Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

UNIVERSITY OF TRANSPORT TECHNOLOGY

Mapping Cadmium Contamination Potential in Surface Soil for Civil Engineering Applications: A Comparative Study of Machine Learning and Deep Learning Models in the Gianh River Basin, Vietnam

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2025.en.5.2.48-70

[*]**Corresponding author:**
Email address:
tphong1617@gmail.com

Received: 13/03/2025
Received in Revised Form: 19/04/2025
Accepted: 07/05/2025

Vuong Hong Nhat[1], Phan Trong Trinh[1,2], Lai Vinh Cam[1], Bui Tien Dieu[3], Le Van Hiep[4], Indra Prakash[5], Nguyen Ngoc Anh[6], Nguyen Van Hong[1], Nguyen Duc Thanh[1], Nguyen Phuong Thao[1], Nguyen Thi Thu Hien[1], Tran Thi Nhung[1], Tran Trung Hieu[1,7], Tran Van Phong[1,2*]

[1]Institute of Earth Sciences, Vietnam Academy of Science and Technology, Hanoi 100000, Vietnam

[2]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi 100000, Vietnam

[3]GIS Group, Department of Business and IT, University of South-Eastern Norway, Gullbringvegen 36, 3800 BØ i Telemark, Norway

[4]University of Transport Technology, Hanoi 100000, Vietnam

[5]DDG(R) Geological Survey of India, Gandhinagar, Gujarat 382010, India

[6]Institute of Marine Environment and Resources, Vietnam Academy of Science and Technology, Vietnam

[7]International Environmental Doctoral School, University of Silesia in Katowice, Sosnowiec, Poland

Abstract: Cadmium (Cd) is a toxic heavy metal with significant environmental and human health risks, particularly when accumulated in surface soils. Its presence reduces soil fertility, disrupts microbial ecosystems, and poses long-term ecological threats. This study explores the application of artificial intelligence (AI) models for mapping the potential distribution of Cd contamination in surface soils within the Gianh River Basin, Quang Binh Province, Vietnam. Four machine learning (ML) models Logistic Regression (LR), Radial Basis Function Network (RBFN), Random Forest (RF), and Support Vector Machine (SVM) and four deep learning (DL) model variants (DNN-Opt1 to DNN-Opt4) were developed and compared. The DNN variants differ based on the configuration of hidden layers and neuron counts.

A total of 100 topsoil samples were collected and classified using the Geoaccumulation Index (Igeo), serving as the target variable for supervised learning. Thirteen conditioning factors were used as input variables, including Elevation, Soil Type, Slope, Curvature, proximity to roads and rivers, and seven Landsat 8 spectral bands. The dataset was divided into training (70%) and testing (30%) subsets. Model performance was evaluated using multiple metrics, including the area under the ROC curve (AUC), accuracy (ACC), Kappa coefficient, root mean square error (RMSE), and confusion matrix.

Published online 22/05/2025

Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Abstract (continued) Among the tested models, the DNN-Opt2 variant demonstrated the highest predictive performance with AUC = 0.858, ACC = 73.33%, Kappa = 0.47, and RMSE = 0.45. The resulting contamination potential map, particularly that derived from the RBFN model, categorized the region into five contamination risk levels: very low, low, moderate, high, and very high. This spatial information is critical not only for environmental management but also for assessing risks to groundwater quality and the structural integrity of buildings located in high-risk zones. The study demonstrates the efficacy of deep learning in enhancing predictive accuracy for heavy metal contamination mapping and underscores its practical relevance in civil and environmental engineering applications.

Keywords: Cd contamination potential map; Igeo; deep learning; machine learning; Gianh river basin.

1. Introduction

Cadmium (Cd) is a hazardous heavy metal known for its persistence in the environment and severe toxic effects on human health, agriculture, groundwater systems, and infrastructure [1, 2]. It enters surface soils through both natural processes and anthropogenic activities such as mining, industrial discharge, and the excessive use of phosphate fertilizers [3-5]. Cd is highly mobile, exhibits significant bioaccumulation potential, and readily enters the food chain through crops [6, 7]. Elevated Cd exposure through dietary intake has become a growing concern in parts of Southeast Asia, often exceeding the safety limits established by the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) [8, 9].

In Vietnam, and particularly in the Gianh River Basin of Quang Binh Province, increasing levels of Cd contamination have been reported due to expanding agricultural and industrial activities [10-14]. Cd accumulation in surface soil not only degrades soil quality and threatens ecosystems but also raises risks to groundwater contamination and the long-term stability of infrastructure, as contaminated soils may chemically interact with construction materials. Accurate spatial mapping of Cd contamination is therefore essential for land-use planning, environmental management, and civil infrastructure safety.

While many current studies focus on source identification, concentration analysis, and remediation technologies such as phytoremediation, bioremediation, and chemical immobilization [15, 16], mapping the spatial distribution of contamination potential has emerged as a critical tool for decision-making in environmental risk management [17]. In recent years, artificial intelligence (AI) methods, particularly machine learning (ML) and deep learning (DL) techniques, have shown promise in improving predictive accuracy for spatial modeling tasks [18-20].

ML techniques have been widely used to model various environmental risks, such as heavy metal accumulation [21, 22], Cd prediction in crops [19], and mapping soil contamination using remote sensing [23-25]. While these applications often focus on regression tasks, classification-based models have successfully been employed in environmental mapping for problems such as groundwater potential [26-28], landslide susceptibility [29, 30], and flood risk [18, 31].

In this study, we apply and compare four widely used ML classification models Logistic Regression (LR), Radial Basis Function Network (RBFN), Random Forest (RF), and Support Vector Machine (SVM) with four variants of Deep Neural Network (DNN) models (DNN-Opt1 to DNN-Opt4) to predict and map the potential of Cd contamination in the Gianh River Basin. The DNN variants differ in their configurations of hidden layers and neuron counts. The input dataset comprises 100 topsoil samples, classified into contaminated and non-contaminated classes based on the Geoaccumulation Index (Igeo). Thirteen influencing factors were used as input variables, including Elevation, Soil Type, Slope, Curvature, Distance to roads, Distance to rivers, and Landsat 8 bands (1–7).

The resulting Cd contamination potential map categorizes the area into five levels-very low to very high. This spatial information provides valuable insights not only for environmental monitoring but also for evaluating groundwater vulnerability and infrastructure safety in high-susceptible zones. By demonstrating the capabilities of ML and DL techniques in this context, this study highlights their practical applications in civil and environmental engineering.

2. Study Area

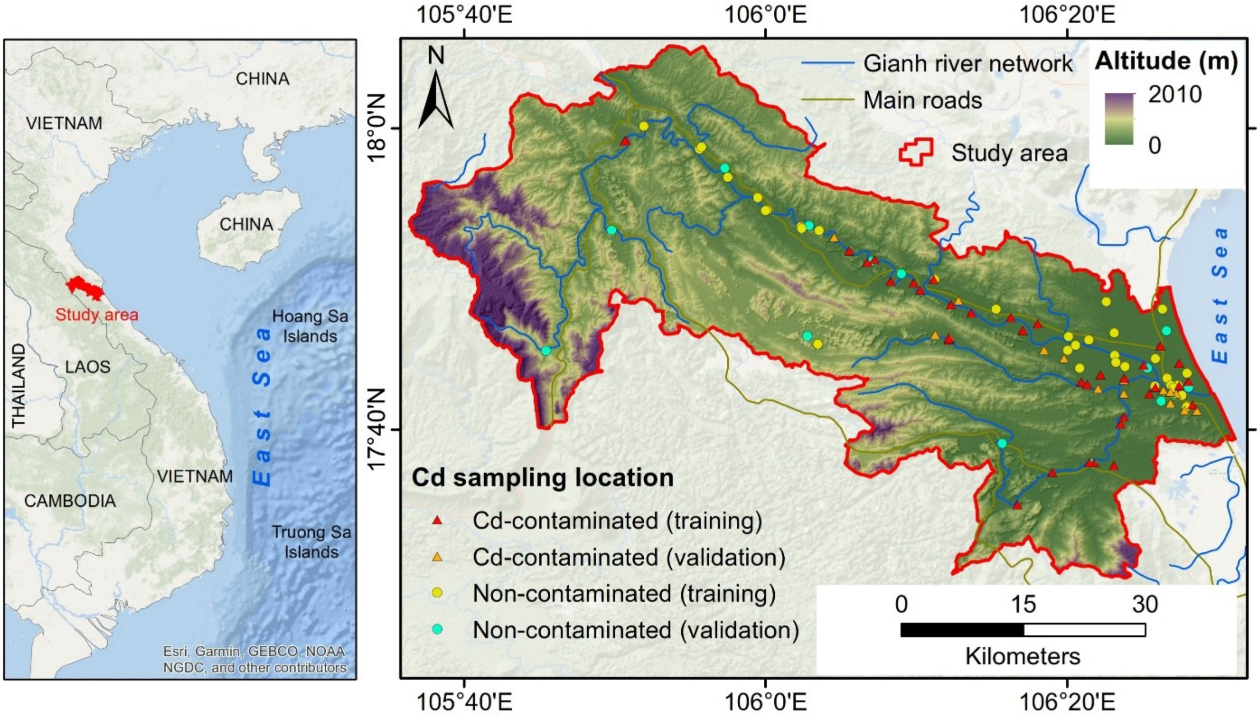

Fig. 1. Study area in the Gianh River basin, Quang Binh Province, Vietnam

The study area encompasses a part of the Gianh River basin, covering an area of 1,808 km² (Fig. 1). The Gianh River is the largest of the five main rivers flowing in the Quang Binh Province (Roòn River, Gianh River, Dinh River, Ly Hoa River, and Nhat Le River). Originating at an elevation of 1,350 m from Co Pi Mountain in the Truong Son Range, it flows through the districts of Minh Hoa, Tuyen Hoa, Quang Trach, Bo Trach, and Ba Don Town before emptying into the East Sea at Gianh Estuary. The river extends 158 km in length, with an average basin width of 38.8 km, a river network density of 1.54 km/km², and a total basin area of 4,680 km².The Gianh River provides significant benefits in terms of fisheries and supplies water for agriculture and domestic use for communities in the riverine plains and the expansive downstream delta region [32]. Additionally, it contributes to the formation of stunning natural landscapes, most notably Phong Nha Cave, a UNESCO World Natural Heritage site, which underscores the need for its protection to prevent water pollution and alterations to the riverbed. With the development of industrial and agricultural activities in the study area, the Gianh River is experiencing an increasing impact from environmental pollution [33]. Organic pollutants also tend to increase toward the river's lower reaches, where population density is higher, and markets, industrial production facilities, and traditional craft villages are concentrated [34]. The Gianh River basin serves as a critical interface between aquatic and terrestrial ecosystems, playing an essential role in maintaining water quality, stabilizing riverbanks, and supporting aquatic biodiversity. This area is considered a potential reservoir for heavy metals, which may accumulate due to flooding or be transported from higher elevations through sedimentation or percolation processes [35-37].

