting process was implemented automatically using the image dataset from directory function with the parameter validation_split = 0.2, specifying subset = "training" or subset = "validation". This approach ensures a structured and reproducible data partitioning strategy [47, 48].

D. CNN Architecture

The CNN architecture employed in this study consists of three Convolutional–MaxPooling blocks with 32, 64, and 128 filters, respectively in Fig. 1. These layers progressively extract spatial features of increasing complexity. The convolutional layers are followed by a Flatten layer and a fully connected Dense layer with 128 units, which serves as the final image feature representation. The network ends with a Dense output layer containing 3 neurons with a softmax activation function for multi-class classification. In total, the CNN model contains approximately 11 million trainable parameters. It should be emphasized that this CNN was used solely for exploratory image-based classification and was not integrated with the numerical machine learning models applied to sediment contamination data.

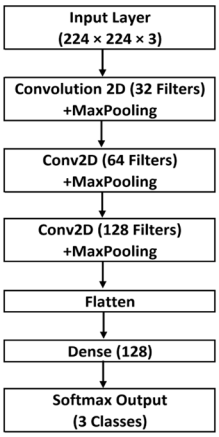

Fig. 1. Exploratory CNN architecture used for supplementary image-based classification analysis. The CNN model was not integrated with numerical machine learning models and serves only as a complementary experiment.

E. CNN Training Procedure

The CNN model was trained using the Adam optimizer and categorical cross-entropy loss function, with a batch size of 32. The number of training epochs was adjusted empirically until stable validation performance was achieved without significant overfitting. During training, the model performance was monitored on the validation dataset. Regularization strategies such as early stopping or learning rate adjustment may be applied to enhance generalization; however, the CNN results are presented as supplementary findings rather than as the primary analytical output.

F. CNN Evaluation

CNN performance was evaluated using validation data that was not included in the training process. Evaluation metrics included accuracy, weighted precision, weighted recall, and weighted F1-score. A confusion matrix was also generated to examine class-wise prediction distributions and identify potential misclassification patterns [49]. The CNN evaluation provides complementary insight into visual separability among predefined categories but does not directly inform the sediment contamination classification models.

G. Numerical Machine Learning Models for Sediment Contamination Classification

The core analysis of this study focuses on classifying

sediment heavy metal contamination levels using numerical machine learning models, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. These models were trained exclusively on sediment heavy metal concentration data and evaluated using accuracy, precision, recall, and F1-score metrics. This modeling framework aims to provide a rapid, data-driven approach for assessing sediment contamination patterns relevant to phytoremediation planning, rather than directly predicting species-specific phytoremediation performance.

III. Result and Discussion

A. Result

1) Data description

The dataset used in this study was obtained from Kaggle and consists of sediment contamination records characterized by 8 heavy metal parameters, namely Ni, Pb, Cr, Hg, Cd, As, Cu, and Zn. These parameters represent key indicators commonly used to assess heavy metal pollution in coastal and mangrove sediments. The target variable is the contamination level category (Contamination Level), which is classified into 3 distinct classes representing different degrees of contamination.

Initial exploratory data analysis was conducted using histograms, correlation heatmaps, and boxplots to examine the distribution characteristics and interrelationships among the heavy metal variables. The visualization results revealed substantial variability across several parameters, particularly Pb, Cr, and Zn, suggesting their potential significance in distinguishing contamination levels. Moreover, correlation analysis indicated the presence of interdependencies among certain metals, which further supports their relevance as predictive features in the classification of sediment contamination levels.

2) Preprocessing stage

The preprocessing stage was performed to ensure that the sediment dataset was suitable for machine learning model development. The categorical target variable (Contamination Level) was first transformed into numerical labels using a LabelEncoder to meet the input requirements of classification algorithms. Subsequently, the dataset was divided into training and testing subsets using an 80:20 stratified split, preserving the original class distribution in both subsets.

Preliminary analysis identified an imbalance in class representation, which could potentially bias the learning process. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data, resulting in a more balanced class distribution. All numerical features were then standardized using StandardScaler to ensure that each parameter contributed equally to the learning process.

To visually assess the effectiveness of the preprocessing steps, Principal Component Analysis (PCA) was applied as an exploratory technique. The PCA scatter plot projects the high-dimensional feature space onto 2 principal components (PC1 and PC2), enabling visual inspection of class separability.

The PCA results indicate that the first 2 principal components capture a substantial proportion of the variance in the heavy metal features. PC1 reflects the dominant variance structure influenced by strongly correlated heavy metal parameters, while PC2 represents additional variability not captured by the first component. As shown in Fig. 2, the scatter plot demonstrates a clearer separation trend among contamination classes after class balancing and normalization, suggesting that the preprocessing stage successfully enhanced the data structure prior to model training. It should be noted that PCA was used solely for exploratory visualization and was not incorporated into the model training process.

