

Machine Learning-Based Spatial Mapping, Source Apportionment and Health Risk Prediction of Heavy Metal Contamination in Zanzan Province, Iran

Hossein Ali Maleki

University of Zanzan

Fatemeh Rajaei

Rajaei_Fatemeh@znu.ac.ir

University of Zanzan

Munna Kumar

Dr Ambedkar Institute of Technology for Divyangjan

Yogesh Kumar Vishwakarma

IIT (BHU)

R. S. Singh

IIT (BHU)

Article

Keywords: Heavy metals, Random Forest, Gradient Boosting, SVR, ANN, health risk, source apportionment

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10896617/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Heavy metal contamination of peri-industrial soils poses severe risks to human health, particularly in mining-intensive regions. This study presents a comprehensive machine learning (ML) analysis of six heavy metals (Zn, Cu, Cd, Ni, Fe, Pb) measured at 357 soil sampling sites across Zanzan Province, northwestern Iran. This study integrates exploratory data analysis, spatial interpolation, principal component analysis (PCA) and K-Means clustering. Where, PCA identified three dominant pollution source factors: PC1 (37.4% variance) associated with mining/smelting (Zn–Cd–Pb); PC2 (20.1%) to industrial/ferrous activities (Cu–Fe); and PC3 (16.0%) to geogenic dust (Ni–Fe). The four ML regression models, namely, Random Forest (RF), Gradient Boosting Regression (GBR), Support Vector Regression (SVR), and Artificial Neural Network (ANN), were used to predict six health risk indices (PLI, CDI, HI Adult, HI Child, TCR Adult, TCR Child). It has been shown that no single ML model is sufficient for all targets; for example, GBR achieved the best R^2 for PLI (0.962), CDI (0.982), and HI (0.981–0.982), whereas SVR uniquely excelled for TCR (0.982–0.985). The permutation importance analysis confirmed Cd as the dominant driver of non-carcinogenic risk and Cu/Ni as primary carcinogenic risk drivers across all four models. A critical child–adult health risk variation was identified that shows 35.0% of sites have a higher child HI threshold (> 1) compare to 2.8% for adults (12.5-fold difference), and 79.8% higher the carcinogenic threshold for children ($TCR > 10^{-4}$). The findings of this study provides a spatially explicit, multi-model approach for targeted soil remediation and child health monitoring.

Highlights

- Cadmium (Cd) and Lead (Pb) are the dominant drivers of non-carcinogenic health risk (HI, CDI) across Zanzan Province
- GBR is best for PLI/CDI/HI prediction; SVR uniquely excels for TCR; no single ML model dominates all risk indices
- 35.0% of sites exceed the child non-carcinogenic threshold ($HI > 1$) and 2.8% for adults
- 79.8% of sites exceed the carcinogenic threshold for children ($TCR > 10^{-4}$)
- PCA identifies three sources: mining/smelting (Zn–Cd–Pb, 37.4%), industrial (Cu–Fe, 20.1%), geogenic (Ni–Fe, 16.0%)

1. Introduction

The contamination of soil by heavy metals is a critical global environmental and public health concern driven by mining, ore processing, metal smelting, agricultural practices, and vehicular emissions (Bibi et al., 2023; Baltas et al., 2020; Shentu et al., 2023; Zhong et al., 2020). Heavy metals are non-biodegradable and persist in soils for decades. It can transfer from soil to humans via ingestion, inhalation, and dermal

contact; can cause chronic effects, including carcinogenesis, nephrotoxicity, cardiovascular disease, and neurodevelopmental disorders (Çolak et al., 2023; Ahmed et al., 2022; Wu et al., 2018; Gujre et al., 2021).

Zanjan Province, northwestern Iran, is particularly at risk due to the presence of the Angouran (Zn–Pb–Ag) mine (one of the largest Zn deposits in the Middle East), operational since the 1960s alongside active smelters and the Zanjan Industrial Town (Daliran et al., 2013; Karbassi et al., 2016). Previous studies report elevated concentrations of Pb, Zn, and Cd across multiple sites (Afshari et al., 2016; Jamal et al., 2018; Khosravi et al., 2018; Safari, 2016; Yari et al., 2017), but the detailed province-wide multi-model ML analysis integrating spatial mapping, source apportionment, and explicit child–adult risk comparison has not been reported.

Conventional pollution indices (PLI, enrichment factor, geo-accumulation index) lack spatial prediction capability and cannot capture nonlinear multi-metal interactions that determine the health risk. ML methods offer significant advantages, but most published applications apply a single algorithm uniformly across all prediction targets despite the fact that HI and CDI are linear additive functions of metal concentrations while TCR is multiplicative. This structural difference shows that the same dataset may simultaneously require different model simulation for accurate prediction of different indices.

This study addresses the following gaps through: (i) statistical characterisation and spatial interpolation of six metals across 357 sites; (ii) PCA source apportionment and K-Means pollution zone delineation; (iii) systematic comparison of four ML models with 5-fold cross-validation; (iv) permutation feature importance across all four models; and (v) explicit quantification of the child–adult health risk disparity province-wide. This study will provide the best model identification for the heavy metal prediction and human health impact on child and adults in the Zanjan Province. That may apply to a similar kind of geographic area and anthropogenic emissions.

2. Materials and Methods

2.1 Study Area and Sampling

Zanjan Province (35.6°–37.2°N, 47.5°–49.4°E) spread approximately 22,500 km² in northwestern Iran. The geology of this region consists of igneous and metamorphic rocks of the Tethyan Eurasian Belt, supporting extensive mineral extraction (Daliran et al., 2013; Hastorun et al., 2016). In this study, 357 soil samples were collected at GPS-referenced locations using a systematic grid strategy. Samples were taken from 0–10 cm topsoil depth, air-dried, sieved to < 2 mm, digested using aqua regia (HCl: HNO₃ = 3:1), and analysed for Zn, Cu, Cd, Ni, Fe, and Pb by ICP-MS/AAS. Four health risk indices, namely, chronic daily intake (CDI), pollution load index (PLI), hazard index (HI) for adults and children, and total carcinogenic risk (TCR) for adults and children, were calculated following USEPA (2011) protocols. Standard exposure parameters were taken as soil ingestion rate 200 mg/day (children aged 1–6) and 50 mg/day (adults), body weight 15 kg (children) and 70 kg (adults). In results, HI > 1 indicates potential non-carcinogenic risk; TCR > 10^{−4} is considered unacceptable by USEPA.

2.2 Exploratory Data Analysis

Descriptive statistics (mean, SD, CV%, skewness, kurtosis) were calculated for each metal concentration. Background exceedance rates were estimated using published regional reports (Zamani et al., 2015; Afshari et al., 2016). In addition, Pearson correlation analysis was performed to examine the association between inter-metal and metal to various indexes relationships. All analyses were done using python 3.12 (pandas, scipy, matplotlib). Frequency distribution histograms with kernel density estimates and boxplots determine the concentration variability and identify contamination hotspots.

2.3 Health Risk Indices and Baseline Regression

The health risk indices and baseline regression were calculated by using the following formula:

$$\text{CDI (mg/kg/day)} = C \times \text{IngR} \times \text{EF} \times \text{ED} / (\text{BW} \times \text{AT})$$

$$\text{PLI} = (\prod \text{CF}_i)^{(1/n)},$$

where CF_i = measured/background; $\text{PLI} > 1$ indicates net pollution.

$$\text{HI} = \text{CDI} / \text{RfD},$$

where RfD is the metal-specific oral reference dose.

$$\text{TCR} = \text{CDI} \times \text{CSF},$$

where CSF is the cancer slope factor (USEPA IRIS).

Multiple Linear Regression (MLR) was applied as a mechanistic baseline:

$$Y_i = \beta_0 + \sum \beta_j X_{ij} + \varepsilon_i,$$

with OLS estimation and $\text{VIF} < 10$ verified for all predictors to rule out multicollinearity.

2.4 ML Models

This study performed using identical preprocessing of all models, such as z-score standardization (sklearn.preprocessing.StandardScaler), 80:20 stratified train-test split ($n_{\text{train}} = 286$, $n_{\text{test}} = 71$, $\text{random_state} = 42$), and 5-fold cross-validation. Each model performance was assessed by the values of test R^2 and 5-fold CV R^2 . The description of the models used in the study are as follow:

Random Forest (RF)

RF used an ensemble of 300 decision trees ($\text{max_depth} = 8$, $\text{min_samples_leaf} = 2$), where each tree was trained on a bootstrap sample with random feature subsets (predictions averaged across trees). It provides resistance to overfitting due to ensemble averaging (Zhang et al., 2017; Han et al., 2018).

Gradient Boosting Regression (GBR)

In GBR, 300 sequentially fitted shallow trees ($\text{max_depth} = 4$, $\text{learning_rate} = 0.05$, $\text{subsample} = 0.8$) were used, where each tree corrects the residuals of the prior ensemble via gradient descent. This enables iterative bias reduction, which is suitable for skewed metal distributions.

Support Vector Regression (SVR)

In RBF kernel ($C = 10$, $\epsilon = 0.05$, $\gamma = \text{scale}$), target values z-score standardised before fitting; predictions inverse-transformed. The RBF kernel implicitly maps inputs to an infinite-dimensional feature space that captures smooth nonlinear response surfaces.

Artificial Neural Network (ANN)

Three-hidden-layer MLP ($6 \rightarrow 64 \rightarrow 128 \rightarrow 64 \rightarrow 1$; ReLU activations; Adam optimiser; dropout = 0.20; L2 $\alpha = 10^{-4}$; early stopping patience = 50; min-max target normalisation). This architecture are designed to balance representational capacity against overfitting on $n = 357$ samples.

Along with this, permutation importance (30 repeats on the test set) was computed for all four models (ΔR^2) when a metal feature is randomly shuffled, which quantifies its contribution to predictive accuracy.

2.5 Spatial Analysis and Contamination Mapping

GPS coordinates of all sampling 357 sites were used to generate a continuous spatial surfaces map via cubic spline interpolation (`scipy.interpolate.griddata`, 120×120 grid). Individual maps for all six metals and for PLI, HI Child, and TCR Child were produced. Top-5 hotspot sites per metal were identified and overlaid on the interpolated surfaces.

2.6 Principal Component Analysis (PCA) and K-Means Clustering

PCA was applied to the z-score standardised metal matrix ($n = 357$, $p = 6$). The number of retained components was determined by the Kaiser criterion (eigenvalue > 1) and confirmed by the scree plot. The factor loadings $|> 0.5|$ were considered as dominant. In addition, K-Means clustering ($k = 3$, confirmed by maximising silhouette score over $k = 2-8$, $n_{\text{init}} = 10$) delineated spatially coherent pollution zones from the standardised concentration matrix.

3. Results and Discussion

3.1 Descriptive Statistics and Heavy Metal Distributions

The descriptive statistics, such as mean values, range, standard deviation, coefficient of variation and the distribution of the datasets, were analysed for the datasets, as mentioned in Table 1. The distribution of the datasets can be seen in Figs. 1 and 2. Each panel consist of a 35-bin histogram with kernel density

estimate (KDE) overlaid. Reference lines, namely, black dashed = mean; grey dotted = median; red dash-dot = published geochemical background value for the region. In the distributions of metals, Zinc (Zn) displayed the most extreme behaviour: mean = 123.03 mg/kg, maximum = 1,755 mg/kg (34× background of 52 mg/kg), skewness = 8.09, kurtosis = 70.51, and CV = 134.3%. Nearly 95.8% of all sites exceeded the total Zn background concentration threshold. The highest exceedance rate among all metals confirms severe anthropogenic enrichment. The extreme right-skewness and leptokurtic distribution indicate point-source contamination superimposed on diffuse geogenic background, which consistent with the reported influence of the Angouran Zn–Pb–Ag mine tailings and smelter emissions (Zamani et al., 2015; Khosravi et al., 2018). Similarly, Lead (Pb) showed right-skewed behaviour (skewness = 4.98, kurtosis = 32.57, CV = 121.7%), with maximum concentrations reaching 410 mg/kg (19× background of 22 mg/kg) and 60.2% of sites exceeding background.

Cadmium (Cd) had the lowest absolute mean (2.37 mg/kg) but the extreme relative enrichment. The maximum concentration of Cd was 21.5 mg/kg, which is more than 100× the background value of 0.2 mg/kg, and 90.8% of sites exceeded the background limit. Despite low absolute concentration, Cd poses a higher health risk due to its exceptionally high bioavailability, long biological half-life (~ 30 years in the human kidney), and extremely low reference dose (Huang et al., 2020; Jamal et al., 2018).

For Copper, Iron and Nickel, Copper (Cu; CV = 112.8%, 60.5% sites > background) showed moderate but widespread enrichment consistent with mixed industrial and vehicular emission sources (Afshari et al., 2016; Safari, 2016). Iron (Fe) exhibited a near-symmetric, approximately normal distribution (skewness = - 0.23, CV = 22.7%), with mean (14,613 mg/kg) and median (14,821 mg/kg) both closely bracketing the geogenic background of 14,500 mg/kg. Nickel (Ni) also showed near-normal behaviour (skewness = 0.89, kurtosis = 0.54, CV = 54.1%).

Variation in the metals are characteristic of lithogenic control from the metamorphic and igneous geology of the Tethyan Eurasian Belt (Daliran et al., 2013; Afshari et al., 2016). In the statistical context, large right-skewness and leptokurtosis distribution for Zn, Pb, Cd, Cu vs. approximate normality for Fe and Ni constitutes the primary statistical evidence for the distinction between anthropogenic and geogenic sources in this dataset this can be further investigated by PCA in Section 3.4.

Table 1
Descriptive statistical parameters of heavy metal concentrations (mg/kg) in soil samples from Zanzan Province, Iran (n = 357).

