Environmental Pollution 282 (2021) 116973

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Development of Artificial Neural Network for prediction of radon dispersion released from Sinquyen Mine, Vietnam☆

Van-Hao Duong [a], Hai-Bang Ly [b], Dinh Huan Trinh [c], Thai Son Nguyen [a, c], Binh Thai Pham [b, *]

[a] Hanoi University of Mining and Geology, 18 Vien Street, Bac Tu Liem District, Hanoi, Viet Nam
[b] University of Transport Technology, Hanoi, 100000, Viet Nam
[c] Radioactive & Rare Minerals Division, Xuan Phuong, Bac Tu Liem, Hanoi, Viet Nam

ARTICLE INFO

Article history:
Received 26 May 2020
Received in revised form
11 March 2021
Accepted 15 March 2021
Available online 23 March 2021

Keywords:
Radon distribution
Radon prediction
ANN
Machine learning
Radon dispersion
Optimization
Sinquen mine

ABSTRACT

Understanding the radon dispersion released from this mine are important targets as radon dispersion is used to assess radiological hazard to human. In this paper, the main objective is to develop and optimize a machine learning model namely Artificial Neural Network (ANN) for quick and accurate prediction of radon dispersion released from Sinquyen mine, Vietnam. For this purpose, a total of million data collected from the study area, which includes input variables (the gamma data of uranium concentration with 3 × 3m grid net survey inside mine, 21 of CR-39 detectors inside dwellings surrounding mine, and gamma dose at 1 m from ground surface data) and an output variable (radon dispersion) were used for training and validating the predictive model. Various validation methods namely coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) were used. In addition, Partial dependence plots (PDP) was used to evaluate the effect of each input variable on the predictive results of output variable. The results show that ANN performed well for prediction of radon dispersion, with low values of error (i.e., $R^2 = 0.9415$, RMSE = 0.0589, and MAE = 0.0203 for the testing dataset). The increase of number of hidden layers in ANN structure leads the increase of accuracy of the predictive results. The sensitivity results show that all input variables govern the dispersion radon activity with different amplitudes and fitted with different equations but the gamma dose is the most influenced and important variable in comparison with strike, distance and uranium concentration variables for prediction of radon dispersion.

1. Introduction

It is well-known that many inhabited areas around the world people are long-term exposure effects and can be get the risks under high levels of natural radionuclides radiation such as natural radioactivity in the soils of Rössing Uranium Mine in western Namibia vary from 45.9 to 1752 Bq/kg for ^{238}U, 70.4–1866 Bq/kg for ^{232}Th and 376–1300 Bq/kg for ^{4} K (Oyedele et al., 2010); the ^{222}Rn concentration in air surrounding rare earth elements mines in North Vietnam from 48 to 920 Bq/m^3 (Le Khanh et al., 2015); the ^{222}Rn concentration in soil range from 670 to 1940 Bq/m^3 and 490–2120 Bq/m^3 at chromite mines of Khanozai and Muslim Bagh, Pakistan (Ahmad et al., 2019); and up to 6000 Bq/kg of ^{238}U and 240,000 Bq/kg of ^{232}Th in Madena Madagasar (Van Hao et al., 2019). One of the natural radionuclides usually concerning is radon gas (^{222}Rn). This is a noble gas, a progeny of the ^{238}U decay series and the longest half-live ($T_{1/2}$ = 3.8 days) in comparison with other natural radioactive gases. Because of the gas state, this ^{222}Rn is easy dispersion and fill in the atmosphere, especially in limited spaces such as caves, inside houses, and underground mines or radioactive element bearing mineral mines which could be contributed a significant internal radiation exposure (Carvalho and Reis, 2006; WHO, 2009; Le Khanh et al., 2015). In natural environment, the ^{222}Rn isotope is one of the most radiotoxic and carcinogenic gas that may affect indoor air quality, the high specific activity of alpha emitting. It is clear linking epidemiological evidence between continued exposure to high ^{222}Rn activity concentrations and lung

