Journal of Radioanalytical and Nuclear Chemistry
https://doi.org/10.1007/s10967-023-09281-w

Machine learning application for radon release prediction from the copper ore mining in Sin Quyen, Lao Cai, North Vietnam

Tran Dinh Bao[1,2] · Trong Vu[3] · Nguyen Tai Tue[4,5] · Tran Dang Quy[4,5] · Thuy Huong Ngo Thi[6] · Gergely Toth[7] · Zsolt Homoki[7] · Tibor Kovacs[7] · Van-Hao Duong[8]

Received: 16 September 2023 / Accepted: 13 November 2023

Abstract

The radon release prediction from radioactive-bearing mines during mineral processing and mining is an essential target. A simple one-hidden-layer artificial neural network (ANN) model was designed with low computation cost to train, reference and get optimum effectiveness in comparison with two-hidden-layer ANN, random forest and support vector machine models which was applied for Sin Quyen copper deposit. The result showed with values of $MAPE = 1.12(\%)$, $RMSE = 2.79(Bq/m^3)$, $MABE = 2.10(\%)$, $R^2 = 0.990$, $r = 0.99$, for training part; $MAPE = 1.12(\%)$, $RMSE = 2.79(Bq/m^3)$, $MABE = 2.09(\%)$, $R^2 = 0.995$, $r = 0.997$ for testing part. The gamma dose and distance were significantly more effective variables for the radon prediction than direction, coordinate, and uranium concentration factors.

Keywords Radon prediction · Sin Quyen · Machine learning · One-hidden-layer · ANN

Introduction

The modern environment is being severely polluted by toxic natural components, recent tectonic activities, mining of mineral resources and other human activities [1–4]. One of the toxic natural components are radionuclides such as ^{222}Rn, ^{220}Rn, ^{210}Po, ^{210}Pb, ^{226}Ra, ^{228}Ra, ^{238}U, ^{40}K and ^{232}Th. Humans are continuously exposed to natural radiation worldwide, moreover, are subjected to health risks when exposed to high levels of natural radiation in the long term [5, 6], especially where mines containing high concentrations of radionuclides are located. For example, the ^{222}Rn (radon) activity concentration was reported to be as high as 920 Bq/m^3 in the air surrounding the rare earth mines Nam

✉ Van-Hao Duong
haodnth@gmail.com

Tran Dinh Bao
trandinhbao@humg.edu.vn

Trong Vu
trongvu@qui.edu.vn

Nguyen Tai Tue
tuenguyentai@hus.edu.vn

Tran Dang Quy
quytrandang@hus.edu.vn

Thuy Huong Ngo Thi
huong.ngothithuy@phenikaa-uni.edu.vn

Gergely Toth
toth.gergely@mk.uni-pannon.hu

Zsolt Homoki
homoki.zsolt@osski.hu

Tibor Kovacs
kovacs.tibor@mk.uni-pannon.hu

[1] Hanoi University of Mining and Geology (HUMG), Duc Thang, Bac Tu Liem, Hanoi, Vietnam

[2] Innovations for Sustainable and Responsible Mining (ISRM) Research Group, Hanoi University of Mining and Geology, Duc Thang, Bac Tu Liem, Hanoi, Vietnam

[3] Faculty of Mining and Construction, Quang Ninh University of Industry, Yen Tho, Dong Trieu, Quang Ninh, Vietnam

[4] VNU Key Laboratory of Geoenvironment and Climate Change Response, University of Science, Vietnam National University, Hanoi, Vietnam

[5] Faculty of Geology, University of Science, Vietnam National University, Hanoi, Vietnam

[6] Faculty of Biotechnology, Chemistry and Environmental Engineering, Phenikaa University, Hanoi 12116, Vietnam

[7] Institute of Radiochemistry and Radioecology, University of Pannonia, Veszprem, Hungary

[8] VNU School of Interdisciplinary Studies, National University, 144 Xuan Thuy Street, Cau Giay District, Hanoi 100000, Vietnam

Published online: 07 December 2023

Xe and Dong Pao in North Vietnam [7]; those of ^{238}U, ^{40}K and ^{232}Th in soils at a uranium mine in the west of Namibia were reported to be up to 1752, 1300 and 1866 Bq/kg, respectively [8]; while those of ^{238}U and ^{232}Th could be up to 6,000 and 24,000 Bq/kg in beach sands at Madena in Madagascar [9, 10]. ^{222}Rn is a noble gas with the longest half-life of all naturally occurring radioactive gases, namely 3.8 days, and a progeny in the ^{238}U decay chain. When released into the atmosphere in confined spaces such as houses, caves and mines, the internal radiation exposures reported are high [7, 11–13]. Radon is well-known as one of the most carcinogenic and radiotoxic radionuclides. The alpha particles emitted from radon are responsible for a significant amount of human exposure to ionizing radiation, consisting of about 52% of the global average annual effective dose from natural radiation of 2.4 mSv [14]. Radon has been associated with epidemiological evidence of lung cancer for individuals exposed to high doses [12].

Radon studies are attractive to many scientists worldwide, e.g. for monitoring, dispersion modelling, predicting and surveying hypothetical models of radon [7, 15–23]. Several studies, including geogenic radon potential mapping studies [24–27], have applied ANNs, machine learning, decision tree models or probabilistic and deep learning algorithms based on radon monitoring and observing anomalies in radon time series [28–30]. Although Van Hao et al. (2021) constructed an ANN model containing two-hidden-layers with promising results and low prediction errors [5], an additional step was required to reduce the amount of data prior to training the ANN model, potentially leading to overfitting due to its complex structure.

In this paper, an attempt is made to improve the efficiency of the ANN model method to predict radon release. The model uses a simple one-hidden-layer structure to reduce the computational costs of training and reference with a higher degree of accuracy but lower prediction errors. The development and optimization of the proposed model was based on three main steps: (1) dataset collection and analysis; (2) testing and training models to optimize an ANN model and model predictive capability; (3) evaluating the model (model comparison) and sensitivity analysis. The study attempts to apply and optimize the machine learning method with a simple one-hidden-layer ANN to predict radon release from areas with high levels of background radiation. The Sin Quyen deposit in North Vietnam is used as a case study (Fig. 1). This method could be faster and more accurate as well as reduce computational costs for monitoring and predicting radon contamination. Furthermore, a methodology to manage radioactive contamination, make a radiation risk assessment, protect human health as well as promote sustainable socioeconomic development during the mining and mineral processing of radionuclide-bearing natural resources in the vicinity of areas exhibiting high levels of natural background radiation in addition to highly radioactive material is proposed.

Materials and method

Dataset

The dataset is comprised of more than 1 million data points which consist of radon concentrations (Bq/m^3) and five input variables, namely X,Y coordinates (m), gamma dose rate (μSv/h), distance (m), direction (degree) and uranium concentration (ppm). The data was mainly collected during the years of 2013–2014 and 2021. The uranium concentration at/or in the vicinity of the deposit reflects the radon emitted from ore/soil/rock which is released as well as distributed to the radon measuring points. The uranium data were measured and surveyed using a gamma spectrometer (Gamma Surveyor of GF Instruments) over a 3 × 3 m grid. Uranium concentrations were measured over an area of 350 × 1250 m at the Sin Quyen mining site denoted by the red rectangle (Fig. 1). The Sin Quyen deposit has been exploited since 2006 with several million tons of the more than 50-million-ton copper ore reserve extracted annually [31, 32]. The radon concentration was cumulatively measured over three months in 21 dwellings surrounding the mine using a CR-39 detector denoted by the pink polygon (radon test area) in the Fig. 1. The CR-39 detector for measuring the cumulative radon concentration in dwellings is very useful and high precise in comparison with other methods for air radon measurement of this methodology. The radon may have originated from local and foreign sources, namely the Sin Quyen deposit, mining and processing activities. The radon measured by the CR-39 detector could reflect, measure and show the approximate cumulative average value over three months. This value was the most suitable to calculate the effective dose originating from radon as well as perhaps avoid influential factors such as the weather, meteorology (rain, humidity, wind, temperature and pressure), topography and geomorphology. The gamma dose rate was also measured at similar points with radon measurement. The input data concerning the gamma dose rate included other technical highlights reflecting the local background radon level and the contribution local influential factors make regarding radon. The other input variables of X,Y coordinates (m), distance (m) and direction (degree) were recorded by a GPS and calculated. The statistical histograms of the input variables are shown in Fig. 2. The linear relationship between each pair of input and output variables is also determined using the heat map in Fig. 3. The darker the color, the more negative the correlation between the variables is and vice versa. Clearly, the radon concentration exhibits a weak correlation with other