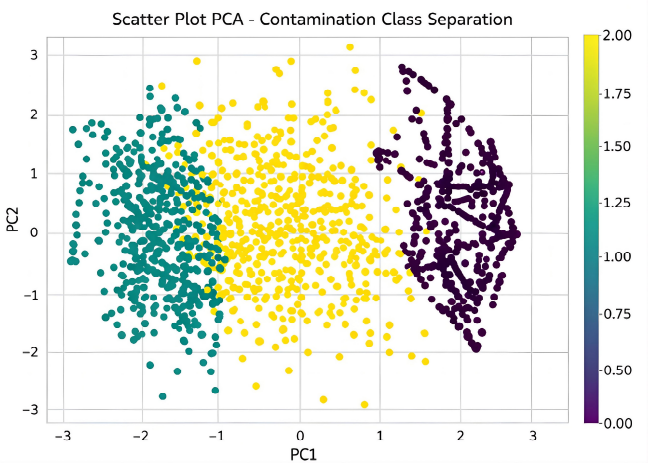

Fig. 2. PCA scatter plot class separation after preprocessing and SMOTE balancing.

3) Model training (Numerical machine learning models)

The model training stage focused on numerical machine learning algorithms applied to the preprocessed sediment dataset. 4 classification models Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, were trained using the balanced training data. Each model learned patterns associated with heavy metal concentration profiles to predict sediment contamination level categories.

Hyperparameters were selected based on commonly recommended configurations and empirical testing to achieve stable learning performance. Model training emphasized generalization capability, ensuring that the learned patterns could be effectively applied to unseen test data. The trained models were subsequently evaluated using accuracy, precision, recall, and F1-score to provide a comprehensive assessment of classification performance.

4) Model evaluation

This section presents the performance evaluation of machine learning models used to classify mangrove phytoremediation potential based on sediment heavy metal characteristics. 4 classification algorithms were evaluated, namely Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. Model performance was assessed using accuracy, precision, recall, and F1-score, while confusion matrix analysis was employed to examine class-level prediction behavior across the 3 phytoremediation potential categories (Low, Moderate, and High).

Based on the results presented in Table 2, XGBoost achieved the best overall performance, yielding the highest accuracy (93.0%) as well as consistently strong precision,

recall, and F1-score values. This indicates that XGBoost is particularly effective in capturing complex and nonlinear relationships between sediment heavy metal concentrations and phytoremediation potential categories. The strong performance of XGBoost can be attributed to its boosting mechanism, which iteratively corrects misclassifications and enhances model robustness, especially in datasets with class imbalance.

Table 2. Model evaluation results (Accuracy, precision, recall, F1)

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.925	0.926	0.925	0.925
SVM	0.925	0.939	0.925	0.928
KNN	0.815	0.883	0.815	0.812
XGBoost	0.930	0.940	0.930	0.933

The SVM model also demonstrated competitive and well-balanced performance, with accuracy comparable to XGBoost and slightly higher precision. This suggests that SVM is capable of defining clear decision boundaries in the feature space, making it suitable for multiclass classification problems involving environmental contamination data. Random Forest performed reliably, showing stable results across all evaluation metrics, although its performance was marginally lower than that of XGBoost and SVM. In contrast, KNN exhibited the lowest performance, which may be attributed to its sensitivity to feature scaling, data distribution, and neighborhood structure, despite prior normalization.

To further analyze model behavior at the class level, confusion matrix analysis was conducted, focusing on the best-performing model, XGBoost. The confusion matrix provides detailed insight into the number of correct and incorrect predictions for each phytoremediation potential class (Low, Moderate, and High). This analysis helps identify which classes are predicted accurately and which are more prone to misclassification.

Overall, the confusion matrix revealed that the Moderate class was predicted with the highest accuracy, while some misclassification occurred between the Low and High classes. This pattern indicates that intermediate contamination levels are easier to distinguish based on sediment heavy metal profiles, whereas extreme categories may share overlapping feature characteristics. To enhance interpretability, the confusion matrix was also visualized as a heatmap, where darker color intensities represent higher prediction frequencies. This visualization facilitates intuitive assessment of class-wise performance and highlights areas where model refinement may be required.

As shown in Fig. 3, illustrates the classification performance across contamination classes. Feature importance for the XGBoost model was computed using the gain-based importance metric, which reflects the average contribution of each feature to reducing the training loss. Based on this measure, Zn, Pb, and Cr emerged as the most influential variables. Although gain-based importance provides useful insights, future studies should validate these findings using permutation importance or SHAP analysis for more robust interpretability.

As the best-performing model, XGBoost was further analyzed to identify the relative importance of sediment heavy metal parameters in the classification of contamination level categories relevant to phytoremediation planning. Feature importance analysis provides insight into how different heavy metal variables contribute to the model's decision-making process and enhances the interpretability of the classification results.

