ECOLOGIA BALKANICA

2025, Vol. 17, Issue 2 December 2025 pp. 084-095

Evaluating heavy metal pollution and health risks in river systems using Random Forest and XGBoost: Evidence from the Shkumbin River

Bederiana Shyti[1,*], Lule Basha[2], Lirim Bekteshi[3]

[1]University of Elbasan "Aleksandër Xhuvani", Faculty of Natural Sciences, Department of Mathematics, postal code 3001, Elbasan, ALBANIA
[2]University of Tirana, Faculty of Natural Science, Department of Applied Mathematics, postal code 1001, Tirana, ALBANIA
[3]University of Elbasan "Aleksandër Xhuvani", Faculty of Natural Sciences, Department of Chemistry, postal code 3001, Elbasan, ALBANIA
*Corresponding author: bederiana.shyti@uniel.edu.al

Abstract. Surface water contamination by heavy metals poses significant ecological and health risks due to their persistence, bioaccumulation, and toxicity. This research evaluated the concentrations of cadmium (Cd), chromium (Cr), copper (Cu), iron (Fe), lead (Pb), and zinc (Zn) in river water samples and assessed their impact on the Heavy Metal Pollution Index (HPI). Descriptive statistics revealed substantial variation among sampling sites, with HPI values ranging from 2.15 to 21.94. Although Cd and Pb were generally present in low concentrations, their localized maxima indicated potential hot spots of contamination, whereas Fe and Zn showed higher overall levels. To identify the most influential predictors of HPI, two machine learning regression models, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), were implemented. The RF model explained more than 90% of the variance in HPI, with Cd, Zn, and Cr emerging as the most critical contributors. The XGBoost model achieved even higher predictive accuracy (R^2 = 0.998, RMSE = 0.76), confirming Cd and Cr as dominant predictors, together accounting for nearly 80% of the model's explanatory power. These findings highlight the pivotal role of Cd and Cr in shaping HPI dynamics and demonstrate the utility of ensemble learning methods for environmental monitoring and risk assessment.

Key words: Heavy Metal Pollution Index, Machine Learning Models, Random Forest, XGBoost Models.

Introduction

Surface water contamination by heavy metals is a pressing environmental issue due to its persistence, bioaccumulative properties, and potential risks to both ecosystems and human health. Industrial effluents, agricultural runoff, and urban discharges are among the primary sources contributing to elevated levels of toxic metals such as cadmium (Cd), chromium (Cr), copper (Cu), iron (Fe), lead (Pb), and zinc (Zn) in river systems. Even at low concentrations, these elements can have detrimental ecological and health impacts, making their monitoring a priority for environmental management strategies. Human activities are a significant contributor to metal pollutants in water sources, making groundwater contamination one of the most pressing environmental issues of our time (Sarkar, 2024). Some heavy metals - such as iron (Fe) and copper (Cu) - are essential for life in trace amounts. Others, like lead (Pb) and chromium (Cr), are toxic even at very low concentrations, making them among the most harmful water contaminants. These toxins accumulate in the bodies of animals and humans, potentially leading to serious diseases such as cancer (Tchounwou et al., 2012).

Numerous studies have examined the water quality of the Shkumbin River, using statistical methods such as the Water Quality Index (WQI) (Basha et al., 2024; Shyti et al., 2024), as well as through direct analysis of dissolved heavy metals in river water. For example, author in (Gjeci et al., 2024) showed that concentrations of nitrate and nitrite in the Shkumbin River were generally below WHO (2004) limits for surface water (50 mg/L for nitrate and 1.0 mg/L for nitrite). However, ammonium levels were found to exceed the recommended maximum of 0.3 mg/L (Gjeci et al., 2024). Similarly, the study of Çomo et al. (2024), concluded that heavy metals in the Shkumbin River were present in the order of Fe > Ni > Cu > Cr > Cd > Pb > Mn > Zn.

Traditional water quality assessments often rely on indices such as the Heavy Metal Pollution Index (HPI), which integrates multiple pollutant concentrations into a single metric, facilitating risk evaluation and site comparison. However, the complex interactions among multiple metals and their nonlinear effects on HPI demand advanced statistical and machine learning approaches for more accurate modeling and interpretation. In recent years, ensemble learning methods such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) have proven effective in environmental applications, owing to their robustness, ability to handle multicollinearity, and strong predictive performance. Machine learning has predicted environmental processes and assessed pollutant toxicity using existing data (Gao et al., 2021; Yang et al., 2021). The unique thing about XGBoost model is its generalization and ability to outperform other models with large datasets and that it could be applied for various tasks as regression and classification and still perform with the same strength (Tao et al., 2021).

Heavy metal contamination of surface waters is a globally recognized environmental problem due to its persistence, non-degradability, and bioaccumulation in aquatic food webs. Numerous studies have reported elevated levels of cadmium (Cd), lead (Pb), chromium (Cr), zinc (Zn), copper (Cu), and iron (Fe) in rivers impacted by industrial discharges, mining activities, and agricultural runoff (Su et al., 2022; Varol & Şen, 2012). These metals, even at trace concentrations, pose serious health risks, including carcinogenicity (Cd, Cr, Pb) and neurotoxicity (Pb, Cu, Zn), and disrupt aquatic biodiversity.