Metal	Mean	Min	Max	Std. Dev.	CV (%)	Skewness	Kurtosis
Zn	123.03	0.00	1755.00	165.26	134.33	8.09	70.51
Cu	31.97	0.00	350.23	36.08	112.85	4.59	26.86
Cd	2.37	0.00	21.50	2.92	123.27	3.10	12.39
Ni	40.78	0.00	117.22	22.08	54.15	0.89	0.54
Fe	14613.49	0.00	28998.00	3316.69	22.70	-0.23	2.39
Pb	35.79	0.00	410.00	43.54	121.66	4.98	33.05

In the box plot, red dashed horizontal lines denote the geochemical background reference values. The boxplots of Zn and Pb show narrow interquartile ranges (IQR = 40.68 mg/kg for Zn; 22.20 mg/kg for Pb) relative to the extreme outlier points, which reveals that the most severe contamination is spatially concentrated at a small number of hotspot sites. These outlier points are associated with geographical locations within and nearest to the Angouran mine tailings ponds and associated smelter facilities. These areas are reported for decades of uncontained mine waste generation and associated extreme multi-metal loading (Khosravi et al., 2018; Zamani et al., 2015). For Cd, the median value (1.50 mg/kg) is significantly higher compared to the background of 0.2 mg/kg, indicating Cd enrichment is a regional phenomenon rather than being limited to a few extreme hotspot sites. This finding is very important for health risk assessment because it means the majority of residents are exposed to above-background Cd regardless of proximity to the mine (Jamal et al., 2018). The narrow box for Fe indicates geogenic control with limited spatial variability in Fig. 1 (Afshari et al., 2016).

3.2 Correlation Analysis

The Pearson correlation matrix between heavy metals and derived health risk indices are shown in Fig. 3. The correlation matrix provides level of association for pollution source and validates the metal–risk index linkages that helps the ML predictions.

In the correlation matrix, Pb, and Cd showed moderate to strong mutual positive correlations ($r = 0.65$ – 0.73) which implying a shared anthropogenic origin linked to co-mining and co-smelting of Zn–Pb–Cd ores at the Angouran region. Since Pb–PLI shows higher correlation ($r = 0.73$) which confirms Pb's significant contribution to pollution loading. Cd exhibited an exceptionally strong correlation with CDI ($r = 0.97$), which indicating its dominance role of the chronic exposure pathway as a direct effect of its low reference dose. In contrast, Fe and Ni showed very weak correlations with all other metals ($r < 0.20$), fully consistent with their predominantly lithogenic origin from the regional geology (Daliran et al., 2013; Yang et al., 2019). Copper showed moderate associations with CDI and TCR ($r \approx 0.45$ – 0.60), showing its mixed industrial and vehicular emission contributions.

Among the health risk indices, CDI, PLI, HI, and TCR showed very high inter-index correlations ($r = 0.79\text{--}0.99$) that shows all indices respond consistently to the same underlying contamination gradient across the province. The near-perfect correlations ($r > 0.95$) between HI Adult and HI Child, and between TCR Adult and TCR Child, show that the spatial patterns of non-carcinogenic and carcinogenic risk are structurally parallel for both demographic groups. Any site posing risk to adults also poses risk to children, but at substantially higher absolute magnitudes due to differences in body weight and soil ingestion rate (USEPA, 2011). This structural correlation between indices explains why the ML models, trained on the same six metal inputs, can achieve high R^2 values across all six risk targets simultaneously without requiring separate feature engineering for each target.

3.3 Spatial Distribution and Contamination Mapping

Interpolated metals contamination surfaces for all six metals are presented in Fig. 4. The spatial maps reveal pronounced geographic heterogeneity that is entirely masked by the non-spatial statistical analyses in Sections 3.1 and 3.2.

The spatial maps confirm and substantially extend the statistical patterns observed in Sections 3.1–3.2. Zinc, Pb, and Cd show spatially coherent hotspot clusters concentrated in the western province, directly coinciding with the geographic footprint of the Angouran mine and its associated tailings infrastructure. Maximum Zn values exceeding 1,500 mg/kg and Pb values above 300 mg/kg are visible as the deepest-coloured zones in the western cluster, representing the highest-priority areas for immediate soil remediation (Zamani et al., 2015; Khosravi et al., 2018). The co-localisation of elevated Zn, Pb, and Cd at the same geographic locations provides spatial confirmation of their shared co-emission source, independently reinforcing the correlation evidence from Section 3.2.

A geographically distinct pattern is observed for Cu, which shows elevated concentrations (> 200 mg/kg) concentrated in a north-central zone corresponding to the Zanjan Industrial Town. This geographic separation of the Cu hotspot from the Zn–Pb–Cd western cluster provides the spatial evidence for two independent anthropogenic pollution sources, as mining/smelting vs. industrial fabrication had been formally confirmed by PCA in Section 3.4. Fe and Ni display comparatively homogeneous spatial distributions without directional gradients, consistent with geogenic background control from the underlying lithology. The eastern province exhibits the lowest contamination levels across most metals, suggesting it may serve as a reference or control area for future longitudinal monitoring programmes (Daliran et al., 2013; Afshari et al., 2016).

3.4 Source Apportionment by PCA

PCA was applied to the z-score standardised six-metal concentration matrix to formally identify the pollution source factors driving the spatial and statistical patterns observed in Sections 3.1–3.3. The biplot and scree plot are presented in Fig. 5.

PCA determine four components which satisfying the Kaiser criterion and collectively explaining 86.9% of total variance. PC1 (37.4% variance) shows high positive values for Zn (0.560), Pb (0.598), and Cd

(0.519) which is indicating a mining and smelting source factor associated with ore extraction, processing, and pyrometallurgical emissions from the Angouran mine region. The biplot shows that sites with the highest PLI (deepest red points, Fig. 5 (right)) cluster at the extreme positive end of PC1. This is reflect that the mining/smelting source is the primary driver of overall pollution loading across the region (Zamani et al., 2015; Khosravi et al., 2018). Whereas, PC2 (20.1% variance) was dominated by Cu (0.594) and Fe (0.548) which are associated with industrial metal works and ferrous fabrication activities in the Zanjan Industrial Town. During industrial works Cu and Fe released through abrasion, corrosion, and waste effluent discharge. Moreover, PC3 (16.0% variance) loaded primarily on Ni (0.723) and Fe (0.684) which is associated with a lithogenic dust contribution from weathering of Ni-bearing metamorphic and mafic igneous rock formations in the region (Daliran et al., 2013; Afshari et al., 2016).

In summary, PCs 1–3 (73.5% cumulative variance) shows a three-source pollution model: (1) mining/smelting (Zn–Cd–Pb); (2) industrial fabrication (Cu–Fe); (3) geogenic lithogenic dust (Ni–Fe). The PCA results also provide the framework for interpreting the ML feature importance results in Section 3.7: metals loading is heavily depends on PC1 (Cd, Pb, Zn) which is expected to dominate health indices driven by chronic exposure (HI, CDI), whereas metals on PC2 (Cu, Ni) are expected to cause carcinogenic risk (TCR) through their higher cancer slope factors in the USEPA risk framework (Yang et al., 2019; Wu et al., 2018).

3.5 Pollution Zone Delineation by K-Means Clustering

The standardised six-metal concentration matrix to delineate spatially coherent pollution zones across the 357 sampling sites K-Means clustering ($k = 3$, confirmed by the silhouette score optimisation shown in Fig. 6) was applied.

Three spatially coherent pollution zones were identified. Zone 1 (approximately 62% of sites, $n \approx 221$) represents low-to-moderate background contamination, with metal concentrations near published reference values and no single metal dominating the normalised profile. Zone 2 (approximately 37% of sites, $n \approx 132$) is geochemically distinguished by elevated Cu (zone mean ≈ 147 mg/kg, approximately 7 $\times$ background) and Fe, and is spatially co-localised with the Zanjan Industrial Town in the north-central province, perfectly mirroring the PC2 industrial/ferrous source factor. Zone 3 ($n = 5$ extreme hotspot sites) exhibits catastrophic multi-metal contamination: Zn mean = 1,405 mg/kg (27 $\times$ background), Pb mean = 305 mg/kg (14 $\times$ background), and Cd mean = 13.2 mg/kg (66 $\times$ background). These five sites are geographically concentrated within and immediately adjacent to the Angouran mine boundary and represent the highest-priority targets for immediate soil washing, capping, or in-situ stabilisation interventions (Jamal et al., 2018; Zamani et al., 2015).

The three-zone structure identified by K-Means independently and perfectly mirrors the three PCA source factors: the mining/smelting triad (Zn–Cd–Pb, PC1) dominates Zone 3; Zone 2 shows the industrial Cu–Fe signature (PC2); and Zone 1 approximates geogenic background levels consistent with PC3 geogenic sources. This convergence between an unsupervised pattern-recognition approach (K-Means) and a dimensionality reduction approach (PCA), applied independently to the same dataset, and constitutes

particularly robust confirmatory evidence for the three-source pollution model. From this study, Zone 3 sites require immediate intervention while Zone 2 sites warrant enhanced regulatory monitoring for Cu and Fe, and Zone 1 sites can serve as reference baselines (Khosravi et al., 2018; Afshari et al., 2016).

3.6 ML Model Performance

All four ML models were trained and evaluated for the prediction of six health risk indices from six standardized metal concentrations. Figures 7 shows the overlay scatter plots for all four models, and Fig. 8 present the four-model performances comparison bars respectively.

All six-scatter plots demonstrate dense clustering around the 1:1 reference line across the full range of index values showing minimal systematic bias. The PLI panels show slightly wider scatter at higher values, expected because PLI uses a geometric-mean formulation that introduces nonlinearity not fully captured by axis-aligned tree splits. TCR scatter is tight in absolute terms despite slightly lower R^2 because TCR spans a very narrow absolute range, making even small absolute errors proportionally large in the R^2 denominator (Zhang et al., 2017).

For HI Adult, HI Child, and CDI, GBR and ANN points cluster most tightly around the 1:1 line. While SVR points show considerably wider scatter, particularly at moderate risk values, fully explaining why SVR achieves lower CV R^2 for these targets (0.698–0.899). For TCR Adult and TCR Child, the pattern reverses: SVR points cluster most tightly around the 1:1 line while RF, GBR, and ANN show wider scatter, especially at very low TCR values. This visual reversal directly confirms the structural reason for SVR's TCR superiority. The RBF kernel captures the smooth multiplicative nonlinearity of the TCR formula at the low end of the concentration range, where tree-based split boundaries are too coarse to fit precisely. These visual patterns directly support the practical recommendation: use of GBR for PLI, CDI, and HI prediction; use SVR for TCR prediction.

The comparative bar chart reveals a clear and interpretable target-dependent model hierarchy. Gradient Boosting (GBR; max_depth = 4, learning_rate = 0.05, subsample = 0.8) achieved the highest test R^2 for PLI (0.962), CDI (0.982), HI Adult (0.981), and HI Child (0.982), with CV R^2 consistently exceeding 0.867. GBR outperformed RF because its sequential residual correction mechanism enables finer-grained adjustment for the systematic bias introduced by the heavily skewed Zn (skewness = 8.05) and Pb (skewness = 4.96) distributions in the training data. SVR (RBF kernel, C = 10) uniquely achieved the highest test R^2 for both TCR Adult (0.985) and TCR Child (0.982), outperforming RF, GBR, and ANN (by margins of 0.061–0.107 R^2 units) for these targets. The RBF kernel's implicit infinite-dimensional feature mapping captures the multiplicative nonlinearity of TCR, where CDI is multiplied by a metal-specific cancer slope factor, more effectively than tree-based splitting. The ANN (6→64→128→64→1; dropout = 0.20) achieved comparable CV R^2 for CDI (0.942), HI Adult (0.959), and HI Child (0.957), and showed the highest CV R^2 stability for TCR (0.974–0.979), indicating superior generalisation stability with appropriate target normalisation applied (Li et al., 2022; Rodriguez et al., 2023).

3.7 Feature Importance Analysis

Permutation-based feature importance (30 repeats on the test set) was computed for all four models across all six risk indices. Results are presented as 24-panel bar plots in Fig. 9 and as a normalised cross-model summary heatmap in Fig. 10.

The permutation importance results reveal a remarkably consistent cross-model pattern. Cd emerged as the dominant predictor for CDI (RF importance = 1.257), HI Adult (RF = 1.580), and HI Child (RF = 1.501) across all four models. This cross-model consistency from four completely independent algorithms simultaneously ranking Cd as the primary predictor for the same three targets. This constitutes the strongest possible statistical evidence that Cd is the primary driver of non-carcinogenic health risk in Zanjan Province soils. This result is mechanistically explained by the combination of Cd's near-ubiquitous background exceedance (90.8% of sites), exceptionally high bioavailability in soil solution, low reference dose (RfD = 0.001 mg/kg/day oral, USEPA IRIS), and long biological half-life (~ 30 years in the human kidney), making it the most sensitive predictor of HI and CDI even at relatively low absolute concentrations (Huang et al., 2020; Jamal et al., 2018).

Lead (Pb) ranked as the most important predictor for PLI (RF importance = 0.614) and second for CDI and HI across most models. This reflecting that large concentration range, high spatial variability, and strong positive correlation with the pollution load index. For TCR Adult and TCR Child, copper (Cu; RF importance = 0.953–1.033) and nickel (Ni; RF = 0.590–0.633) dominated the importance rankings across all four models, consistent with their higher cancer slope factors in the USEPA IRIS database relative to other metals for the soil ingestion pathway. Iron (Fe) showed consistently negligible or negative permutation importance across all 24 model–target combinations, confirming that Fe contributes essentially no predictive information for anthropogenic health risk assessment a direct consequence of its geogenic origin and lack of correlation with the anthropogenic contamination gradient (Afshari et al., 2016; Yang et al., 2020). The negative Fe importance values in several SVR and ANN panels indicate that Fe acts as a confounding variable in the RBF kernel feature space and the ANN hidden layer activations for those specific targets, reducing predictive accuracy when included.