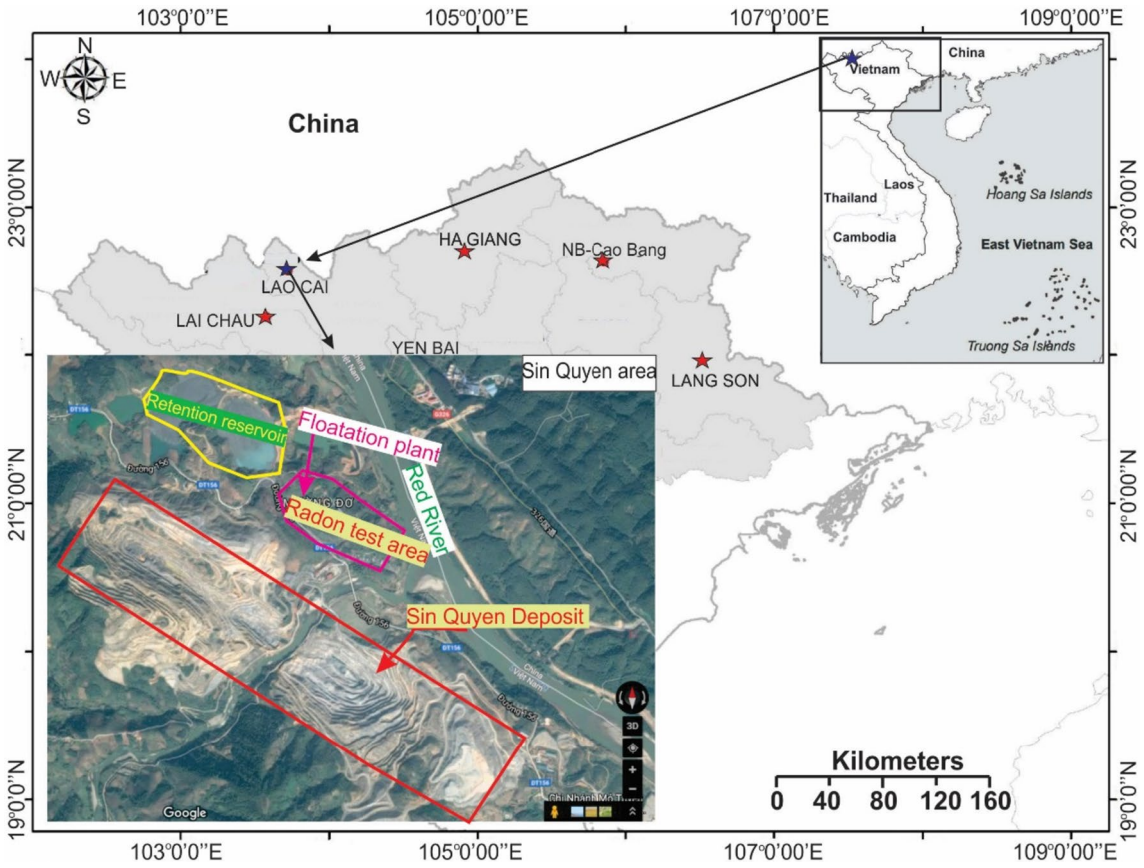


Fig. 1 A map of the study area. The area denoted by a red rectangle is where uranium measurements were made in the Sin Quyen deposit (modified from [5])

variables, suggesting a non-linear relationship between the given input and output data.

Since the scaling and distribution of input data are mutually different to each other, they are scaled between 0 and 1 using Eq. (1):

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where x (x_1, …, x_n) denotes the set of measured values and z_i represents the *i-th* normalized data.

Artificial neural network

Machine learning is widely used in many fields such as when studying geological hazards, climate change and remote sensing as well as assessing environmental pollution [6, 33–38]. The ANN functions as a human brain consisting of billions of highly connected neurons or nodes [39, 40]. Each neuron takes input signals from other neurons and sends the output signals to others. Even though usually an ANN is a complex, layered structure consisting of both an input and output layer with multiple hidden layers, an ANN with only one-hidden-layer is sufficient to approximate any continuous function uniformly [41]. In this study, the radon concentration is considered to be a function of six independent parameters in the ANN model. The proposed architecture of the ANN by denoting the optimum number of neurons in the hidden layer as S is illustrated in Fig. 4. The next section deals with the determination of S in detail:

The input vector of the network in Fig. 4 can be assumed to be vector P in Eq. (2):

$$P = (p_1, p_2, p_3, p_4, p_5, p_6) \tag{2}$$

Each element p_i of the input vector P is connected to each neuron n_j in a hidden layer through the weight w_{ij}. The total weight of this neuron is comprised of the input layer and bias b_j as shown in Eq. 3 before the sum is passed through a transfer function f_h to produce the output a_j given in Eqs. 4 and 5, respectively.

$$s_j = \sum_{i=1}^{6} w_{ij} p_i + b_j \quad j = 1,\ 2,\ ..S, \tag{3}$$

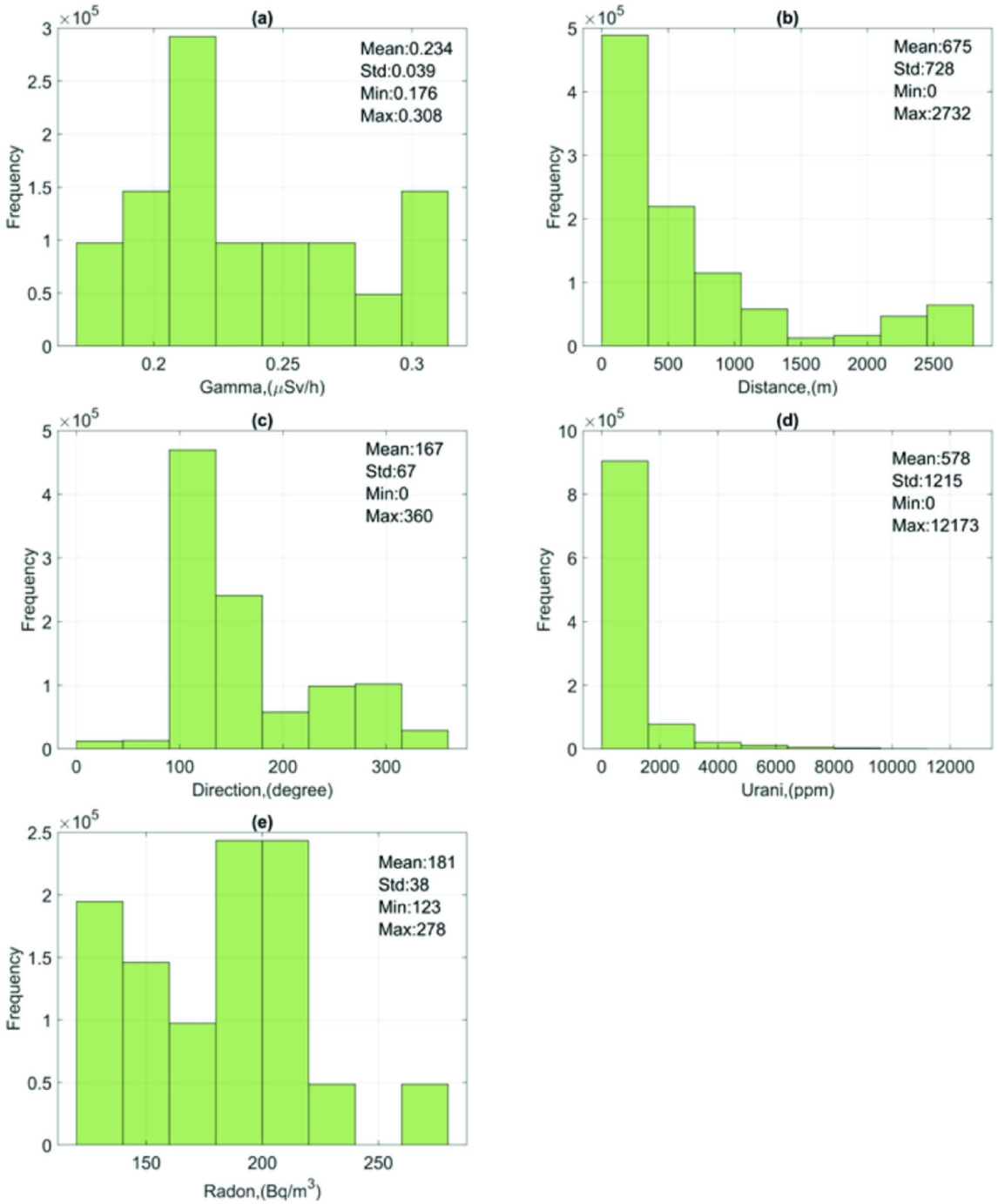


Fig. 2 Histograms with the statistical values of gamma dose rate (**a**), distance (**b**), direction (**c**), uranium (**d**) and radon concentrations (e)

$$f_h(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{4}$$

$$a_j = f_h(s_j) \quad j = 1,\ 2,\ \ldots S, \tag{5}$$

In the output layer, the output neuron again produces a product Y_t which is the total of the connected weight w_{jt} and bias b_t passing through the transfer function f_o. This operation is presented in Eqs. (6)–(8):