To assess overall water quality, researchers have developed and applied composite indices such as the Heavy Metal Pollution Index (HPI), which integrates multiple metal concentrations into a single risk indicator (Mohan et al., 1996). HPI has been widely applied to rivers across Asia, Europe, and Africa, providing a robust framework for comparing pollution levels across sites and seasons (Backman et al., 1998; Prasad & Bose, 2001). For instance, Varol (2011) reported spatio-temporal variations in HPI across the Tigris River, highlighting seasonal peaks linked to agricultural runoff, while Olowojuni et al. (2025) assessed heavy metal contamination in water and sediments from five sampling stations in Asejire Reservoir, Oyo State, Nigeria. Traditional statistical approaches, such as correlation analysis and multiple linear regression, have been used to identify relationships between heavy metals and HPI (Singh et al., 2004). However, these methods often struggle with nonlinearities, multicollinearity among metals, and site-specific variations. In response, machine learning (ML) techniques have increasingly been applied in environmental monitoring. Random Forest is a machine learning method that combines the predictions of multiple decision trees to improve performance and the RF model has the capacity to predict the unsampled areas with higher accuracy Random Forest (RF) has been used successfully to predict groundwater contamination and surface water quality due to its robustness against overfitting and ability to capture nonlinear relationships (Apogba et al., 2024). Recent studies integrating ML models with heavy metal monitoring confirm their potential. Authors as Shahed et al. (2022), develop and apply data-driven models using Random Forest (RF), a machine learning approach, to predict Total Nitrogen (TN), Total Phosphorus (TP), Total Suspended Solids (TSS), and Ortho-Phosphorus (Ortho-P) EMCs in urban runoff.

Similarly, Extreme Gradient Boosting (XGBoost) has gained attention for its superior predictive accuracy and interpretability in hydrological modeling and water quality assessments. Authors as Zhang et al. (2024) demonstrated that XGBoost could outperform linear models in predicting water quality indices in complex datasets with mixed pollution sources. Moreover, hybrid

approaches combining HPI with ML algorithms have emerged as promising tools for both risk assessment and management decision-making (Liu et al., 2021).

Despite these advances, relatively few studies have applied ensemble learning models directly to HPI prediction in European river systems, particularly in the Western Balkans. This gap underscores the significance of the present study, which evaluates heavy metal contributions to HPI using both RF and XGBoost, offering new insights into the predictive roles of Cd, Zn, and Cr. By integrating advanced modeling approaches with environmental indices, the research contributes to more accurate pollution diagnostics and supports targeted mitigation strategies.

The present study aims to (i) analyze the concentrations of selected heavy metals in river water samples, (ii) evaluate spatial variation in HPI, and (iii) apply RF and XGBoost regression models to identify the most influential predictors of HPI. By integrating descriptive statistics with machine learning approaches, this research provides new insights into the relative importance of specific heavy metals, with implications for targeted pollution control and water resource management.

Materials and methods

Sampling and Data Collection

In southeast Albania, the Shkumbin River rises in the Valamara mountain range, and from its source to its delta in the Adriatic Sea, this significant watercourse extends for 181 kilometers, draining an area of 2,444 square kilometers in an east–west direction (Basha et al., 2024).

To assess heavy metal pollution in the Shkumbini River, water samples were collected from five strategic locations along its flow: Proptisht, Librazhd, Elbasan, Peqin and Bashtovë. Samples were gathered during four representative months (February, May, August and November) in 2024 to account for seasonal variations (Table 1).

Table 1. Geographic coordinates of selected locations in Central Albania.

Place	Latitude	Longitude
Proptisht	40.98	20.51
Librazhd	41.19	20.32
Elbasan	41.11	20.08
Peqin	41.05	19.75
Bashtovë	41.00	19.47

The collection of reliable water samples is a critical step in assessing heavy metal contamination in river systems. In this study, the sampling design was carefully structured to ensure that spatial and temporal variability could be adequately represented. Sampling sites were selected to include upstream reference points, midstream sections, and downstream areas subject to agricultural, industrial, and urban influences. This distribution provided a comprehensive view of both background conditions and pollution hot spots. To account for seasonal variability, field campaigns were conducted at multiple times during the year, covering both high- and low-flow periods.

Water sampling was carried out in accordance with the UNEP/WHO guidelines (UNEP/WHO, 1996). All samples were collected using pre-cleaned, acid-washed high-density polyethylene (HDPE) bottles, which were transported and handled according to trace-metal protocols to avoid contamination. At each site, grab samples were obtained from the main current, typically at mid-stream, to minimize local disturbances caused by bank effects or eddies. For sites deeper than 0.5 meters, composite samples were prepared by combining aliquots collected from the surface, mid-depth, and near-bottom layers, thereby capturing vertical variability. In all cases, strict "clean hands/dirty hands" procedures were followed to minimize contamination during collection and handling.

Field parameters, including temperature, pH, electrical conductivity (EC), dissolved oxygen (DO), and turbidity, were measured in situ using portable multi-parameter probes. For dissolved heavy metals, water samples were immediately filtered on-site through 0.45 µm membrane filters, after which both filtered (dissolved fraction) and unfiltered (total fraction) samples were preserved by acidification with ultrapure nitric acid (HNO_3)

to a pH below 2. Samples were stored in coolers at ≤ 6°C and transported to the laboratory under dark conditions to ensure sample integrity.

To maintain quality assurance and quality control (QA/QC), a set of control samples was collected alongside field samples. These included field blanks prepared by processing ultrapure water through the sampling equipment, field duplicates collected independently at the same site, and equipment blanks obtained after cleaning procedures. Such measures were implemented to detect possible contamination, verify reproducibility, and confirm the effectiveness of decontamination. In addition, triplicate samples were collected periodically to evaluate analytical precision.

