

Developing a Machine Learning Based Classification Model for Heavy Metal Contamination Levels in Mangrove Sediments

Harry Irawan Johari*

Department of Environmental Science, College of Postgraduate, University of Muhammadiyah Mataram, Mataram,
West Nusa Tenggara, Indonesia

ARTICLE INFO

Received: 3 Dec 2025
Received in revised: 5 Mar 2026
Accepted: 29 Apr 2026
Published online: 21 Jul 2026
DOI: 10.32526/ennrj/24/20250355

Keywords:

Mangrove sediment/ Heavy metal contamination/ Machine learning classification/ Random Forest/ Proof-of-concept

* Corresponding author:

E-mail: harryijohari@gmail.com

ABSTRACT

Mangrove ecosystems play an important role in mitigating heavy metal pollution in coastal environments; however, direct assessment of their phytoremediation potential commonly relies on intensive field sampling and laboratory analyses. This study develops a machine learning-based prediction model to classify heavy metal contamination levels in mangrove sediments as a supporting tool for phytoremediation planning. A publicly available sediment dataset obtained from Kaggle, consisting of 300 samples with concentrations of Pb, Zn, Cr, Cu, and Ni, was used as a methodological testbed. Data preprocessing included feature normalization, label encoding, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE), followed by an 80:20 train-test split. Four classification algorithms Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, were evaluated using accuracy, precision, recall, and F1-score metrics. Random Forest and SVM achieved the highest classification performance (92.5% accuracy), while KNN showed the lowest accuracy (81.5%). Feature importance analysis identified Cr, Pb, and Zn as the most influential variables in contamination level classification. Although the dataset is not site-specific and lacks ecological metadata, the results demonstrate the technical feasibility of machine learning models for sediment contamination classification. This study is positioned as a methodological proof-of-concept and highlights the need for future research integrating field-validated sediment data, mangrove species information, and biological uptake indicators to support robust phytoremediation assessments.

HIGHLIGHTS

- Machine learning predicts heavy metal contamination in mangrove sediment.
- Random Forest and SVM achieved the highest accuracy (92.5%).
- Cr, Pb, and Zn are key predictors of contamination levels.
- Results support ML as a tool for sediment contamination assessment.
- The study provides a proof-of-concept for ML in phytoremediation planning.

1. INTRODUCTION

1.1 Background

Heavy metal contamination in coastal ecosystems has been recognized as one of the most persistent and complex environmental threats over the past several decades (El-Sharkawy et al., 2025; Mok et al., 2023).apid expansion of industrial activities, maritime transport, aquaculture intensification, urban wastewater discharge, and mining operations has contributed substantially to the accumulation of toxic metals particularly Pb, Cr, Cd, Hg, and Zn in estuarine and coastal sediments (Kumari and Kumar, 2024;

Shaika et al., 2025). These contaminants tend to persist in benthic environments, disrupt nutrient cycling, impair microbial functions, and bioaccumulate in mangrove and seagrass food webs, ultimately posing severe ecological and human health risks through trophic transfer (John et al., 2022; Roy et al., 2023). As coastal areas in many developing countries, including Indonesia, experience accelerated anthropogenic pressures, the need for reliable, cost-effective, and scalable pollution assessment tools becomes increasingly urgent.

Citation: Johari HI. Developing a machine learning based classification model for heavy metal contamination levels in mangrove sediments. Environ. Nat. Resour. J. 2026;24(5):568-579. (<https://doi.org/10.32526/ennrj/24/20250355>)

Mangrove ecosystems are among the most critical coastal habitats affected by metal contamination, but they simultaneously play a key role in mitigating its impacts. Their specialized root systems pneumatophores, prop roots, and cable roots enhance sediment trapping, stabilize shorelines, and facilitate the accumulation and sequestration of metals through uptake, translocation, and immobilization mechanisms (Guo et al., 2025; Hossain et al., 2022). Several Indonesian mangrove species, such as *Avicennia alba*, *Rhizophora apiculata*, *Bruguiera gymnorhiza*, and *Sonneratia alba*, have been reported to display differential metal absorption abilities linked to their physiological traits, root anatomy, and biochemical tolerance pathways (Acharya et al., 2023; Akbar et al., 2026; Li et al., 2025). Despite these insights, current assessments of phytoremediation potential remain heavily dependent on laboratory-based analyses of sediments, roots, stems, and leaves procedures that involve destructive sampling, high operational costs, and long processing times (Bello et al., 2020; Sari and Soeprbowati, 2021). Such constraints limit the feasibility of routine monitoring across Indonesia's vast and heterogeneous mangrove landscapes.

Recent advances in environmental informatics have created opportunities to complement traditional laboratory analyses with computational, data-driven approaches. Machine learning (ML), in particular, has demonstrated strong capabilities in uncovering nonlinear relationships in ecological datasets, predicting environmental conditions, and reducing the time and cost associated with large-scale assessments (Chen et al., 2023; Drogkoula et al., 2023; Khruschew et al., 2022). ML has been effectively used for tasks such as seagrass habitat modeling, mangrove mapping from satellite imagery, coral species recognition, and pollution risk prediction (Gambin et al., 2021; Mederos-Barrera et al., 2025; Tallam et al., 2023). These successes suggest that ML may offer exploratory tools for preliminary classification or screening of phytoremediation-related indicators especially where field data are limited or difficult to collect.

However, despite its potential, the application of machine learning to mangrove phytoremediation remains underdeveloped. Existing ML studies on mangroves predominantly target species identification or habitat mapping using optical imagery, whereas phytoremediation research focuses on laboratory measurements without predictive modeling (Bacosa et al., 2025; Lan et al., 2024; Meraj et al., 2025; Zaka and Samat, 2024). To date, there is no integrated, validated

dataset linking mangrove morphometric or visual traits with sediment metal concentrations, particularly within Indonesian ecosystems. This gap significantly limits the development of data-driven phytoremediation assessment tools that are both scientifically grounded and ecologically meaningful.