$$s_t = \sum_{j=1}^{S} w_{jt} a_j + b_t \quad t = 1, \tag{6}$$

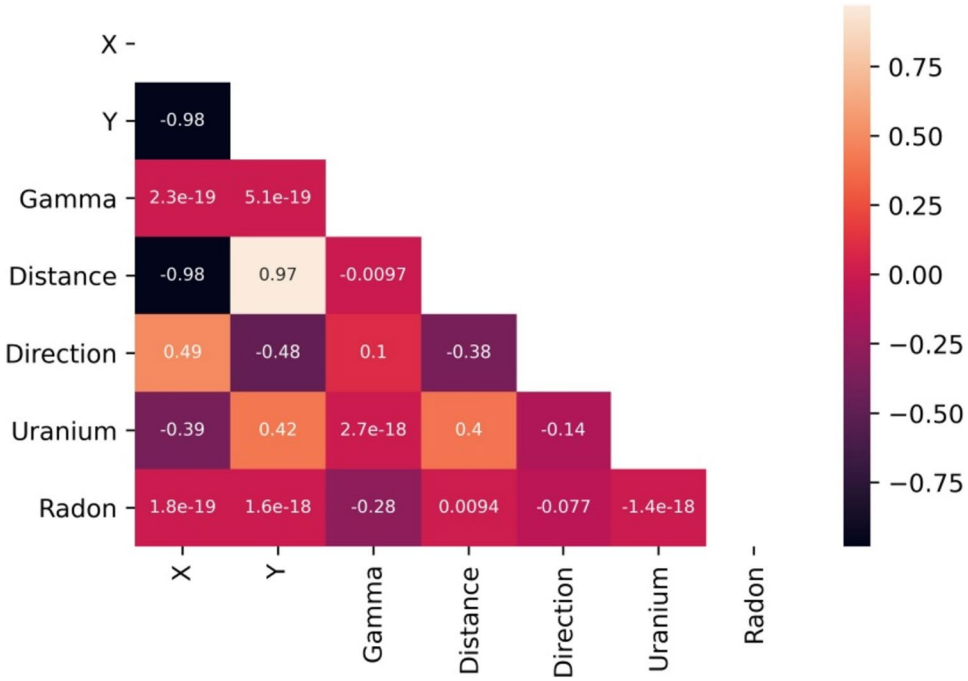


Fig. 3 Correlation matrix of input and output variables

$$f_o(x) = \frac{1}{1+e^{-x}} \tag{7}$$

$$Y_t = f_o(s_t) \quad t = 1, \tag{8}$$

The training error between the predicted output (Y_t) and the measured data (y_t) is determined by Eq. (9) where T denotes the number of training data points:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t)^2 \tag{9}$$

The error gradient for the output layer is measured by Eq. (10):

$$\delta_t = \left(Y_t - y_t\right) f_o\prime(s_t) \tag{10}$$

The backpropagation algorithm is used to adjust the weights and biases of the ANN to minimize the objective function in Eq. (9). The functions for adjusting the weights and biases between the hidden and output layers are given by Eqs. (11) and (12), respectively:

$$w_{jt}(k+1) = w_{jt}(k) + \alpha\left(y_t - Y_t\right)Y_t(1-Y_t)a_j \tag{11}$$

$$b_t(k+1) = b_t(k) + \alpha\left(y_t - Y_t\right)Y_t(1-Y_t) \tag{12}$$

Equations (13) and (14) are used to update the weights and biases between the input and hidden layers, respectively:

$$w_{ij}(k+1) = w_{ij}(k) + \beta\left[\sum_{t=1}^{1}\left(y_t - Y_t\right)Y_t(1-Y_t)w_{jt}\right]a_j(1-a_j)p_i \tag{13}$$

$$b_j(k+1) = b_j(k) + \beta\left[\sum_{t=1}^{1}\left(y_t - Y_t\right)Y_t(1-Y_t)w_{jt}\right]a_j(1-a_j) \tag{14}$$

where α and β ($0<\alpha, \beta<1$) are the learning rates between the layers with k constants for the kth adjustment. The learning rates represent the rate of network convergence.

MATLAB software was used to establish the proposed ANN. In the first epoch of the training process, weights (w_{ij}, w_{jt}) and biases (b_j, b_t) were randomly initialized. The networks were trained using the Levenberg–Marquardt algorithm (Moré, 1978) for many cycles (epochs) until the network reached a stable *MSE* value (Eq. (9)).

To quantify and compare the accuracy of the proposed models, five common metrics, including, namely the root mean square error (*RMSE*) (Eq. (15)), mean absolute percentage error (*MAPE*) (Eq. (16)), mean absolute bias error (*MABE*) (Eq. (17)), correlation coefficient (r) (Eq. (18)) and coefficient of determination (R^2) (Eq. (19)) were used:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t - Y_t\right)^2} \tag{15}$$

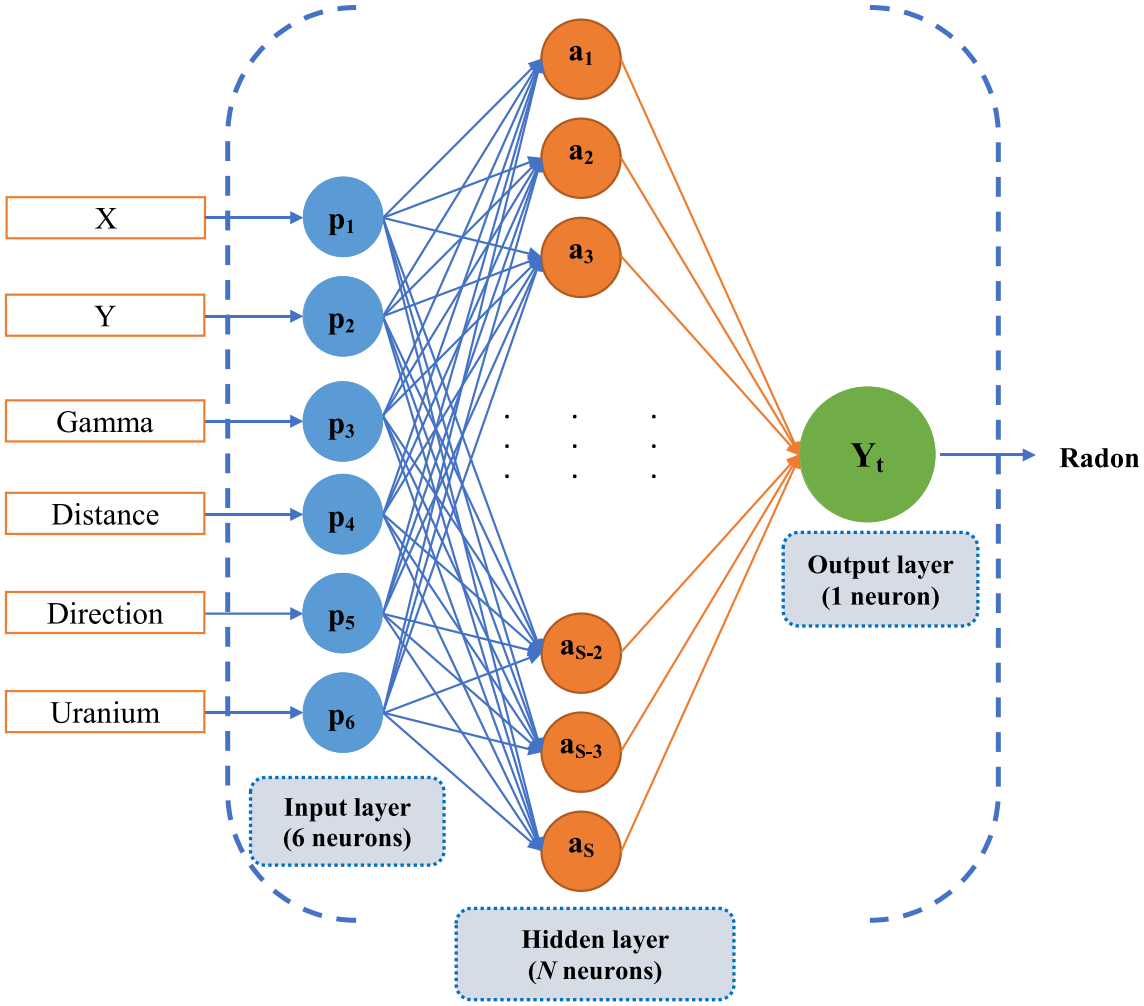


Fig. 4 ANN architecture of 6-*N*-1

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\frac{|y_t - Y_t|}{y_t} \times 100 \tag{16}$$

$$MABE = \sum_{t=1}^{T}\frac{|y_t - Y_t|}{T} \tag{17}$$

$$r = \frac{\sum_{t=1}^{T}\left(y_t - \overline{y}_t\right)\left(Y_t - \overline{Y}_t\right)}{\sqrt{\sum_{t=1}^{T}\left(y_t - \overline{y}_t\right)^2 \sum_{t=1}^{T}\left(Y_t - \overline{Y}_t\right)^2}} \tag{18}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\left(y_t - Y_t\right)^2}{\sum_{t=1}^{T}\left(y_t - \overline{y}_t\right)^2} \tag{19}$$

where y_t and Y_t denote the measured and predicted radon values, respectively; $\overline{y}_t$ and $\overline{Y}_t$ represent the average measured and predicted radon values, respectively; and T stands for the number of training data points.