Water quality evaluation requires a combined evaluation of physical, chemical and biological parameters. We are focused on the evaluation of chemical pollution mainly from heavy metals, which have a significant impact on the health of the aquatic living world (Singh et al., 2022). In the laboratory, concentrations of cad-mium (Cd), chromium (Cr), copper (Cu), iron (Fe), lead (Pb), and zinc (Zn) were determined using inductively coupled plasma mass spectrometry (ICP-MS) for trace-level elements and inductively coupled plasma optical emission spectrometry (ICP-OES) for metals with higher expected concen-trations. Analytical quality was assured through the use of calibration standards, matrix spikes, and certified reference materials. Data were ex-pressed in milligrams per liter (mg/L) to ensure consistency across all analyses. Values below the limit of detection (LOD) were treated according to a standard substitution approach (LOD/2), and sensitivity analyses were conducted to assess the impact of this treatment.

By following this standardized protocol, the study ensured that the data collected were representative of spatial and seasonal variation while maintaining the accuracy and reliability required for subsequent statistical analyses and modeling of the Heavy Metal Pollution Index (HPI).

Analytical methods

- ***HPI calculation***

The Heavy Metal Pollution Index (HPI) was calculated using the formula proposed by Mohan et al. (1996):

$$\boldsymbol{HPI} = \frac{\sum \boldsymbol{W_i Q_i}}{\sum \boldsymbol{W_i}}$$

where: W_i is the unit weightage for the i^{th} parameter, Q_i is the sub-index of the i^{th} parameter.

The sub-index Q_i is calculated as:

$$\boldsymbol{Q_i} = \frac{(\boldsymbol{M_i} - \boldsymbol{I_i})}{(\boldsymbol{S_i} - \boldsymbol{I_i})} \times \mathbf{100}$$

where: M_i is the measured concentration of the i^{th} metal, I_i is the ideal concentration (usually zero for heavy metals), S_i is the standard permissible value.

The unit weight W_i is inversely proportional to the standard permissible value:

$$\boldsymbol{W_i} = \frac{\mathbf{1}}{\boldsymbol{S_i}}$$

- ***Reference Standards***

Permissible limits for each metal were taken from WHO guidelines (WHO, 2004), and international environmental quality standards, adjusted for surface water use. These include maximum admissible concentrations (MAC) relevant for the protection of aquatic life.

The World Health Organization (WHO) provides guidelines primarily for drinking water quality, rather than for surface water (e.g., rivers, lakes, and streams). For surface water quality, we refer to international environmental agencies, including:

- European Union Water Framework Directive (EU-WFD)
- United States Environmental Protection Agency (US EPA)
- Canadian Water Quality Guidelines
- Other regional standards based on intended water use (e.g., for aquatic life, irrigation, industrial, or recreational purposes).

Below is a table summarizing the typical maximum allowable concentrations of selected heavy metals in surface water, primarily to protect aquatic life (Table 2). These reference values provide a benchmark against which the concentrations found in the Shkumbin River can be evaluated.

- ***Random Forest model***

The Random Forest (RF) algorithm, introduced by Breiman (2001), is an ensemble learning method based on the aggregation of multiple decision trees. Given a dataset $= \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i$denotes the feature vector and y_i the response, RF constructs *B*ootstrap samples from *A*. For each bootstrap sample, a decision tree is grown using recursive partitioning. At each node, instead of

considering all p predictors, a random subset of $m<<p$ variables is selected, and the optimal split is determined by minimizing an impurity measure such as the Gini index for classification:

$$G(t) = \sum_{k=1}^{K} p_k(t)\,(1 - p_k(t))$$

where: $p_k(t)$ is the proportion of class k instances in node t. For regression, the variance is minimized. The final prediction is obtained through majority voting in classification:

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_B(x)\}$$

or averaging in regression:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

By combining bootstrap aggregation (bagging) with random feature selection, Random Forests reduce variance, mitigate overfitting, and maintain high predictive accuracy even in high-dimensional and noisy datasets. Additionally, RF provides variable importance measures, such as the Mean Decrease in Impurity (MDI) and the Mean Decrease in Accuracy (MDA), which are widely used for feature selection and interpretation (Hastie et al., 2009).

Table 2. Typical maximum concentrations of selected heavy metals in aquatic environments (µg/L) and associated toxicological notes.

Heavy metal	Typical max value (µg/L)	Notes
Lead (Pb)	2 – 10 µg/L	Toxic to fish; bioaccumulates
Cadmium (Cd)	0.1 – 5 µg/L	Highly toxic even at low levels; varies with hardness
Chromium (Cr)	50 µg/L (for Cr (VI))	Hexavalent chromium is more toxic
Mercury (Hg)	0.05 – 0.1 µg/L	Extremely toxic; bio accumulative
Copper (Cu)	2 – 10	Toxic to aquatic organisms at elevated concentrations
Zinc (Zn)	30 – 100 µg/L	Affects fish at higher concentrations
Nickel (Ni)	25 – 70 µg/L	Limits vary based on pH and water hardness
Arsenic (As)	10 µg/L	Inorganic arsenic is highly toxic and carcinogenic

- ***Extreme Gradient Boosting (XGBoost)***

Extreme Gradient Boosting (XGBoost), proposed by Chen & Guestrin (2016), is an advanced implementation of gradient boosting algorithms designed for high efficiency, scalability, and predictive accuracy. Gradient boosting builds an additive model in a forward stage-wise manner, where at each iteration a new weak learner $f_t(x)$ (typically a decision tree) is added to minimize the objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

where $l(\cdot)$is a differentiable loss function.