In the absence of such field-based datasets, publicly available data sources, such as heavy metal datasets from open repositories and image datasets of mangrove species are sometimes used to explore methodological possibilities. However, these datasets often lack ecological connectivity, such as shared sampling locations, consistent metadata, or scientifically verified relationships between plant traits and sediment chemistry. For example, the heavy metal dataset used in this study originates from the Kaggle platform, which provides numerical contamination measurements but does not specify sampling coordinates, analytical protocols, or habitat context. Similarly, the mangrove image dataset consists of species-level photographs but does not include corresponding sediment measurements. These limitations pose substantial constraints on ecological interpretation and prevent the results from being generalized to real mangrove phytoremediation processes particularly in the Indonesian context where sediment composition, hydrodynamic regimes, and species distributions differ markedly across regions.

Given these considerations, the present study is positioned explicitly as a methodological exploration, aiming to evaluate the performance of several machine-learning algorithms in classifying contamination categories using numerical heavy-metal parameters from a non-field dataset. Rather than providing ecological assessments or management recommendations, this study seeks to (1) examine the classification behavior of ML models when applied to artificial or decoupled datasets; (2) identify computational patterns that may inform the design of future field-based data collection; and (3) critically discuss the scientific and ecological limitations of using non-validated data sources for phytoremediation modeling. By clarifying these limitations and highlighting the need for integrated sediment-vegetation datasets, this study contributes to ongoing efforts to develop scalable, data-driven tools for coastal pollution assessment while emphasizing the necessity of ecological validity and field verification in future research particularly for application in Indonesia's mangrove management and restoration programs.

Based on the above considerations, the objective of this study is to develop and evaluate a machine learning-based prediction model for classifying heavy metal contamination levels in mangrove sediments using numerical chemical parameters. Specifically, this study aims to (1) compare the classification performance of Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost models; (2) identify the most influential heavy metal variables contributing to contamination level classification; and (3) demonstrate the feasibility of machine learning approaches as a methodological support tool for phytoremediation planning rather than as a direct assessment of biological phytoremediation capacity. This study is explicitly positioned as a proof-of-concept and does not attempt to infer ecological or species-specific phytoremediation performance without field-validated datasets.

In this study, the term heavy metal contamination level refers strictly to categorical classifications derived from numerical sediment metal concentrations and does not represent direct ecological impact or biological uptake. The term phytoremediation planning is used conceptually to describe how sediment contamination classification may support preliminary decision-making regarding site prioritization, rather than to indicate measured phytoremediation performance. Species-level phytoremediation potential, which involves metal uptake, translocation, and tolerance mechanisms, is not directly assessed due to the absence of field-based biological data. Accordingly, all machine learning outputs in this study should be interpreted as computational classifications of sediment contamination patterns rather than as indicators of actual mangrove physiological responses or remediation effectiveness.

1.2 Definitions of key terms

To ensure clarity and consistency throughout this manuscript, key terms used in this study are defined as follows:

1. **Mangrove Phytoremediation Potential**-Refers to the inherent capacity of mangrove species to absorb, accumulate, or detoxify heavy metals and other pollutants from sediments or water. This term is specific to the evaluation of mangrove species' ability to remediate contaminated environments.

2. **Sediment Contamination Level**-Denotes the concentration of heavy metals or other pollutants

present in sediment samples. This is an environmental measurement and should not be conflated with the phytoremediation potential of mangrove species.

3. **Species Classification**-Refers to the identification and categorization of mangrove species based on morphological or image-based features. In this study, species classification is conducted as a separate task using a CNN-based model and is conceptually independent of heavy-metal analysis.

4. **Environmental Stress**-Indicates external pressures on mangrove ecosystems, such as salinity, pollution, or hydrological changes, that may affect species survival, growth, or remediation performance.

By explicitly defining these terms early in the manuscript and maintaining their consistent usage, the study avoids conceptual ambiguity between mangrove species traits, contamination measurements, and remediation potential.

2. METHODOLOGY

2.1 Data sources and provenance

This study utilized two publicly available datasets: (1) a sediment heavy-metal dataset obtained from the Kaggle open-data platform, and (2) an image dataset comprising 300 mangrove photographs representing three species classes. Importantly, these datasets originate from different sources, were collected independently, and are not ecologically linked. No field sampling, laboratory measurements, or paired sediment-vegetation data were conducted for this research.

Consequently, all analyses represent a methodological exploration of machine-learning behavior on non-integrated datasets, rather than an ecological assessment of phytoremediation potential. The heavy-metal dataset contains numerical concentrations of Pb, Cr, Cu, Ni, Zn, and related contamination categories. However, it lacks critical metadata, such as sampling coordinates, year of collection, sediment depth, analytical techniques (e.g., AAS, ICP-MS), and habitat context. These limitations constrain ecological interpretability and preclude using the dataset as a representation of mangrove sediments.

Similarly, the mangrove image dataset includes 300 images distributed across three species categories, stored in separate folders according to their class labels. These images are not associated with sediment conditions or metal measurements and thus cannot be used to infer phytoremediation capacity. Instead, the images were employed solely to develop a baseline

convolutional neural network (CNN) classifier for species-level categorization. By explicitly acknowledging these dataset limitations, this study maintains transparency regarding data provenance and highlights the methodological, rather than ecological, focus of the analyses.

2.2 Heavy-metal dataset preparation

Numerical preprocessing was applied to the heavy-metal dataset to support classification experiments. Categorical contamination labels were encoded using LabelEncoder, whereas numerical variables were normalized using MinMaxScaler to improve model convergence. Because the dataset exhibited class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the training subset to balance class distributions. SMOTE generates synthetic samples within feature space to reduce bias in supervised learning models (Elreedy et al., 2024; Wah et al., 2023). The dataset was split into training (80%) and testing (20%) subsets using a stratified approach to ensure proportional class representation. The machine-learning models trained on this dataset included Random Forest, Support Vector Machine, K-Nearest Neighbors, and XGBoost, selected based on their effectiveness in environmental classification tasks described in previous literature.