RMSE represents the differences between the predicted radon values determined by the models and the measured ones. In general, a lower RMSE is preferable. MAPE is a statistical measure to accurately assess the proposed ANN models and its minimal value indicates that the model is highly accurate. MABE is another metric to estimate how close the predicted radon values are to the measured ones and when it is low, the model is highly accurate. The correlation coefficient r ($0 \le r \le 1$) is used to quantify the correlation

between the model and observations. If $r = 1$, an exact linear relationship between the predicted and measured values was determined. Finally, the coefficient of determination displays information about the variation in the predicted radon values of the model and when close to 1, the prediction is reliable.

ANN training for radon prediction

The ANN training process follows the flowchart presented in Fig. 5. The first step involves data collection and analysis as described in Section 'Dataset', while the second one concerns data partitioning. 80% of the database is used for training, while the remainder is used for validation. The third step consists of training where a one-hidden-layer ANN model is optimized and applied to study optimization of the ANN structure. The predictive capacity of the model is evaluated in the fourth step using the test dataset and various standard metrics (RMSE, MAPE, MABE, r and R^2) to optimize the configuration of the model by mainly determining the optimum number of hidden neurons. The fifth step compares the proposed model with some benchmark machine learning models such as the two-hidden-layer ANN model, Support Vector Machine (SVM) and Random Forest (RF) to evaluate its prediction efficiency. Finally, a sensitivity analysis is performed to determine the features that have the most significant impact on the predictions of the proposed model.

Results and discussion

Optimizing the ANN structure

The proposed ANN used only one-hidden-layer, as mentioned in Section 'Artificial neural network'. This section deals with the number of neurons S in the hidden layer. To the best of our knowledge, no accepted procedure or formula has been published in the literature to determine the optimum number of hidden neurons. Based on Kolmogorov's theorem, Hecht-Nielsen [42] proposed that $2n + 1$ (n denotes the number of predictor parameters) should determine the maximum number of neurons in the one-hidden-layer ANN. According to this suggestion, the maximum number of hidden neurons in this study was 13 ($n = 6$) (Table 1), while the minimum number was 2.

MATLAB software initializes the ANNs using random weights and biases. The ANN trained the model ten times for

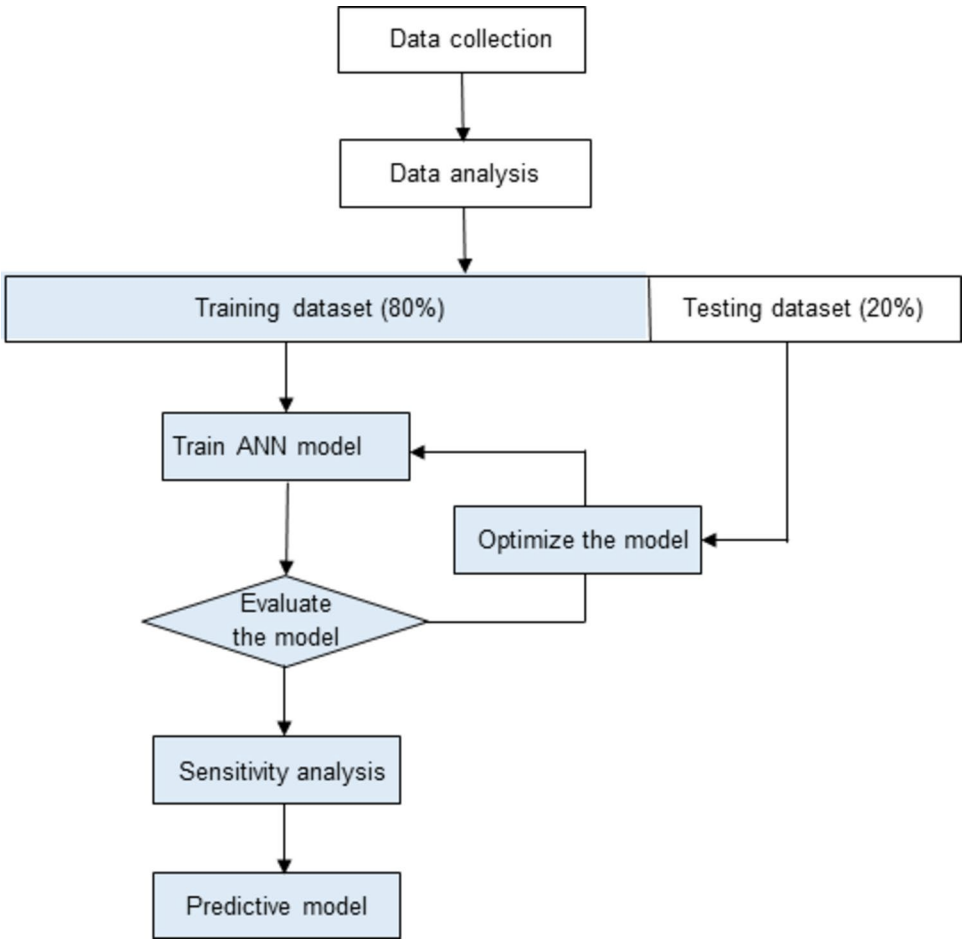


Fig. 5 Flowchart of ANN training for radon prediction

Table 1 Performance of ANN models with different numbers of neurons in the hidden layer

Model No	Number of hidden neurons	*RMSE* (Bq/m^3)	*MAPE(%)*	*MABE(%)*	*r*	R^2
Training						
1	2	31.1	12.7	23.6	0.579	0.336
2	3	21.5	9.01	17.0	0.826	0.683
3	4	18.1	7.47	14.1	0.880	0.775
4	5	16.0	6.78	12.8	0.907	0.823
5	6	12.4	4.06	7.44	0.946	0.894
6	7	11.3	3.72	6.92	0.952	0.906
7	8	10.7	3.86	7.20	0.958	0.917
8	9	7.77	3.26	6.14	0.976	0.952
9	10	6.97	2.54	4.97	0.980	0.960
10	11	2.96	1.12	2.03	0.991	0.982
11	12	3.03	1.15	2.17	0.994	0.988
12	13	2.79	1.12	2.10	0.995	0.990
Testing						
1	2	31.1	12.6	23.5	0.580	0.336
2	3	21.5	9.01	17.0	0.826	0.682
3	4	18.1	7.48	14.1	0.880	0.774
4	5	16.0	6.76	12.8	0.908	0.824
5	6	12.4	4.06	7.43	0.946	0.894
6	7	11.3	3.71	6.89	0.955	0.913
7	8	10.7	3.85	7.18	0.960	0.921
8	9	7.74	3.25	6.12	0.979	0.959
9	10	6.94	2.53	4.94	0.983	0.967
10	11	2.97	1.12	2.03	0.997	0.994
11	12	3.02	1.14	2.17	0.997	0.994
12	13	2.79	1.12	2.09	0.997	0.995

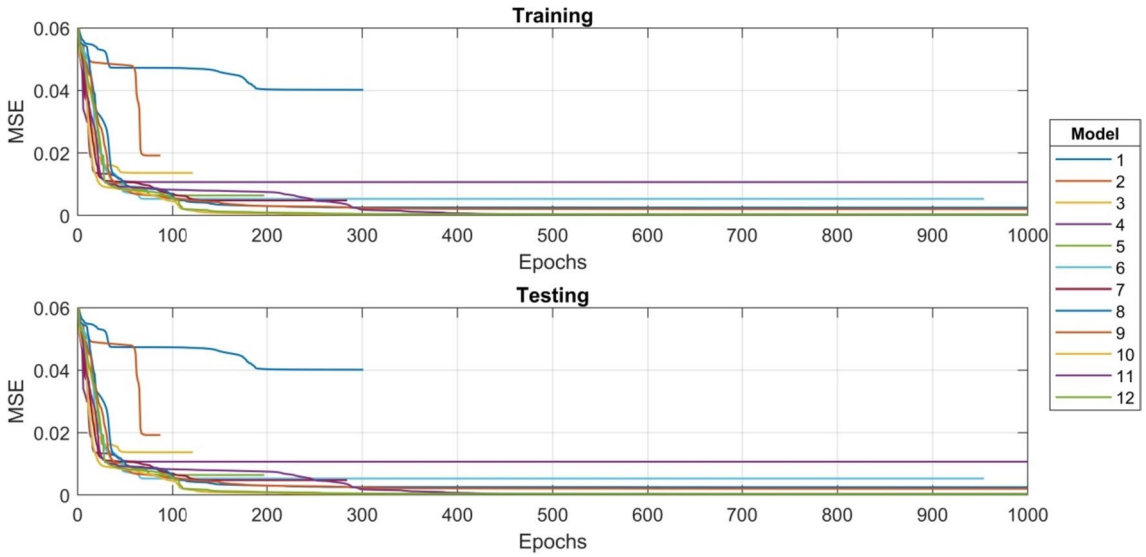


Fig. 6 MSE of different ANN models in the training and test dataset versus the number of training epochs

each of the ANN configurations (number of hidden neurons) to achieve the optimum results. The training and evaluation MSEs of ANN models with different numbers of hidden neurons are illustrated in Fig. 6. 13 hidden neurons were used in Model 12 yielding the best MSE value at epoch 300 (Table 1). In general, the performances of the ANN models with regard to training and test data are similar, demonstrating that the models did not memorize the training data but learned the actual relationship between the value points.