The regularization term is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

where T is the number of leaves, w_j is the weight of leaf j, and γ, λ are regularization parameters.

XGBoost employs a second-order Taylor expansion of the loss function for efficient optimization (Friedman, 2001). By integrating shrinkage (learning rate), column subsampling, and advanced regularization, XGBoost achieves superior generalization performance compared to traditional gradient boosting. It has become one of the most widely used algorithms in data science competitions and applied research across domains such as finance, bioinformatics, and environmental modeling.

Results and Discussion

Assessment of Heavy Metal Contamination and Water Quality using HPI

Over the years, numerous studies have analyzed water samples from the Shkumbin River to determine variations in heavy metal concentrations across different months.

Among these:

Iron (Fe). Iron is an essential micronutrient for living organisms. In the collected data, concentrations ranged from a minimum of 350 µg/L in November at Proptisht to a maximum of 1240 µg/L

in May at Elbasan. Potential sources of iron in river water include:

- Natural sources: weathering of rocks and soils along the river course.
- Anthropogenic sources: industrial activities, mining residues, sewage discharges, and leaking pipelines.

Excessive concentrations of iron may cause undesirable reddish coloration of the water, the formation of iron oxide deposits that damage aquatic infrastructure, and toxic effects on fish and microorganisms under oxygen-depleted conditions. Elevated iron levels are often associated with the presence of other heavy metals or organic pollutants. Conversely, lower iron concentrations are an indicator of improved water quality, reflecting the absence of industrial or sewage-related inputs, and therefore a greater suitability for agricultural use.

Zinc (Zn). Zinc plays an essential role in aquatic ecosystems, influencing the health of fish, aquatic plants, and other organisms. While necessary in trace amounts, elevated concentrations may result in neurological and immunological damage, growth and metabolic disorders, reduced water quality, and biodiversity loss. In the studied samples, zinc concentrations ranged from 4.2 mg/L in February at Proptisht to 24.5 mg/L in August at Elbasan. The peak in summer is likely driven by industrial discharges (notably from Elbasan's metal-processing zones), seasonal increases in human activity, reduced river flow due to low rainfall, and agricultural water consumption. Elevated temperatures during August also enhance pollutant mobility and accumulation.

Copper (Cu). Copper is a vital trace element required for various biological processes. However, excess levels can cause environmental pollution, primarily from industrial effluents, fertilizers, pesticides, and soil erosion. In the Shkumbin River, mining areas near Elbasan and Librazhd are potential contributors. Measured values ranged from 0.52 µg/L in November at Elbasan to 2.88 µg/L in August at Elbasan. Excess copper adversely affects aquatic life, impairing the nervous and respiratory systems of fish, destroying plankton and microorganisms, and reducing agricultural suitability when contaminated water is used for irrigation.

Chromium (Cr). Chromium contamination in the Shkumbin River likely originates from historical mining and industrial activities in the Elbasan region, as well as urban runoff and natural erosion. Sample analysis revealed values ranging from 1.81 µg/L in November to 7.12 µg/L in August at Elbasan. While trivalent chromium (Cr^{3+}) is a micronutrient, elevated concentrations become toxic and environmentally hazardous.

Lead (Pb). Lead is among the most toxic pollutants detected in the river, with concentrations ranging from 0.70 µg/L in February at Proptisht to 2.81 µg/L in August at Elbasan. Industrial activity in Elbasan (metallurgy, car battery production, and electronic waste) is the main contributor, compounded by urban runoff during rainfall events. Lead contamination affects soil fertility, poses severe risks to aquatic ecosystems, and remains a long-term threat to public health.

Cadmium (Cd). Cadmium, though naturally present in trace amounts, is highly toxic and primarily linked to industrial operations, mining, fertilizers, and urban waste. In the samples, Cd levels varied between 0.036 µg/L in November at Proptisht and 0.880 µg/L in August at Peqin. Although below the WHO limit of 5 µg/L for surface waters, cadmium is concerning due to its ability to bioaccumulate and disrupt aquatic biodiversity.

The descriptive statistics of the studied variables demonstrate significant variation across both heavy metals and the HPI indicator (Table 3). The HPI values ranged from 2.15 to 21.94, with a median of 6.04, showing that while the majority of observations are quite low, a few places have significantly higher pollution levels. Cd and Pb showed low amounts, with median values of 8.8×10^{-5} mg/L and 0.00147 mg/L, respectively. However, their maximum values were 8.8×10^{-4} mg/L and 0.00281 mg/L, indicating localized contamination. Cr and Cu were moderately concentrated, with median values of 0.0035 mg/L and 0.0016 mg/L, respectively, whereas Fe and Zn were more concentrated, with median values of 0.68 mg/L and 10.7 mg/L, respectively, and Zn reached up to 24.5 mg/L. Overall, the data indicate that, although most sites maintain low to moderate levels of heavy metals, certain elements and locations show elevated values, which could have implications for environmental and public health monitoring.