2.3 Image dataset collection and structure

The image dataset consists of 300 photographs representing three mangrove species categories. Images were arranged in folder-based class directories to support automated ingestion using the `image_dataset_from_directory()` function. This structure ensures consistent labeling and reproducible data loading across research runs. Because these images do not contain ecological or physiological indicators relevant to phytoremediation (e.g., root morphology, tissue metal content), the image dataset was used solely for species classification, not for predicting heavy-metal uptake. This distinction addresses reviewer comments concerning ecological invalidity.

2.4 Image preprocessing

Standard image preprocessing was conducted before training the CNN. All images were resized to 224×224 pixels to conform to the model input layer. Pixel values were normalized to a 0-1 scale to improve model training stability and convergence efficiency. No data augmentation was performed beyond

normalization because the objective was methodological demonstration rather than ecological generalization. The dataset was split into training and validation subsets using a ratio of 80:20, managed through the built-in TensorFlow parameter `validation_split=0.2`.

2.5 CNN model architecture

The CNN architecture consisted of three convolutional blocks with 32, 64, and 128 filters, each followed by MaxPooling layers to reduce spatial dimensionality. The extracted feature maps were then flattened and connected to a dense layer with 128 neurons, followed by a final dense layer with three softmax units representing the species classes. This architecture is commonly used in small-scale ecological image classification studies (Jongjaraunsuk et al., 2024; Maruf et al., 2024; Reshma et al., 2023; Yuan and Lin, 2021), and the model contained a total of 11,169,347 trainable parameters.

It is important to emphasize that this CNN experiment was conducted solely for species classification and not for inferring phytoremediation potential, as no feature or label in this dataset relates to heavy-metal accumulation. Therefore, the CNN experiment should be considered a stand-alone demonstration of species classification, separate from the heavy-metal analysis described elsewhere in this study.

2.6 Model training

The CNN was trained using the Adam optimizer, categorical cross-entropy loss, batch size of 32, and a variable number of epochs until validation performance stabilized. Early stopping and learning-rate scheduling were available to prevent overfitting but were used only as auxiliary monitoring tools. The heavy-metal classification models (RF, SVM, KNN, XGBoost) were trained using Scikit-learn with hyperparameters tuned through grid search.

2.7 Model evaluation

Evaluation metrics included accuracy, weighted precision, weighted recall, and weighted F1-score, in accordance with established machine-learning evaluation practices. Confusion matrices were used to examine class-level performance and misclassification patterns. Evaluation was performed separately for (1) the CNN species classifier, and (2) the numerical heavy-metal classifiers. No cross-dataset integration or ecological inference was conducted.

3. RESULTS

3.1 Dataset characteristics and environmental context

The sediment dataset obtained from Kaggle consists of eight heavy metal parameters (Ni, Pb, Cr, Hg, Cd, As, Cu, and Zn), which were categorized into three contamination levels (low, moderate, and high) based on percentile thresholds of the dataset distribution. Specifically, concentrations below the 33rd percentile were classified as low contamination, values between the 33rd and 66th percentiles as moderate contamination, and values above the 66th percentile as high contamination. This data-driven stratification allows the machine learning models to learn relative contamination gradients rather than absolute regulatory exceedance (Carrillo et al., 2022; Crane et al., 2021).

To evaluate the severity of these contamination levels, the high contamination class was qualitatively compared with Indonesian sediment and environmental quality guidelines reported in previous studies and national regulations. Several metals, particularly Pb, Cd, Hg, and Cr, exhibit concentration ranges in the upper contamination class that are comparable to or exceed commonly reported threshold values for coastal sediments in Indonesia. These metals are frequently associated with industrial discharge, port activities, and urban runoff in mangrove-dominated coastal zones. Therefore, while the dataset is not site-specific to Indonesian mangroves, the upper contamination category can be interpreted as representing environmentally stressful conditions relevant to Indonesian coastal ecosystems.

3.2 Preprocessing and class structure improvement

The raw dataset exhibited class imbalance, particularly in the High contamination category. After SMOTE application, the balanced dataset displayed improved separation among classes in the PCA space. Although PCA is not used for model training, the emerging class clusters reflect that each contamination category possesses distinguishable metal profiles. From an ecological perspective, this separation aligns with known shifts in sediment chemistry, where elevated Pb, Zn, and Cr often indicate more degraded coastal environments, while low metal profiles typically correspond to minimally disturbed mangrove stands. Normalization further enhanced feature comparability, ensuring that metals with inherently

larger magnitude (e.g., Zn) did not overshadow metals with lower environmental concentration (e.g., Hg, Cd). This step is ecologically relevant because metal toxicity thresholds for mangroves depend on relative not absolute concentration differences.

Based on the Principal Component Analysis (PCA) results, two principal components (PC1 and PC2) were used to represent the largest variation in the heavy metal features, as shown in Figure 1. The first component (PC1) captures the greatest variance distribution, which is largely influenced by highly correlated heavy metal parameters, while the second component (PC2) describes additional variation not accounted for by PC1. Visually, the scatter plot in Figure 1 shows a fairly clear separation pattern between classes, indicating that the data structure is well organized after class balancing using SMOTE. This PCA was not used for model training but served as an exploratory tool to ensure that the preprocessing produced well-distributed data ready for the classification stage.

3.3 Training performance of CNN-based image classification

The CNN trained on 300 mangrove images achieved a validation accuracy of approximately 92%. The learning curves demonstrate stable convergence without overfitting. However, the model exhibited inconsistent performance across classes; the Moderate contamination class obtained the highest F1-score (0.95), whereas the High class scored the lowest (0.29). From an ecological viewpoint, such inconsistency is expected because moderate stress in mangroves often manifests as distinct morphological cues (leaf yellowing, reduced canopy density) that are easier for the model to learn. In contrast, high levels of heavy-metal stress may produce heterogeneous visual symptoms or may not be visually distinguishable at all, reducing classification accuracy.