3. Role of Cadmium (Cd) Mapping in Assessing Groundwater Use and Infrastructure Suitability on Contaminated Soil

Cadmium (Cd) mapping plays a pivotal role in environmental planning and public health protection, especially in regions where industrial, agricultural, or geological activities contribute to soil contamination. By visualizing the spatial distribution of cadmium concentrations, this approach enables researchers and policymakers to assess potential risks to groundwater and evaluate land suitability for infrastructure development. In areas where cadmium is present at hazardous levels, mapping helps identify contamination hotspots, trace pollution pathways, and prioritize zones for remediation or restricted land use [38].

When it comes to groundwater use, Cd mapping is essential for determining water safety and sustainability. It provides critical insights into how cadmium migrates through soil and potentially reaches aquifers, aiding in the protection of drinking water sources [39]. Similarly, infrastructure planning on contaminated soil benefits from this spatial analysis, as it highlights areas where structural materials may corrode or where health risks for construction workers and future occupants may be elevated. Ultimately, cadmium mapping informs safer, more sustainable land-use decisions by integrating environmental data with technological tools such as remote sensing, machine learning, and geostatistical modeling.

4. Methodology and data used

4.1. Methodology

The research methodology follows a structured six-step workflow, as illustrated in Fig. 2. The core objective is to develop a cadmium (Cd) contamination potential map by solving a binary classification problem, where soil samples are categorized as either contaminated (1) or non-contaminated (0) based on Cd concentration levels.

To achieve this, eight machine learning (ML) and deep learning (DL) models were evaluated to identify the most suitable approach for accurate spatial prediction. These models include four ML algorithms-Logistic Regression (LR), Radial Basis Function Network (RBFN), Random Forest (RF), and Support Vector Machine (SVM)-and four optimized Deep Neural Network (DNN) variants (DNN-Opt1 to DNN-Opt4), each configured with different combinations of hidden layers and neurons.

Model development and training were conducted using Weka 3.8.6, an open-source data mining software developed by the University of Waikato. Detailed configurations and hyperparameters for all models are provided in Table 1.

Soil sampling and laboratory analysis procedures are shown in Fig. 3. A total of 100 topsoil samples were collected from the study area and analyzed for Cd concentrations. Measurements were conducted using the Agilent ICP-MS 7900, an inductively coupled plasma mass spectrometry system equipped with 4th-generation Octopole Reaction System (ORS) technology [40]. The analytical work was performed at the Institute of Earth Sciences, Vietnam Academy of Science and Technology.

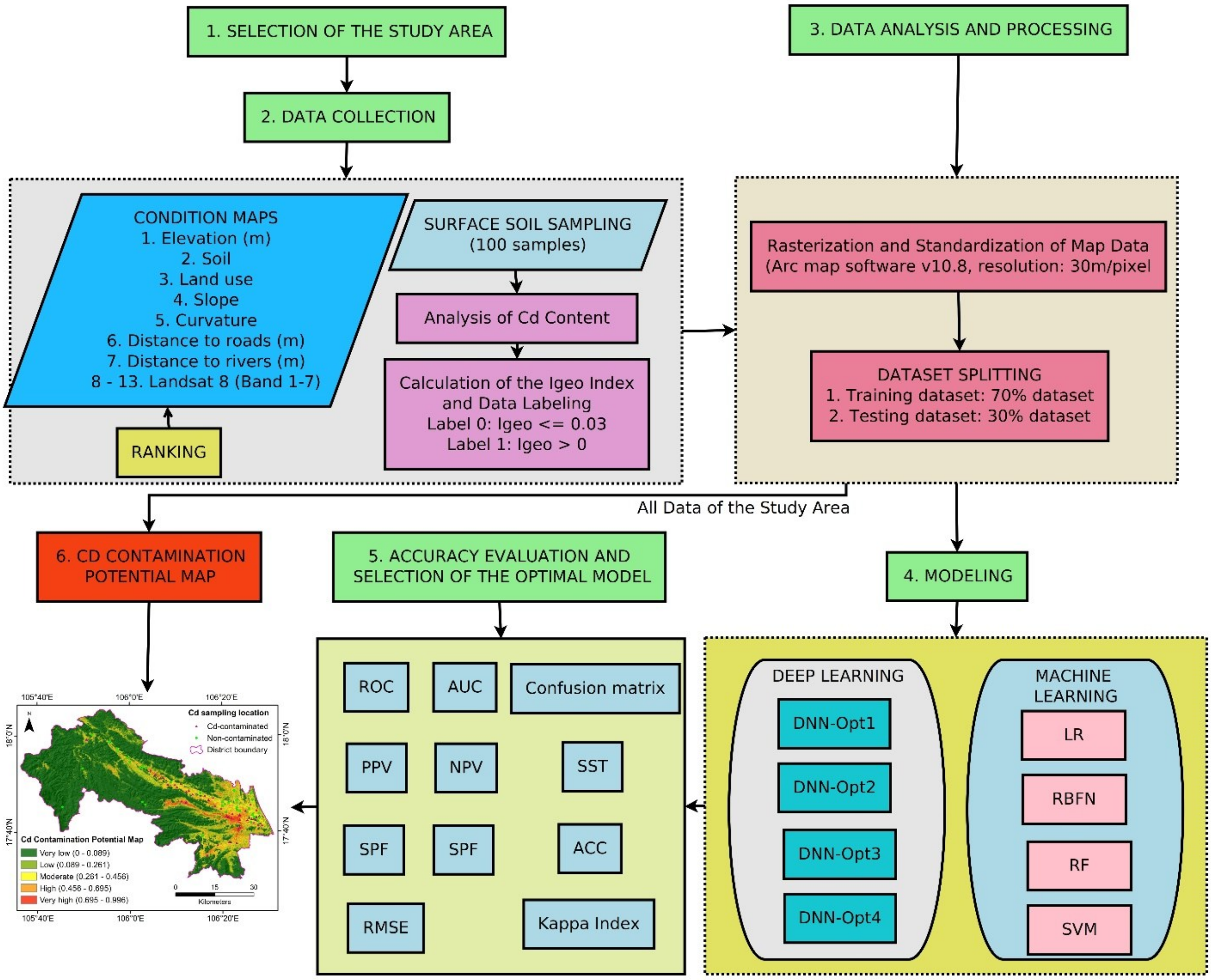

Fig. 2. Flow chart of methodology in this research

Sunday, 14 January 2024 at 10:02
N 17° 45' 24.993" E 106° 23' 7.485"
Quảng Bình,H. Quảng Trạch,X. Quảng Phương
Altitude:0.4meter
Speed:2.4km/h
MARS

Fig. 3. Soil Sampling in the Field and Cadmium Analysis in the Laboratory

Table 1. Hyperparameters of machine learning and deep learning models

No.	Hyperparameter	Models							
		DNN-Opt1	**DNN-Opt2**	**DNN-Opt3**	**DNN-Opt4**	**LR**	**RBFN**	**RF**	**SVM**
	Number of Epochs	10				-	-	-	-
	Batch size	100				100	100	100	100
	Activation function	Softmax				-	-	-	-
	Loss function	MCXENT				-	-	-	-
	Number of hidden layers	0	1	2	3	-	-	-	-
	Number of neurons	0	8	16	24	-	-	-	-
	Optimization algorithm	Stochastic gradient descent				-	-	-	-
	Updater	Adam				-	-	-	-
	Learning rate	0.001				-	-	-	-
	Weight initialization method	XAVIER				-	-	-	-
	Bias initialization	0.0				-	-	-	-
	Gradient normalization threshold	1.0				-	-	-	-
	Number of decimal places	2				3	2	2	2
	Maximum of interations	-	-	-	-	-1	-1	100	-
	Ridge	-	-	-	-	1E-8	1E-8	-	-
	Clustering seed	-	-	-	-	-	1	-	-
	Minimum standard deviation	-	-	-	-	-	0.1	-	-
	Number of clusters for K-mean	-	-	-	-	-	2	-	-
	Maximum depth of the tree	-	-	-	-	-	-	0	-
	Number of Execution Slots	-	-	-	-	-	-	1	-
	SVM type	-	-	-	-	-	-	-	C-SVC
	Coefficient	-	-	-	-	-	-	-	0.0
	Cost	-	-	-	-	-	-	-	1.0
	Degree	-	-	-	-	-	-	-	3
	Tolerance	-	-	-	-	-	-	-	0.001
	Gamma	-	-	-	-	-	-	-	0.0
	Loss	-	-	-	-	-	-	-	0.1
	Seed	-	-	-	-	-	-	1	1

4.1.1. Selection of conditioning factors (parameters)

The accumulation and distribution of heavy metals in soils are governed by a complex interplay of geospatial and environmental factors, with key influences including soil physicochemical properties, topographic features, and land use patterns [41]. Among these, cadmium (Cd) contamination is particularly associated with soil type, elevation, slope, hydrology, and anthropogenic activities. Cd can also originate from external sources such as industrial emissions, oil refineries, sewage sludge, and the application of chemical fertilizers [42].

To effectively model and map Cd contamination potential, this study selected 13 conditioning factors as input variables for the machine learning models. These include six environmental variables soil type, elevation, slope, curvature, distance to roads, and distance to rivers alongside seven spectral bands from Landsat 8 imagery (Bands 1–7) (Fig. 4). The inclusion of Landsat 8 data compensates for the unavailability of certain direct indicators of Cd pollution, as the satellite's wide spectral coverage offers detailed surface reflectance information useful for capturing variations in vegetation cover, moisture, and land surface conditions [43].

Topographic parameters such as elevation, slope, and curvature play a vital role in influencing runoff, erosion, and soil accumulation processes, which directly impact the mobility and concentration of Cd in surface soils. Distance to rivers serves as a proxy for hydrological transport potential, while proximity to roads is considered an indirect indicator of vehicular emissions and related pollutant deposition. Soil types are critical due to their inherent physical and chemical properties-such as pH, organic matter content, texture, and grain size distribution, that affect the retention and movement of heavy metals [42, 44].

To evaluate the relative importance of these 13 variables, the OneR attribute evaluation method was applied, providing insights into the individual contribution of each factor to Cd contamination potential within the study area [30].