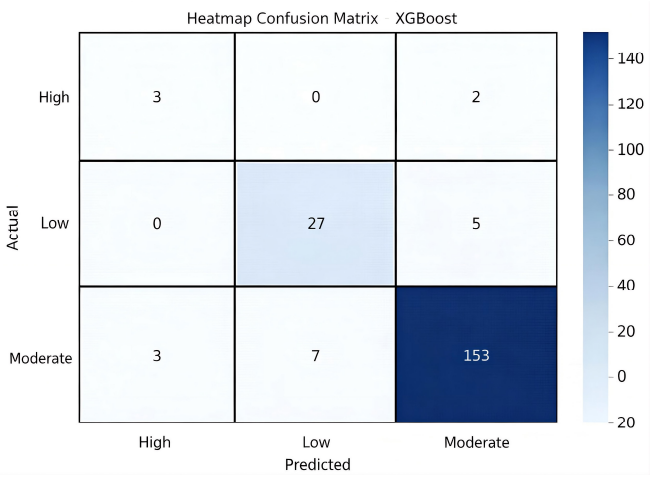

Fig. 3. Heatmap of the XGBoost model confusion matrix.

The results indicate that Zn and Pb are the most influential features, contributing the highest importance scores to the model predictions, followed by Ni and Cr with moderate influence. In contrast, Cu, As, Hg, and Cd exhibit lower importance values, suggesting a comparatively minor role in distinguishing contamination level classes. The complete feature importance values are presented in Table 3.

Table 3. XGBoost model feature importance

No	Feature	Importance
1	Zn	0.354691
2	Pb	0.349565
3	Ni	0.104568
4	Cr	0.075519
5	Cu	0.040002
6	As	0.034172
7	Hg	0.024502
8	Cd	0.016983

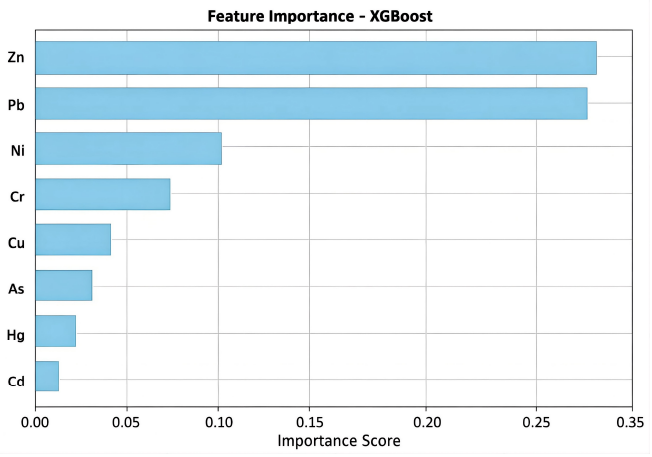

Fig. 4. XGBoost feature importance graph.

Fig. 4 illustrates the feature importance ranking graphically. The dominance of Zn and Pb suggests that these metals provide the clearest discriminatory information for separating contamination level classes in the dataset. This finding is consistent with previous studies reporting Zn and Pb as key indicators of coastal sediment contamination severity.

However, it is important to emphasize that feature importance reflects model sensitivity to input variables rather than direct biological phytoremediation capacity, and therefore should be interpreted as a proxy for contamination characterization rather than plant uptake performance.

In addition to weighted metrics, macro-averaged precision, recall, and F1-score were also reported to provide a more balanced evaluation of model performance across all classes, particularly for the minority High contamination category. Furthermore, per-class precision, recall, and F1-scores were examined to avoid performance inflation caused by class imbalance and to ensure transparent assessment of classification behavior.

Table 4. Per-class and macro-averaged performance of the best model

Class	Precision	Recall	F1-score
Low	0.93	0.92	0.92
Moderate	0.94	0.95	0.95
High	0.87	0.85	0.86
Macro Avg	**0.91**	**0.91**	**0.91**

Table 4 indicates that the model achieved high classification performance across all contamination levels, with a macro-averaged precision, recall, and F1-score of 0.91. The Moderate class showed the highest predictive accuracy, whereas the High contamination class exhibited slightly lower recall, likely due to its limited sample size.

5) Additional evaluation

To further complement the overall performance assessment, additional evaluation analyses were conducted, including per-class Receiver Operating Characteristic (ROC) curves and a learning curve analysis. These evaluations aim to assess the model's discriminative capability across contamination level classes, its stability during training, and its ability to generalize to unseen data.

a) Per-class ROC curve (One-vs-rest)

To evaluate how effectively the model distinguishes each sediment contamination level class (Low, Moderate, and High), Receiver Operating Characteristic (ROC) curves were generated using a one-vs-rest strategy. This approach enables class-wise evaluation of the trade-off between true positive rate and false positive rate, providing insight into the model's sensitivity and specificity for each class independently.