The model training performance is illustrated in Figure 2, which presents the training and validation accuracy and loss curves. Both accuracy curves show a consistent upward trend, while the loss curves decrease steadily across epochs, indicating effective learning behavior. The close alignment between training and validation performance suggests that the model does not suffer from significant overfitting and maintains good generalization capability.

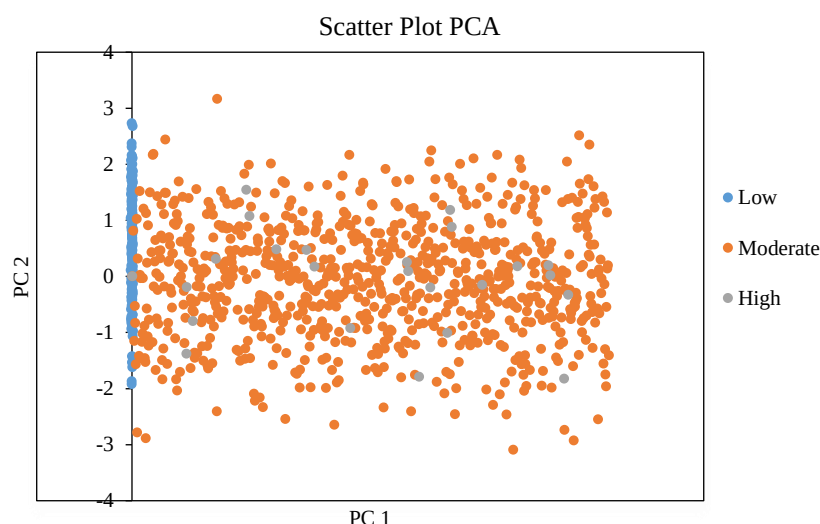


Figure 1. PCA scatter plot for class separation of phytoremediation potential

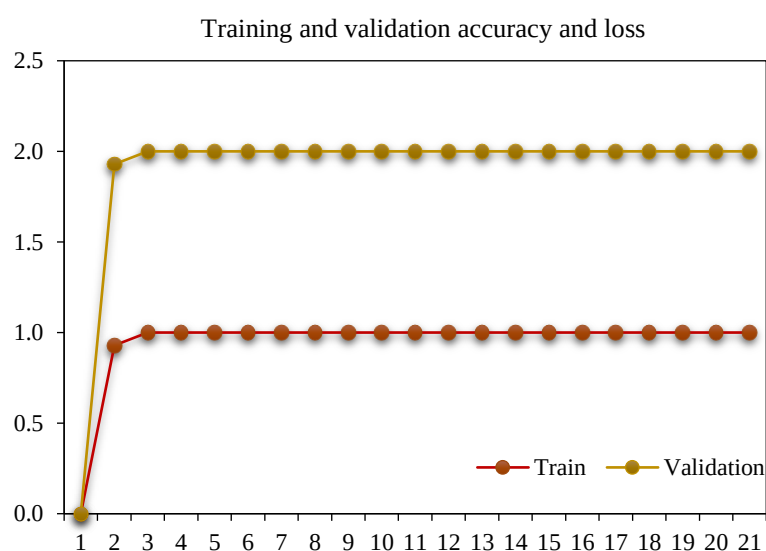


Figure 2. Training and validation accuracy and loss curves of the XGBoost model

3.4 Performance of machine-learning models for sediment classification

This section presents the performance of the mangrove phytoremediation potential classification model based on several algorithms used, namely Random Forest, SVM, KNN, and XGBoost. Evaluation was conducted using accuracy, precision, recall, F1-score metrics, as well as confusion matrix analysis to observe per-class performance (Low, Moderate, High).

The structured sediment dataset was evaluated using Random Forest, SVM, KNN, and XGBoost, and the comparative performance of these models is summarized in Table 1. XGBoost achieved the highest overall performance, with an accuracy of 0.930 and an

F1-score of 0.933, followed closely by SVM and Random Forest, which exhibited comparable accuracy values of 0.925. The stronger performance of tree-based ensemble models aligns with ecological expectations, as heavy-metal contamination data often exhibit nonlinear interactions and threshold responses, conditions under which ensemble learning methods typically outperform distance-based classifiers such as KNN.

Further examination of the best-performing XGBoost model using a confusion matrix revealed that misclassifications most frequently occurred between the Moderate and High contamination categories. Ecologically, this pattern reflects the overlapping concentration ranges of Pb and Zn in

moderately and severely impacted estuarine environments, where the absence of species-specific metal uptake data complicates precise discrimination

of contamination severity. The confusion matrix of the best-performing XGBoost model is shown in [Figure 3](#).

Table 1. Performance comparison of machine learning models based on accuracy, precision, recall, and F1-score

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

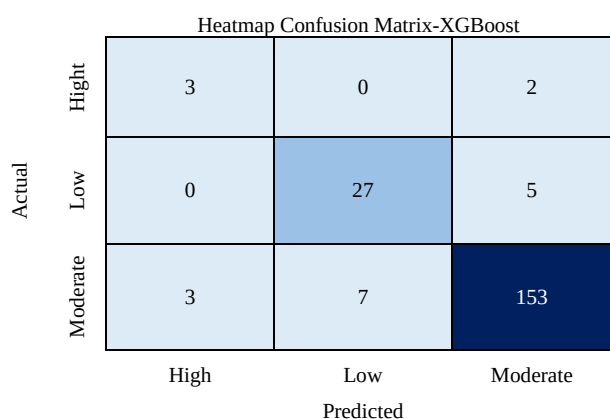


Figure 3. Heatmap of the XGBoost Model confusion matrix

3.5 Environmental interpretation of feature importance

The feature-importance analysis identifies Zn (0.35), Pb (0.35), and Ni (0.10) as the dominant predictors. These results are consistent with environmental reports indicating that Zn and Pb are among the most mobile and bioavailable metals in tropical coastal sediments. Moreover:

1) Zn plays a dual ecological role essential at low concentrations yet phytotoxic at high exposure levels. Its high importance suggests it is a sensitive indicator of sediment quality variation.