4.1.2. Cd contamination data labeling

Initially, the analysis of Cd content was conducted on soil samples collected from the field. Subsequently, the Geoaccumulation Index (Igeo) was analyzed and calculated to determine locations with Cd contamination and those without. Accordingly, samples identified as Cd-contaminated were assigned a value of 1, while non-contaminated samples were assigned a value of 0. Table 3 provides a detailed presentation of the results of Cd analysis and the data labeling for the machine learning models. A summary of the Igeo index is presented below:

The Igeo is a widely utilized method for assessing the degree of heavy metal pollution in the environment, particularly in soil and sediments [45, 46]. This index was first proposed by the German geochemist Müller in 1969 to compare the concentrations of heavy metals in environmental samples with their natural background values [47]. Its advantages include simplicity and ease of application; however, a limitation is its dependence on background values [47]. It is primarily applied in the assessment of soil, water, and sediments for pollution management purposes [48]. The formula for calculating Igeo [47] is as follows:

$$\mathrm{Igeo}=\log_2\left(\frac{H_n}{1.5 \text{ x } K_n}\right) \quad (1)$$

Where: H_n is the concentration of Cd in the environmental sample under evaluation. K_n is the natural background concentration of Cd.

Classification of pollution levels according to Igeo [47]:

- Igeo ≤ 0: No pollution (or natural level).
- 0 < Igeo ≤ 1: Low pollution level.
- 1 < Igeo ≤ 2: Moderate pollution level.
- 2 < Igeo ≤ 3: High pollution level.
- 3 < Igeo ≤ 4: Very high pollution level.
- Igeo > 4: Extremely high pollution level.

4.1.3. Deep Neural Network (DNN)

The DNN operates based on a structure comprising multiple neuron layers, including an input layer, several hidden layers, and an output layer [49]. Data passes through these layers in a feedforward process to generate predictions [50]. The learning process involves calculating the error between the predicted and actual outcomes, followed by adjusting the weights and biases through backpropagation and optimization algorithms to minimize the error [51]. Techniques such as regularization help prevent overfitting, enabling the DNN to learn complex features from the data due to its deep structure [52]. Experimentation with variations in the number of layers and neurons in the hidden layers impacts the model's predictive capability [53]. In this study, we modified the neuron architecture in the hidden layers of the DNN model according to four scenarios. These models, corresponding to the four scenarios, were designated DNN-Opt1, DNN-

Opt2, DNN-Opt3, and DNN-Opt4. Detailed parameters of these models are presented in Table 1.

4.1.4. Logistic Regression (LR)

LR is a regression model used to predict the probability of an event occurring, commonly applied to binary classification problems [54]. It employs the sigmoid function on the weighted sum of input variables multiplied by coefficients, transforming this sum into a value between 0 and 1, which represents the probability [54]. LR is trained by optimizing a loss function, such as cross-entropy, and adjusting the coefficients to best fit the data [55]. LR is a widely used machine learning model that has been applied in Earth science fields, such as mapping landslide susceptibility [56] and flood susceptibility mapping [57]. Table 1 provides detailed hyperparameters of the LR model.

4.1.5. Radial Basis Function Networks (RBFN)

RBFN is a type of artificial neural network that employs radial basis functions to process data [58]. RBFN consists of three layers: an input layer, a hidden layer with neurons activated by radial basis functions (typically Gaussian), and an output layer [59]. Each neuron in the hidden layer measures the distance from the input to a specific 'center,' and then transforms this distance into an output value through the RBF [59]. The output layer is typically a linear combination of the outputs from the hidden layer, enabling the model to perform nonlinear classification or prediction [59]. The hyperparameters of the RBFN model are detailed in Table 1.

4.1.6. Random Forest (RF)

RF is a machine-learning model that utilizes multiple decision trees to make predictions [60]. Each tree in RF is trained on a random subset of the data, and these trees also consider only a random subset of features when splitting nodes. The final output of the model is an aggregation of predictions from all the trees, which helps mitigate overfitting [60]. For classification tasks, the most frequently selected class becomes the result; for regression tasks, it is the average of the predictions [61]. Consequently, RF offers high accuracy and good interpretability. The parameters of the RF model used in this study are presented in Table 1.

4.1.7. Support Vector Machine

SVM is a widely used machine learning model for classification and regression tasks [62]. SVM seeks to construct an optimal hyperplane to separate data classes [62]. The objective is to maximize the margin between this hyperplane and the nearest data points from each class, known as support vectors [62]. For data that cannot be linearly separated, SVM employs kernel functions to transform the data into a higher-dimensional space. SVM excels in efficiently handling binary classification problems but can also be extended to multiclass classification [63]. The model is highly regarded for its ability to reduce overfitting through the optimization of the margin between classes [64].

4.1.8. Assessment of model accuracy

The model's performance was evaluated using standard classification metrics, including Area Under the Curve (AUC), Accuracy (ACC), Sensitivity (SST), Specificity (SPF), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Root Mean Square Error (RMSE), and the Kappa coefficient. These metrics provide a comprehensive assessment of the model's predictive capabilities in terms of both correctness and reliability. Definitions and mathematical formulations of these evaluation metrics are presented in detail by [30].

4.2. Data used

The dataset used in this study consists of two main components: (1) data derived from the analysis of cadmium (Cd) content and the geoaccumulation index (Igeo) of 100 soil samples representing different soil types within the study area (Table 3), and (2) baseline spatial data comprising 13 factor maps, as detailed in Table 2 and Fig. 4. The 100 soil samples were randomly divided into training and testing sets in a 70:30 ratio. Accordingly, the training dataset includes 70 samples, with 35 labeled as '1' and 35 as '0,' while the testing dataset comprises 30 samples, with 15 labeled as '1' and 15 as '0.' The sample division was carried out using a random sampling algorithm implemented in ArcMap 10.8.

Different soil types identified in the study area are categorized from S1 to S22. Their descriptions are provided in Table 4.

Table 2. Data Sources for Maps of Factors Influencing Cd Contamination in the Gianh River Basin

No.	Factors	Scales/Resolution	Sources
1	Elevation (m)	30 m/pixel	DEM SRTM 1 arc from NASA (link: https://www.earthdata.nasa.gov/data/instruments/srtm)
2	Soil type	1:50,000	National Institute of Agricultural Planning and Projection
3	Slope (degree)	30m/pixel	Generate from DEM
4	Curvature	30m/pixel	Generate from DEM
5	Distance to roads (m)	30m/pixel	Generate from the main road map (digitized from Google Earth)
6	Distance to rivers (m)	30m/pixel	Generate from the main river map (digitized from Google Earth)
7	Landsat 8 Images (Band 1–7)	30m/pixel	USGS Explorer (https://earthexplorer.usgs.gov/)

Table 3. Results of Cd content analysis, Igeo, and labeling of soil-type samples in the study area

No.	Sample code	Cd (mg/kg)	Igeo (cd)	Label
Mangrove forest				
1	SG70	0.21	0.574	1
2	SG76	0.24	1.766	1
3	SG79	0.07	-0.148	0
4	SG80	0.16	1.341	1
5	SG86	0.24	1.612	1
6	SG89	0.21	0.650	1
7	SG90	0.15	1.217	1
8	SG91	0.11	1.537	1
9	SG92	0.13	1.129	1
Aquaculture				
10	SG68	0.37	1.332	1
11	SG75	0.08	-0.878	0
12	SG78	0.14	-0.070	0
13	SG93	0.11	-0.418	0
14	SG94	0.23	0.646	1
15	SG95	0.35	1.252	1
16	SG96	0.09	-0.708	0
Crops				
17	SG11	0.24	0.707	1
18	SG12	0.21	0.515	1
19	SG16	0.12	-0.293	0
20	SG19	0.09	-0.708	0
21	SG28	0.10	-0.556	0
22	SG29	0.03	-2.293	0
23	SG49	0.15	0.029	0
24	SG59	0.22	0.582	1
25	SG50	0.23	0.646	1
26	SG77	0.03	-2.293	0
27	SG37	0.13	-0.177	0
28	SG30	0.21	0.515	1
29	SG40	0.11	-0.418	0
30	SG41	0.13	-0.177	0
31	SG42	0.15	0.029	0
32	SG48	0.17	0.210	0
33	SG58	0.52	1.823	1
34	SG55	0.64	2.122	1
35	SG57	1.10	2.904	1
36	SG65	0.26	0.823	1
37	SG62	0.07	-1.070	0
38	SG43	0.09	-0.708	0
39	SG69	0.69	2.231	1
40	SG53	0.17	0.210	0
41	SG46	0.2	0.444	1
42	SG39	0.07	-1.070	0
43	SG52	0.17	0.210	0
44	SG25	0.09	-0.708	0
45	SG24	0.08	-0.878	0
Rice seedlings				
46	SG02	0.03	-2.293	0
47	SG03	0.02	-2.878	0
48	SG04	0.14	-0.070	0
49	SG06	0.09	-0.708	0
50	SG07	0.04	-1.878	0
51	SG08	0.06	-1.293	0
52	SG09	0.16	0.122	0
53	SG10	0.17	0.210	0
54	SG13	0.32	1.122	1
55	SG15	0.10	-0.556	0
56	SG61	0.15	0.029	0
57	SG81	0.09	-0.708	0
58	SG60	0.3	1.029	1
59	SG17	0.11	-0.418	0
60	SG14	0.16	0.122	0
61	SG05	0.1	-0.556	0
62	SG18	0.24	0.707	1

Table 3. (continued)

No.	Sample code	Cd (mg/kg)	Igeo (cd)	Label
63	SG21	0.39	1.408	1
64	SG23	0.42	1.515	1
65	SG82	0.98	2.737	1
Rice				
66	SG31	0.23	0.646	1
67	SG36	0.06	-1.293	0
68	SG38	0.15	0.029	0
69	SG47	0.24	0.707	1
70	SG51	0.28	0.930	1
71	SG63	0.07	-1.070	0
72	SG67	0.26	0.823	1
73	SG71	0.07	-1.070	0
74	SG74	0.04	-1.878	0
75	SG87	0.07	-1.070	0
76	SG64	1.38	3.231	1
77	SG66	0.34	1.210	1
78	SG85	0.26	0.823	1
79	SG97	1.02	2.795	1
80	SG98	0.25	0.766	1
1	SG99	0.63	2.100	1
82	SG100	0.47	1.677	1
83	SG72	0.06	-1.293	0
84	SG84	0.18	0.292	0
85	SG88	0.08	-0.878	0
86	SG73	0.42	1.515	1
Other soils				
87	SG35	0.09	-0.708	0
88	SG44	0.08	-0.878	0
89	SG45	0.53	1.850	1
90	SG33	0.03	-2.293	0
91	SG34	0.08	-0.878	0
92	SG27	0.15	0.029	0
93	SG26	0.09	-0.708	0
94	SG01	0.21	0.515	1
95	SG20	0.27	0.877	1
96	SG22	0.34	1.210	1
97	SG54	0.09	-0.708	0
98	SG56	0.47	1.677	1
99	SG83	0.37	1.332	1
100	SG32	0.11	-0.418	0