☆ This paper has been recommended for acceptance by Pavlos Kassomenos

* Corresponding author.
E-mail addresses: duongvanhao@humg.edu.vn (V.-H. Duong), banglh@utt.edu.vn (H.-B. Ly), huan.trinhdinh@gmail.com (D.H. Trinh), nguyenthaisondvl@gmail.com (T.S. Nguyen), binhpt@utt.edu.vn (B.T. Pham).

https://doi.org/10.1016/j.envpol.2021.116973
0269-7491/

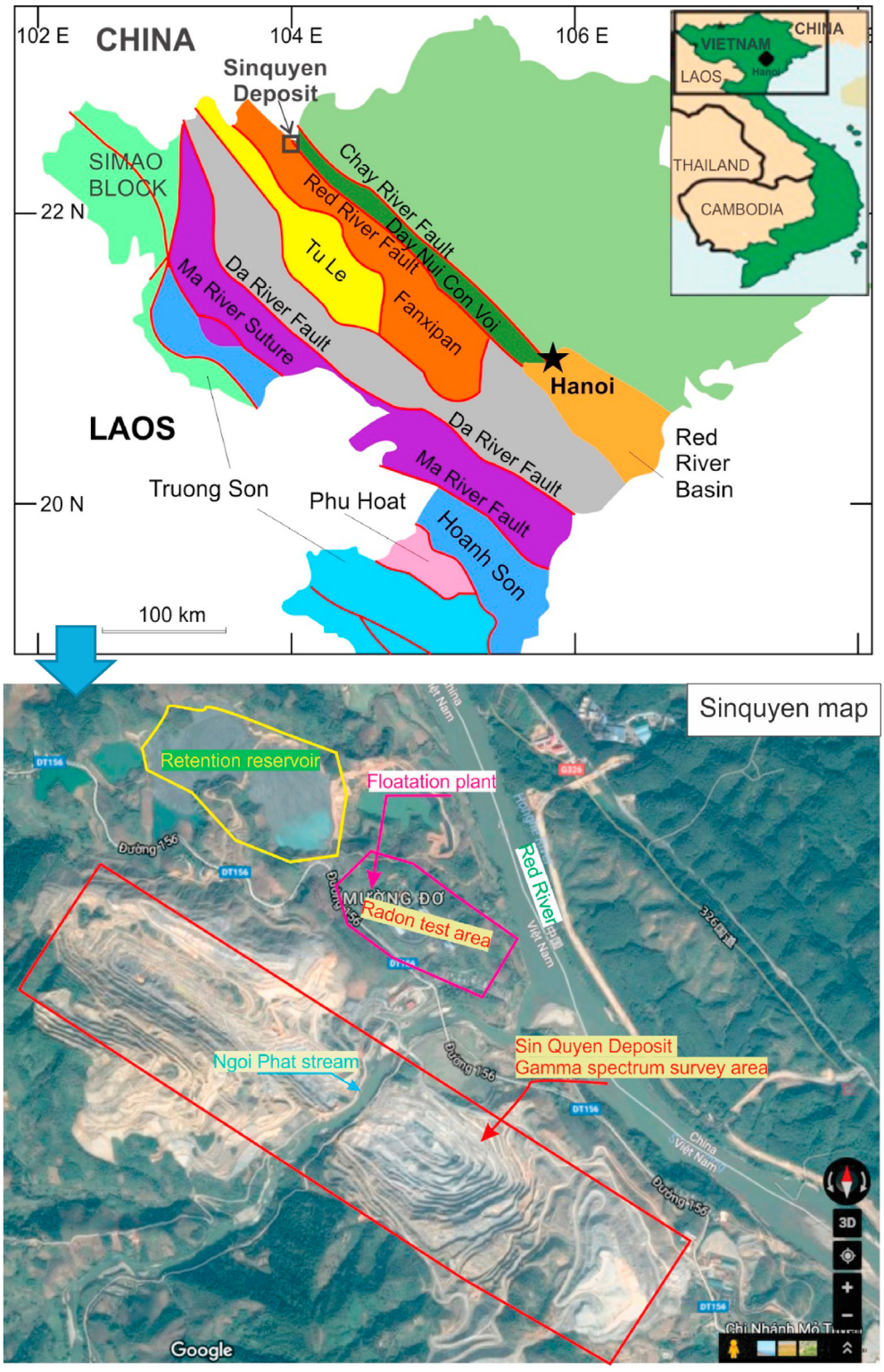


Fig. 1. Studied zone inside red line with 3 × 3m grid of gamma spectrum survey data, 21 points of radon test and gamma dose inside pink line. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

cancer (WHO, 2009). The ^{222}Rn was known a natural radionuclide are constituents of the earth crust (Banzi et al., 2017), the one can find everywhere in the environment. In some of special regions are prone to higher ^{222}Rn concentrations in air, especially such as granite, uranium, phosphate mine, and rare earth element mine areas, and certain ^{238}U, ^{226}Ra containing mineral or placer deposits (Carvalho et al., 2007; Chalupnik and Wysocka, 2008; Hilton, 2008; Nguyen et al., 2016; Van Hao et al., 2019; Thanh Duong et al., 2020).

Traditionally, many studies have been carried out to monitor and determine the radon concentration (Ramola et al., 2005; Hadad

Table 1
Statistical analysis of inputs and output used in this study.

Values	Uranium concentration (ppm)	Gamma dose (μSv/h)	Distance (m)	Strike (degree)	Dispersion radon activity (Bq/m^3)
Role	Input	Input	Input	Input	Output
Min	0	0.176	0.094	0	123
Average	578.067	0.234	675.447	166.710	181.333
Max	12,173	0.308	2731.801	359.980	278.000
Std	1214.939	0.039	727.703	66.681	38.123