Figure 1 presents the variation of the calculated HPI across different sampling locations and

months (February, May, August, and November). It highlights both seasonal fluctuations and spatial differences. Peqin and Bashtova exhibit significant fluctuations, particularly in the months of May and August. Both locations show a significant increase in HPI during May and August, which may be related to seasonal factors such as agricultural runoff or industrial activity.

In November, the HPI drops noticeably, indicating a temporary improvement or a decrease in pollution sources. Proptisht, Librazhd, and Elbasan show lower and more stable values.

Table 3. Descriptive statistics of HPI and selected heavy metals in sampling sites.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
HPI	2.150	4.513	6.035	8.918	12.120	21.940
Cd	2.80e-05	5.025e-05	8.80e-05	2.681e-04	4.935e-04	8.80e-04
Cr	0.001180	0.002785	0.003500	0.003798	0.004925	0.007120
Cu	0.000520	0.001200	0.001600	0.001635	0.002065	0.002880
Fe	0.2800	0.5375	0.6800	0.6945	0.7900	1.2400
Pb	0.000650	0.001143	0.001470	0.001553	0.001850	0.002810
Zn	4.20	9.25	10.70	12.53	16.95	24.50

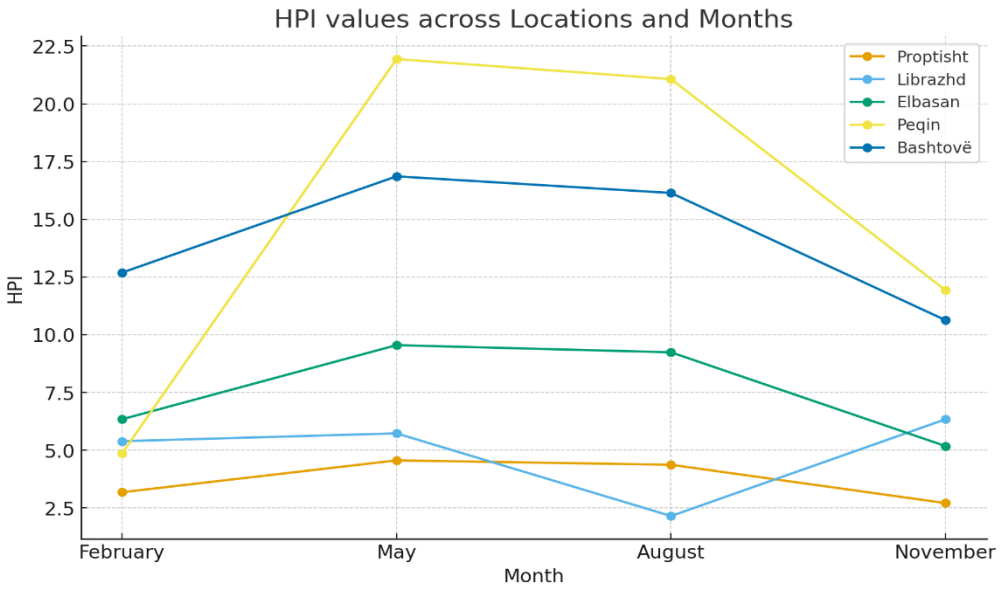

Fig. 1. HPI values across sampling locations and months.

Machine Learning Models for HPI Prediction

- ***Random Forest regression model***

A Random Forest regression model was created to investigate the effect of specific heavy metals (Cd, Cr, Cu, Fe, Pb, and Zn) on the Health Pollution Index (HPI). The model was trained on 500 trees, with two variables examined at each split. The findings show high model performance, with a mean squared residual of 6.78 and an explained variance of around 80.67%, implying that the selected variables account for a significant percentage of the variability in HPI. Variable importance measurements revealed that cadmium (Cd) had the greatest predictive impact, followed by zinc (Zn) and chromium (Cr), as demonstrated by the increase in mean squared error (%IncMSE) when each variable was changed. These findings highlight the critical role of certain heavy metals, particularly Cd, in shaping pollution-related health risks.

The Random Forest model was tuned using 5-fold cross-validation in order to identify the optimal number of predictors sampled at each split (mtry) (Fig. 2). The results showed that as mtry increased, the model's performance improved consistently. The Root Mean Squared Error

(RMSE) fell from 3.06 to 2.11 at mtry = 5, while the coefficient of determination (R^2) improved from 0.77 to 0.90. The Mean Absolute Error (MAE) followed a similar pattern, decreasing from 2.79 to 1.66. While mtry = 6 resulted in a slightly higher R^2 value (0.93), the RMSE (2.13) was somewhat greater. As a result, using the lowest RMSE criterion, the best model with mtry = 5 was chosen. This specification demonstrated high predictive accuracy, accounting for more than 90% of the variance in HPI and providing a solid foundation for subsequent analysis

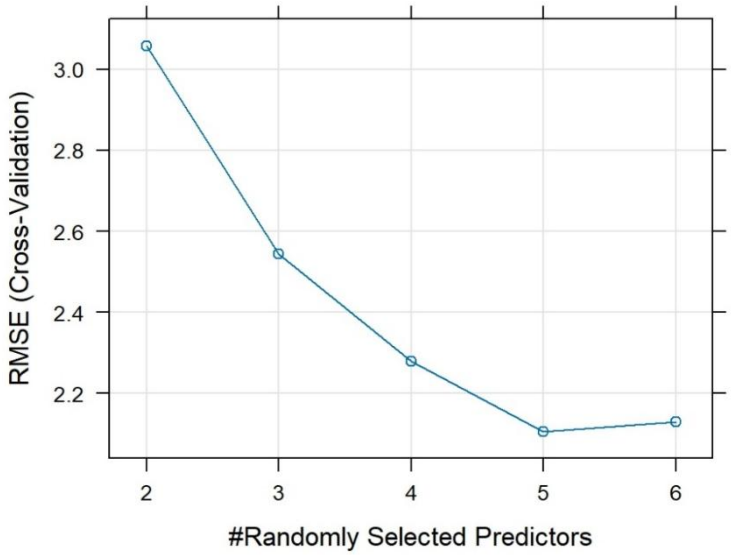

Fig. 2. Hyperparameter tuning of Random Forest based on cross-validated RMSE.