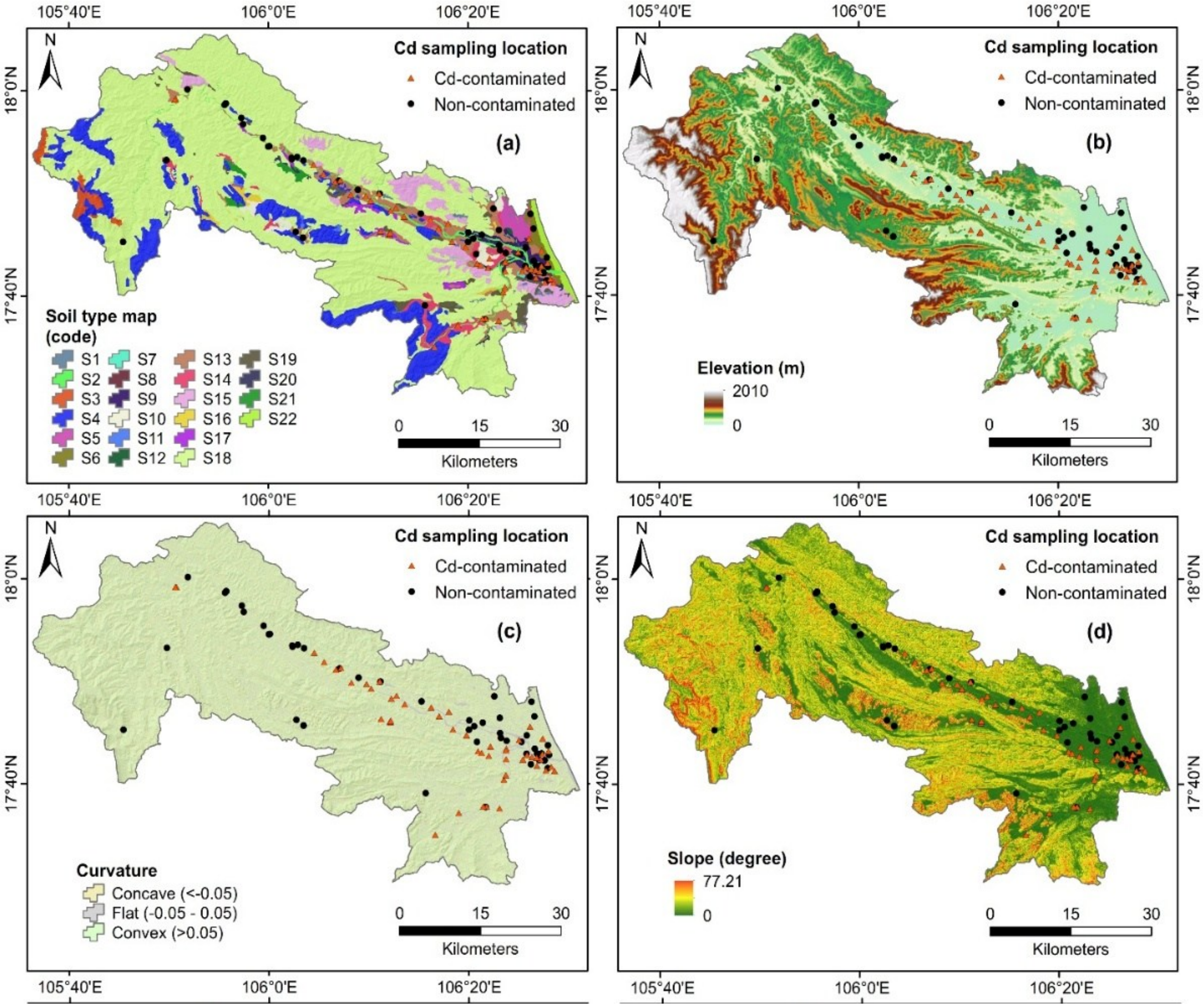

Fig. 4. The parameter condition maps represent key environmental and anthropogenic factors influencing Cd contamination in the study area, including: (a) soil type, (b) elevation, (c) curvature, (d) slope, (e) distance to roads, (f) distance to rivers, and (g–m) spectral reflectance from Landsat 8 imagery (Bands 1–7).

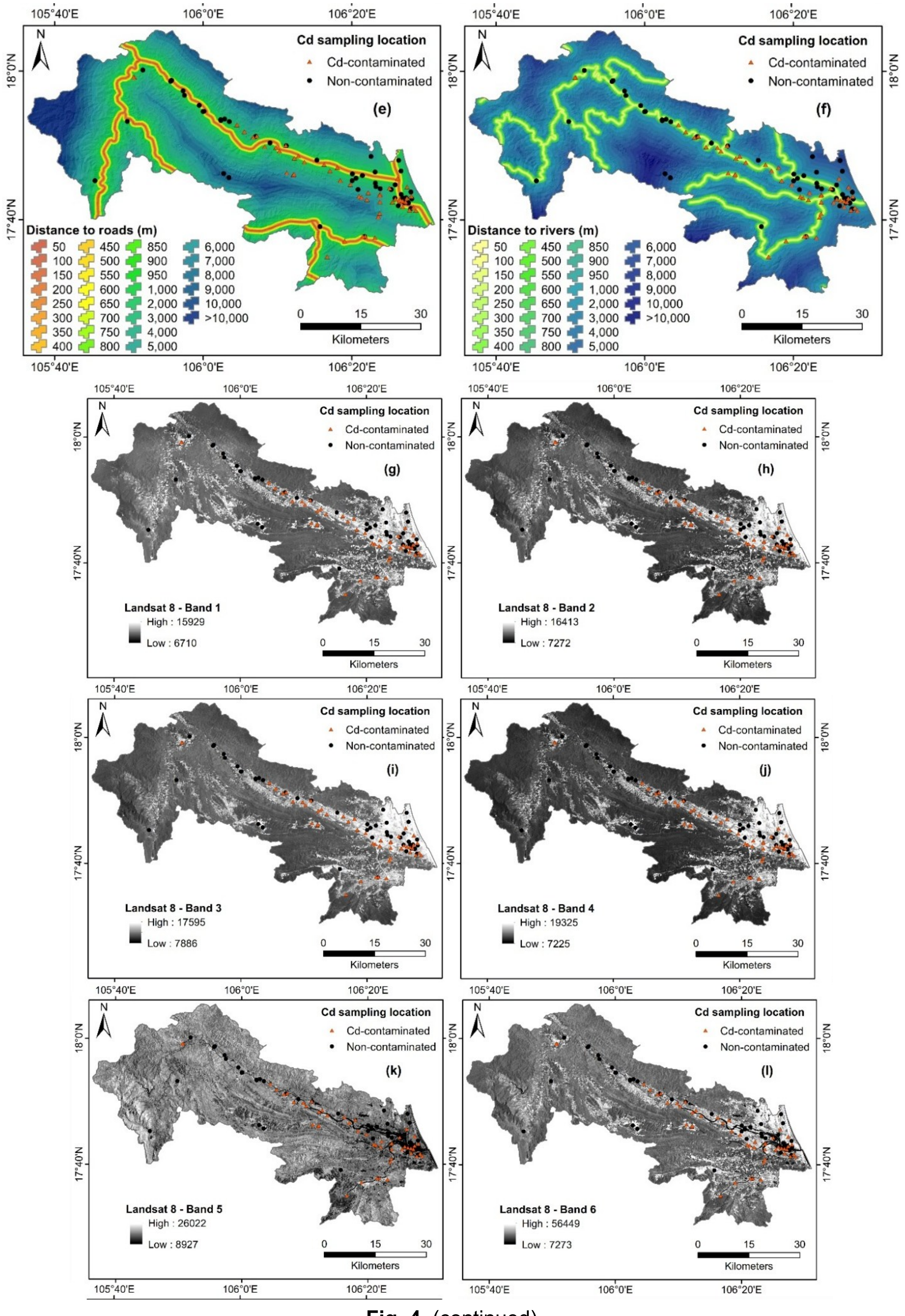

Fig. 4. (continued)

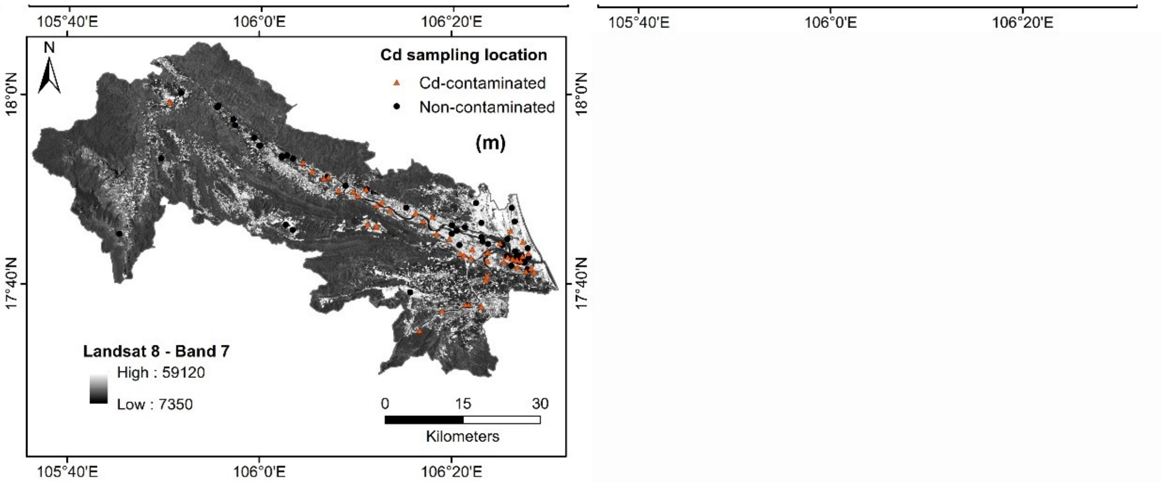

Fig. 4. (continued)

Table 4. Description of Soil Types (S1–S22) Used in the Soil Classification Map

No	Codes of Soil Types	Description
1	S1	Lakes, reservoirs
2	S2	Rivers, streams
3	S3	Humus gray soil on mountains
4	S4	Limestone
5	S5	Acidic coastal sandy soil
6	S6	Slightly acidic to neutral coastal sandy soil
7	S7	Acidic gley soil
8	S8	Highly saline soil
9	S9	Moderately to slightly saline soil
10	S10	Newly transformed acidic soil
11	S11	Typical yellow-brown soil
12	S12	Active acid sulfate soil
13	S13	Acidic alluvial soil
14	S14	Slightly acidic to neutral alluvial soil
15	S15	Thin-layered acidic soil
16	S16	Degraded gray soil
17	S17	Light-textured gray soil
18	S18	Feralitic gray soil
19	S19	Indurated gray soil
20	S20	Mottled gray soil
21	S21	Rocky gray soil
22	S22	Typical white-yellow dunes

5. Results and discussion

5.1. Ranking of conditioning factors

Fig. 5 illustrates the ranking results of the 13 conditioning factors influencing cadmium (Cd) contamination in the Gianh River Basin and surrounding areas, based on the OneR evaluation method. Among all factors, soil type was identified as the most influential, with a score of 58.57, followed by distance to rivers with a score of 55.71. The remaining factors, ranked in descending order of influence, are Band 6, Band 2, Elevation, Band 4, Band 1, Curvature, Slope, Band 5, Band 3, Band 7, and distance to roads.