The performances of the ANN models with the metrics RMSE, MAPE, MABE, r and R^2, which were calculated using Eqs. (15)–(19), are presented in Table 1 and Fig. 7. The best performances—*RMSE* of less than 5 Bq/m^3, *MAPE* and *MABE* of less than 3% as well as excellent *r* and R^2 values in excess of 99%—are presented in Models 10–12. Although a slight difference in the performance of the three models was observed, Model 12 exhibited the best performance with the highest R^2 value. RMSE, MAPE, MABE and r on the test dataset were equal to 2.791, 1.117, 2.094 and 0.997, respectively. The scatter plots of the predicted radon values by the ANN models for the training and test datasets are also depicted in Appendix 1.

The performances of Model 12 along with the other ones are compared using the Taylor diagram in Fig. 8 [43]. The main advantage of this diagram is that the performances of the models in the groups according to their RMSE, standard deviation and correlation coefficient are shown. The distance of each model from the observed point denoted by black triangles plotted on the horizontal axis quantifies how closely that particular model matches the measurements. It can be seen that Models 1–9 exhibited the worst performance since their data points are scattered over a great distance from the observed point. Generally speaking, Models 10–12 correlated well with the observations, exhibiting the same RMSE, high pattern correlation and standard deviation illustrated by the dashed line at radial distance 0.246. A slightly higher degree of correlation and RMSE was observed in Model 12. All results hereafter refer to Model 12.

Predictive capability of the proposed ANN model

The predictive capacity of the proposed ANN model using the configuration selected is presented in this section. The density plots on the right-hand side of Fig. 9 compare the probability density of the measured and predicted radon values for the test, training and whole datasets. The cumulative distributions of the three datasets are also compared on the left-hand side of Fig. 9. A highly significant correlation was observed between the measured and predicted values. Most deviations occurred when a model predicted the radon value to be approximately 200 Bq/m^3, moreover, the shapes of the predicted distribution for the three datasets are similar, indicating the generalized capability of the proposed ANN model with regard to the new dataset.

Another meaningful predictive capacity is the prediction error per quantile generated by the proposed ANN model. In Fig. 9, the quantiles over equal ranges from 10 to 90% are plotted to evaluate the deviation between their measured and

Fig. 7 Performance graph for the 12 investigated models

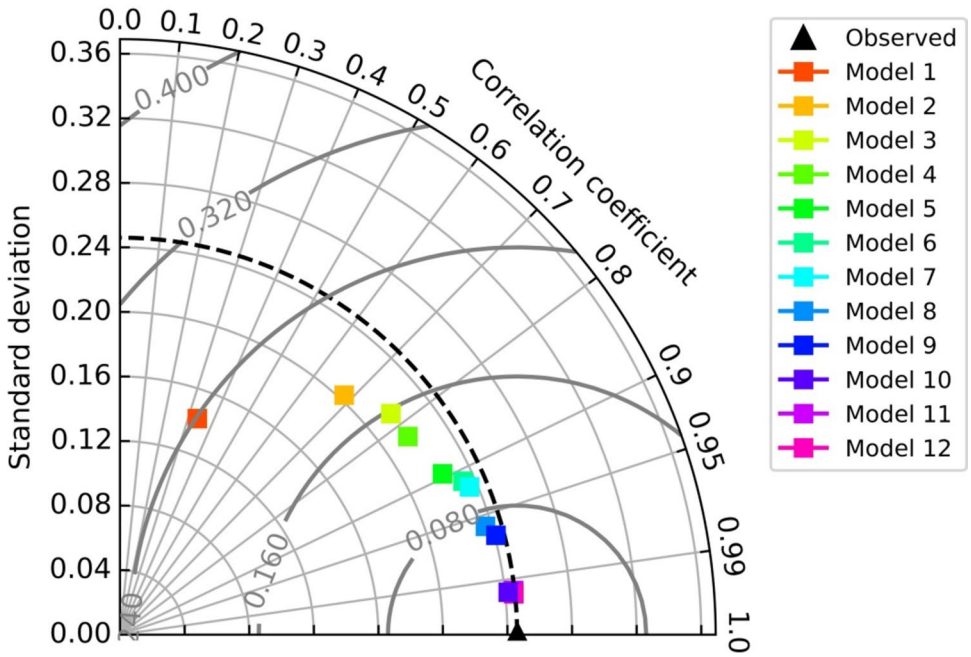


Fig. 8 Taylor diagram comparing the performances of Models 1–12

predicted values. In general, the expected quantile values are similar to those measured, demonstrating the capability of the proposed model with regard to radon prediction locally. A detailed evaluation is given in Table 2 in which the error (E) is computed for each quantile, the deviations of which are highest in the 10, 20, 60 and 90% quantiles. Furthermore, the average deviation for the test dataset is slightly larger than for the training and whole datasets (Fig. 10).

Finally, the proposed model in terms of predicting the radon values at 21 dwellings according to the specific uranium measurement points from the test dataset is examined in Fig. 11. It can be seen that the predicted radon values closely resemble the measured ones. Obviously, the proposed model in this study can be employed for estimating radon dispersion with a high degree of accuracy.

Comparison of the models

To illustrate its prediction efficiency, the proposed ANN model was compared with the previous benchmark machine learning models, including the two-hidden-layer ANN model with 20 neurons [5]; Support Vector Machine (SVM) and Random Forest (RF). In order to develop the SVM model, the radial basis function kernel was deployed to train the model, moreover, two parameters, namely cost (C) and sigma (δ), were tuned to control its accuracy. For the RF model, the minimum leaf size (m) and the number of trees (nt) were chosen to evaluate its performance. A "trial and error" procedure was conducted by experimenting with C, δ, m and nt in various ways to determine the optimum parameters for both models. Based on the *RMSE* values, the best SVM and RF models were defined when $C = 51.623$, $\delta = 0.024$ as well as $nt = 800$ and 4, respectively.

As is shown in Table 3, the proposed ANN model outperforms the others producing a lower RMSE and MAPE as well as a higher r and R^2. In contrast, the SVM model yielded the poorest performance in terms of both the training and test dataset. It can be seen that by increasing the number of hidden layers, as is the case in the ANN model with two hidden layers, the prediction accuracy is reduced and overfitting results when the performance of the test dataset is better than the training one.

Sensitivity analysis

The proposed model predicts the radon concentration values based on the distance along the X & Y axes, direction, gamma dose rate and uranium concentration. This section attempts to determine which features influence the predictions the most. For this purpose, a simple and popular method, namely Permutation Importance [44], is implemented as follows:

(1) Obtain the trained ANN model;

(2) Shuffle the values in a single feature column and make predictions using the generated dataset. The predicted and measured values are used to calculate how much the loss *RMSE* was affected by shuffling and estimate the importance of the shuffled features.

(3) Put the data back into its original order and repeat step (2) with the next feature.

Due to the random nature of the shuffling, step (2) was repeated five times to obtain an average result. According to