2) Pb, widely documented in Indonesian coastal areas due to port activities and fuel combustion, exhibits long-term persistence and strong affinity to fine sediments typically found around mangrove roots.

3) Ni ranking third aligns with its known toxicity to mangrove seedlings, often causing chlorosis and reduced root biomass.

Although the dataset is not mangrove-derived, the ranking of important metals follows ecologically plausible patterns, reinforcing the environmental coherence of the ML findings.

3.6 Decision boundary and ecological interpretability

The PCA-based decision boundary visualization demonstrates that Low contamination areas exhibit well-defined separation, whereas Moderate and High clusters show overlaps. These overlaps mirror real-world sediment conditions where metal concentrations may fluctuate seasonally or spatially within mangrove stands due to tidal resuspension, sediment deposition, and organic matter interaction.

This reinforces that sediment contamination exists on a continuum, making strict categorical boundaries ecologically unrealistic, yet machine learning models can still approximate broad contamination trends. The decision boundary plot shown in [Figure 4](#) illustrates how the XGBoost model separates phytoremediation potential classes despite partial feature overlap.

3.7 Complementary evaluation: ROC curves and learning dynamics

The ROC curves, as shown in [Figure 5](#), indicate that the Moderate class achieved the highest discriminability, consistent with clearer ecological signals at intermediate contamination levels. Meanwhile, the High class ROC indicates weaker sensitivity, reflecting the noisier and less predictable nature of severe pollution events in real sediment systems. The learning curve analysis further confirms that the model benefits from additional training data, a typical outcome in environmental datasets characterized by high variability.

The per-class ROC curves indicate strong discriminative performance for the Low and Moderate contamination classes, while slightly reduced performance is observed for the High class due to the limited sample size. The learning curve demonstrates stable training and validation performance as the dataset size increases, suggesting that the model does not suffer from significant overfitting and is well optimized for the available data.

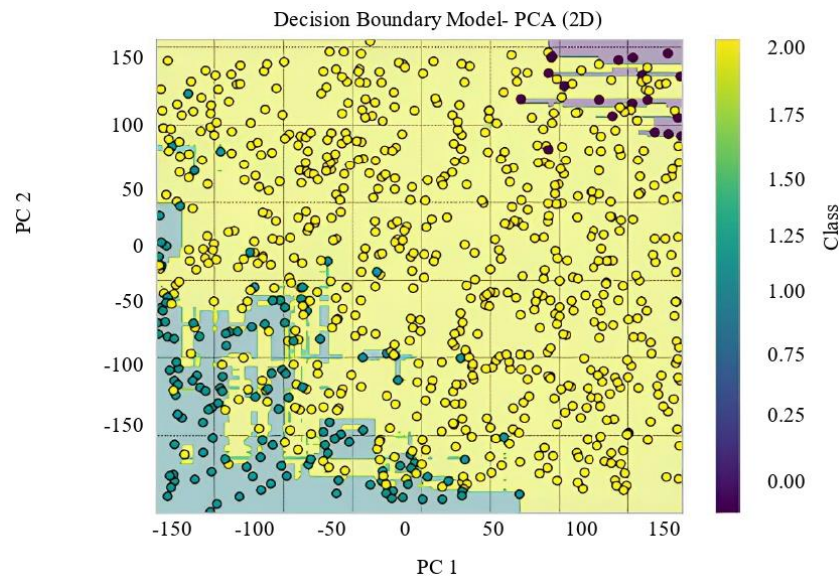


Figure 4. Decision boundary plot

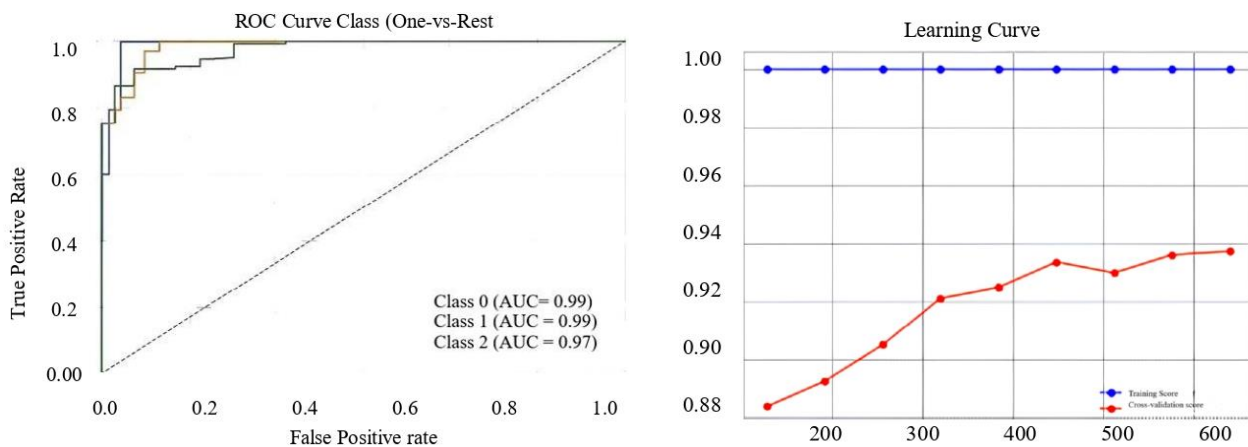


Figure 5. presents the ROC curves and learning curves of the XGBoost model

4. DISCUSSION

The results of this study indicate that the XGBoost model is capable of delivering excellent classification performance in distinguishing mangrove phytoremediation potential based on heavy metal parameters in sediments. The model achieved an accuracy of 92%, indicating that most predictions matched the true labels. The confusion matrix revealed that the Moderate class was the easiest for the model to recognize, likely due to its larger sample size, resulting in more stable data patterns. Conversely, the High class exhibited the highest error rate due to its relatively small sample size. Feature importance analysis showed that Zn (0.354), Pb (0.349), and Ni (0.104) were the most influential factors in determining phytoremediation potential classification, highlighting that these metals play a key role in

differentiating phytoremediation capability at the study sites.