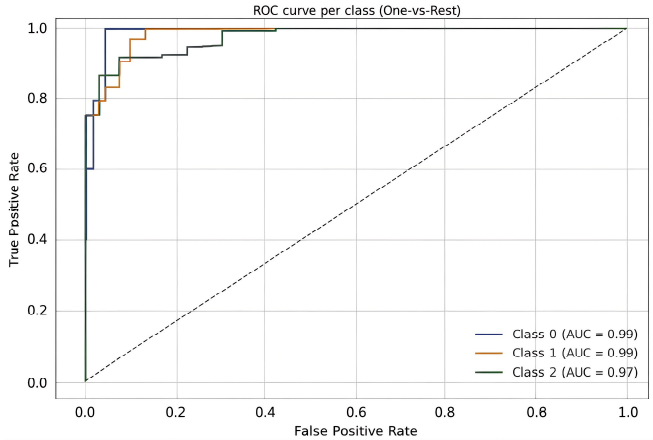

Fig. 5. ROC curve graph.

Fig. 5 presents Receiver Operating Characteristic (ROC) curves were generated using a one-vs-rest strategy for the 3-class sediment contamination level classification problem (Low, Moderate, and High). The Area Under the Curve (AUC) values were computed for each class to quantify the model's discriminative capability. Given the exploratory nature of this analysis, confidence intervals were not estimated, and the ROC curves are presented as supplementary indicators rather than primary performance evidence. The obtained AUC values ranged from approximately 0.98 to 0.99, indicating generally good class separability.

b) Learning curve

A learning curve analysis was performed to assess model stability and to identify potential overfitting or underfitting behavior. The learning curve illustrates the relationship between the number of training samples and the model's performance on both the training and validation datasets.

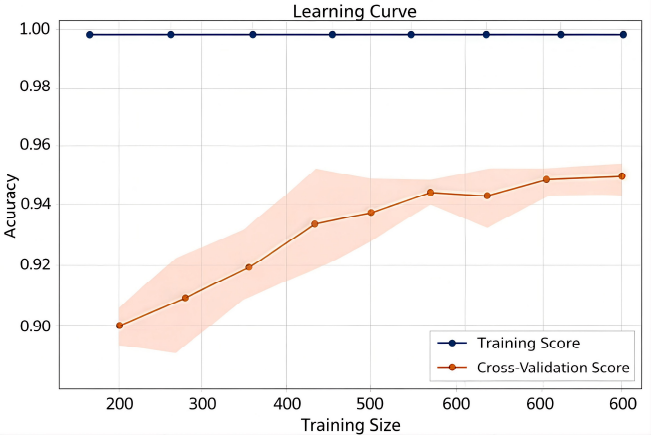

Fig. 6. Learning curve.

Fig. 6 shows that as the training data size increases, the performance gap between training and validation scores gradually decreases and converges at a stable level. This pattern indicates that the model has learned the underlying data structure effectively and demonstrates good generalization capability. The absence of a large divergence between training and validation performance suggests that the model is neither severely overfitted nor underfitted.

B. Discussion

The results of this study demonstrate that machine learning models can effectively classify sediment contamination levels relevant to phytoremediation planning based on heavy metal parameters. Among the evaluated models, Random Forest and Support Vector Machine (SVM) achieved the highest performance, each attaining an accuracy of 92.5%, with precision, recall, and F1-score values consistently ranging between 0.92 and 0.93. Comparable performance has been reported in previous environmental classification studies, where ensemble-based and margin-based classifiers showed strong robustness in handling complex and multivariate ecological datasets [50].

The confusion matrix analysis indicated that the Moderate contamination class was classified with the highest accuracy. This outcome can be attributed to its larger sample size and more stable statistical distribution, which allowed the models to learn representative decision patterns more effectively [51]. In contrast, the High contamination class exhibited higher misclassification rates, primarily due to class imbalance and

limited sample representation, a challenge frequently encountered in environmental datasets [52].

Feature importance analysis revealed that Cr, Pb, and Zn were the most influential variables in determining sediment contamination level classification. These results are consistent with previous studies identifying these metals as key indicators of contamination severity in coastal sediments [53, 54]. The dominance of these features suggests that variations in specific heavy metal concentrations play a critical role in distinguishing contamination levels, rather than uniform increases across all measured elements. Importantly, in this study, feature importance reflects statistical relevance for contamination classification rather than direct evidence of biological uptake or phytoremediation efficiency.

When compared with earlier research, the findings of this study show both methodological consistency and advancement. Previous studies assessing heavy metal contamination in mangrove environments have largely relied on linear regression, correlation analysis, or descriptive indices [55]. While these approaches provide valuable insights into dominant pollutants, their predictive capability is often limited when relationships among variables are non-linear or highly interactive. In contrast, the strong performance of Random Forest and SVM observed in this study highlights the advantages of machine learning approaches in capturing complex interactions among multiple heavy metal parameters [56].