This ranking helps highlight the most critical variables affecting Cd distribution and supports the refinement of the model by identifying and potentially eliminating low-impact factors. The evaluation results demonstrate that all 13 selected factors contribute to Cd pollution in the study area. Even Distance to roads, the least influential factor, has an evaluation score of 38.57. Soil type was ranked as the most influential factor (Fig. 5). Analysis of the soil type map (Fig. 4a) revealed that contaminated sites were predominantly located in the following soil classes: 12 out of 50 samples were found in the 'Typical white-yellow dunes' class, 11 out of 50 in the 'Active acid sulfate soil' class, and 10 out of 50 in the 'Newly transformed acidic soil' class. The remaining contaminated sites were scattered across other soil types. Several reasons may explain why Cd contamination tends to concentrate in these specific soil classes: The 'Typical white-yellow dunes' are characterized by low organic matter content and poor pH buffering capacity, making them prone to Cd deposition from rainwater or groundwater transported from external sources. The 'Active acid sulfate soil' type exhibits extremely low pH values (pH < 4), which increases the solubility of heavy metals, particularly Cd. The 'Newly transformed acidic soil' contains reactive forms of Fe and Mn, which are easily oxidized or reduced, creating ideal conditions for Cd to be released from mineral structures and adsorbed onto organic matter. Additionally, environmental factors such as hydrology, topography, land use practices, and anthropogenic activities in the study area may further contribute to the spatial variability of Cd distribution across different locations.

The spectral bands (Bands 1–7) of Landsat 8 imagery were selected as input variables due to their advantages in objectively capturing surface information, including geological characteristics, land cover, and soil properties. Band 1, within the violet spectral region with wavelengths ranging from 0.43 to 0.45 μm, facilitates the differentiation of shallow water and supports atmospheric correction [65, 66]. Band 2, in the blue spectral region with wavelengths from 0.45 to 0.51 μm, enables the discrimination of soil, water, and vegetation, as well as the assessment of water turbidity [65, 66]. Band 3, in the green spectral region with wavelengths from 0.53 to 0.59 μm, aids in identifying vegetation, urban areas, and sediments [65, 66]. Cadmium tends to accumulate in sediments and shallow water, particularly in areas impacted by mining, industrial, or agricultural activities [38, 67]. Bands 1 and 2 can reflect variations in water turbidity and suspended matter content, which are often associated with the presence or transport of cadmium in aquatic environments [66]. Changes in reflectance in Bands 1 and 2 may indicate soil erosion processes that carry cadmium into rivers and lakes [68]. Bare soil, sediments, or anthropogenic surfaces (e.g., urban areas, industrial waste sites) potential sources or accumulation zones for cadmium typically exhibit strong reflectance in the blue and green bands [69]. Consequently, Bands 2 and 3 can assist in classifying land use types, thereby supporting the spatial assessment of cadmium pollution risks. Band 4, in the red spectral region with wavelengths from 0.64 to 0.67 μm, is sensitive to chlorophyll absorption and is used to evaluate plant health [65, 66]. Band 5, in the near-infrared (NIR) region with wavelengths from 0.85 to 0.88 μm, exhibits strong reflectance from healthy vegetation and is employed to assess biomass density [65, 66]. Cadmium inhibits photosynthesis and induces physiological stress in plants, leading to leaf yellowing, reduced biomass, and tissue degradation [70-72]. These changes result in increased absorption in the red region (Band 4) and reduced reflectance in the NIR region (Band 5). The combined use of Bands 4 and 5 can reflect the degree of plant health deterioration due to cadmium pollution. Band 6, in the shortwave infrared 1 (SWIR1) region with wavelengths from 1.57 to 1.65 μm, provides information on soil moisture, minerals, and construction materials [65, 66]. Band 7, in the shortwave infrared 2 (SWIR2) region with wavelengths from 2.11 to 2.29 μm, supports the identification of minerals and the classification of soil and rock types [65, 66]. Bands 6 and 7 are highly sensitive to hydroxyl-bearing minerals, soil moisture, and fine-grained materials [66]. Cadmium tends to accumulate in soils with high clay content, organic matter, or iron oxides [73, 74]. These mineral phases may produce characteristic absorption signals in the SWIR region, thereby aiding in the identification of areas with high potential for cadmium accumulation.

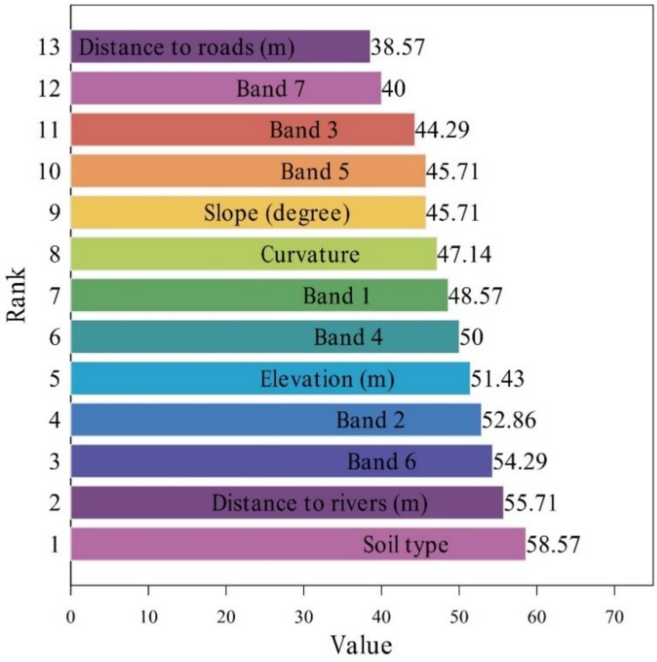

Fig. 5. Ranking of the conditioning factors based on the OneR method

The integration of Landsat 8 data thus serves to compensate for missing ground information while leveraging the 'black-box' learning capabilities of deep learning models. Feature importance ranking results indicate that all Landsat 8 bands contribute to the prediction of Cd contamination in the study area (Fig. 5). Consequently, all 13 factors were selected as input parameters for developing machine learning models.

5.2. Evaluation of the accuracy of the models

The performance of the models was evaluated using a multi-criteria assessment method. Specifically, this involved a comprehensive analysis of evaluation metrics (AUC, ACC, PPV, NPV, SST, SPF, RMSE, and Kappa) across both the training and testing datasets. On the training dataset (Fig. 6a, Table 5), Random Forest (RF) exhibited the highest performance (AUC = 1.000, ACC = 100%, PPV = 100%, NPV = 100%, SST = 100%, SPF = 100%, RMSE = 0.19, Kappa = 1), while Support Vector Machine (SVM) showed the lowest performance (AUC = 0.298, ACC = 37.14%, PPV = 37.14%, NPV = 37.14%, SST = 37.14%, SPF = 37.14%, RMSE = 0.53, Kappa = -0.26). On the testing dataset (Fig. 6b, Table 6), DNN-Opt2 demonstrated the best performance (AUC = 0.858, ACC = 73.33%, PPV = 73.33%, NPV = 73.33%, SST = 73.33%, SPF = 73.33%, RMSE = 0.45, Kappa = 0.47), with SVM again recording the lowest performance. Although DNN-Opt2 exhibited lower performance than RF on the training dataset, an examination of the ROC curve (Fig. 6a) reveals that DNN-Opt2 displayed greater stability. Additionally, the RF model exhibited evaluation metrics indicative of overfitting (AUC = 1.000, ACC = 100%, PPV = 100%, NPV = 100%, SST = 100%, SPF = 100%, Kappa = 1.0) (Table 5). Consequently, we conclude that the RF model lacks consistent performance in this study. This may be attributed to the small dataset size (100 samples, with 70 used for training), which could explain why commonly high-performing models such as LR, SVM, RF, and RBFN did not achieve high performance. For small datasets, identifying a suitable model based on evaluation metrics is essential. The DNN-Opt2 model demonstrated the most optimal performance across both the training and testing datasets. Therefore, DNN-Opt2 was selected as the model for predicting and establishing the Cd contamination potential map in the Gianh River basin.

A comparison of the variants of the deep learning model DNN (DNN-Opt1, DNN-Opt2, DNN-Opt3, DNN-Opt4) with machine learning models (LR, SVM, RF, RBFN) reveals that the DNN-Opt2 model outperforms the others (Fig. 6a, 6b, Tables 5, 6). Variations in the number of hidden layers and neurons within those layers lead to differing performance levels among the DNN models (Table

1). On the training dataset, an analysis of the metrics indicates that the DNN-Opt1 model, with 0 hidden layers and 0 neurons in the hidden layer (Table 1), shows improved performance when modified to DNN-Opt2, which has 1 hidden layer and 8 neurons in the hidden layer (Table 1) (Fig. 6a, Table 5). However, further increasing the number of hidden layers and neurons in DNN-Opt3 and DNN-Opt4 results in a decline in performance (Fig. 6a, Table 5). On the testing dataset, DNN-Opt2 exhibits the best performance among the DNN variants (DNN-Opt1, DNN-Opt3, DNN-Opt4).