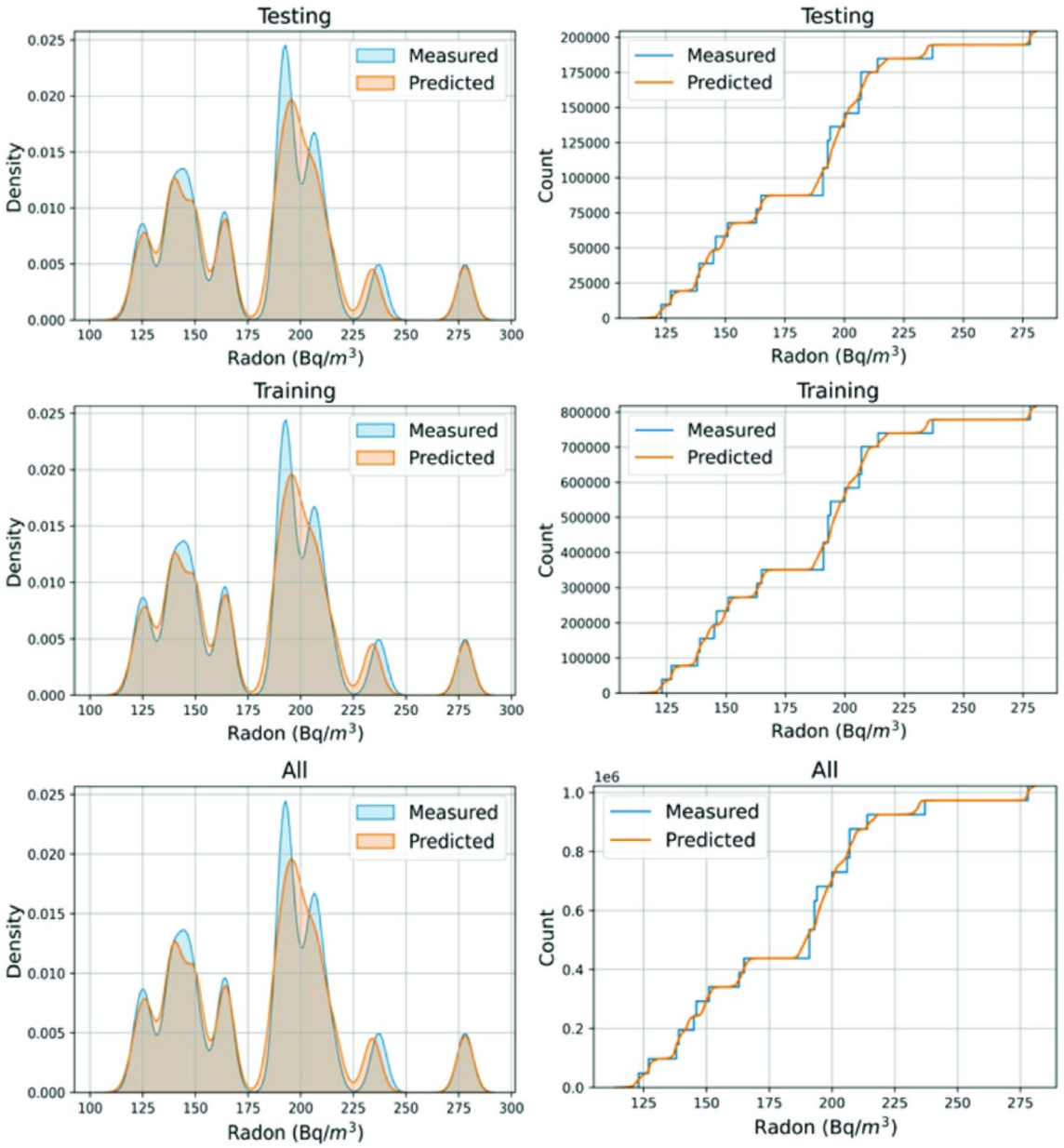


Fig. 9 Density (right) and cumulative distribution (left) plots between the measured and predicted radon values for the test, training and whole dataset

the results, the gamma dose rate as well as distance are the most critical and correlated features when predicting radon values, followed by the uranium concentration, the coordinates of the uranium measurement along the x and y axes as well as direction.

Conclusions

Radon is one of the most toxic radioactive gases presenting radiological hazards to humans. In this study,

Table 2 Comparison of the measured and predicted radon values for the test, training and whole datasets in each quantile

Quantile (%)	Testing			Training			All		
	M	P	E	M	P	E	M	P	E
10	138	135.9	−2.07	138	135.6	−2.41	138	135.6	−2.34
20	145	142.1	−2.93	145	142.1	−2.93	145	142.1	−2.93
30	151	150.8	−0.19	151	150.7	−0.33	151	150.7	−0.29
40	165	165.3	0.29	165	165.3	0.26	165	165.3	0.27
50	191	190.2	−0.81	191	190.1	−0.87	191	190.1	−0.86
60	193	195.3	2.31	193	195.3	2.28	193	195.3	2.29
70	200	200.5	0.51	200	200.5	0.50	200	200.5	0.50
80	207	207.3	0.25	207	207.2	0.22	207	207.3	0.23
90	214	217.7	3.66	214	217.6	3.65	214	217.6	3.65
	Average		0.11	Average		0.04	Average		0.06

*M = measured (Bq/m^3); P = predicted (Bq/m^3); E = error (Bq/m^3)

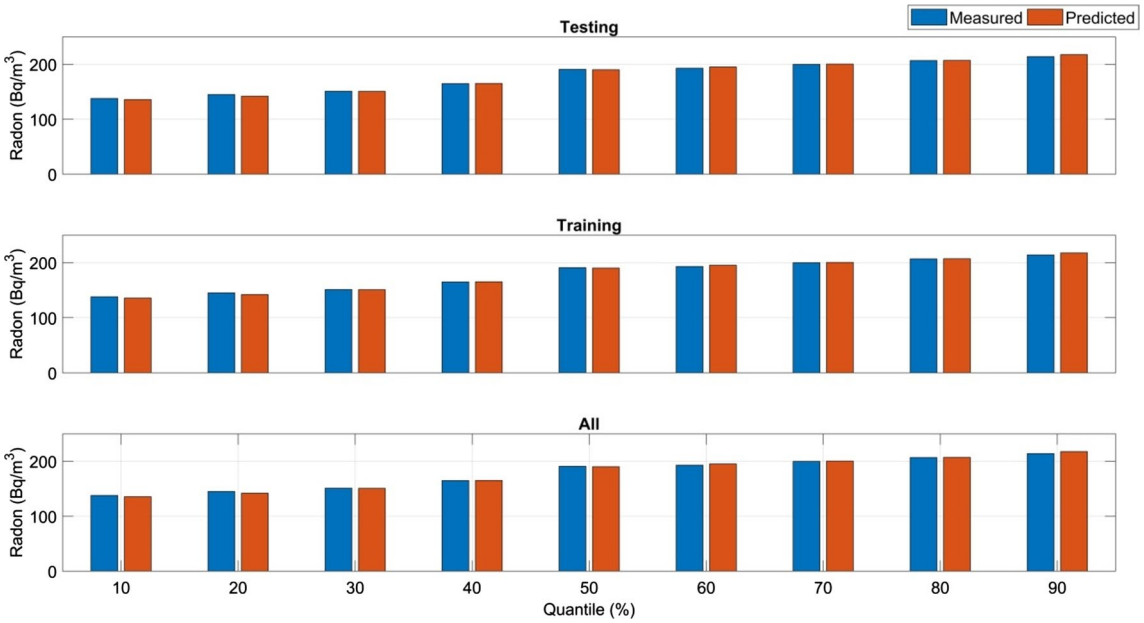


Fig. 10 Quantile of the measured and predicted radon values for the test, training and whole datasets

a predictive model of radon release was built using a large dataset at the Sin Quyen deposit. After optimizing its structure and evaluating its predictive capability as well as conducting a comparison and sensitivity analysis, the proposed model was constructed using a simple one-hidden-layer ANN requiring lower computational costs for training and referencing, which can be trained without needing to reduce the amount of input data. The model could also reduce overfitting as the training (RMSE = 2.793) and testing errors (RMSE = 2.791) are rather similar. The proposed model is not only a simple modelling approach but also an accurate prediction model yielding small errors as far as both the training and test dataset is concerned. A highly significant correlation and low deviation were observed between the measured and predicted values. The predicted values for the training, test and whole datasets suggest that the proposed model generalized the unseen data well. In comparison with other machine learning models of two-hidden-layer ANNs, Support Vector Machines (SVM) and Random Forests (RF), the proposed model is advantageous given that it predicts more accurately. Permutation Importance was performed on the underlying mechanism of the proposed model, revealing the gamma dose rate and distance