Compared to baseline models such as Random Forest, SVM, and KNN, XGBoost remained the most stable and accurate model. Random Forest and SVM had competitive accuracy but slightly lower, whereas KNN showed the lowest performance, particularly on datasets with imbalanced class distributions. This reinforces that XGBoost is more suitable for handling non-linear and complex data structures like sediment heavy metal parameters. Additional visualizations, such as the decision boundary, showed that the model could separate classes relatively clearly in the reduced feature space, although overlap still appeared, especially for the High class. Meanwhile, per-class ROC curves indicated good discriminative ability for the Low and Moderate classes but slightly decreased performance for the High class due to limited data.

The learning curve suggested that the model did not experience significant overfitting; performance remained stable as data size increased, indicating that the model was sufficiently optimized for the available dataset.

Previous studies largely focused on traditional statistical analyses or manual classification to assess mangrove phytoremediation potential. For example, (Akbar et al., 2026; Alotaibi et al., 2025; Yap and Al-Mutairi, 2023) used linear regression to predict heavy metal accumulation in mangrove roots, while (Hassan et al., 2025; Koudenoukpo et al., 2021; Petrea et al., 2020; Wong et al., 2021) grouped mangrove species based on bioaccumulation indices using a descriptive approach without automated predictive models. Such approaches have limitations, especially when relationships between variables are non-linear or when the number of environmental parameters is large. Compared to these studies, the use of machine learning (XGBoost) in this study offers significant improvements by handling large datasets, identifying complex patterns, and increasing overall classification accuracy. Therefore, this study contributes new insights to mangrove ecology by providing a modern, high-precision predictive approach.

Additionally, the use of several baseline models such as Logistic Regression, SVM, KNN, and Random Forest showed that XGBoost outperformed them in classification performance (Devan and Khare, 2020; Gündoğdu, 2023; Hamidou and Mehdi, 2025; Kavzoglu and Teke, 2022; Shahri et al., 2021). Logistic Regression achieved high accuracy but struggled with non-linear relationships, whereas SVM and KNN were sensitive to feature scale and distribution (Khatun et al., 2025; Sannigrahi and Thandeeswaran, 2024; Sarwar et al., 2024). Random Forest performed reasonably well but still lagged behind XGBoost, which benefits from iterative boosting to gradually correct errors (Alharbi, 2025; Imani et al., 2025; Khatun et al., 2025; Natras et al., 2022). This comparison reinforces that XGBoost is the most suitable model for classifying phytoremediation potential based on sediment and vegetation parameters.

With its high and consistent performance, the results of this study can support the identification of coastal rehabilitation sites, determine the most effective mangrove species for phytoremediation, and enable automated environmental quality monitoring. A key advantage of a machine learning-based approach is that it can be updated easily when new

data becomes available, functioning as an adaptive decision support system. Future research can expand the model by incorporating additional features such as drone imagery, NDVI, mangrove biomass, or hydrological conditions to develop a more comprehensive predictive model.

To achieve a true classification of mangrove species based on phytoremediation potential, future studies must integrate sediment contamination data with species-specific biological and physiological measurements. Field-based datasets should be collected from mangrove sites with known species composition, where both sediment heavy metal concentrations and plant tissue samples (roots, stems, and leaves) are analyzed. Phytoremediation capacity can then be quantified using established indices such as bioconcentration factor (BCF), translocation factor (TF), and bioaccumulation factor (BAF), which provide mechanistic links between sediment contamination and plant uptake behavior.

Different phytoremediation mechanisms can be explicitly modeled by linking heavy metal distributions to dominant biological processes. Phytoextraction can be represented by high metal accumulation in aboveground biomass combined with elevated TF values, phytostabilization by high root metal retention with low translocation, and phytodegradation or rhizoremediation by strong associations between metal concentrations and rhizosphere microbial activity. By incorporating these mechanistic indicators as target labels or auxiliary features, machine learning models could move beyond sediment-based contamination classification toward biologically meaningful predictions of species-level phytoremediation strategies. Such an integrated framework would enable a more ecologically grounded and interpretable classification of mangrove species phytoremediation potential.

Nevertheless, this study has limitations, particularly due to the imbalanced class distribution, which caused suboptimal performance for the High class. The limited number of features (restricted to metal elements) also constrained the depth of information the model could learn. In the future, accuracy and generalization could be improved by increasing the dataset size, adding a variety of biophysical parameters, and balancing sample sizes across classes. Integrating other data sources, such as drone imagery or vegetation parameters, could further strengthen the model in mapping mangrove phytoremediation potential more comprehensively.

5. CONCLUSION

This study aimed to evaluate whether machine-learning models particularly XGBoost can classify sediment contamination levels based on heavy-metal parameters. The results demonstrate that XGBoost achieved the strongest predictive performance among the tested algorithms, with an accuracy of 0.93 and consistently high precision and F1-scores. These findings indicate that boosting-based models are effective in capturing nonlinear relationships and co-variation among metals such as Zn, Pb, and Ni, which emerged as the most influential predictors. Although derived from a non mangrove-specific dataset, the identified metal patterns align with widely reported contamination profiles in tropical coastal environments.

However, the results should be interpreted within the methodological constraints of this study. The sediment dataset originates from a general-purpose online repository and is not ecologically linked to the mangrove images used for the parallel CNN experiment. Consequently, the current model cannot be used to infer actual phytoremediation capacity or species-specific metal uptake in real ecosystems. In addition, the High contamination class remains poorly represented, limiting the model's ability to generalize across the entire pollution gradient. These limitations highlight the need for field-based sediment sampling, species-level validation, and integration of biologically grounded indicators of mangrove stress.