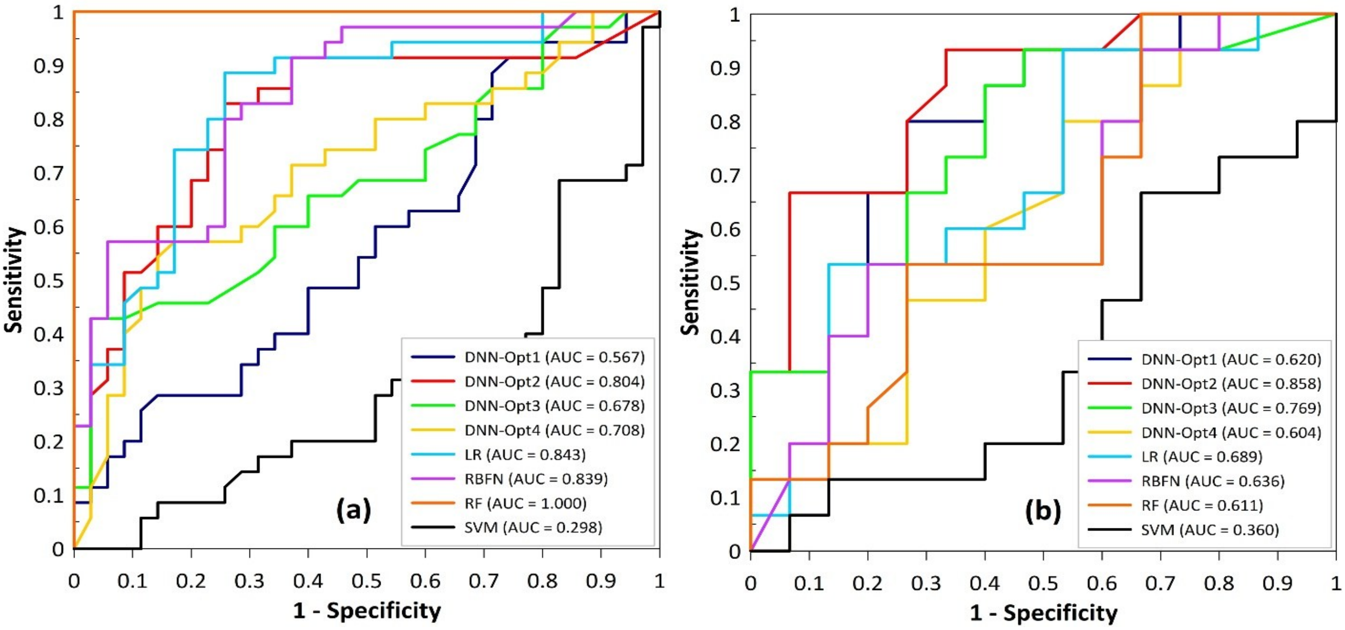

Fig. 6. Performance of the models based on ROC curve and AUC evaluation: (a) training dataset, (b) testing dataset

Table 5. The evaluation results of the training dataset

No	Parameters	Models							
		DNN-Opt1	DNN-Opt2	DNN-Opt3	DNN-Opt4	LR	RBFN	RF	SVM
1	TP	19	15	34	32	29	26	35	13
2	TN	17	32	8	14	26	28	35	13
3	FP	16	20	1	3	6	9	0	22
4	FN	18	3	27	21	9	7	0	22
5	PPV (%)	54.29	42.86	97.14	91.43	82.86	74.29	100.00	37.14
6	NPV (%)	48.57	91.43	22.86	40.00	74.29	80.00	100.00	37.14
7	SST (%)	51.35	83.33	55.74	60.38	76.32	78.79	100.00	37.14
8	SPF (%)	51.52	61.54	88.89	82.35	81.25	75.68	100.00	37.14
9	ACC (%)	51.43	67.14	60.00	65.71	78.57	77.14	100.00	37.14
10	Kappa	0.03	0.34	0.20	0.31	0.57	0.54	1.00	-0.26
11	RMSE	0.49	0.48	0.49	0.49	0.41	0.41	0.19	0.53

Typically, increasing the number of hidden layers and neurons in a deep learning model introduces more trainable parameters, which can lead to improved predictive performance. However, in this study, an opposite trend was observed: models with more complex architectures exhibited reduced performance. This outcome suggests that, under conditions of limited training data, increasing model complexity may lead to overfitting rather than performance gains. The use of multiple parameter configurations was useful in identifying an optimal model structure suited to small-sample scenarios.

Given the practical constraints in heavy metal pollution studies, particularly the high cost of soil sample collection and laboratory analysis—the findings highlight the importance of tailoring model complexity to dataset size. Future research could further explore the scalability of deep learning approaches by incorporating larger datasets to assess their potential for improved generalizability and accuracy in cadmium contamination mapping.

Table 6. The evaluation results of the testing dataset

No	Parameters	Models							
		DNN-Opt1	**DNN-Opt2**	**DNN-Opt3**	**DNN-Opt4**	**LR**	**RBFN**	**RF**	**SVM**
1	TP	10	11	7	11	8	6	8	6
2	TN	12	11	14	5	9	11	8	5
3	FP	5	4	8	4	7	9	7	9
4	FN	3	4	1	10	6	4	7	10
5	PPV (%)	66.67	73.33	46.67	73.33	53.33	40.00	53.33	40.00
6	NPV (%)	80.00	73.33	93.33	33.33	60.00	73.33	53.33	33.33
7	SST (%)	76.92	73.33	87.50	52.38	57.14	60.00	53.33	37.50
8	SPF (%)	70.59	73.33	63.64	55.56	56.25	55.00	53.33	35.71
9	ACC (%)	73.33	73.33	70.00	53.33	56.67	56.67	53.33	36.67
10	Kappa	0.47	0.47	0.40	0.07	0.13	0.13	0.07	-0.27
11	RMSE	0.48	0.45	0.48	0.49	0.52	0.52	0.49	0.53

5.3. Cd Contamination Potential map

In Vietnam, the national technical regulation on soil quality, QCVN 03:2023/BTNMT [75], stipulates maximum permissible concentrations of various soil contaminants, including cadmium (Cd). While this regulation provides a legal threshold for assessing soil contamination, the Igeo, originally proposed by Müller (1969), offers a complementary approach by evaluating contamination levels relative to local geochemical background concentrations [47]. In this study, Igeo values in the Gianh River basin were rigorously computed based on region-specific baseline data. Accordingly, Igeo was adopted as a key indicator for classifying sampling locations into 'contaminated' and 'non-contaminated' categories, providing labeled data for supervised learning model development.

Based on the model's performance evaluations presented in Section 4.2, the cadmium contamination potential map for the Gianh River Basin and its surrounding areas was generated using the DNN-Opt2 model (Fig. 7). The map is divided into five prediction classes ranging from low to high, determined by the natural breaks method [76]. In the study area, the 'very low' potential class accounts for 65.46% of the area, the 'low' class for 13.71%, the 'moderate' class for 10.52%, the 'high' class for 7.44%, and the 'very high' class for 2.87% (Fig. 7). The map results indicate that areas with high contamination potential are predominantly concentrated in urban zones and regions with high population density. The Cd contamination potential map was validated against the locations of analyzed soil samples, demonstrating highly reliable effectiveness. For the Cd-contaminated soil sample locations, 14 out of 50 samples fall within the 'very high' potential class, 16 out of 50 within the 'high' class, 20 out of 50 within the 'moderate' class, and none within the 'low' or 'very low' classes (Fig. 7). For the non-contaminated soil sample locations, 2 out of 50 samples are in the 'very low' class, 14 out of 50 in the 'low' class, 25 out of 50 in the 'moderate' class, 8 out of 50 in the 'high' class, and 1 out of 50 in the 'very high' class (Fig. 7).

There remain instances where analyzed non-contaminated soil samples are located in areas predicted as 'high' or 'very high.' This discrepancy is attributed to the prediction model's error and could be improved by increasing the sample size.

Nevertheless, the prediction classes for Cd-contaminated areas appear accurate, as no contaminated samples fall within the 'low' or 'very low' classes. Additionally, the Cd contamination potential map, constructed using the deep learning model DNN-Opt2, offers a novel approach to predicting and mapping Cd pollution risks compared to previous studies. Earlier studies on Cd pollution risk mapping typically relied on interpolating Cd content samples and then delineating contaminated zones based on Igeo index analysis. Consequently, the approach of applying machine learning to map Cd contamination potential provides a higher degree of objectivity. The Cd contamination potential map for the Gianh River basin could achieve greater accuracy with a larger number of analyzed soil samples and further testing of more optimized machine learning and deep learning models. The Cd contamination potential map for the Gianh River area serves as an effective tool for environmental managers in monitoring and mitigating Cd pollution and also provides critical spatial information for groundwater quality assessment and informed decision-making in infrastructure development, particularly in areas susceptible to heavy metal accumulation.

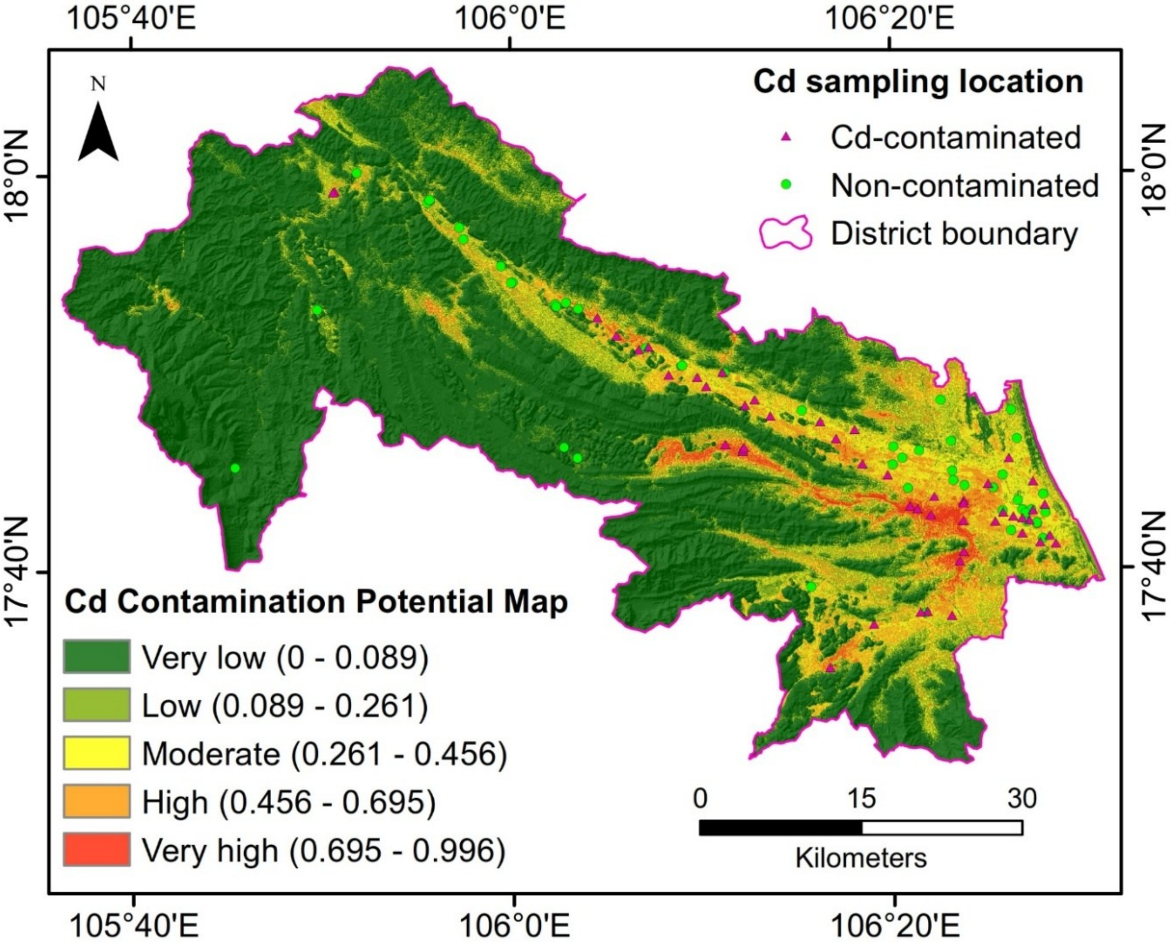

Fig. 7. Cadmium contamination potential map in surface soil of Gianh river basin using DNN-Opt2 model

6. Conclusion

This study presents a robust and adaptable AI-driven framework for assessing cadmium contamination potential in surface soils, with a focus on applications relevant to civil engineering. Through a comparative analysis of machine learning and deep learning models, we demonstrated that structural optimization especially in terms of hidden layers and neuron configuration significantly affects model performance. The DNN-Opt2 model emerged as the most effective approach for the study area.