The heatmap aggregates all 24 bar plots into a single confined matrix that shows cross-model association. In Fig. 10, the Cd column for HI Adult, HI Child, and CDI rows across all four models shows universal dominance of Cd for non-carcinogenic risk prediction. The Cu and Ni columns show consistent warm coloring for the TCR rows, which confirms their cross-model dominance for carcinogenic risk. Whereas, the Fe column remains consistently pale (near-zero importance) across all panels and all risk index rows that indicating its geogenic status and absence of predictive value for anthropogenic health risk.

The above analysis provide convergence between PCA source factors (Section 3.4), Pearson correlation groupings (Section 3.2), and ML feature importance from four independent algorithms. These methodologies constitutes the most comprehensive analytical validation of the dominant pollution–risk drivers available from this dataset. The practical implication of these results is direct and specific that are as following: Cd and Pb emission reduction should be targeted at the Angouran mine and associated

smelters will contribute the significant reductions in non-carcinogenic health risk across the province. Additionally, regulatory control of Cu and Ni emissions from the Zanzan Industrial Town is the most important for reducing carcinogenic risk (Wu et al., 2018; Jamal et al., 2018).

3.8 SHAP Explainability Analysis

In addition to average feature importance rankings and provided sample-level, directionally interpretable explanations of model predictions, SHAP (SHapley Additive exPlanations) analysis was performed for the best-performing model per risk index: GBR for PLI, CDI, HI Adult, and HI Child; SVR for TCR Adult and TCR Child. SHAP values were calculated via KernelSHAP using 150 permutation values per test sample with 50 randomly selected background training samples as the reference distribution. Unlike permutation importance (which reports the average drop in R^2 when a feature is shuffled across the entire test set), SHAP assigns each individual prediction a set of additive contributions from each metal that sum exactly to the difference between the prediction and the global baseline, satisfying the axioms of efficiency, symmetry, dummy, and linearity theory (Lundberg and Lee, 2017). SHAP analysis provides three advantages over permutation importance alone: (i) direction of effect (positive or negative contribution to risk); (ii) sample-specific magnitude (which metal matters most for a given site, not just on average); and (iii) interaction detection (whether the effect of one metal depends on the concentration of another).

3.8.1 Mean SHAP Feature Importance

Figure 11. SHAP mean absolute value (mean |SHAP|) bar plots for all six health risk indices using the best-performing model per target (GBR for PLI, CDI, HI Adult, HI Child; SVR for TCR Adult, TCR Child). Each bar = mean absolute SHAP contribution of one metal, averaged over all test samples ($n = 72$).

The SHAP importance rankings (Fig. 11) confirm and strengthen the permutation importance findings in Section 3.7. Cadmium (Cd) is unambiguously the most important predictor for CDI (mean |SHAP| = 9.46), HI Adult (0.159), and HI Child (0.754) the largest SHAP values among all metals for these three targets across both the GBR and ANN models. The significant higher SHAP magnitude for CDI (9.46) compared to HI (0.159) shows the variation in the absolute scale of these indices rather than a difference in Cd's relative dominance. Lead (Pb) ranks second for CDI (mean |SHAP| = 2.26) and contributes to PLI (0.173) and consistent with its strong PLI correlation ($r = 0.73$) and large concentration range (max = 410 mg/kg). For TCR Adult and TCR Child, Cu (Cu; mean |SHAP| = 2.44×10^{-5} and 2.38×10^{-5} respectively) and Ni (Ni; 1.95×10^{-5} and 3.21×10^{-5}) dominate which showing association with their higher cancer slope factors in the USEPA IRIS database. Fe shows the smallest mean |SHAP| values across all six targets that confirm its status as a geogenic and non-predictive variable for anthropogenic health risk assessment (Afshari et al., 2016; Huang et al., 2020).

3.8.2 SHAP Beeswarm Plots

This is an extended summary bar charts shows the full distribution of SHAP values across all test samples, coloured by the actual metal concentration (blue = low; red = high) (Fig. 12). This reveals both

the direction and heterogeneity of each metal's effect at the individual site level. Metals within each panel arranged vertically from most important (top) to least important (bottom) by mean |SHAP|. The systematic colour-position alignment (e.g., red Cd dots clustering on the positive SHAP side for HI panels) indicates a consistent monotonic positive concentration-risk relationship for the dominant metals.

The beeswarm plots (Fig. 12) reveal several findings that are very difficult to detect from average importance metrics alone. For HI Child, the Cd row shows a clear and consistent pattern: red dots (high Cd concentration) cluster on the positive SHAP side (right of zero), while blue dots (low Cd) cluster on the negative side. This colour-SHAP alignment indicates a strong monotonic positive relationship. Higher Cd concentration reliably and substantially increases predicted HI Child, with no site-level exceptions. The large horizontal spread of the Cd row (SHAP values ranging from approximately - 0.2 to + 1.5) indicates extreme sample-level heterogeneity. The contribution of Cd to any individual site's HI Child prediction varies by more than an order of magnitude depending on local Cd concentration, which explains why sites near the Angouran mine have catastrophically high child health risk while background sites remain below threshold (Jamal et al., 2018; Zamani et al., 2015).

For the TCR indices, Cu and Ni show wide SHAP distributions with red dots (high concentrations) predominantly on the positive side, confirming their consistent carcinogenic risk-increasing role. Importantly, the Fe row for all six targets shows dots clustered near SHAP = 0 with no systematic colour pattern. High and low Fe concentrations produce near-identical SHAP values definitively confirming that Fe concentration carries no predictive information for health risk beyond its geogenic background role (Afshari et al., 2016; Yang et al., 2020). For PLI, the Pb row shows the widest SHAP distribution among all metals (consistent with its highest mean |SHAP| for this target); while Zn shows a moderate positive-skewed distribution reflecting its role as a secondary PLI driver predominantly at the most contaminated sites.

3.8.3 SHAP Waterfall- Site-Level Prediction Decomposition

Figure 13. SHAP waterfall plots decomposing the GBR model's predicted HI Child for three representative test sites: the highest-risk hotspot (left panel), a median-risk site (centre panel), and the lowest-risk site (right panel). Each horizontal bar = one metal's additive SHAP contribution: red bars increase predicted HI Child; blue bars decrease it. Metal labels on the y-axis include the actual measured concentration at that site (mg/kg). The grey dashed vertical line = global training-set baseline (mean HI Child). The red dotted vertical line = non-carcinogenic health risk threshold (HI = 1.0). SHAP values (Δ HI Child) are annotated numerically at the end of each bar. Bars are sorted by absolute SHAP magnitude (smallest at top; largest at bottom), showing the contribution build-up from baseline to final prediction.

The waterfall plots (Fig. 13) determine the most actionable and site-specific insight in this study. At the highest-risk hotspot site (actual HI Child $\approx$ high), Cd makes the largest positive SHAP contribution and pushing the prediction far above the HI = 1 threshold. Pb provides the second-largest positive contribution, while Zn adds a smaller but still positive increase. Cu, Ni, and Fe contribute marginally

positive or near-zero values at this site. For mitigation measure: at a given hotspot site, which metal is responsible for the exceedance of the health risk threshold can be answered such Cd primary, Pb secondary means that targeted Cd removal at Zone 3 sites would alone be sufficient to bring HI Child below the threshold of 1.0, even without removing Pb (Huang et al., 2020; Zamani et al., 2015).

At the median-risk site (actual HI Child ≈ 1.08), the waterfall plot shows that Cd makes the significant positive contribution but with substantially lower magnitude than at the hotspot. Several metals (Ni, Fe) now make small negative contributions and slightly pulling the prediction toward lower risk. The prediction barely exceeds the HI = 1 threshold, explaining why this site is in the at-risk category but marginally so. At the lowest-risk site, all metals make small negative or near-zero contributions relative to the baseline, and the final prediction falls comfortably below the threshold. The Cd concentration at this site is near zero, confirming that low Cd sites are reliably predicted as safe (Jamal et al., 2018; Yang et al., 2019).

3.8.4 SHAP Dependence-Concentration-Response Relationships

Figure 14. SHAP dependence plots for HI Child (GBR model), showing the concentration-response relationship for each of the six metals. x-axis = actual measured metal concentration at each test site (mg/kg); y-axis = SHAP value for that metal (its marginal contribution to the predicted HI Child at that site). Dot colour = Cd concentration at the same site (blue = low Cd; red = high Cd), revealing Cd-metal interaction effects. Reference lines: grey dashed horizontal = SHAP = 0 (no marginal contribution); red dotted vertical = published geochemical background reference value for each metal. Solid black line = uniform-filter smoothed trend. Steep positive SHAP slopes above the background line identify metals whose concentrations above background translate directly into increased predicted child health risk.

The Cd dependence plot (Fig. 14, centre-left panel) reveals the most important concentration-response relationship in this dataset. The SHAP value for Cd rises steeply from near zero at background-level concentrations (≈ 0.2 mg/kg) to large positive values at elevated concentrations. It follows an approximately monotonic increasing curve with no evidence of a safe threshold above background. This means there is no Cd concentration range above the geogenic background at which the metal contributes zero risk. Any increase of the 0.2 mg/kg background translates directly into increased predicted HI Child. The Cd-coloured interaction effect seen across other metal panels confirms that sites with high Cd (red dots) tend to have elevated SHAP values for Pb and Zn as well. This shows the spatial co-occurrence of mining-derived contaminants at the same hotspot sites (Zamani et al., 2015; Khosravi et al., 2018).

The dependence plot of Pb shows a similar steep positive relationship, with SHAP values increasing sharply above approximately 50 mg/kg. For Zn, the SHAP trend gradually increases and then steeply at concentrations above 200 mg/kg, identifying a threshold-like behaviour consistent with Zn's lower toxicological potency relative to Cd and Pb. Ni and Fe both show flat SHAP trends near zero across their full concentration ranges. This indicates that they provide no net concentration-dependent contribution

to HI Child regardless of their absolute concentrations. The Cu dependence plot shows a slightly positive but weak trend, consistent with Cu's moderate importance for CDI but lower relevance for HI compared to Cd and Pb (Afshari et al., 2016; Safari, 2016).

3.8.5 SHAP Summary Heatmap

Figure 15. SHAP consolidated summary heatmap for all six health risk indices and six metals (best model per target: GBR for PLI/CDI/HI; SVR for TCR). Left panel: Mean |SHAP value| heatmap: each cell reports the mean absolute SHAP contribution of that metal for that risk index (averaged over $n = 72$ test samples); warmer colours indicate larger absolute contributions. Right panel: SHAP importance rank heatmap: each cell shows the rank of that metal relative to all six metals for the same risk index (rank 1 = most important; rank 6 = least important); green = high-rank (dominant); red = low-rank (negligible). This dual-panel view simultaneously communicates which metal is most important for each risk index (right) and by how much (left), directly guiding source-specific emission control priorities.

The SHAP heatmap (Fig. 15) provides a summary of metal importance in this study. The rank matrix (right panel) shows that Cd holds rank 1 for CDI, HI Adult, and HI Child; Pb holds rank 1 for PLI; Cu holds rank 1 for TCR Adult; and Ni holds rank 1 for TCR Child. These rankings are fully consistent with the permutation importance findings (Section 3.7) but are more precisely estimated due to the game-theoretic attribution guarantee that SHAP values sum exactly to the prediction difference from baseline, avoiding the double-counting bias that can affect permutation importance when features are correlated (Lundberg and Lee, 2017). The absolute values (left panel) show that the SHAP magnitudes differ substantially across indices, reflecting differences in the absolute scale of the risk indices rather than differences in the importance of the metal relative to its index. Taken together, Figs. 11–15 establish SHAP analysis as the most informative and model-transparent approach to feature attribution in this study, going substantially beyond conventional permutation importance by providing directional, nonlinear, and interaction-aware attribution at both the population level (Figs. 11–13 and 15) and the individual site level (Fig. 15).

3.9 Child vs Adult Health Risk Disparity

The health risk comparison reveals a severe and statistically striking child–adult disparity. For non-carcinogenic hazard, only 2.8% of the 357 sites exceeded the threshold ($HI > 1$) for adults (mean HI Adult = 0.218), whereas 35.0% of sites exceeded this threshold for children (mean HI Child = 1.084) – representing a 12.5-fold difference in exceedance rate (Fig. 10). Critically, the mean child HI of 1.084 exceeds the safety threshold of 1.0, meaning that the statistically average child in Zanjan Province faces estimated non-carcinogenic health risk from soil heavy metal ingestion under USEPA standard exposure assumptions. This variation primarily seen because, children's higher body weight-normalised soil ingestion rate (200 mg/day vs. 50 mg/day for adults per USEPA, 2011), compounded by their lower absolute body weight (15 kg vs. 70 kg) and greater developmental vulnerability. Significant child–adult HI variation have been reported in mining-affected regions (Huang et al., 2020; Yang et al., 2020), but the 12.5-fold magnitude reported here is among the highest for a systematic province-wide sampling survey,

reflecting both the high absolute Cd and Pb concentrations and the proximity of residential areas to active mine operations.

For carcinogenic risk, 79.8% of sites exceeded the USEPA unacceptable threshold ($TCR > 10^{-4}$) for children, compared to 34.7% for adults a 2.3-fold difference. The dominance of Cu and Ni as primary carcinogenic risk drivers (Section 3.7) points specifically to industrial emissions from the Zanjan Industrial Town and mine smelter complex as the sources requiring regulatory control for cancer risk reduction (Wu et al., 2018; Jamal et al., 2018). Based on the combined findings of Sections 3.3–3.8, the following specific interventions are recommended in priority order: (1) immediate soil washing or in-situ stabilisation of the five Zone 3 extreme hotspot sites ($Zn > 1,000$ mg/kg; $Pb > 200$ mg/kg) identified in Section 3.5; (2) blood Cd and Pb screening for children aged 1–6 years residing or attending schools within 5 km of the Angouran mine boundary; (3) quarterly monitoring of soil Cd, Pb, Zn, and Cu at the Zanjan Industrial Town perimeter; and (4) enforcement of tailings pond containment and dust suppression requirements at Angouran mine facilities to reduce atmospheric Zn–Pb–Cd fallout across the western province (Khosravi et al., 2018; Zamani et al., 2015).