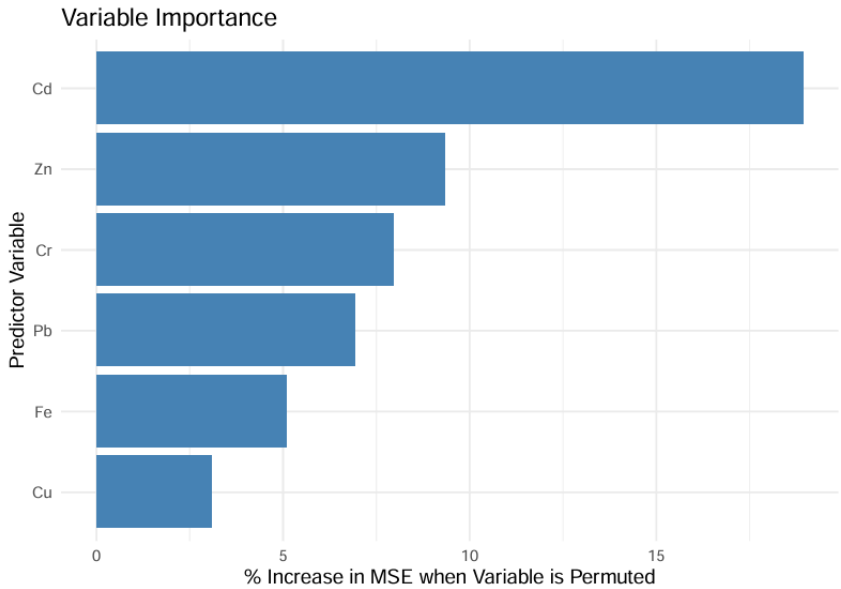

Fig. 3. Variable importance of predictors in the Random Forest model.

To determine the relative contribution of each predictor variable to the Random Forest model, we used the permutation approach to obtain variable significance scores. In this method, the values of each predictor are randomly permuted while all other variables remain constant. The subsequent increase in the model's mean squared error (MSE) is then measured. A higher MSE suggests that the variable is more crucial for accurate prediction, because randomization reduces predictive ability. The important scores were calculated as a percentage increase in MSE, resulting in a

standardized measure of each heavy metal's relative predictive influence on HPI. The examination of variable importance inside the Random Forest model revealed that heavy metals contributed differently to the prediction of HPI (Fig. 3). Cadmium (Cd) was by far the most powerful predictor, with each permutation resulting in the greatest increase in prediction error. Zinc (Zn) and chromium (Cr) also made significant contributions, while lead (Pb) and iron (Fe) had a moderate impact. Copper (Cu) had the lowest predictive importance, indicating a lesser connection with HPI. These findings indicate that Cd, Zn, and Cr are the primary causes of HPI changes in the dataset.

- ***Extreme Gradient Boosting regression model***

An Extreme Gradient Boosting (XGBoost) regression model was used to predict the HPI variable with six predictors (Cd, Cr, Cu, Fe, Pb, and Zn). The model was trained over 200 boosting iterations (nrounds = 200), with the squared error loss function (objective = "reg: squarederror") and the root mean squared error (RMSE) as the evaluation measure. The hyperparameters were configured as follows: a maximum tree depth of 4 (max_depth = 4), a learning rate of 0.1 (eta = 0.1), a subsample ratio of 0.8, and 0.8 column sampling per tree. The training method demonstrated a significant drop in error during the first iterations, which stabilized as the number of trees rose. In the final iteration, the model had a training RMSE of 0.0016 and a test RMSE of 0.762. This indicates that the model fits the training data very closely while maintaining a relatively low error on the independent test set, suggesting good predictive performance without substantial overfitting.

The XGBoost regression model's prediction ability was tested on an independent test set using root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The model showed an RMSE of 0.762, MAE of 0.560, and R^2 of 0.998. These findings show that the model fits the data well, explaining more than 99% of the variance in the response variable (HPI). The extremely low RMSE and MAE values demonstrate that prediction errors are small, indicating that the XGBoost model accurately represents the underlying link between heavy metal concentrations (Cd, Cr, Cu, Fe, Pb, and Zn) and the HPI index. The learning curve illustrates the evolution of the training and testing root mean squared error (RMSE) across 200 boosting iterations in the XGBoost model (Fig. 4). During the first 50 iterations, both training and testing errors fell significantly, demonstrating that the model quickly captured the data's main structure. The test RMSE reached its lowest point at about 50 iterations, after which it stayed pretty steady at 0.76, whereas the training RMSE continued to fall toward zero. This pattern indicates that the model provided a strong fit to the data while minimizing overfitting. The stability of the test error over subsequent iterations suggests that the model generalizes well; nevertheless, an early halting approach around 50-70 iterations could yield a more parsimonious model without sacrificing predictive performance.