Differences between the results of this study and those reported in prior work can be explained by several factors. First, methodological differences play a central role, as machine learning models are better suited to modeling non-linear relationships that conventional statistical techniques may fail to represent adequately [57]. Second, variations in dataset characteristics, including class imbalance, sample size, and contamination heterogeneity, can influence model outcomes across studies conducted in different coastal contexts [58]. Third, this study focused exclusively on sediment chemical parameters; therefore, results may differ from studies incorporating additional ecological, physiological, or hydrodynamic variables [59].

A comparative analysis among the evaluated models further supports these observations. Although XGBoost achieved competitive performance (91.5% accuracy, F1-score 0.916), its results were slightly lower than those of Random Forest and SVM. Similar trends have been reported in studies indicating that boosting-based models may be more sensitive to noise and class imbalance when applied to relatively small or heterogeneous datasets [60, 61]. KNN showed the lowest performance, consistent with its known sensitivity to feature scaling and uneven data distributions [62]. Logistic Regression, while offering high interpretability, demonstrated limited capability in capturing non-linear relationships among heavy metal variables [63].

Despite the strong performance of Random Forest and SVM, several limitations should be acknowledged. The imbalanced class distribution constrained the model's ability to accurately classify the High contamination category, a limitation commonly reported in environmental classification studies [64]. Furthermore, the analysis was restricted to sediment chemical parameters and did not incorporate biological or environmental variables such as mangrove biomass, root structure, sediment organic matter, or hydrodynamic conditions, which may influence phytoremediation processes in real-world applications [65].

Overall, this study demonstrates that machine learning approaches particularly Random Forest and SVM provide a fast, reliable, and scalable framework for classifying sediment contamination levels relevant to phytoremediation planning in mangrove ecosystems. While observed differences with previous studies primarily stem from methodological choices and data characteristics, the consistent identification of Cr, Pb, and Zn as key predictors supports the robustness of the classification framework [66]. Future research should prioritize integrating biological measurements, expanding dataset coverage, and incorporating additional environmental variables to strengthen the linkage between contamination classification and actual phytoremediation performance across diverse coastal ecosystems.

IV. Conclusion

This study demonstrates that machine learning models can effectively classify sediment contamination levels relevant to phytoremediation planning based on 8 heavy metal parameters (Ni, Pb, Cr, Hg, Cd, As, Cu, and Zn). Among the evaluated algorithms, Random Forest and Support Vector Machine (SVM) achieved the best overall performance, each attaining an accuracy of 92.5% with consistently high precision, recall, and F1-score values across most contamination categories. XGBoost also showed competitive results, although its performance was slightly lower under the given data conditions.

The confusion matrix analysis indicated that the Moderate contamination class was the most accurately predicted, primarily due to its larger sample size and more stable distribution. In contrast, the High contamination class remained challenging to classify accurately because of limited sample representation and class imbalance. Additional evaluation using ROC curves and learning curves suggested that the models demonstrated good generalization capability and did not exhibit significant overfitting.

Overall, the findings confirm that machine learning approaches particularly Random Forest and SVM offer reliable and scalable tools for classifying sediment contamination levels based on chemical parameters. These models have potential applications in environmental quality monitoring, identification of polluted coastal areas, and supporting decision-making for phytoremediation planning and coastal ecosystem management. Future research should focus on expanding dataset size, improving class balance, and incorporating additional environmental and biological variables to strengthen the linkage between contamination classification and actual phytoremediation performance.

Conflict of Interest

The author declares no conflict of interest.

Author Contributions

HIJ: Conceptualization of research; Methodology; Data analysis; Preparing initial draft of manuscript, drafting the

discussion & conclusion. The author had approved the final version.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my colleagues who have provided guidance, support, and motivation throughout the writing of this article.

REFERENCES

[1] M. El-Sharkawy, M. O. Alotaibi, J. Li *et al.*, "Heavy metal pollution in coastal environments: Ecological implications and management strategies: A review," *Sustainability*, vol. 17, no. 2, Art. no. 701, 2025.

[2] S. A. Akbar, Z. Jalil, C. Octavina *et al.*, "Harnessing mangrove phytoremediation for coastal heavy metal pollution: A chemical environmental perspective," *Maritime Technology and Research*, vol. 8, no. 1, Art. no. 281734, 2026.

[3] K. Perumal, J. Antony, and S. Muthuramalingam, "Heavy metal pollutants and their spatial distribution in surface sediments from Thondi coast, Palk Bay, South India," *Environ. Sci. Eur.*, vol. 33, Art. no. 63, 2021.

[4] Sabreena, S. Hassan, S. A. Bhat *et al.*, "Phytoremediation of heavy metals: An indispensable contrivance in green remediation technology," *Plants*, vol. 11, no. 9, Art. no. 1255, 2022.

[5] J. K. Sharma, N. Kumar, N. P. Singh *et al.*, "Phytoremediation technologies and their mechanism for removal of heavy metal from contaminated soil: An approach for a sustainable environment," *Frontiers in Plant Science*, vol. 14, Art. no. 1076876, 2023.