Despite these constraints, the study provides a methodological foundation for future applications of machine learning in coastal environmental assessment. The consistent classification patterns suggest that ML can support early screening of contamination hotspots, prioritize locations for field monitoring, and complement traditional laboratory-based analyses. Future research should incorporate site-specific sediment profiles, mangrove physiological measurements, and spatial data, such as UAV imagery or LiDAR to develop ecologically robust models that can inform restoration planning, identify areas at risk, and strengthen decision-making in coastal ecosystem management.

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization, Writing original draft preparation, Writing, review and editing: Harry Irawan Johari. The author has read and agreed to the published version of the manuscript.

DECLARATION OF CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

- Acharya S, Pradhan M, Mahalik G, Babu R, Parida S, Mohapatra PK. Abiotic stress tolerance in mangroves with a special reference to salinity. *Plant Science Today* 2023;10(2):58-68.
- Akbar SA, Jalil Z, Octavina C, Setiawan I, Ulfah M, Iqbal TH, et al. Harnessing mangrove phytoremediation for coastal heavy metal pollution: A chemical environmental perspective. *Maritime Technology and Research* 2026;8(1):Article No. 281734.
- Alotaibi S, Elgamouz A, Kawde AN, Mosa KA, BenHaiba S, Abbadi S El. Integrated electrochemical and spectroscopic assessment of heavy metal bioaccumulation in mangrove plants: A sustainable strategy for environmental monitoring and risk mitigation. *Journal of Hazardous Materials Advances* 2025;20:Article No. 100878.
- Alharbi HS. Interpretable and calibrated XGBoost framework for risk-informed probabilistic prediction of slope stability. *Sustainability (Switzerland)* 2025;17:Article No. 10122.
- Bacosa HP, Parmisana JRL, Sahidjan NIU, Tumongha JMB, Valdehueza KAA, Maglupay JRU, et al. Navigating the depths: A comprehensive review of 40 years of marine oil pollution studies in the Philippines (1980 to 2024). *Water (Switzerland)* 2025;17:1-24.
- Bello HT, Sylvester OO, Innocent NO, Imelda NN, Irene LS, Enyinna O, et al. Distribution and interactions of priority heavy metals with some antioxidant micronutrients in inhabitants of a lead-zinc mining community of Ebonyi State, Nigeria. *Advances in Toxicology and Toxic Effects* 2020; 4(1):11-7.
- Carrillo KC, Rodríguez-Romero A, Tovar-Sánchez A, Ruiz-Gutiérrez G, Fuente JRV. Geochemical baseline establishment, contamination level and ecological risk assessment of metals and As in the Limoncocha lagoon sediments, Ecuadorian Amazon region. *Journal of Soils and Sediments* 2022;22:293-315.
- Chen L, Han B, Wang X, Zhao J, Yang W, Yang Z. Machine learning methods in weather and climate applications: A survey. *Applied Sciences (Switzerland)* 2023;13(21):Article No. 12019.
- Crane JL, Bijak AL, Maier MA, Nord MA. Development of current ambient background threshold values for sediment quality parameters in U.S. lakes on a regional and statewide basis. *Science of the Total Environment* 2021;793:Article No. 148630.
- Devan P, Khare N. An efficient XGBoost-DNN-based classification model for network intrusion detection system. *Neural Computing and Applications* 2020;32:12499-514.
- Drogkoula M, Kokkinos K, Samaras N. A comprehensive survey of machine learning methodologies with Emphasis in water resources management. *Applied Sciences (Switzerland)* 2023;13(22):Article No. 12147.