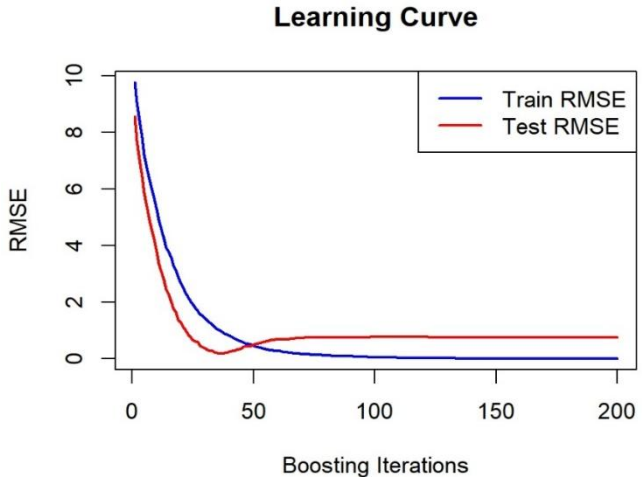

Fig. 4. Learning curve of the XGBoost regression model showing RMSE values over boosting iterations.

The XGBoost models feature importance analysis found that Cadmium (Cd) was the dominant predictor, accounting for the maximum gain (51.8%) and being selected in around 37% of the splits (Table 4). Chromium (Cr) was the second most impactful variable, with a gain of 26.7% and a comparable selection frequency (26%). Cd and Cr contributed roughly 80% of the model's predictive power. Other factors had smaller contributions: lead (Pb) (13.0% increase), copper (Cu) (4.1%), iron (Fe) (3.5%), and zinc (Zn) (0.8%). Although these predictors emerged in the trees (particularly Cu and Zn with considerable coverage values), their predictive power was small in comparison to Cd and Cr. Overall, the XGBoost model identified Cd and Cr as the most critical variables for predicting HPI, suggesting that these two heavy metals play the strongest role in determining index variation.

Table 4. Feature importance of predictor variables in the XGBoost regression model for HPI.

Variable	Gain	Cover	Frequency
Cd	0.518	0.354	0.375
Cr	0.267	0.250	0.261
Pb	0.130	0.037	0.038
Cu	0.041	0.159	0.153
Fe	0.035	0.091	0.084
Zn	0.008	0.108	0.090

Conclusions

This study assessed heavy metal contamination in river water samples by combining descriptive statistics, the Heavy Metal Pollution Index (HPI), and advanced machine learning models, specifically Random Forest and XGBoost. The findings indicate that water quality across the study area shows considerable spatial variation. Although overall HPI values ranged from low to moderate (2.15–21.94), certain locations exhibited higher levels, signaling localized risks of pollution. While cadmium (Cd) and lead (Pb) were generally detected in trace amounts, occasional peaks pointed to possible point-source contamination, whereas iron (Fe) and zinc (Zn) tended to occur at elevated background concentrations across multiple sites.

Among the analyzed metals, cadmium consistently emerged as the most critical contributor to pollution risk, with both Random Forest and XGBoost models confirming Cd as the strongest predictor of HPI. Zinc and chromium (Cr) also played significant roles, while copper (Cu) showed the weakest association. The Random Forest model achieved high predictive performance ($R^2 \approx 0.90$, RMSE = 2.11), effectively capturing nonlinear relationships between heavy metals and HPI, whereas XGBoost further enhanced prediction accuracy ($R^2 = 0.998$, RMSE = 0.76), underlining its suitability for water quality assessment.

The identification of cadmium and chromium as dominant drivers of HPI underscores the importance of targeted monitoring and mitigation strategies focused on these metals. Given the relatively higher levels of zinc and iron, continuous surveillance of these elements is also warranted to reduce potential ecological and health risks. Beyond its empirical findings, this research highlights the methodological advantage of integrating ensemble machine learning techniques with traditional indices such as HPI. Such an approach not only improves the accuracy of pollutant identification but also strengthens predictive assessments of river water quality. Ultimately, this study demonstrates that combining HPI with machine learning offers a powerful framework for advancing river pollution monitoring and supporting evidence-based environmental management policies.

References

Apogba, J.N., Anornu, G.K., Koon, A.B., Dekongmen, B.W., Sunkari, E.D., Fynn, O.F., & Kpiebaya, P. (2024). Application of machine learning techniques to predict groundwater quality in the Nabogo Basin, Northern Ghana.

Heliyon, 10(7), e28527. doi: 10.1016/j.heliyon.2024.e28527

Backman, B., Bodiš, D., Lahermo, P., Rapant, S., & Tarvainen, T. (1998). Application of a groundwater contamination index in Finland and Slovakia. *Environmental Geology*, 36(1–2), 55–64. doi: 10.1007/s002540050320

Basha, L., Shyti, B., & Bekteshi, L. (2024). Evaluating the performance of machine learning approaches in predicting Albanian Shkumbini River's waters using water quality index model. *Journal of Environmental Engineering and Landscape Management*, 32(2), 117–127. doi: 10.3846/jeelm.2024.20979

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. doi: 10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. doi: 10.1145/2939672.2939785

Çomo, E., Hasimi, A., Tako, E., & Ormeni, Rr. (2024). Analysis of heavy metal distribution in Albanian rivers. *Environmental Chemistry Letters*, 22(2), 133–148. doi: 10.18178/ijesd.2024.15.3.1479