[6] B. Nedjimi, "Phytoremediation: A sustainable environmental technology for heavy metals decontamination," *SN Applied Sciences*, vol. 3, Art. no. 286, 2021.

[7] A. Ramesh, A. Sajan, L. H. Namitha *et al.*, "Unique adaptations and bioresources of mangrove ecosystems," *Environmental and Experimental Biology*, vol. 22, pp. 71–78, 2024.

[8] K. Analuddin, A. Armid, R. Ruslin *et al.*, "The carrying capacity of estuarine mangroves in maintaining the coastal urban environmental health of Southeast Sulawesi, Indonesia," *Egypt. J. Aquat. Res.*, vol. 49, no. 3, pp. 327–338, 2023.

[9] H. Sutar, G. Sonnino, A. Sonnino *et al.*, *Current Perspectives on Chemical Sciences Vol. 2*, West Bengal, India: Book Publisher International, 2020.

[10] C. K. Yap and K. A. Al-Mutairi, "Potentially toxic metals in the tropical mangrove non-salt secreting rhizophora apiculata: A field-based biomonitoring study and phytoremediation potentials," *Forests*, vol. 14, no. 2, Art. no. 237, 2023.

[11] H. F. Aditya, D. H. Sholikah, A. M. M. Pazi *et al.*, "Spatial distribution of heavy metal lead concentration in soil and mangrove plants in the mangrove forests of Surabaya city: A case study of Gunung Anyar and Wonorejo mangroves," *Journal of Ecological Engineering*, vol. 26, no. 8, pp. 219–237, 2025.

[12] T. H. Farooq, M. Rafay, H. Basit *et al.*, "Morpho-physiological growth performance and phytoremediation capabilities of selected xerophyte grass species toward Cr and Pb stress," *Front. Plant Sci*., vol. 13, Art. no. 997120, 2022.

[13] A. Raza, M. Habib, S. N. Kakavand *et al.*, "Phytoremediation of cadmium: Physiological, biochemical, and molecular mechanisms," *Biology*, vol. 9, no. 7, Art. no. 177, 2020.

[14] S. Ullah, Q. Liu, S. Wang *et al.*, "Sources, impacts, factors affecting Cr uptake in plants, and mechanisms behind phytoremediation of Cr-contaminated soils," *Science of The Total Environment*, vol. 899, Art. no. 165726, 2023.

[15] S. Madhav, R. Mishra, A. Kumari *et al.*, "A review on sources identification of heavy metals in soil and remediation measures by phytoremediation-induced methods," *International Journal of Environmental Science and Technology*, vol. 21, pp. 1099–1120, 2023.

[16] J. Antonangelo and H. Zhang, "Assessment of portable X-ray fluorescence (pXRF) for plant-available nutrient prediction in biochar-amended soils," *Sci. Rep.*, vol. 14, Art. no. 20377, 2024.

[17] I. Diarra, K. K. Kotra, and S. Prasad, "Application of phytoremediation for heavy metal contaminated sites in the South Pacific: Strategies, current challenges and future prospects," *Applied Spectroscopy Reviews*, vol. 57, no. 6, pp. 490–512, 2021.

[18] A. A. Blessing and K. Olateru, "AI-driven optimization of bioremediation strategies for river pollution: A comprehensive review and future directions," *Frontiers in Microbiology*, vol. 16, Art. no. 1504254, 2025.

[19] O. Lo-Thong-Viramoutou, P. Charton, X. F. Cadet *et al.*, "Non-linearity of metabolic pathways critically influences the choice of machine learning model," *Frontiers in Artificial Intelligence*, vol. 5, Art. no. 744755, 2022.

[20] M. Pichler and F. Hartig, "Machine learning and deep learning—A review for ecologists," *Methods in Ecology and Evolution*, vol. 14, no. 4, pp. 994–1016, 2023.

[21] D. Li, M. Tian, W. Ding *et al.*, "Dissecting possible correlations between leaf functional traits and heavy metal accumulation in two contrasting mangrove species across tidal gradients," *Environ. Exp. Bot.*, vol. 238, Art. no. 106234, 2025.

[22] A. H. Mohammed, A. M. Khalifa, H. M. Mohamed *et al.*, "Assessment of heavy metals at mangrove ecosystem, applying multiple approaches using in-situ and remote sensing techniques, Red Sea, Egypt," *Environmental Science and Pollution Research*, vol. 31, pp. 8118–8133, 2024.

[23] S. Moazzem, "Modelling the nature-based treatment systems to improve the water quality in port Phillip Bay catchment," Ph.D. dissertation, School of Engineering, RMIT University, Melbourne, Australia, 2025.

[24] J. B. M. d. Oliveira, D. C. Junior, C. E. T. Parente *et al.*, "Fungi in mangrove: Ecological importance, climate change impacts, and the role in environmental remediation," *Microorganisms*, vol. 13, no. 4, Art. no. 878, 2025.