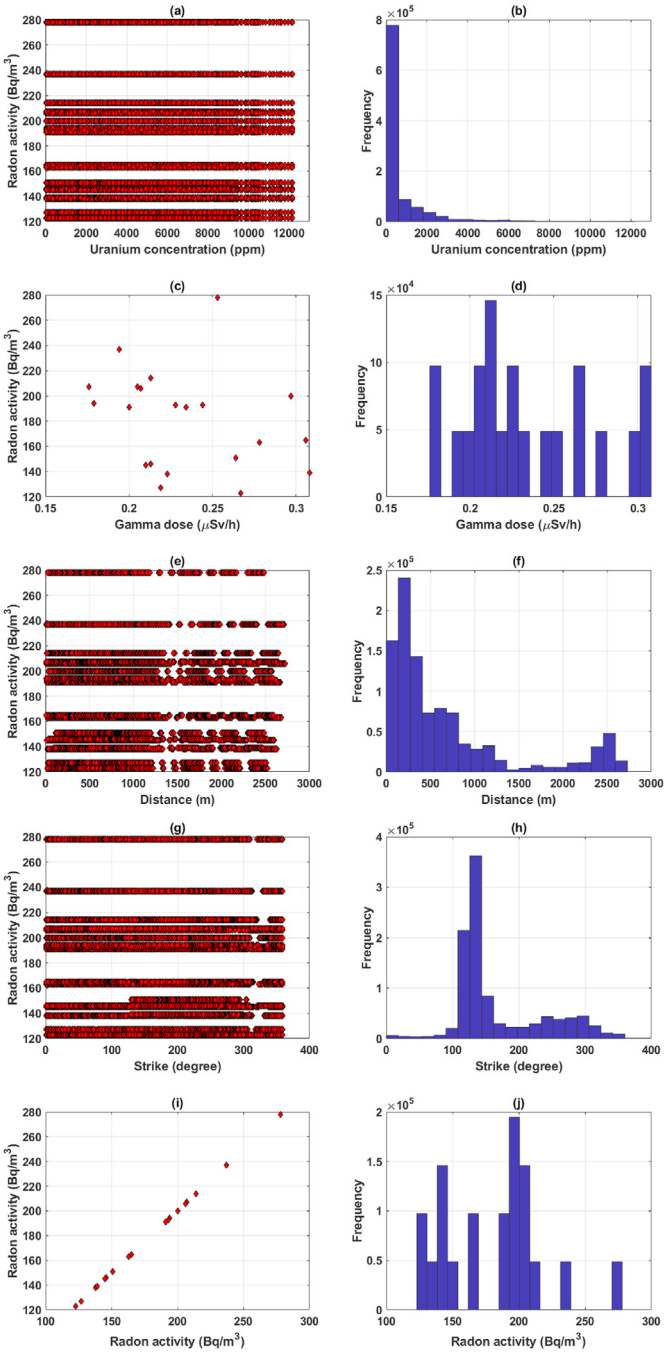


Fig. 2. Correlation graphs and histograms with respect to the dispersion radon activity for (a, b) uranium concentration; (c, d) gamma dose; (e, f) distance; (g, h) strike and (i, j) dispersion radon activity.

et al., 2007; Heidary et al., 2011; Grant et al., 2012; Laiolo et al., 2012; Le Khanh et al., 2015; Jilani et al., 2017), which have to determine in situ at real time and using average of few single measurements, need multi-measure to reach to real value, and for huge area will need much more time and finance or developing of radon dispersion modeling or predicting based on conditions are assumed to be standard form or hypothetical models (Doering et al., 2018; Wu et al., 2014, 2014, 2014; Xie et al., 2012). Wu et al. (2014) built the theoretical model of release radon concentration in environment air by diffusion and advection flux of radon on unit area by Fick's law and Darcy's law. The method only concern to the effected parameters of the radon emanation and prediction for narrow area which is similar background of local radon value and maybe true for narrow test space; Xie et al. (2012) was reported the dispersion of radon released from a uranium mine by using Computational fluid dynamics for atmospheric Rn physical model and mathematical models (Wind features with main wind direction; Atmosphere stability and Surface roughness is constant) then compared with field measurements. But this model maybe true for very standard condition with very narrow area (250 m of distance) it is same condition area and did not concern to the background so applying for complex huge area will be difficult; Doering et al. (2018) was developed a traditional modelling of the dispersion of radon and develop contour maps from a landform covered by low uranium grade.