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. doi: 10.1214/aos/1013203451

Gao, F., Shen, Y., Sallach, J.B., Li, H., Liu, C., & Li, Y. (2021). Direct prediction of bioaccumulation of organic contaminants in plant roots from soils with machine learning models based on molecular structures. *Environmental Science & Technology*, 55(24), 16358–16368. doi: 10.1021/acs.est.1c02376

Gjeci, N., Hamiti, Xh., Qarri, F., & Lazo, P. (2024). Assessment of surface water quality in the Shkumbin River. *Journal of Environmental Monitoring*, 12(1), 45–53. doi: 10.62638/ZasMat1270

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York, 745 p. doi: 10.1007/978-0-387-84858-7

Mohan, S.V., Nithila, P., & Reddy, S.J. (1996). Estimation of heavy metals in drinking water and development of heavy metal pollution index. *Journal of Environmental Science and Health, Part A*, 31(2), 283–289. doi: 10.1080/10934529609376357

Olowojuni, O.A., Amulejoye, F.D., Ikuesan, B.B., Maulu, S., Bwalya, H., & Hasimuna, O.J. (2025). Water quality, heavy metal contamination, and ecological risk assessment in Asejire reservoir, Nigeria. *Journal of Freshwater Ecology*, 40(1), 2516505. doi: 10.1080/02705060.2025.2516505

Prasad, B., & Bose, J.M. (2001). Evaluation of the heavy metal pollution index for surface and spring water near a limestone mining area of the lower Himalayas. *Environmental Earth Sciences*, 1, 183–188. doi: 10.1007/s002540100380

Sarkar, L. (2024). Assessing the heavy metal contamination prevailing in groundwater at Rishipur Village, West Bengal, India. *Current World Environment*, 19(2), 664–678. doi: 10.12944/CWE.19.2.12

Shahed, M., Nayeb Yazdi, M., & Sample, D.J. (2022). Using random forest, a machine learning approach to predict nitrogen, phosphorrus, and sediment event mean concentrations in urban runoff. *Journal of Environmental Management*, 317, 115412. doi: 10.1016/j.jenvman.2022.115412

Shyti, B., Bekteshi, L., Paraloj, S., & Hila, E. (2024). Remodeling of the WQI index for the evaluation of the Shkumbini River's water quality in Albania using the statistical method. *Ecologia Balkanica*, 16(1), 58–67.

Singh, K.P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for evaluation of spatial and temporal variations in water quality of Gomti River (India). *Water Research*, 38(18), 3980–3992. doi: 10.1016/j.watres.2004.06.011

Singh, A., Sharma, A., K. Verma, R., L. Chopade, R., P. Pandit, P., Nagar, V., Aseri, V., Choudhary, S.K., Awasthi, G., Awasthi, K.K., & Sankhla, M.S. (2022). Heavy Metal Contamination of Water and Their Toxic Effect on Living Organisms. In Junqueira, D.D., & Palma de Oliveira, D. (Eds), *The Toxicity of Environmental Pollutants*. IntechOpen, 302 p. doi: 10.5772/intechopen.105075

Su, X., Ling, H., Wu, D., Xue, Q., & Xie, L. (2022). Spatial-temporal variations, ecological risk assessment, and source identification of heavy metals in the sediments of a shallow eutrophic lake, China. *Toxics*, 10(1), 16. doi: 10.3390/toxics10010016

Tao, H., Habib, M., Aljarah, I., Faris, H., Afan, H.A., & Yaseen, Z.M. (2021). An intelligent evolutionary extreme gradient boosting algorithm for modeling scour depths under submerged weir. *Information Sciences*, 570, 172–184. doi: 10.1016/j.ins.2021.04.063

Tchounwou, P.B., Yedjou, G.C., Patlolla, A.K., & Sutton, D.J. (2012). Heavy metal toxicity and the environment. EXS, 101, 133–164. doi: 10.1007/978-3-7643-8340-4_6

UNEP/WHO. (1996). *Water quality monitoring: A practical guide to the design and implementation of freshwater quality studies and monitoring programmes*. World Health Organization, United Nations Environment Programme, 348 p. Retrieved from: https://www.who.int/

Varol, M., & Şen, B. (2012). Assessment of nutrient and heavy metal contamination in surface water and sediments of the upper Tigris River, Turkey. *Catena*, 92, 1–10. doi: 10.1016/j.catena.2011.11.011

Varol, M. (2011). Assessment of heavy metal contamination in sediments of the Tigris River using pollution indices and multivariate statistical techniques. *Journal of Hazardous Materials*, 195, 355–364. doi: 10.1016/j.jhazmat.2011.08.05

World Health Organization (WHO). (2004). *Guidelines for drinking-water quality*. World Health Organization, 631 p. Retrieved from: https://www.who.int/

Yang, H., Huang, K., Zhang, K., Weng, Q., Zhang, H., & Wang, F. (2021). Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities. *Environmental Science & Technology*, 55(20), 14316–14328. doi: 10.1021/acs.est.1c02479

Zhang, K., Wang, X., Liu, T., Wei, W., Zhang, F., Huang, M., & Liu, H. (2024). Enhancing water quality prediction with advanced machine learning techniques: An extreme gradient boosting model based on long short-term memory and autoencoder. *Journal of Hydrology*, 644, 132115. doi: 10.1016/j.jhydrol.2024.132115

Received: 14.08.2025
Accepted: 19.11.2025