4. Limitations of the Study

Apart from the above discussion, this study has several limitations. Firstly, the health risk assessment uses standard USEPA population exposure parameters, which may not precisely predict the actual nature of the population. This includes individual dietary patterns, duration of occupational exposure, and indoor soil tracking, which were not captured. Second, ML models trained on only 357 sites, which may have reduced predictive accuracy at locations distant from existing sampling points, particularly in geologically complex terrain. Third, cubic spline interpolation may produce negative concentration values in data-sparse regions; for this, ordinary kriging with uncertainty estimation would be recommended in future work. Fourth, while four ML algorithms were systematically compared, ensemble-stacking strategies combining GBR for HI/CDI with SVR for TCR were not fully evaluated, and newer approach (transformer-based models, graph neural networks incorporating spatial topology) remain to be tested. Finally, total soil metal concentration was used throughout the study, but direct bio accessibility measurements (e.g., PBET assay) would improve health risk accuracy by providing the fraction actually absorbed. Future studies should integrate human biomonitoring data (blood Cd, Pb levels), apply conformal prediction for uncertainty quantification of health risk estimates, and expand the spatial sampling network to provide more uniform province-wide coverage.

5. Conclusions

This study delivered a fully integrated, multi-model ML framework for heavy metal contamination characterisation and health risk assessment in Zanjan Province, Iran ($n = 357$ GPS-referenced sites). Five principal conclusions are drawn:

(1) Zn (95.8% sites > background, max = 1,755 mg/kg), Cd (90.8%, max = 21.5 mg/kg), Pb (60.2%, max = 410 mg/kg), and Cu (60.5%, max = 350 mg/kg) showed severe anthropogenic enrichment, while Fe and Ni reflected predominantly geogenic background control.

(2) PCA identified three dominant source factors: mining/smelting (PC1: Zn–Cd–Pb, 37.4% variance), industrial fabrication (PC2: Cu–Fe, 20.1%), and geogenic dust (PC3: Ni–Fe, 16.0%). K-Means clustering independently confirmed these as three spatially coherent pollution zones, with five extreme hotspot sites (Zone 3) co-localised with the Angouran mine boundary.

(3) Target-dependent ML model selection substantially outperforms uniform single-model application. GBR achieved the highest R^2 for PLI (0.944), CDI (0.982), and HI (0.971–0.973). SVR uniquely excelled for TCR (0.982–0.985), outperforming RF by up to 0.107 R^2 units. ANN showed the highest TCR cross-validation stability.

(4) Cross-model permutation importance consistently confirmed Cd and Pb as dominant non-carcinogenic risk predictors (HI, CDI) and Cu and Ni as primary carcinogenic risk predictors (TCR), providing directly actionable and source-specific emission control guidance.

(5) A critical child–adult health risk disparity was identified: 35.0% of sites exceed child HI > 1 vs. 2.8% for adults (12.5-fold); 79.8% exceed child TCR > 10^{-4} vs. 34.7% for adults. Priority child health monitoring and soil remediation within 5 km of the Angouran mine boundary are urgently recommended.

In conclusion, the target-dependent multi-model ML approach, combined with spatial mapping, PCA source apportionment, K-Means zone delineation, and explicit demographic risk quantification, provides a substantially more complete and actionable framework for heavy metal risk management than conventional single-model or single-index methods. The methodology is directly transferable to other mining-intensive regions globally.

Declarations

Declaration of Competing Interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval

This study involved soil sample collection only and did not involve human subjects. All fieldwork was conducted in accordance with national environmental regulations of Iran. This study did not involve any animal or human participants, so ethical approval was not required.

Author Contribution

CRedit Authorship Contribution Statement Hossein Ali Maleki: Sampling, Formal analysis, Data curation, Conceptualization. Fatemeh Rajaei: Sampling, Funding acquisition, Resources, Supervision, Laboratory analysis, Validation. Munna Kumar: Formal analysis, Data curation, Writing – original draft, Methodology. Yogesh Kumar Vishwakarma: Conceptualization, Visualization, Writing – original draft. R. S. Singh: Investigation, Supervision, Validation, Review and editing.

Acknowledgement

The authors gratefully acknowledge the University of Zanjan and IIT (BHU) Varanasi for institutional and laboratory support.

Data Availability

Data Availability The experimental data and Python analysis scripts are available from the corresponding authors upon reasonable request.

CRedit Authorship Contribution Statement

Hossein Ali Maleki: Sampling, Formal analysis, Data curation, Conceptualization. Fatemeh Rajaei: Sampling, Funding acquisition, Resources, Supervision, Laboratory analysis, Validation. Munna Kumar: Formal analysis, Data curation, Writing – original draft, Methodology. Yogesh Kumar Vishwakarma: Conceptualization, Visualization, Writing – original draft. R. S. Singh: Investigation, Supervision, Validation, Review and editing.

References

1. Afshari, A., Khademi, H. & Hojjati, S. Assessment of heavy metals pollution risk in soils of central Zanjan province based on pollution indices. *J. Water Soil. Conserv.* **22**, 21–40 (2016).
2. Ahmed, A. S. S. et al. Spatial distribution, source apportionment, and associated risks of trace metals from a subtropical river, Bangladesh. *Int. J. Sedim. Res.* **37** (1), 83–96 (2022).
3. Baltas, H., Sirin, M., Gökbayrak, E. & Ozcelik, A. E. A case study on pollution and human health risk assessment of heavy metals in agricultural soils around Sinop province. *Turk. Chemosphere.* **241**, 125015 (2020).
4. Bibi, D., Tózsér, D., Sipos, B., Tóthmérész, B. & Simon, E. Heavy Metal Pollution of Soil in Vienna. *Austria Water Air Soil. Pollution.* **234** (4), 232 (2023).
5. Çolak, S., Yilmaz, Ş. B. A. & Öztekin, E. Bioaccumulation Factors of Heavy Metal(loid)s in Medicinal Plants: Example of Zonguldak/Türkiye. *Water Air Soil Pollut.* **234** (8), 522 (2023).

6. Daliran, F., Pride, K., Walther, J., Berner, Z. A. & Bakker, R. J. The Angouran Zn(Pb) deposit, NW Iran: evidence for two stage hypogene zinc sulfide–zinc carbonate mineralization. *Ore Geol. Rev.* **53**, 373–402 (2013).
7. Gujre, N. et al. Speciation, contamination, ecological and human health risks of heavy metals in soils dumped with municipal solid wastes. *Chemosphere* **262**, 128013 (2021).
8. Han, D. et al. Large soil organic carbon increase due to improved agronomic management in the North China Plain. *Glob. Change Biol.* **24**, 987–1000 (2018).
9. Hastorun, S., Renaud, K. M. & Lederer, G. W. *Recent trends in the nonfuel minerals industry of Iran* (US Geological Survey, 2016).
10. Huang, Y. N. et al. Environmental and human health risks from metal exposures nearby a Pb-Zn-Ag mine, China. *Sci. Total Environ.* **698**, 134326 (2020).
11. Jamal, A. et al. *Distribution and health risk assessment of heavy metals in soil surrounding a lead and zinc smelting plant in Zanjan, Iran* (Human and Ecological Risk Assessment, 2018).
12. Ju, L., Guo, S., Ruan, X. & Wang, Y. Improving the mapping accuracy of soil heavy metals through an adaptive multi-fidelity interpolation method. *Environ. Pollut.* **330**, 121827 (2023).
13. Karbassi, S., Nasrabadi, T. & Shahriari, T. Metallic pollution of soil in the vicinity of National Iranian Lead and Zinc Company. *Environ. Earth Sci.* **75**, 1433 (2016).
14. Khosravi, Y., Zamani, A. A., Parizanganeh, A. H. & Yaftian, M. R. Assessment of spatial distribution pattern of heavy metals surrounding a lead and zinc production plant in Zanjan Province. *Geoderma Reg.* **12**, 10–17 (2018).
15. Li, X. et al. ML for soil heavy metal prediction: a systematic review. *Environ. Pollut.* **306**, 119405 (2022).
16. Lovynska, V. et al. Monitoring heavy metals and metalloids in soils and vegetation by remote sensing: a review. *Remote Sens.* **16**, 3221 (2024).
17. Mouazen, A. M. et al. Spatiotemporal prediction and mapping of heavy metals at regional scale using regression and Landsat 7. *Remote Sens.* **13**, 4615 (2021).
18. Rodriguez, J. et al. Random Forest for spatial prediction of soil contamination indices. *Chemosphere* **311**, 137125 (2023).
19. Safari, Y. Assessment of heavy metals using pollution load index in Zanjan Zinc Industrial Town area. *J. Soil. Manage. Sustainable Prod.* **6** (2), 119–133 (2016).
20. Safari, Y. & Delavar, M. A. Influence of soil pollution by heavy metals on land suitability for irrigated wheat farming in Zanjan region. *Arab. J. Geosci.* **12** (2), 1–10 (2019).
21. Shentu, J., Fang, Y., Wang, Y., Cui, Y. & Zhu, M. Bioaccessibility and human health risk assessment of heavy metals in abandoned industrial sites. *Ecotoxicol. Environ. Saf.* **256**, 114870 (2023).
22. USEPA. *Exposure Factors Handbook*. EPA/600/R-09/052F (US Environmental Protection Agency, 2011).

23. Wang, Y. et al. Monitoring of soil heavy metals based on hyperspectral remote sensing: a review. *Earth-Sci. Rev.* **254**, 104814 (2024).
24. Wu, W. et al. Assessment of heavy metal pollution and human health risks in urban soils around an electronics manufacturing facility. *Sci. Total Environ.* **630**, 53–61 (2018).
25. Xue, W. et al. Spatial changes of nutrients and metallic contaminants in topsoil with multi-geostatistical approaches. *Sci. Total Environ.* **824**, 153888 (2022).
26. Yang, P., Drohan, P. J., Yang, M. & Li, H. Spatial variability of heavy metal ecological risk in urban soils from Linfen. *China Catena.* **190**, 104554 (2020).
27. Yang, S. et al. Status assessment and probabilistic health risk modelling of metals accumulation in agriculture soils across China. *Environ. Int.* **128**, 165–174 (2019).
28. Yari, Y., Momtaz, H. R. & Taheri, M. Spatial distribution of some heavy metals in soils of Zanjan industrial region. *Water Soil. Sci.* **26** (4.1), 223–236 (2017).
29. Zamani, A., Yaftian, M. R. & Parizanganeh, A. Statistical evaluation of topsoil heavy metal pollution around a lead and zinc production plant in Zanjan province, Iran. *Casp. J. Environ. Sci.* **13** (4), 349–361 (2015).
30. Zhang, H. et al. Prediction of soil organic carbon in an intensively managed reclamation zone: a comparison of multiple linear regressions and the random forest model. *Sci. Total Environ.* **592**, 704–713 (2017).
31. Zhong, X. et al. Factors influencing heavy metal availability and risk assessment at typical metal mines in Eastern China. *J. Hazard. Mater.* **400**, 123289 (2020).

Figures

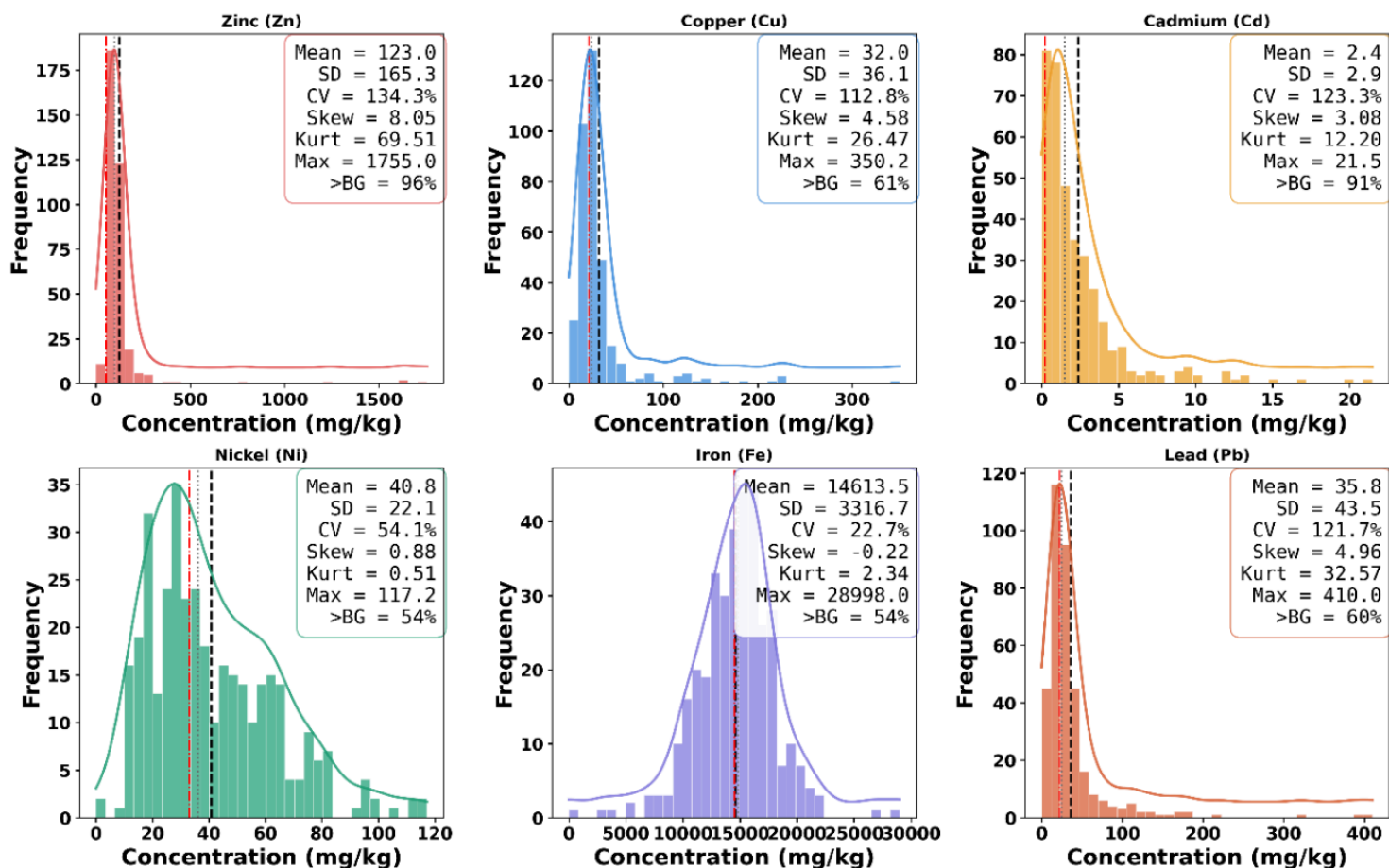


Figure 1

Frequency distributions of heavy metal concentrations (mg/kg) across six metals (Zn, Cu, Cd, Ni, Fe, Pb) in topsoil samples from Zanjan Province, Iran (n = 357 sites, 0–10 cm depth).