The resulting contamination potential map has practical implications for civil and environmental engineers, particularly in guiding the assessment of risks related to groundwater contamination and the durability of infrastructure in affected zones. Cd infiltration into groundwater can compromise potable water supplies, while prolonged soil contamination may accelerate structural degradation due to chemical interactions with foundations and construction materials.

Acknowledgments

This research was financially supported by the Department of Science and Technology of Quang Binh province under a grant project “Research and evaluation of the current status of heavy metal distribution in the Gianh riverside area, Quang Binh province”.

References

[1] B.J. Alloway. (2013). Sources of Heavy Metals and Metalloids in Soils. In: B.J. Alloway (Ed.), Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability. *Dordrecht: Springer*, pp. 11-50.

[2] M. Jaishankar, T. Tseten, N. Anbalagan, B.B. Mathew, K.N. Beeregowda. (2014). Toxicity, mechanism and health effects of some heavy metals. *Interdiscip Toxicol*, 7(2), 60-72.

[3] B. Ma, W. Song, X. Zhang, M. Chen, J. Li, X. Yang, L. Zhang. (2023). Potential application of novel cadmium-tolerant bacteria in bioremediation of Cd-contaminated soil. *Ecotoxicology and Environmental Safety*, 255, 114766.

[4] S.H.J. Sari, M.-F. Chien, C. Inoue. (2022). Subcellular localization and chemical speciation of Cd in Arabidopsis halleri ssp. gemmifera to reveal its hyperaccumulating and detoxification strategies. *Environmental and Experimental Botany*, 203, 105047.

[5] H. Zhao, Y. Wu, X. Lan, Y. Yang, X. Wu, L. Du. (2022). Comprehensive assessment of harmful heavy metals in contaminated soil in order to score pollution level. *Scientific Reports*, 12, 3552.

[6] X. Li, Y. Li, X. Zhu, X. Gui, C. Ma, W. Peng, Y. Li, Y. Zhang, W. Huang, D. Hua, S. Jia, M. Wu. (2022). Evaluation of the cadmium phytoextraction potential of tobacco (Nicotiana tabacum) and rhizosphere micro-characteristics under different cadmium levels. *Chemosphere*, 286, 131714.

[7] Q. Liu, Y. Sheng, W. Wang, X. Liu. (2021). Efficacy and microbial responses of biochar-nanoscale zero-valent during in-situ remediation of Cd-contaminated sediment. *Journal of Cleaner Production*, 287, 125076.

[8] H. Chen, X. Yang, P. Wang, Z. Wang, M. Li, F.-J. Zhao. (2018). Dietary cadmium intake from rice and vegetables and potential health risk: A case study in Xiangtan, southern China. *Science of The Total Environment*, 639, 271-277.

[9] A.A. Meharg, G. Norton, C. Deacon, P. Williams, E.E. Adomako, A. Price, Y. Zhu, G. Li, F.-J. Zhao, S. McGrath, A. Villada, A. Sommella, P.M.C.S. De Silva, H. Brammer, T. Dasgupta, M.R. Islam. (2013). Variation in Rice Cadmium Related to Human Exposure. *Environmental Science & Technology*, 47(11), 5613-5618.

[10] T.T. Le, K.-W. Kim, D.Q. Nguyen, H.T.T. Ngo. (2023). Trace element contamination in rice and its potential health risks to consumers in North-Central Vietnam. *Environmental Geochemistry and Health*, 45, 3361-3375.

[11] H.T.T. Ngo, L.A.T. Tran, D.Q. Nguyen, T.T.H. Nguyen, T.T. Le, Y. Gao. (2021). Metal Pollution and Bioaccumulation in the Nhue-Day River Basin, Vietnam: Potential Ecological and Human Health Risks. *International Journal of Environmental Research and Public Health*, 18(24), 13425.

[12] H.M. Thang, P.Q. Ha. (2016). Effect of Cadmium (Cd) Content in Soil on the Accumulation of Cadmium in Mustard Greens Grown on Red River Alluvial Soil (in Vietnam). *Journal of Vietnam Agricultural Science and Technology*, 10, 63-66.

[13] V.V. Thiep, T.T. Hung, N.T.H. Binh. (2021). Cadmium Content in Mottled Spinefoot (Siganus fuscescens) and Initial Risk Assessment to Consumer's Health at the Coastal Zone in Quang Binh Province (in Vietnam). *Vietnam Journal of Agricultural Sciences*, 19, 923-931.

[14] T.S. Tran, V.C. Dinh, T.A.H. Nguyen, K.-W. Kim. (2022). Soil contamination and health risk assessment from heavy metals exposure near mining area in Bac Kan province, Vietnam. *Environmental Geochemistry and Health*, 44, 1189-1202.

[15] F. Huang, J. Hu, L. Chen, Z. Wang, S. Sun, W. Zhang, H. Jiang, Y. Luo, L. Wang, Y. Zeng, L. Fang. (2023). Microplastics may increase the environmental risks of Cd via promoting Cd uptake by plants: A meta-analysis. *Journal of Hazardous Materials*, 448, 130887.

[16] J. Zhang, X. Cao, Z. Yao, Q. Lin, B. Yan, X. Cui, Z. He, X. Yang, C.-H. Wang, G. Chen. (2021). Phytoremediation of Cd-contaminated farmland soil via various Sedum alfredii-oilseed rape cropping systems: Efficiency comparison and cost-benefit analysis. *Journal of Hazardous Materials*, 419, 126489.

[17] A. Larrose, A. Coynel, J. Schäfer, G. Blanc, L. Massé, E. Maneux. (2010). Assessing the current state of the Gironde Estuary by mapping priority contaminant distribution and risk potential in surface sediment. *Applied Geochemistry*, 25(12), 1912-1923.

[18] B.T. Pham, C. Luu, D.V. Dao, T.V. Phong, H.D. Nguyen, H.V. Le, J. von Meding, I. Prakash. (2021). Flood risk assessment using deep learning integrated with multi-criteria decision analysis. *Knowledge-Based Systems*, 219, 106899.

[19] B.-Y. Huang, Q.-X. Lü, Z.-X. Tang, Z. Tang, H.-P. Chen, X.-P. Yang, F.-J. Zhao, P. Wang. (2024). Machine learning methods to predict cadmium (Cd) concentration in rice grain and support soil management at a regional scale. *Fundamental Research*, 4(5), 1196-1205.

[20] B. Zhao, W. Zhu, S. Hao, M. Hua, Q. Liao, Y. Jing, L. Liu, X. Gu. (2023). Prediction heavy metals accumulation risk in rice using machine learning and mapping pollution risk. *Journal of Hazardous Materials*, 448, 130879.

[21] K.N. Palansooriya, J. Li, P.D. Dissanayake, M. Suvarna, L. Li, X. Yuan, B. Sarkar, D.C.W. Tsang, J. Rinklebe, X. Wang, Y.S. Ok. (2022). Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. *Environmental Science & Technology*, 56(7), 4187-4198.

[22] X. Lu, L. Sun, Y. Zhang, J. Du, G. Wang, X. Huang, X. Li, X. Wang. (2024). Predicting Cd accumulation in crops and identifying nonlinear effects of multiple environmental factors based on machine learning models. *Science of The Total Environment*, 951, 175787.

[23] X. Jia, D. O'Connor, Z. Shi, D. Hou. (2021). VIRS based detection in combination with machine learning for mapping soil pollution. *Environmental Pollution*, 268, 115845.

[24] S. Zhong, K. Zhang, M. Bagheri, J.G. Burken, A. Gu, B. Li, X. Ma, B.L. Marrone, Z.J. Ren, J. Schrier, W. Shi, H. Tan, T. Wang, X. Wang, B.M. Wong, X. Xiao, X. Yu, J.-J. Zhu, H. Zhang. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science & Technology*, 55(19), 12741-12754.

[25] Z.M. Yaseen. (2021). An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere*, 277, 130126.

[26] T.V. Phong, B.T. Pham. (2023). Performance of Naïve Bayes Tree with ensemble learner techniques for groundwater potential mapping. *Physics and Chemistry of the Earth, Parts A/B/C*, 132, 103503.

[27] T.X. Bien, P.T. Trinh, L.T. Duong, T.V. Phong, V.H. Nhat, H.V. Le, D.D. Nguyen, I. Prakash, P.T. Tam, B.T. Pham. (2024). Groundwater potential zoning using Logistics

Model Trees based novel ensemble machine learning model. *Vietnam Journal of Earth Sciences*, 46(2), 272-281.

[28] T.V. Phong, B.T. Pham, P.T. Trinh, H.-B. Ly, Q.H. Vu, L.S. Ho, H.V. Le, L.H. Phong, M. Avand, I. Prakash. (2021). Groundwater Potential Mapping Using GIS-Based Hybrid Artificial Intelligence Methods. *Groundwater*, 59(5), 745-760.

[29] D.V. Dao, A. Jaafari, M. Bayat, D. Mafi-Gholami, C. Qi, H. Moayedi, T.V. Phong, H.-B. Ly, T.-T. Le, P.T. Trinh, C. Luu, N.K. Quoc, B.N. Thanh, B.T. Pham. (2020). A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *CATENA*, 188, 104451.

[30] T.A. Tuan, P.V. Hong, T.T. Tam, N.T.A. Nguyet, N.V. Dung, P.T. Huy, T.V. Phong. (2024). Landslide susceptibility in Phuoc Son, Quang Nam: A deep learning approach. *Vietnam Journal of Earth Sciences*, 47(1), 39-57.

[31] C. Luu, B.T. Pham, T.V. Phong, R. Costache, H.D. Nguyen, M. Amiri, Q.D. Bui, L.T. Nguyen, H.V. Le, I. Prakash, P.T. Trinh. (2021). GIS-based ensemble computational models for flood susceptibility prediction in the Quang Binh Province, Vietnam. *Journal of Hydrology*, 599, 126500.

[32] T. Tuat, N.D. Nhat. (1981). Overview of hydrogeography of rivers in Vietnam, Part one, North (in Vietnam). *Science and Technology Publishing House*.

[33] PCQBP. (2017). Report on management of main wastewater sources in river basins in Quang Binh province, Quang Binh, Vietnam.