[25] B. M. Alharbi, A. M. Abdulmajeed, A. A. Jabbour *et al.*, "Eco-physiological responses of Avicennia Marina (Forssk.) Vierh. to trace metals pollution via intensifying antioxidant and secondary metabolite contents," *Metabolites*, vol. 13, no. 7, Art. no. 808, 2023.

[26] A. B. Alhassan and M. O. Aljahdali, "Sediment metal contamination, bioavailability, and oxidative stress response in mangrove avicennia marina in central Red Sea," *Front. Environ. Sci.*, vol. 9, Art. no. 691257, 2021.

[27] O. Aydin, C. Osorio-Murillo, K. A. Butler *et al.*, "Conservation planning implications of modeling seagrass habitats with sparse absence data: A balanced random forest approach," *J. Coast Conserv.*, vol. 26, Art. no. 22, 2022.

[28] K. Maurya, S. Mahajan, and N. Chaube, "Remote sensing techniques: Mapping and monitoring of mangrove ecosystem—A review," *Complex & Intelligent Systems*, vol. 7, pp. 2797–2818, 2021.

[29] B. Reshma, B. Rahul, K. R. Sreenath *et al.*, "Taxonomic resolution of coral image classification with convolutional neural network," *Aquat. Ecol.*, vol. 57, pp. 845–861, 2022.

[30] Z. M. Yaseen and F. L. Alhalimi, "Heavy metal adsorption efficiency prediction using biochar properties: A comparative analysis for ensemble machine learning models," *Sci. Rep.*, vol. 15, Art. no. 13434, 2025.

[31] P. Mukube, M. Hitzman, L. Machogo-Phao *et al.*, "Geochemistry of terrestrial plants in the central African copperbelt: Implications for sediment hosted copper-cobalt exploration," *Minerals*, vol.14, no. 3, Art. no. 294, 2024.

[32] L. Guo, X. Xu, C. Niu *et al.*, "Machine learning-based prediction and experimental validation of heavy metal adsorption capacity of bentonite," *Science of the Total Environment*, vol. 926, Art. no. 171986, 2024.

[33] M. M. Matsa, T. Dube, and O. Mupepi, "Effectiveness of phytoremediation in waste-water treatment: A case of Karoi water supply station, Zimbabwe," *International Journal of Environmental Science and Technology*, vol. 22, pp. 7013–7024, 2024.

[34] A. Kumar, Tripti, D. Raj *et al.*, "Soil pollution and plant efficiency indices for phytoremediation of heavy metal(loid)s: Two-decade study (2002–2021)," *Metals*, vol. 12, no. 8, Art. no. 1330, 2022.

[35] K. M. Yang, "Recent trend in phytoremediation of petroleum hydrocarbon contaminated soil: A bibliometric review," *International Journal of Phytoremediation*, vol. 28, pp. 19–27, 2025.

[36] M. G. Hughes, T. M. Glasby, D. J. Hanslow *et al.*, "Random forest classification method for predicting intertidal wetland migration under sea level rise," *Front. Environ. Sci.*, vol. 10, Art. no. 749950, 2022.

[37] M. M. Erandi, T. A. Rubicel, T. V. Joséet *al.*, "An approach for accurate identification and monitoring of species in mangrove forests based on multi-source spectral data and deep learning," *Ecol. Inform.*, vol. 85, Art. no. 102961, 2025.

[38] Y. Yang, Z. Meng, J. Zu *et al.*, "Fine-scale mangrove species classification based on UAV multispectral and hyperspectral remote sensing using machine learning," *Remote Sens.*, vol. 16, no. 16, Art. no. 3093, 2024.

[39] Y. Zeng, T. Shi, Q. Liu *et al.*, "A geographically weighted neural network model for digital soil mapping of heavy metal copper in coastal cities," *J. Hazard. Mater.*, vol. 480, Art. no. 136285, 2024.

[40] X. Sun, J. Li, M. Saberian *et al.*, "Mechanistic insights into water use strategies and heat tolerance of Australian native trees in saline soils: Implications for phytoremediation under climate change," *Plant Soil*, vol. 516, pp. 173–194, 2025.