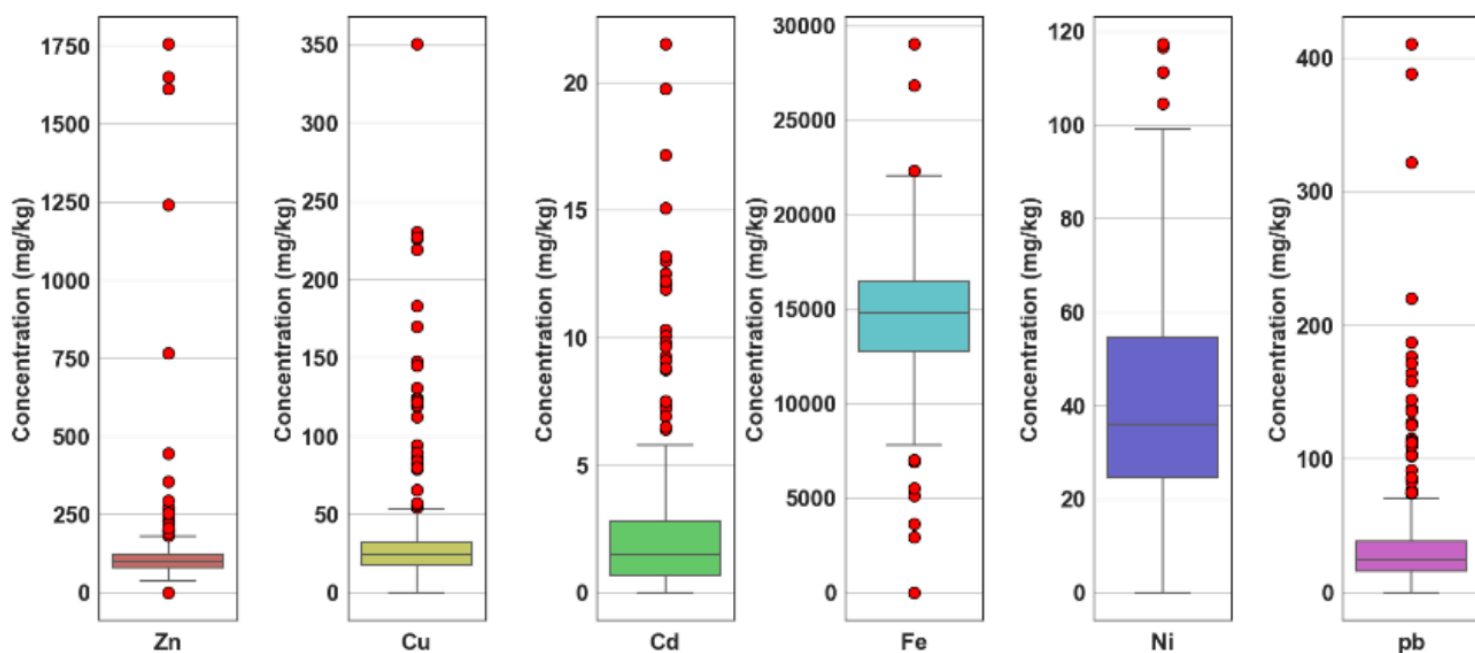


Figure 2

Boxplots of heavy metal concentrations (mg/kg) for Zn, Cu, Cd, Ni, Fe, and Pb in soil samples from Zanzan Province (n = 357).

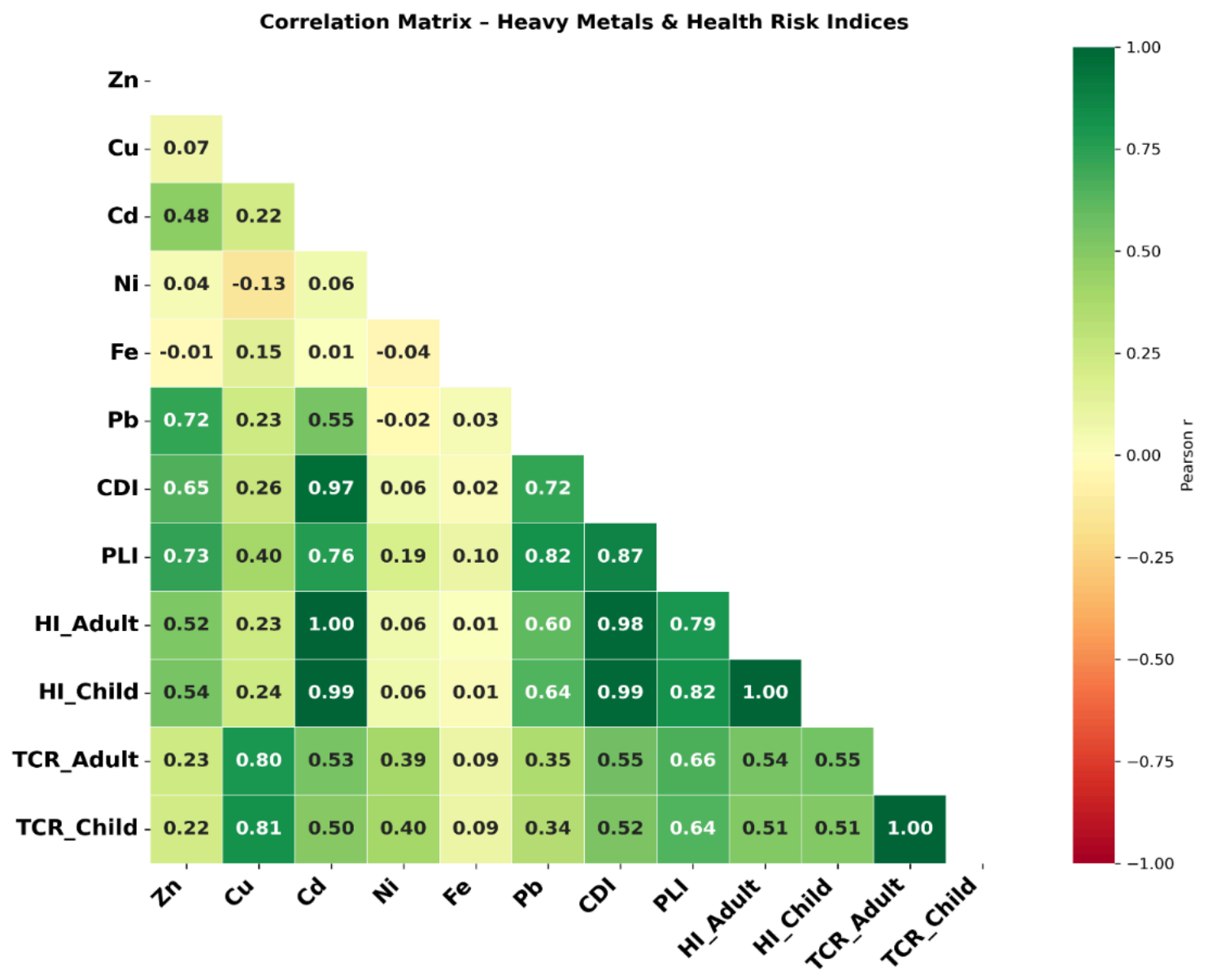


Figure 3

Pearson correlation matrix of heavy metal concentrations (Zn, Cu, Cd, Ni, Fe, Pb) and health risk indices (CDI, PLI, HI Adult, HI Child, TCR Adult, TCR Child) in topsoil samples from Zanzan Province, Iran (n = 357).

Interpolated Heavy Metal Contamination Maps - Zanjan Province
(Cubic spline interpolation, 120×120 grid, n = 357)

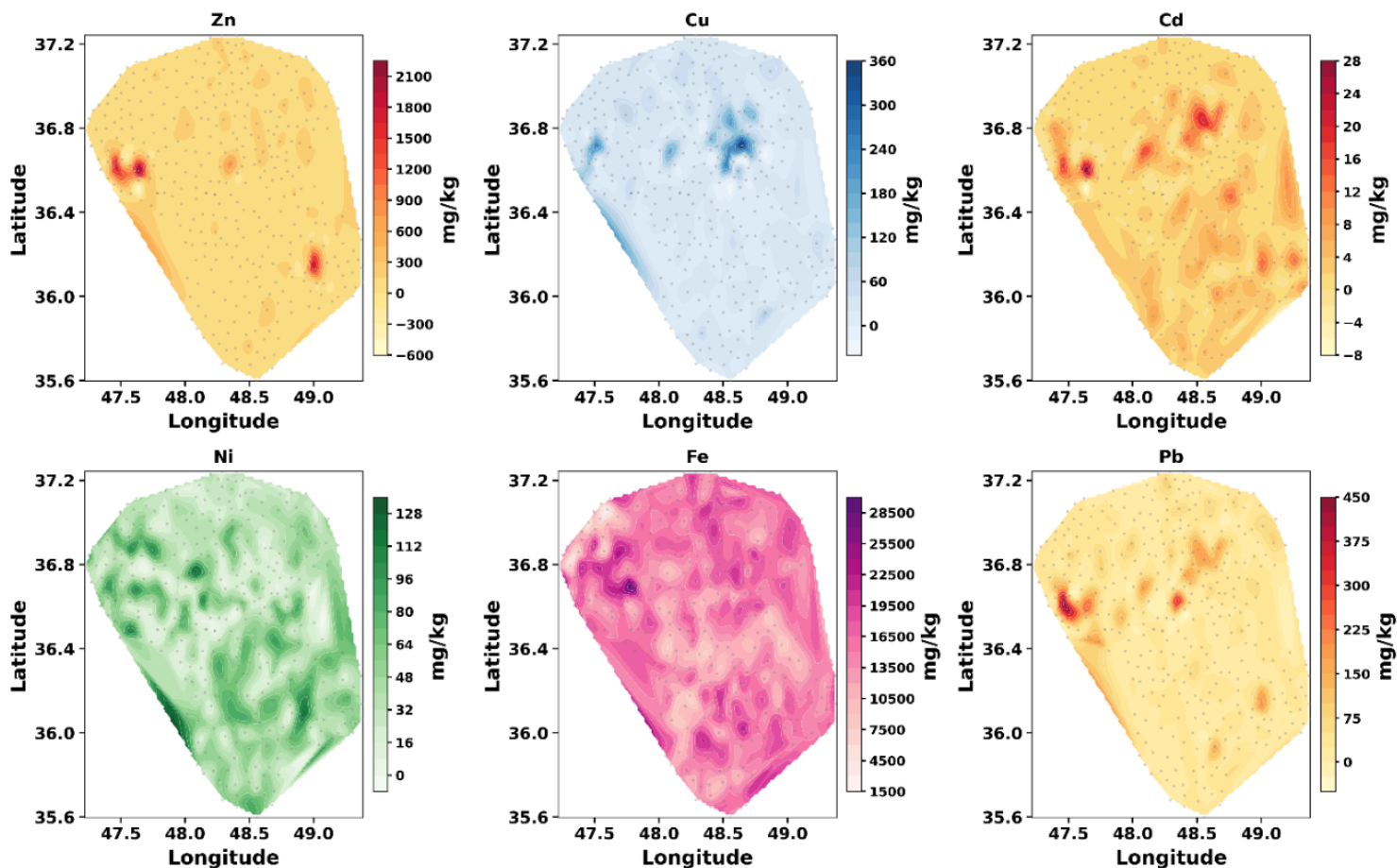


Figure 4

Interpolated spatial distribution maps of Zn, Cu, Cd, Ni, Fe, and Pb concentrations (mg/kg) across Zanjan Province, Iran (n = 357 GPS-referenced topsoil sites). Method: cubic spline interpolation (on a 120×120 grid). Colour intensity: proportional to metal concentration with individual colour scales per panel. Black filled circles = actual sampling locations; white contour lines = isoconcentration boundaries. The western hotspot cluster of elevated Zn, Pb, and Cd (approx. 47.5°–48.3°E, 36.4°–36.8°N) coincides with the Angouran Zn–Pb–Ag mine complex. Elevated Cu and Fe in the north-central zone correspond to the Zanjan Industrial Town. The spatial homogeneity of Fe and Ni confirms their geogenic origin.