[34] H.T. Trang, N.D. Luyen, D.D. Huyen. (2019). Research on Gianh River water quality assessment (in Vietnam). *Journal of Science, College of Education, Hue University*, 03, 93-100.

[35] J. Bai, B. Cui, X. Xu, Q. Ding, H. Gao. (2009). Heavy Metal Contamination in Riverine Soils Upstream and Downstream of a Hydroelectric Dam on the Lancang River, China. *Environmental Engineering Science*, 26(5), 941-946.

[36] P. Pavlović, M. Mitrović, D. Đorđević, S. Sakan, J. Slobodnik, I. Liška, B. Csanyi, S. Jarić, O. Kostić, D. Pavlović, N. Marinković, B. Tubić, M. Paunović. (2016). Assessment of the contamination of riparian soil and vegetation by trace metals — A Danube River case study. *Science of The Total Environment*, 540, 396-409.

[37] C. Ye, S. Li, Y. Zhang, Q. Zhang. (2011). Assessing soil heavy metal pollution in the water-level-fluctuation zone of the Three Gorges Reservoir, China. *Journal of Hazardous Materials*, 191(1-3), 366-372.

[38] A. Kubier, R.T. Wilkin, T. Pichler. (2019). Cadmium in soils and groundwater: A review. *Applied Geochemistry*, 108, 104388.

[39] WHO. (2021). Cadmium. In Chemical hazards in drinking-water (WHO Guidelines for Drinking-water Quality).

[40] Ed Mccurdy. (2014). Introducing the New Agilent 7900 Quadrupole ICP-MS: performance and technology. *Agilent ICP-MS Journal,* 2014, 2-3.

[41] I. Suhani, S. Sahab, V. Srivastava, R.P. Singh. (2021). Impact of cadmium pollution on food safety and human health. *Current Opinion in Toxicology*, 27, 1-7.

[42] A. Mahar, A. Ali, A.H. Lahori, F. Wahid, R. Li, M. Azeem, S. Fahad, M. Adnan, Rafiullah, I.A. Khan, Z. Zhang. (2020). Promising Technologies for Cd-Contaminated Soils: Drawbacks and Possibilities. *In: Fahad, S., Hasanuzzaman, M., Alam, M., Ullah, H., Saeed, M., Ali Khan, I., Adnan, M. (Eds.), Environment, Climate, Plant and Vegetation Growth. Springer International Publishing, Cham*, pp. 63-91.

[43] T.R. Loveland, J.R. Irons. (2016). Landsat 8: The plans, the reality, and the legacy. *Remote Sensing of Environment*, 185, 1-6.

[44] M.B. Kirkham. (2006). Cadmium in plants

on polluted soils: Effects of soil factors, hyperaccumulation, and amendments. *Geoderma*, 137(1-2), 19-32.

[45] N.A. Shafie, A.Z. Aris, M.P. Zakaria, H. Haris, W.Y. Lim, N.M. Isa. (2013). Application of geoaccumulation index and enrichment factors on the assessment of heavy metal pollution in the sediments. *Journal of Environmental Science and Health, Part A*, 48(2), 182-190.

[46] G.M.A. Bermudez, R. Jasan, R. Plá, M.L. Pignata. (2012). Heavy metals and trace elements in atmospheric fall-out: Their relationship with topsoil and wheat element composition. *Journal of Hazardous Materials*, 213-214, 447-456.

[47] G. Muller. (1969). Index of Geoaccumulation in Sediments of the Rhine River. *GeoJournal*, 2, 108-118.

[48] M.J. Nasir, A. Wahab, T. Ayaz, S. Khan, A.Z. Khan, M. Lei. (2023). Assessment of heavy metal pollution using contamination factor, pollution load index, and geoaccumulation index in Kalpani River sediments, Pakistan. *Arabian Journal of Geosciences*, 16, 143.

[49] N.T. Hoan, H.L. Thu, N.V. Dung, H.T. Quynh, N.K. Anh, L.D. Hanh, P.V. Duan, T.T. Phan, T.V. Phong. (2025). Forest cover change mapping based on Deep Neuron Network, GIS, and High-resolution Imagery. *Vietnam Journal of Earth Sciences*, 151-175.

[50] H. Hussain, P.S. Tamizharasan, C.S. Rahul. (2022). Design possibilities and challenges of DNN models: a review on the perspective of end devices. *Artificial Intelligence Review*, 55, 5109-5167.

[51] S. Wang, T. Zhou, J. Bilmes. (2019). Bias Also Matters: Bias Attribution for Deep Neural Network Explanation. *In: Kamalika, C., Ruslan, S. (Eds.), Proceedings of the 36th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research*, pp. 6659-6667.

[52] A. Ashiquzzaman, A.K. Tushar, R. Islam MD, Shon D, Kichang LM, J.-H. Park, D.-S. Lim, J. Kim. (2018). Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network. In: K.J. Kim, H. Kim, N. Baek, (Eds.), IT Convergence and Security 2017. *Springer, Singapore*, pp. 35-43.

[53] M. Ibnu Choldun R, J. Santoso, K. Surendro. (2020). Determining the Number of Hidden Layers in Neural Network by Using Principal Component Analysis. *In: Bi, Y., Bhatia, R., Kapoor, S. (Eds.), Intelligent Systems and Applications. Springer International Publishing, Cham*, pp. 490-500.

[54] X. Zou, Y. Hu, Z. Tian, K. Shen. (2019). Logistic Regression Model Optimization and Case Analysis. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 135-139.

[55] D.J. Foster, S. Kale, H. Luo, M. Mohri, K. Sridharan. (2018). Logistic Regression: The Importance of Being Improper, *In: Sébastien, B., Vianney, P., Philippe, R. (Eds.), Proceedings of the 31st Conference On Learning Theory. PMLR, Proceedings of Machine Learning Research*, pp. 167-208.

[56] M.E.A. Budimir, P.M. Atkinson, H.G. Lewis. (2015). A systematic review of landslide probability mapping using logistic regression. Landslides, 12, 419-436.

[57] A.E.M. Al-Juaidi, A.M. Nassar, O.E.M. Al-Juaidi. (2018). Evaluation of flood susceptibility mapping using logistic regression and GIS conditioning factors. *Arabian Journal of Geosciences*, 11, 765.

[58] J. Ghosh, A. Nag. (2001). An Overview of Radial Basis Function Networks. *In: Howlett, R.J., Jain, L.C. (Eds.), Radial Basis Function Networks 2: New Advances in Design. Physica-Verlag HD, Heidelberg*, pp. 1-36.

[59] Y. Wu, H. Wang, B. Zhang, K.-L. Du. (2012). Using Radial Basis Function Networks for Function Approximation and Classification. *International Scholarly Research Notices*, 2012, 324194.

[60] L. Breiman. (2001). Random Forests.

Machine Learning, 45, 5-32.

[61] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43(6), 1947-1958.

[62] S. Suthaharan. (2016). Support Vector Machine. *In: Suthaharan, S. (Ed.), Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Springer US, Boston, MA*, pp. 207-235.

[63] S. Cheong, S.H. Oh, S.-Y. Lee. (2004). Support Vector Machines with Binary Tree Architecture for Multi-Class Classification. *Neural Information Processing - Letters and Reviews*, 2(3), 47-51.

[64] R. Burbidge, B. Buxton. (2001). An introduction to support vector machines for data mining.

[65] D.P. Roy, M.A. Wulder, T.R. Loveland, C.E. Woodcock, R.G. Allen, M.C. Anderson, D. Helder, J.R. Irons, D.M. Johnson, R. Kennedy, T.A. Scambos, C.B. Schaaf, J.R. Schott, Y. Sheng, E.F. Vermote, A.S. Belward, R. Bindschadler, W.B. Cohen, F. Gao, J.D. Hipple, P. Hostert, J. Huntington, C.O. Justice, , A. Kilic, V. Kovalskyy, Z.P. Lee, L. Lymburner, J.G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R.H. Wynne, Z. Zhu. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145, 154-172.

[66] USGS. (2020). Landsat 8 (L8) Data Users Handbook. U.S. Geological Survey, Sioux Falls, SD.

[67] Q. Tang, L. Chang, Q. Wang, C. Miao, Q. Zhang, L. Zheng, Z. Zhou, Q. Ji, L. Chen, H. Zhang. (2023). Distribution and accumulation of cadmium in soil under wheat-cultivation system and human health risk assessment in coal mining area of China. *Ecotoxicology and Environmental Safety*, 253, 114688.

[68] V. Lovynska, B. Bayat, R. Bol, S. Moradi, M. Rahmati, R. Raj, S. Sytnyk, O. Wiche, B. Wu, C. Montzka. (2024). Monitoring Heavy Metals and Metalloids in Soils and Vegetation by Remote Sensing: A Review. *Remote Sensing*, 16(17), 3221.

[69] D.I. Rukhovich, P.V. Koroleva, A.D. Rukhovich, M.A. Komissarov. (2023). Informativeness of the Long-Term Average Spectral Characteristics of the Bare Soil Surface for the Detection of Soil Cover Degradation with the Neural Network Filtering of Remote Sensing Data. *Remote Sensing*, 15(1), 124.

[70] F.U. Haider, C. Liqun, J.A. Coulter, S.A. Cheema, J. Wu, R. Zhang, M. Wenjun, M. Farooq. (2021). Cadmium toxicity in plants: Impacts and remediation strategies. *Ecotoxicology and Environmental Safety*, 211, 111887.

[71] C. Loix, M. Huybrechts, J. Vangronsveld, M. Gielen, E. Keunen, A. Cuypers. (2017). Reciprocal Interactions between Cadmium-Induced Cell Wall Responses and Oxidative Stress in Plants. *Frontiers in Plant Science*, 8, 1867.

[72] S. Soni, A.B. Jha, R.S. Dubey, P. Sharma. (2024). Mitigating cadmium accumulation and toxicity in plants: The promising role of nanoparticles. *Science of The Total Environment*, 912, 168826.

[73] J. Liu, R. Zhu, L. Ma, H. Fu, X. Lin, S.C. Parker, M. Molinari. (2021). Adsorption of phosphate and cadmium on iron (oxyhydr)oxides: A comparative study on ferrihydrite, goethite, and hematite. *Geoderma*, 383, 114799.

[74] H.U. Rahim, W.A. Akbar, J.M. Alatalo. (2022). A Comprehensive Literature Review on Cadmium (Cd) Status in the Soil Environment and Its Immobilization by Biochar-Based Materials. *Agronomy*, 12(4), 877.

[75] MONRE. (2023). National technical

regulation on Soil quality. QCVN 03:2023/BTNMT. *In: Environment, M.o.N.R.a. (Ed.), Hanoi*.

[76] J. Chen, S.T. Yang, H.W. Li, B. Zhang, J.R. Lv. (2013). Research on Geographical Environment Unit Division Based on the Method of Natural Breaks (Jenks). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. XL-4/W3, 47-50.