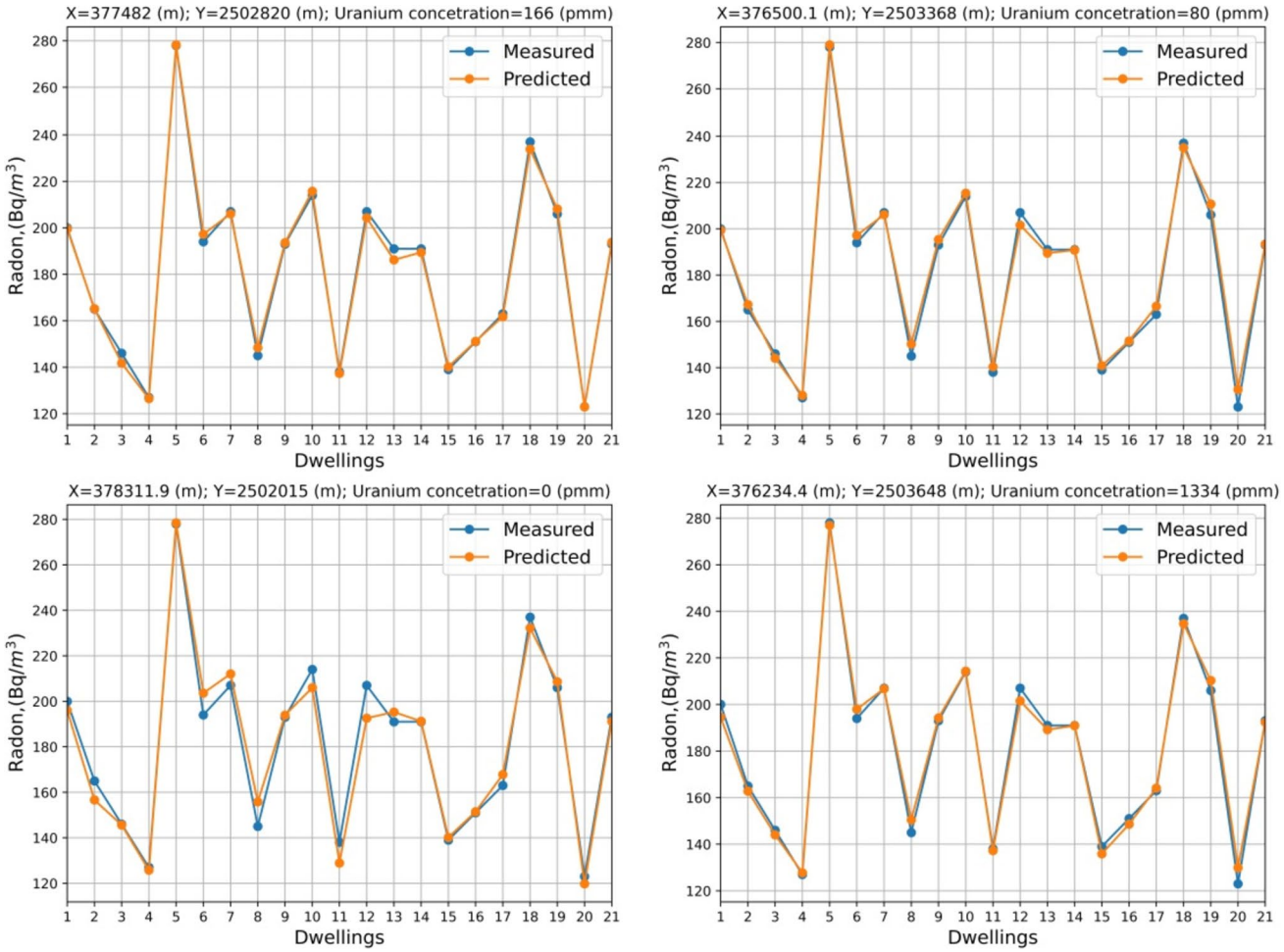


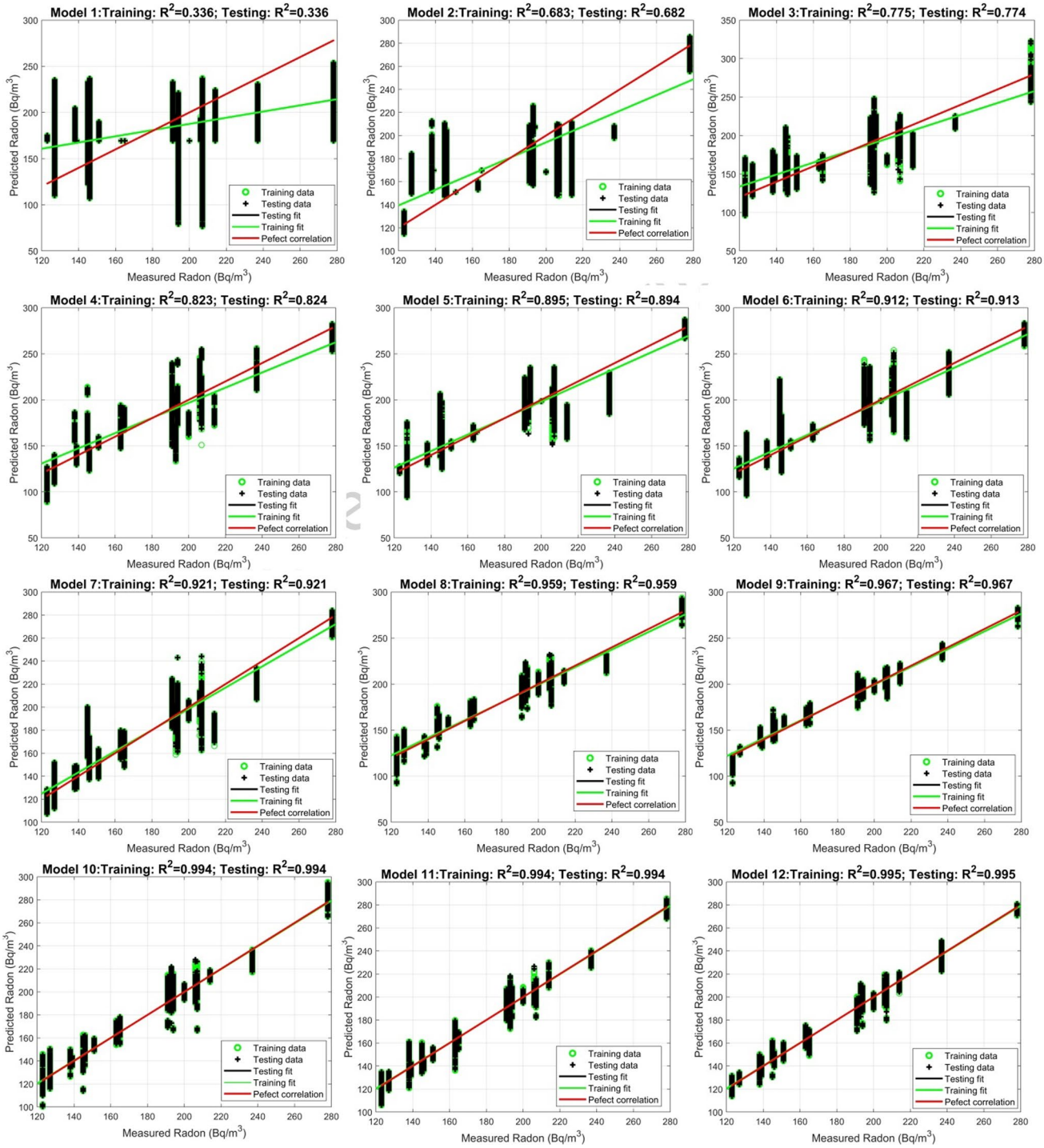
Fig. 11 Examples of predicting the radon dispersion at 21 dwellings according to uranium concentration measurements

Table 3 Performance of some benchmark machine learning models applied using the present dataset

Performance	Proposed ANN		Two-hidden-layers ANN		SVM		RF	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
RMSE (Bq/m^3)	2.79	2.79	8.69	9.13	12.0	13.1	6.65	9.61
MAPE (%)	1.12	1.12	2.02	2.03	5.28	7.03	2.21	3.22
r	0.995	0.997	0.974	0.970	0.830	0.824	0.987	0.970
R^2	0.990	0.995	0.949	0.942	0.812	0.805	0.974	0.941

as the strongest predictors of radon release when compared to the parameters of uranium concentration, uranium measurement coordinates and direction. Although the proposed model with a simple one-hidden-layer ANN optimizes and is more accurate, it could be improved in further studies by carrying out real-time monitoring of multiple inputs of dataset with different parameters as well as a radon survey when using this proposed model.

Appendix 1 Measured versus predicted Radon values resulted from 12 ANN models.



Acknowledgements The authors are grateful to VNU University of Science and Hanoi University of Mining and Geology staff for their help with sampling and measuring in the fieldwork.

Author contributions Conception: V-H.D., N.T.T., D-B.T., T.V.; Study design, methods used: V-H.D., N.T.T., D-B.T., T.V.; Acquisition and collation of data: V-H.D., D-B.T., T.D.Q., N.T.T., T.V.; Analysis and interpretation of data: V-H.D., N.T.T., D-B.T., T.V., H.T.T.N., G.T., Z.H., T.K.; Writing the manuscript: V-H.D., N.T.T., T.V., T.K.; Critical

revision of paper: V-H.D., N.T.T., T.V., T.D.Q., H.T.T.N., T.K.; All authors reviewed the manuscript.

Funding The research leading to these results received funding from Vietnam National University, Hanoi under Grant Agreement No QG.21.19 "[Research and application of artificial intelligence in monitoring and predicting of radioactive release in mining areas, a case study in Sin Quyen copper mine]".