Principal Component Analysis - Heavy Metal Source Apportionment

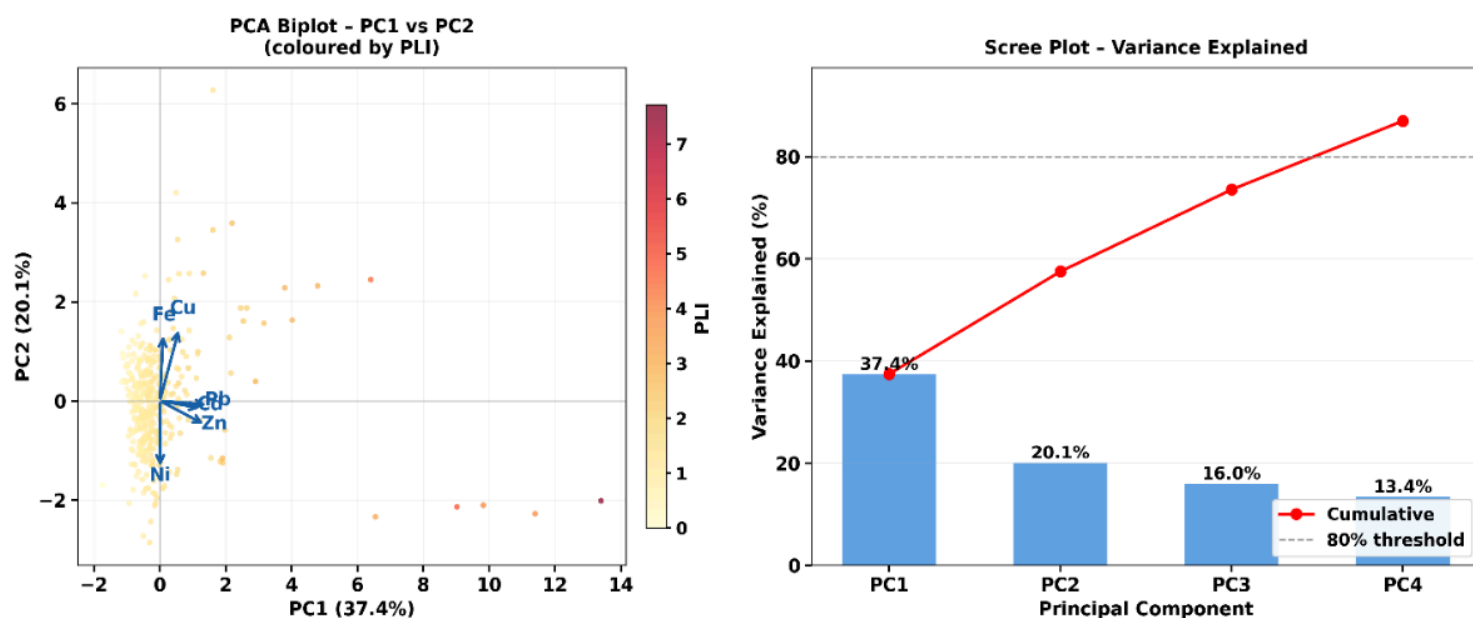


Figure 5

PCA for the six-metal concentration matrix from Zanjan Province ($n = 357$, $p = 6$). Left panel PCA biplot of PC1 vs PC2: each point = one sampling site.

K-Means Cluster Analysis - Pollution Zone Identification

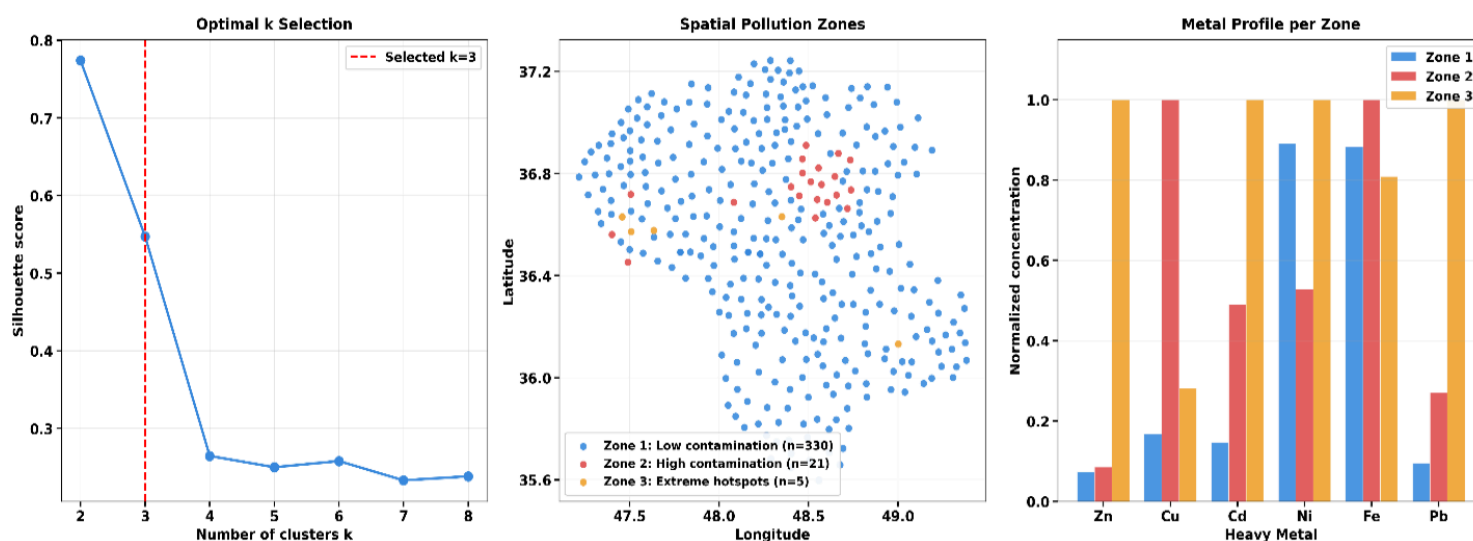


Figure 6

K-Means cluster analysis results for topsoil samples in Zanjan Province ($n = 357$; $k = 3$ optimised by silhouette score). Silhouette score vs number of clusters k ($k = 2-8$): $k = 3$ achieves the highest silhouette score, confirming it as the statistically optimal partition (left panel). Normalised metal concentration profile per zone (scaled to zone maximum), showing the geochemical signature of each zone and its correspondence to the three PCA source factors (Right panel).

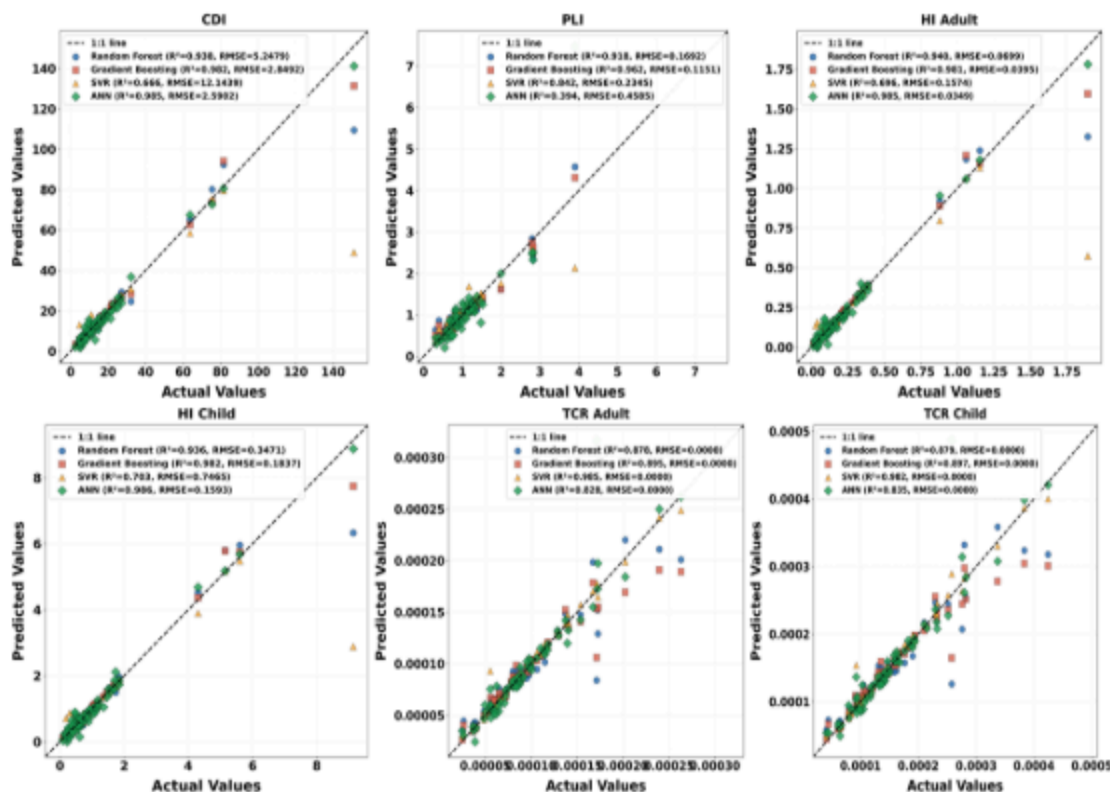


Figure 7

Scatter plots of predicted vs actual values for all four ML models on a common axis per health risk index (six panels, one per target). (n = 71).

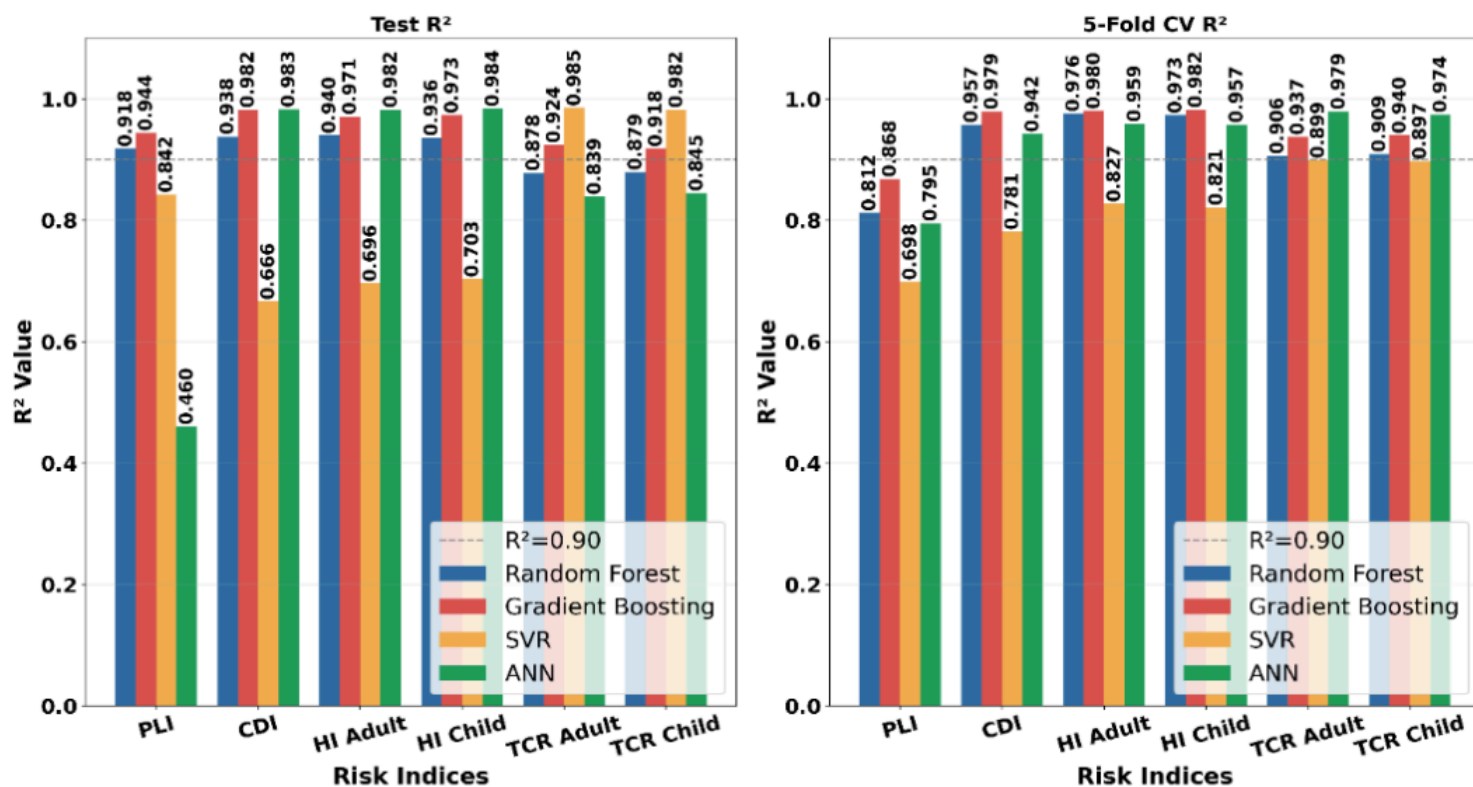


Figure 8

Comparative performance of all four ML models (RF, GBR, SVR, ANN) for six health risk indices. Left panel: Test R^2 on the held-out 20% test set (n test = 71).