[41] M. Aasim, S. A. Ali, S. Aydin *et al.*, "Artificial intelligence–based approaches to evaluate and optimize phytoremediation potential of in vitro regenerated aquatic macrophyte *Ceratophyllum demersum L*.," *Environmental Science and Pollution Research*, vol. 30, pp. 40206–40217, 2023.
[42] Y. Zhai, L. Zhou, H. Qi *et al.*, "Application of visible/near-infrared spectroscopy and hyperspectral imaging with machine learning for high-throughput plant heavy metal stress phenotyping: A review," *Plant Phenomics*, vol. 5, Art. no. 0124, 2023.
[43] T. Miao, L. Shen, H. Zhao *et al.*, "Multi-level driving mechanisms: Cascading relationships among physical factors, nutrient cycling, and biological responses in the Yangtze River–lake ecosystems," *Sustainability*, vol. 17, no. 22, Art. no. 9928, 2025.
[44] B. E. Mahrad, A. Newton, J. D. Icely *et al.*, "Contribution of remote sensing technologies to a holistic coastal and marine environmental management framework: A review," *Remote Sens.*, vol. 12, no. 14, Art. no. 2313, 2020.
[45] D. Muenzel, A. W. Anggoro, D. E. Bulan *et al.*, "Optimising site selection for ecosystem approaches to shrimp aquaculture in mangrove systems," *Aquaculture International*, vol. 33, Art. no. 632, 2025.
[46] S. Nuyts, M. D. de P. Costa, P. I. Macreadie *et al.*, "A decision support tool to help identify blue carbon sites for restoration," *J. Environ. Manage.*, vol. 367, Art. no. 122006, 2024.
[47] S. F. Ahmed, M. S. B. Alam, M. Hassan *et al.*, "Deep learning modelling techniques: Current progress, applications, advantages, and challenges," *Artif. Intell. Rev.*, vol. 56, pp. 13521–13617, 2023.
[48] H. R. Maier, F. Zheng, H. Gupta *et al.*, "On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization," *Environmental Modelling and Software*, vol. 167, Art. no. 105779, 2023.
[49] S. Sathyanarayanan and B. R. Tantri, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, vol. 27, pp. 4023–4031, 2024.
[50] M. B. Hossain, Z. Masum, M. S. Rahman *et al.*, "Heavy metal accumulation and phytoremediation potentiality of some selected mangrove species from the world's largest mangrove forest," *Biology*, vol. 11, no. 8, Art. no. 1144, 2022.
[51] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models," *Data Science in Finance and Economics*, vol. 3, no. 4, pp. 354–379, 2023.
[52] S. B. S. Lai, N. H. N. B. M. Shahri, M. B. Mohamad *et al.*, "Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data," *Mathematics and Statistics*, vol. 9, no. 3, pp. 379–385, 2021.
[53] Y. Rimal, N. Sharma, S. Paudel *et al.*, "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy," *Sci. Rep.*, vol. 15, Art. no. 13444, 2025.
[54] S. Lin, Y. Wang, H. Wei *et al.*, "Hybrid method for oil price prediction based on feature selection and XGBOOST-LSTM," *Energies*, vol. 18, no. 9, Art. no. 2246, 2025.
[55] M. Grebovic, L. Filipovic, I. Katnic *et al.*, "Machine learning models for statistical analysis," *International Arab Journal of Information Technology*, vol. 20, no. 3A, pp. 505–514, 2023.
[56] T. Kavzoglu and A. Teke, "Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost)," *Bulletin of Engineering Geology and the Environment*, vol. 81, Art. no. 201, 2022.
[57] E. K. Sahin, "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest," *SN Appl. Sci.*, vol. 2, Art. no. 1308, 2020.
[58] A. Pourzangbar, M. Jalali, and M. Brocchini, "Machine learning application in modelling marine and coastal phenomena: A critical review," *Frontiers in Environmental Engineering*, vol. 2, Art. no. 1235557, 2023.
[59] Y. Lin, J. Gao, Y. Tu *et al.*, "Estimating low concentration heavy metals in water through hyperspectral analysis and genetic algorithm-partial least squares regression," *Science of the Total Environment*, vol. 916, Art. no. 170225, 2024.
[60] T. Ma, L. Wu, S. Zhu *et al.*, "Multiclassification prediction of clay sensitivity using extreme gradient boosting based on imbalanced dataset," *Applied Sciences*, vol. 12, no. 3, Art. no. 1143, 2022.
[61] M. H. L. Louk and B. A. Tama, "Revisiting gradient boosting-based approaches for learning imbalanced data: A case of anomaly detection on power grids," *Big Data and Cognitive Computing*, vol. 6, no. 2, Art. no. 41, 2022.
[62] S. Zhang, "Challenges in KNN classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, pp. 4663–4675, 2022.
[63] A. M. Jibrin, S. I. Abba, J. Usman *et al.*, "Tracking the impact of heavy metals on human health and ecological environments in complex coastal aquifers using improved machine learning optimization," *Environmental Science and Pollution Research*, vol. 31, pp. 53219–53236, 2024.
[64] J. Guo, X. Lin, and Y. Xiao, "Integration of smart sensors and phytoremediation for real-time pollution monitoring and ecological restoration in agricultural waste management," *Front. Plant Sci.*, vol. 16, Art. no. 1550302, 2025.
[65] L. D. de Lacerda, R. D. Ward, R. Borges *et al.*, "Mangrove trace metal biogeochemistry response to global climate change," *Front. For. Glob. Chang*, vol. 5, Art. no. 817992, 2022.
[66] M. C. Ogwu, S. C. Izah, W. E. Sawyer *et al.*, "Environmental risk assessment of trace metal pollution: A statistical perspective," *Environmental Geochemistry and Health*, vol. 47, Art. no. 94, 2025.