- El-Sharkawy M, Alotaibi MO, Li J, Du D, Mahmoud E. Heavy metal pollution in coastal environments: Ecological implications and management strategies: A review. *Sustainability (Switzerland)* 2025;17(2):Article No. 701.
- Elreedy D, Atiya AF, Kamalov F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning* 2024; 113:4903-23.
- Gambin AF, Angelats E, Gonzalez JS, Miozzo M, Dini P. Sustainable Marine Ecosystems: Deep Learning for Water Quality Assessment and Forecasting. *IEEE Access* 2021;9:121344-65.
- Guo H, Song Z, Wang S, Yan S, Wang Y, Gao Y, et al. Assessment of heavy metal contamination and ecological risk in mangrove marine sediments inside and outside Zhanjiang Bay: Implications for conservation. *Journal of Marine Science and Engineering* 2025;13:1-15.
- Gündoğdu S. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimedia Tools and Applications* 2023; 82:34163-81.
- Hamidou ST, Mehdi A. Enhancing IDS performance through a comparative analysis of Random Forest, XGBoost, and Deep Neural Networks. *Machine Learning with Applications* 2025;22:Article No. 100738.
- Hassan MSADH, Al Naqeeb NA, Al-Musawi MR, Banoon SR, Mohammed K, Bilal M, et al. Using artificial intelligence to predict aquatic pollution: A comprehensive review. *International Journal of Aquatic Biology* 2025;13:50-67.
- Hossain MB, Masum Z, Rahman MS, Yu J, Noman MA, Jolly YN, et al. Heavy metal accumulation and phytoremediation potentiality of some selected mangrove species from the world's largest mangrove forest. *Biology (Basel)* 2022;11(8):Article No. 1144.
- Imani M, Beikmohammadi A, Arabnia HR. Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. *Technologies (Basel)* 2025;13:Article No. 88.
- John J, Nandhini AR, Chellam PV, Sillanpää M. Microplastics in mangroves and coral reef ecosystems: A review. *Environmental Chemistry Letters* 2022;20:397-416.
- Jongjaraunsuk R, Taparhudee W, Sirisuay S, Kaewnern M, Dulyapark V, Janekitkarn S. Transfer learning model application for *Rastrelliger brachysoma* and *R. kanagurta* image classification using smartphone-captured images. *Fishes* 2024;9(3):Article No. 103.
- Kavzoglu T, Teke A. Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal for Science and Engineering* 2022;47:7367-85.
- Khatun P, Umam S, Razzak RB, Shamsuddin IB, Salma N. A study on the effectiveness of machine learning models for hepatitis prediction. *Scientific Reports* 2025a;15:Article No. 30659.
- Khruschev SS, Plyusnina TY, Antal TK, Pogosyan SI, Ryznichenko GY, Rubin AB. Machine learning methods for assessing photosynthetic activity: Environmental monitoring applications. *Biophysical Reviews* 2022;14:821-42.
- Koudenoukpo ZC, Odountan OH, Agboho PA, Dalu T, Van Bocxlaer B, de Bistoven LJ, et al. Using self-organizing maps and machine learning models to assess mollusc community structure in relation to physicochemical variables in a West Africa river-estuary system. *Ecological Indicators* 2021;126:Article No. 107706.
- Kumari P, Kumar P. Metal(loid) source and effects on Peri-Urban agriculture/aquaculture sediments. In: Kumar P, Aishwarya, editors. *Technological Approaches for Climate Smart Agriculture*. Cham: Springer International Publishing; 2024. p. 133-64.
- Lan J, Liu P, Hu X, Zhu S. Harmful algal blooms in eutrophic marine environments: Causes, monitoring, and treatment. *Water (Switzerland)* 2024;16(17):Article No. 2525.
- Li D, Tian M, Ding W, Cisse EHM, Miao L, Ye B, et al. Dissecting possible correlations between leaf functional traits and heavy metal accumulation in two contrasting mangrove species across tidal gradients. *Environmental and Experimental Botany* 2025;238:Article No. 106234.
- Maruf A Al, Fahim SH, Bashar R, Romy RA, Chowdhury SI, Aung Z. Classification of freshwater fish diseases in Bangladesh using a novel ensemble deep learning model: Enhancing accuracy and interpretability. *IEEE Access* 2024;12:96411-35.
- Mederos-Barrera A, Albors L, Marques F, Marcello J, Martinez G, Eugenio F. Comparison of conventional machine learning and convolutional deep learning models for seagrass mapping using satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2026;19:3888-907.
- Meraj G, Abouleish MY, Ali T, Hashimoto S, Marazi A, Chakraborty R, et al. Middle Eastern mangroves at the arid limit (Red Sea and Arabian/Persian Gulf): Eco-biophysical dynamics, blue-carbon MRV, climate-risk pathways, and governance for resilient restoration: A comprehensive review. *Frontiers in Marine Science* 2025;12:1-31.
- Mok WJ, Ghaffar MA, Noor MIM, Lananan F, Azra MN. Understanding climate change and heavy metals in coastal areas: A macroanalysis assessment. *Water (Switzerland)* 2023;15(5):Article No. 891.
- Natras R, Soja B, Schmidt M. Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting. *Remote Sensing* 2022;14:Article No. 3547.
- Petrea SM, Costache M, Cristea D, Strungaru SA, Simionov IA, Mogodan A, et al. A machine learning approach in analyzing bioaccumulation of heavy metals in turbot tissues. *Molecules* 2020;25:Article No. 4696.
- Reshma B, Rahul B, Sreenath KR, Joshi KK, Grinson G. Taxonomic resolution of coral image classification with Convolutional Neural Network. *Aquatic Ecology* 2023; 57:845-61.
- Roy T, Dey TK, Jamal M. Microplastic/nanoplastic toxicity in plants: An imminent concern. *Environmental Monitoring and Assessment* 2023;195:Article No. 27.
- Sannigrahi M, Thandeeswaran R. Predictive analysis of network-based attacks by hybrid machine learning algorithms utilizing Bayesian optimization, logistic regression, and random forest algorithm. *IEEE Access* 2024;12:142721-32.
- Sari K, Soeprbowati TR. Impact of water quality deterioration in mangrove forest in Semarang Coastal Area. *Indonesian Journal of Limnology* 2021;2(2):37-48.
- Sarwar J, Khan SA, Azmat M, Khan F. A comparative analysis of feature selection models for spatial analysis of floods using hybrid metaheuristic and machine learning models.

- Environmental Science and Pollution Research 2024; 31:33495-514.
- Shahri NHNB, Lai SBS, Mohamad MB, Rahman HABA, Rambli A Bin. Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data. Mathematics and Statistics 2021;9(3):79-85.
- Shaika NA, Khan S, Awal S, Haque MM, Bashar A, Simsek H. Aquatic pollution in the Bay of Bengal: Impacts on fisheries and ecosystems. Hydrology 2025;12(7):Article No. 191
- Tallam K, Nguyen N, Ventura J, Fricker A, Calhoun S, O'Leary J, et al. Application of deep learning for classification of intertidal Eelgrass from drone-acquired imagery. Remote Sensing 2023;15:1-13.
- Wah YB, Ismail A, Azid NNN, Jaafar J, Aziz IA, Hasan MH, et al. Machine learning and synthetic minority oversampling techniques for imbalanced data: Improving machine failure prediction. Computers, Materials and Continua 2023; 75(3):4821-41.
- Wong WY, Al-Ani AKI, Hasikin K, Khairuddin ASM, Razak SA, Hizaddin HF, et al. Water, soil and air pollutants' interaction on mangrove ecosystem and corresponding artificial intelligence techniques used in decision support systems: A review. IEEE Access 2021;9:105532-63.
- Yap CK, Al-Mutairi KA. Potentially toxic metals in the tropical mangrove non-salt secreting *Rhizophora apiculata*: A field-based biomonitoring study and phytoremediation potentials. Forests 2023;14(2);Article No. 237.
- Yuan Y, Lin L. Self-supervised pretraining of transformers for satellite image time series classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2021;14:474-87.
- Zaka MM, Samat A. Advances in remote sensing and machine learning methods for invasive plants study: A comprehensive review. Remote Sensing 2024;16;Article No. 3781.