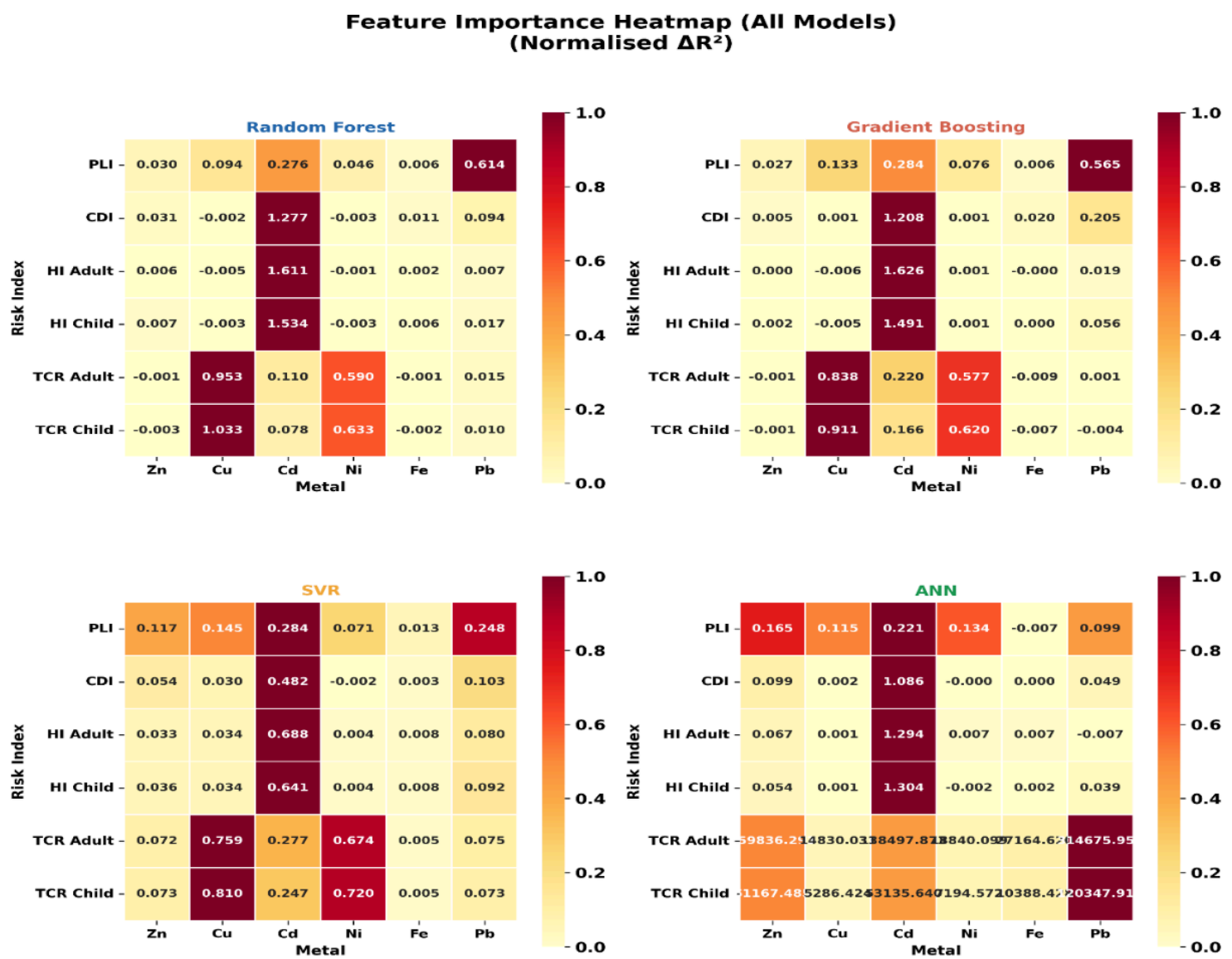


Figure 9

Permutation-based feature importance bar plots (ΔR^2 , 30 permutation repeats; mean $\pm$ 1 SD error bars) for all four ML models across all six health risk indices: 24 panels (6 rows = PLI, CDI, HI Adult, HI Child, TCR Adult, TCR Child; $\times$ 4 columns = RF, GBR, SVR, ANN). Within each panel, metals are sorted by mean $|\Delta R^2|$ in ascending order (least important at left; most important at right). The two most important metals per panel are shown in full colour; remaining metals are faded. Positive ΔR^2 indicates that metal contributes positively to predictive accuracy while negative ΔR^2 shows metal acts as a confounding or noise variable for that specific model–target combination.

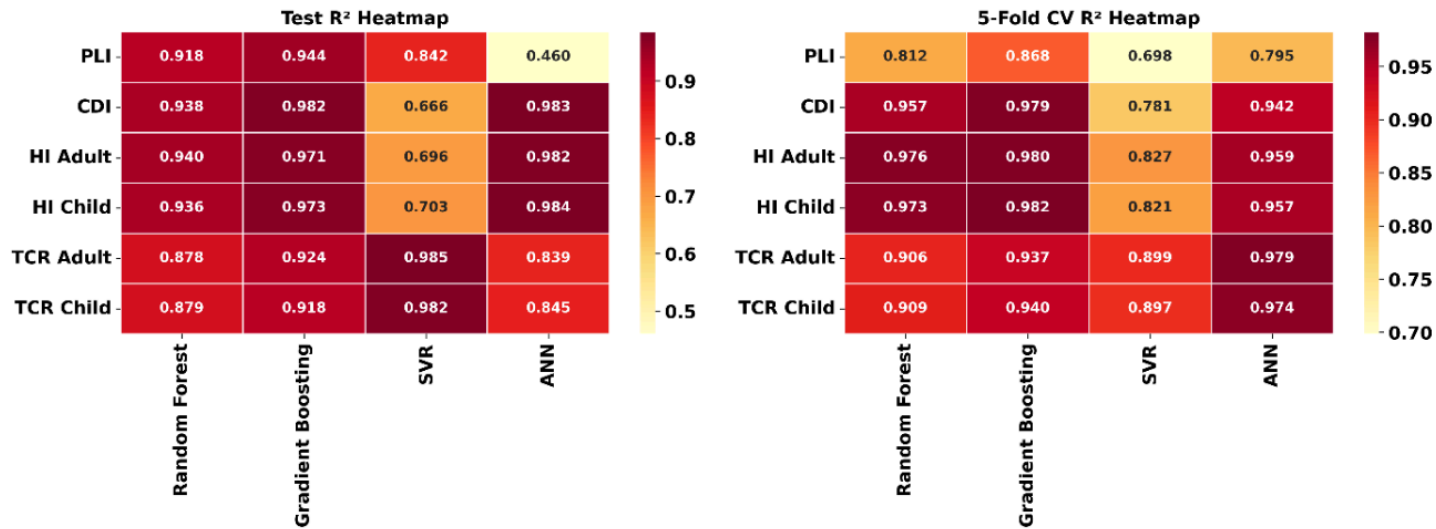


Figure 10

Cross-model summary heatmap of permutation feature importance (mean ΔR^2 over 30 repeats) for all four ML models (RF, GBR, SVR, ANN) and all six health risk indices.

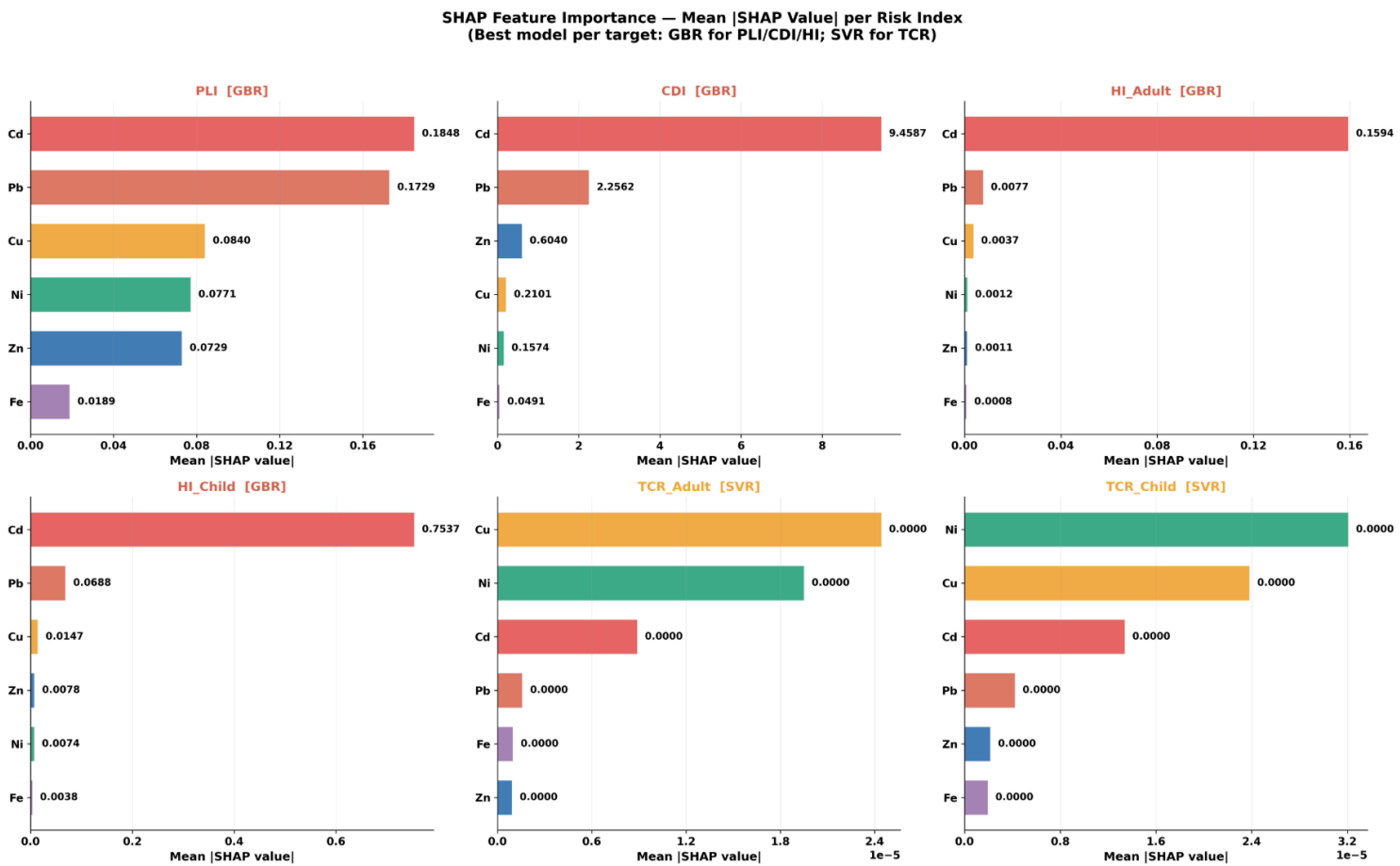


Figure 11

SHAP mean absolute value (mean |SHAP|) bar plots for all six health risk indices using the best-performing model per target (GBR for PLI, CDI, HI Adult, HI Child; SVR for TCR Adult, TCR Child). Each bar = mean absolute SHAP contribution of one metal, averaged over all test samples (n = 72).

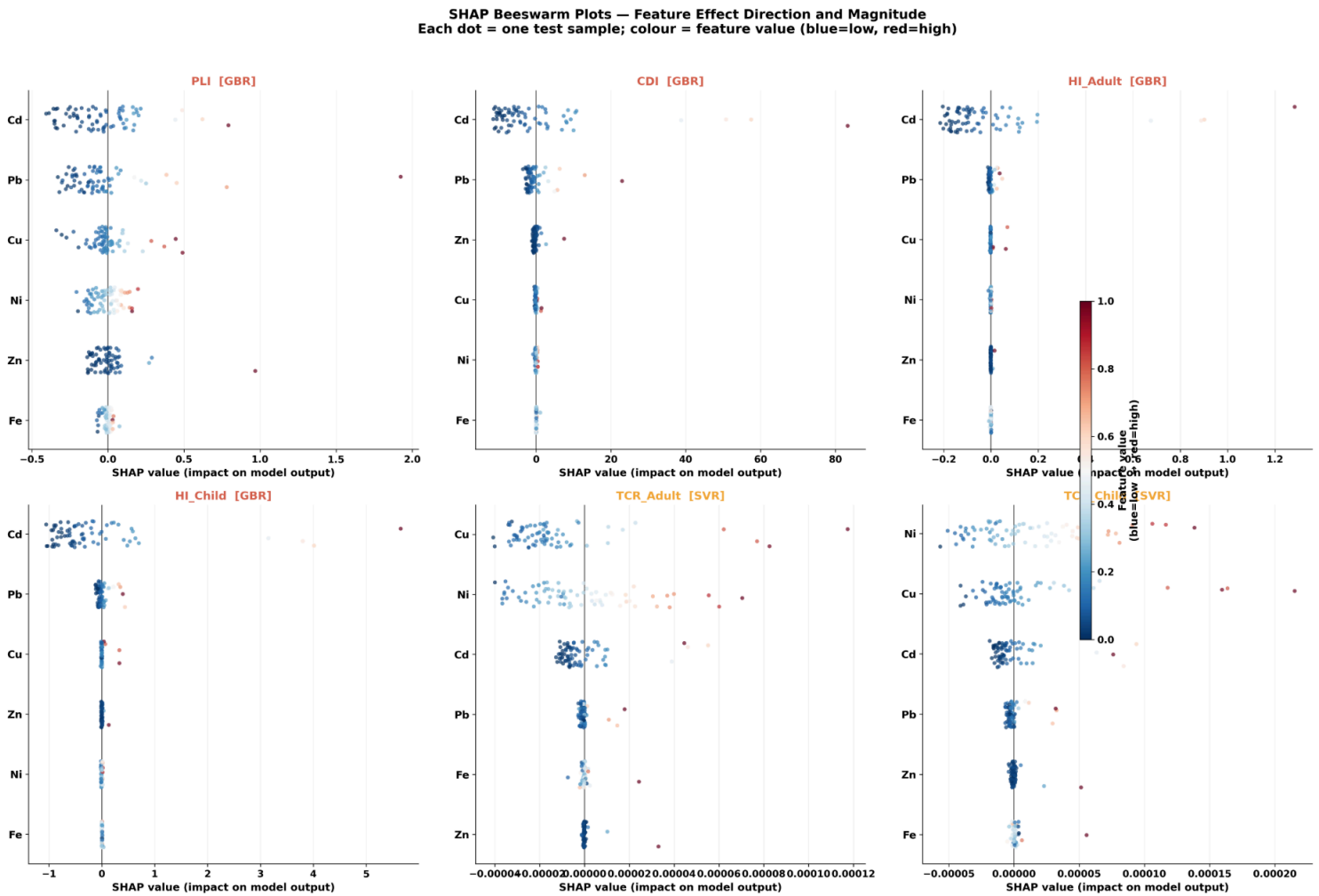


Figure 12

SHAP beeswarm plots for all six health risk indices (best model per target). Each dot = one test sample (n = 72). Horizontal position = SHAP value: positive (right of zero) = increases predicted risk; negative (left of zero) = decreases predicted risk. Dot colour = actual measured concentration of that metal at the corresponding site: blue = low concentration; red = high concentration.