References

1. Mai H, Maeghtb J-L, Valentin C (2020) Assessment of heavy metal concentrations and its potential eco-toxic effects in soils and sediments in Dong Cao catchment, Northern Vietnam. Vietnam J Earth Sci 42:187–204. https://doi.org/10.15625/0866-7187/42/2/15046
2. Trinh PT, Van Liem N, Van Huong N, Vinh HQ, Van Thom B, Thao BT, Tan MT, Hoang N (2012) Late quaternary tectonics and seismotectonics along the Red River fault zone, North Vietnam. Earth Sci Rev 114:224–235. https://doi.org/10.1016/j.earscirev.2012.06.008
3. Van Hoang N, Van DT, Hoa PL (2020) Heavy metal contamination of soil based on pollution, geo-accumulation indices and enrichment factor in Phan Me coal mine area, Thai Nguyen province, Vietnam. Vietnam J Earth Sci 42:105–117. https://doi.org/10.15625/0866-7187/42/2/14950
4. Van LH, Tien TN, Tat TV, Thanh TN, Lam AN, Bui DD, Le Van D, Ngoc DT, Huu HN (2021) Holocene sedimentation offshore Southeast Vietnam based on geophysical interpretation and sediment composition analysis. Vietnam J Earth Sci 43:336–379. https://doi.org/10.15625/2615-9783/16268
5. Duong V-H, Ly H-B, Trinh DH, Nguyen TS, Pham BT (2021) Development of Artificial Neural Network for prediction of radon dispersion released from Sinquyen Mine. Vietnam Environ Pollut 282:116973. https://doi.org/10.1016/j.envpol.2021.116973
6. Van HD, Lantoarindriaka A, Piestrzyński A, Trinh PT (2020) Fort-Dauphin beach sands, south Madagascar: Natural radionuclides and mineralogical studies. Vietnam J Earth Sci 42:118–129. https://doi.org/10.15625/0866-7187/42/2/14951
7. Phon LK, Dung BD, Chau ND, Kovacs T, Van Nam N, Van Hao D, Son NT, Luan VTM (2015) Estimation of effective dose rates caused by radon and thoron for inhabitants living in rare earth field in northwestern Vietnam (Lai Chau province). J Radioanal Nucl Chem 306:309–316. https://doi.org/10.1007/s10967-014-3881-8
8. Oyedele J, Shimboyo S, Sitoka S, Gaoseb F (2010) Assessment of natural radioactivity in the soils of rössing uranium mine and its satellite town in western Namibia, southern Africa. Nucl Instrum Methods Phys Res Sect A 619:467–469. https://doi.org/10.1016/j.nima.2010.01.068
9. Hao DV (2018) Measurement of natural radionuclides for samples very rich in thorium by gamma spectrometry-Mandena Deposit, South Madagascar. In: Proceedings of the fourth international conference on application of radiotracers and energetic beams in sciences 50, 895
10. Van Hao D, Nguyen Dinh C, Jodłowski P, Kovacs T (2019) High-level natural radionuclides from the Mandena deposit, South Madagascar. J Radioanal Nucl Chem 319:1331–1338. https://doi.org/10.1007/s10967-018-6378-z
11. Carvalho F, Reis M (2006) Radon in Portuguese houses and workplaces. In: International Conference Healthy Buildings, HB
12. Organization WH (2009) WHO handbook on indoor radon: a public health perspective. World Health Organization
13. Tracy BL (2010) Radon In Atwood DA (ed) Radionuclides in the Environment. Wiley, pp. 191–206
14. UNSCEAR Sources and Effects of Ionizing Radiation: 1993 Report to the General Assembly, with Scientific Annexes.
15. Doering C, McMaster SA, Johansen MP (2018) Modelling the dispersion of radon-222 from a landform covered by low uranium grade waste rock. J Environ Radioact 192:498–504. https://doi.org/10.1016/j.jenvrad.2018.07.024
16. Grant C, Lalor G, Balcázar M (2012) Radon monitoring in sites of economical importance in Jamaica. Appl Radiat Isot 71:96–101. https://doi.org/10.1016/j.apradiso.2012.07.007
17. Hadad K, Doulatdar R, Mehdizadeh S (2007) Indoor radon monitoring in Northern Iran using passive and active measurements. J Environ Radioact 95:39–52. https://doi.org/10.1016/j.jenvrad.2007.01.013
18. Heidary S, Setayeshi S, Ghannadi-Maragheh M, Negarestani A (2011) Monitoring and measurement of radon activity in a new design of radon calibration chamber. Radiat Meas 46:694–700. https://doi.org/10.1016/j.radmeas.2011.06.014
19. Jilani Z, Mehmood T, Alam A, Awais M, Iqbal T (2017) Monitoring and descriptive analysis of radon in relation to seismic activity of Northern Pakistan. J Environ Radioact 172:43–51. https://doi.org/10.1016/j.jenvrad.2017.03.010
20. Laiolo M, Cigolini C, Coppola D, Piscopo D (2012) Developments in real-time radon monitoring at Stromboli volcano. J Environ Radioact 105:21–29. https://doi.org/10.1016/j.jenvrad.2011.10.006
21. Ramola R, Negi M, Choubey V (2005) Radon and thoron monitoring in the environment of Kumaun Himalayas: survey and outcomes. J Environ Radioact 79:85–92. https://doi.org/10.1016/j.jenvrad.2004.05.012
22. Wu HX, Wei QL, Yang B, Liu QC (2014) Fast prediction method of radon concentration in environment air. Appl Mech Mater 539:819–822. https://doi.org/10.4028/www.scientific.net/AMM.539.819
23. Xie D, Wang H, Kearfott KJ (2012) Modeling and experimental validation of the dispersion of 222Rn released from a uranium mine ventilation shaft. Atmos Environ 60:453–459. https://doi.org/10.1016/j.atmosenv.2012.07.006
24. Panahi M, Yariyan P, Rezaie F, Kim SW, Sharifi A, Alesheikh AA, Lee J, Lee J, Kim S, Yoo J (2022) Spatial modeling of radon potential mapping using deep learning algorithms. Geocarto Int 37:9560–9582. https://doi.org/10.1080/10106049.2021.2022011
25. Petermann E, Meyer H, Nussbaum M, Bossew P (2021) Mapping the geogenic radon potential for Germany by machine learning. Sci Total Environ 754:142291. https://doi.org/10.1016/j.scitotenv.2020.142291
26. Rezaie F, Kim SW, Alizadeh M, Panahi M, Kim H, Kim S, Lee J, Lee J, Yoo J, Lee S (2021) Application of machine learning algorithms for geogenic radon potential mapping in Danyang-Gun. South Korea Front Environ Sci 9:753028. https://doi.org/10.3389/fenvs.2021.753028
27. Rezaie F, Panahi M, Lee J, Lee J, Kim S, Yoo J, Lee S (2022) Radon potential mapping in Jangsu-gun, South Korea using probabilistic and deep learning algorithms. Environ Pollut 292:118385
28. Külahcı F, İnceöz M, Doğru M, Aksoy E, Baykara O (2009) Artificial neural network model for earthquake prediction with radon monitoring. Appl Radiat Isot 67:212–219. https://doi.org/10.1016/j.apradiso.2008.08.003
29. Mir AA, Çelebi FV, Alsolai H, Qureshi SA, Rafique M, Alzahrani JS, Mahgoub H, Hamza MA (2022) Anomalies prediction in radon time series for earthquake likelihood using machine learning-based ensemble model. IEEE Access 10:37984–37999. https://doi.org/10.1109/ACCESS.2022.3163291
30. Zmazek B, Todorovski L, Džeroski S, Vaupotič J, Kobal I (2003) Application of decision trees to the analysis of soil radon data for earthquake prediction. Appl Radiat Isot 58:697–706. https://doi.org/10.1016/S0969-8043(03)00094-0

31. ESCAP U (1992) State of the environment in Asia and the Pacific 1990 https://doi.org/10.1016/j.envpol.2021.118385
32. Ta V (1975) Report of geological surveys and their results performed at the IOCG Sin Quyen deposit in Lao Cai, North vietnam. Main Dept Geol Vietnam 318:49965
33. Luu C, Nguyen DD, Amiri M, Van PT, Bui QD, Prakash I, Pham BT (2022) Flood susceptibility modeling using radial basis function classifier and fisher's linear discriminant function. Vietnam J Earth Sci 45:55–72
34. Ly H-B, Asteris PG, Pham TB (2020) Accuracy assessment of extreme learning machine in predicting soil compression coefficient. Vietnam J Earth Sci 13:228–336. https://doi.org/10.15625/0866-7187/42/3/14999
35. Pham BT, Amiri M, Nguyen MD, Ngo TQ, Nguyen KT, Tran HT, Vu H, Anh BTQ, Van Le H, Prakash I (2021a) Estimation of shear strength parameters of soil using optimized inference intelligence system. Vietnam J Earth Sci 43:189–198
36. Pham BT, Amiri M, Nguyen MD, Ngo TQ, Nguyen KT, Tran HT, Vu H, Anh BTQ, Van Le H, Prakash I (2021b) Estimation of shear strength parameters of soil using optimized inference intelligence system. Vietnam J Earth Sci 43:189–198
37. Thanh DQ, Nguyen DH, Prakash I, Jaafari A, Nguyen V-T, Van Phong T, Pham BT (2020) GIS based frequency ratio method for landslide susceptibility mapping at Da Lat City, Lam Dong province, Vietnam. Vietnam J Earth Sci 42:55–66
38. Van Phong T, Ly H-B, Trinh PT, Prakash I, Btjvjoes P (2020) Landslide susceptibility mapping using forest by penalizing attributes (FPA) algorithm based machine learning approach. Vietnam J Earth Sci 42:237–246
39. Pham BTSS, Ly HB (2020) Using artificial neural network (ANN) for prediction of soil. Vietnam J Earth Sci 42:311–319
40. Tran VQ, Prakash I (2020) Prediction of soil loss due to erosion using support vector machine model. Vietnam J Earth Sci 12:247–254
41. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Math Control Signals Syst 2:303–314. https://doi.org/10.1007/BF02134016
42. Hecht-Nielsen R (1987) Kolmogorov's mapping neural network existence theorem. In: Proceedings of the international conference on Neural Networks, 1987. IEEE press New York, pp 11–14
43. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106:7183–7192
44. Hooker G, Mentch, Lucas (2019) Please stop permuting features: An explanation and alternatives. arXiv preprint arXiv:1905.031512