SHAP Waterfall Plots — HI Child Prediction Breakdown for Three Representative Sites
(Best model: GBR; each bar shows one metal's additive contribution to the prediction)

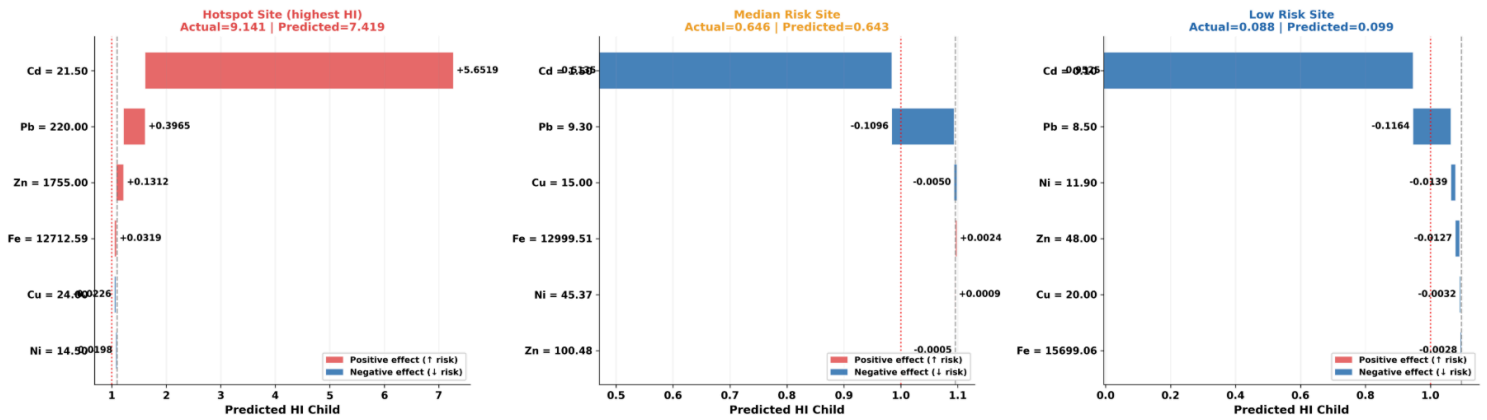


Figure 13

SHAP waterfall plots decomposing the GBR model's predicted HI Child for three representative test sites: the highest-risk hotspot (left panel), a median-risk site (centre panel), and the lowest-risk site (right panel). Each horizontal bar = one metal's additive SHAP contribution: red bars increase predicted HI Child; blue bars decrease it. Metal labels on the y-axis include the actual measured concentration at that site (mg/kg). The grey dashed vertical line = global training-set baseline (mean HI Child). The red dotted vertical line = non-carcinogenic health risk threshold (HI = 1.0). SHAP values (Δ HI Child) are annotated numerically at the end of each bar. Bars are sorted by absolute SHAP magnitude (smallest at top; largest at bottom), showing the contribution build-up from baseline to final prediction.

SHAP Dependence Plots — HI Child (GBR Model)
Each dot = one test sample; colour = Cd concentration (blue=low → red=high)

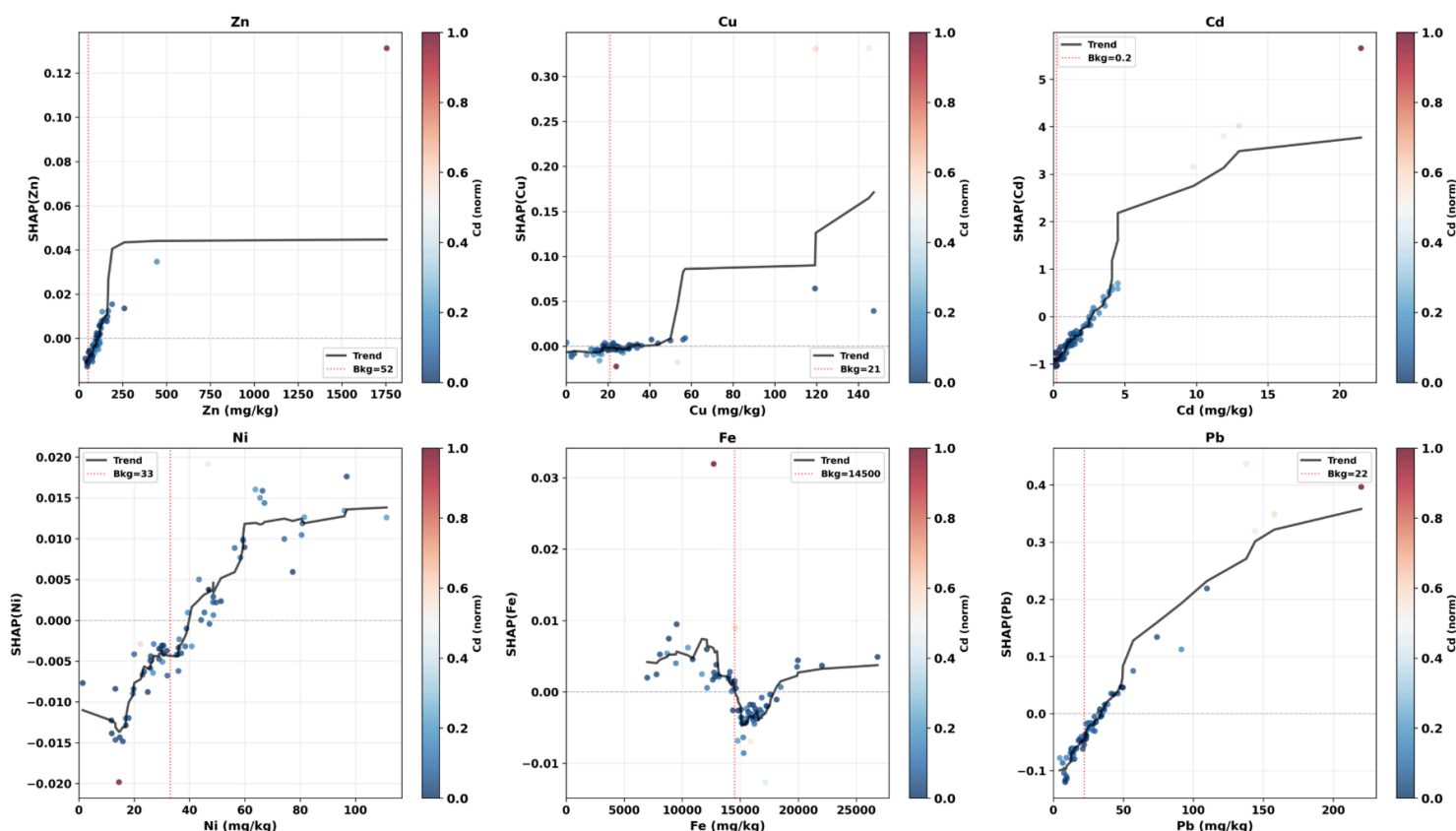


Figure 14

SHAP dependence plots for HI Child (GBR model), showing the concentration-response relationship for each of the six metals. x-axis = actual measured metal concentration at each test site (mg/kg); y-axis = SHAP value for that metal (its marginal contribution to the predicted HI Child at that site). Dot colour = Cd concentration at the same site (blue = low Cd; red = high Cd), revealing Cd–metal interaction effects. Reference lines: grey dashed horizontal = SHAP = 0 (no marginal contribution); red dotted vertical = published geochemical background reference value for each metal. Solid black line = uniform-filter smoothed trend. Steep positive SHAP slopes above the background line identify metals whose concentrations above background translate directly into increased predicted child health risk.

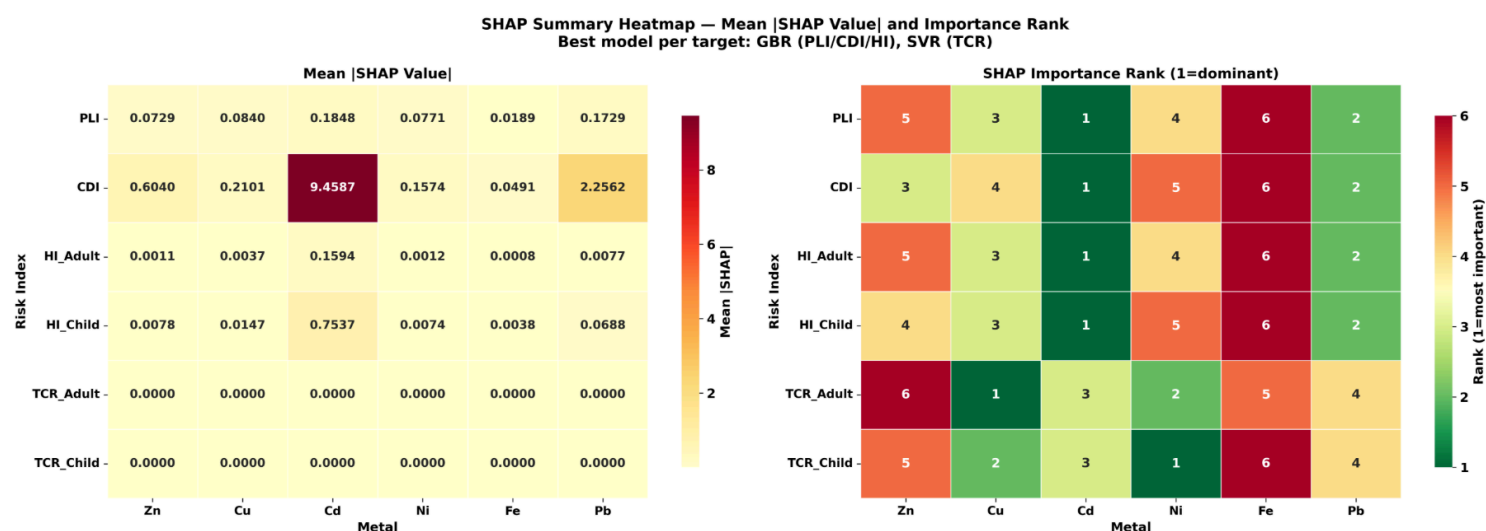


Figure 15

SHAP consolidated summary heatmap for all six health risk indices and six metals (best model per target: GBR for PLI/CDI/HI; SVR for TCR). Left panel: Mean |SHAP value| heatmap: each cell reports the mean absolute SHAP contribution of that metal for that risk index (averaged over $n = 72$ test samples); warmer colours indicate larger absolute contributions. Right panel: SHAP importance rank heatmap: each cell shows the rank of that metal relative to all six metals for the same risk index (rank 1 = most important; rank 6 = least important); green = high-rank (dominant); red = low-rank (negligible). This dual-panel view simultaneously communicates which metal is most important for each risk index (right) and by how much (left), directly guiding source-specific emission control priorities.

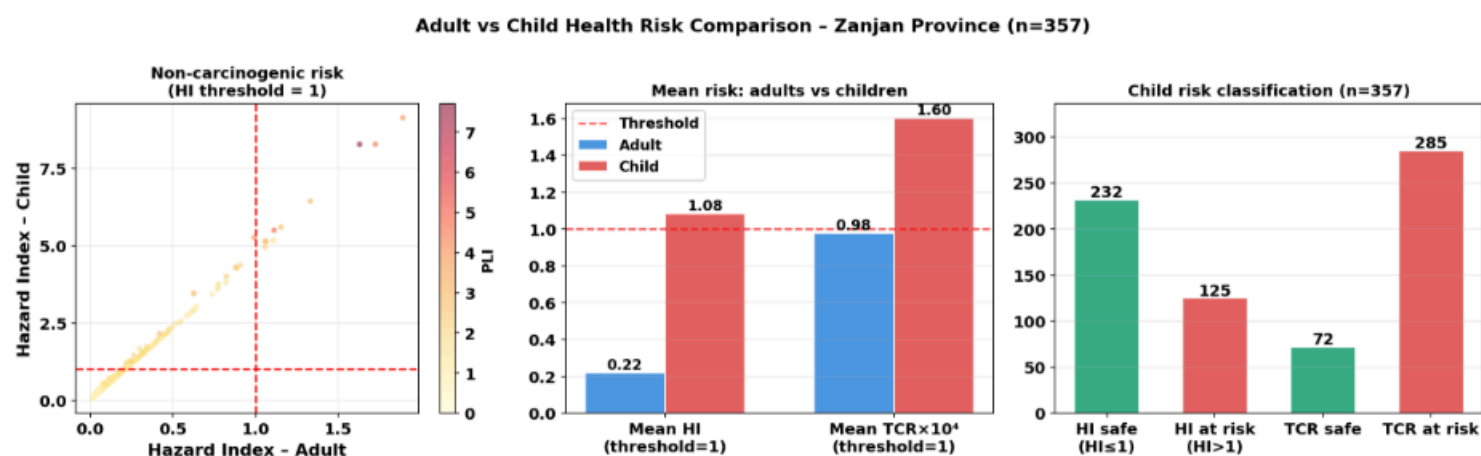


Figure 16

Adult vs child health risk comparison across 357 soil sampling sites in Zanjan Province, Iran. Left panel scatter plot of Hazard Index (HI) for adults (x-axis) vs children (y-axis); red dashed lines at HI = 1.0 mark the non-carcinogenic risk threshold; sites in the upper-right quadrant exceed the threshold for both demographics; sites in the upper-left exceed it only for children. Centre panel comparison of mean HI and mean TCR ($\times 10^4$) values for adults and children with the red threshold reference line at 1.0. Right

panel: bar chart classifying all 357 sites as safe vs at-risk for child non-carcinogenic risk ($HI \leq 1$ or $HI > 1$) and child carcinogenic risk ($TCR \leq 10^{-4}$ or $TCR > 10^{-4}$), with site counts annotated.