In recent years, more advanced techniques called machine learning or artificial intelligence have been developed and applied in prediction problems such as natural hazard assessment (Pourghasemi et al., 2020; Van Dao et al., 2020), soil properties prediction (Pham et al., 2019; Rivera and Bonilla, 2020), predicton of construction material properties (Ly et al., 2019; Rashidi et al., 2016; Wei et al., 2019). Compared with traditional approaches, these techniques are considered as cost and time effective models. In the case of radon dispersion prediction, there is less or rarely studies of prediction of radon dispersion using the machine learning such as application of decision trees to the analysis of soil radon data for earthquake prediction (Zmazek et al., 2003) or artificial neural network (ANN) model for earthquake prediction with radon monitoring (Külahcı et al., 2009). In general, these studies initially showed the high potential of machine learning models for prediction problems which can be applied to quick and accurate the radon dispersion.

In this study, the main aim is to develop an machine learning model namely ANN to predict the radon dispersion at the iron oxide copper and gold – rare earth elements – uranium (IOCG-REE-U) Sinquyen deposit, which is the biggest copper mine which contain high natural radionuclides and the ^{238}U is the main natural radionuclide (Nguyen et al., 2016). For modeling, Input variables (the gamma data of uranium concentration with 3 × 3m grid net survey inside mine, 21 of CR-39 detectors inside dwellings surrounding mine, and gamma dose at 1 m from ground surface data) and an output variable (radon dispersion) were used. Various validation methods namely coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) were used. In addition, Partial dependence plots (PDP) was used to evaluate

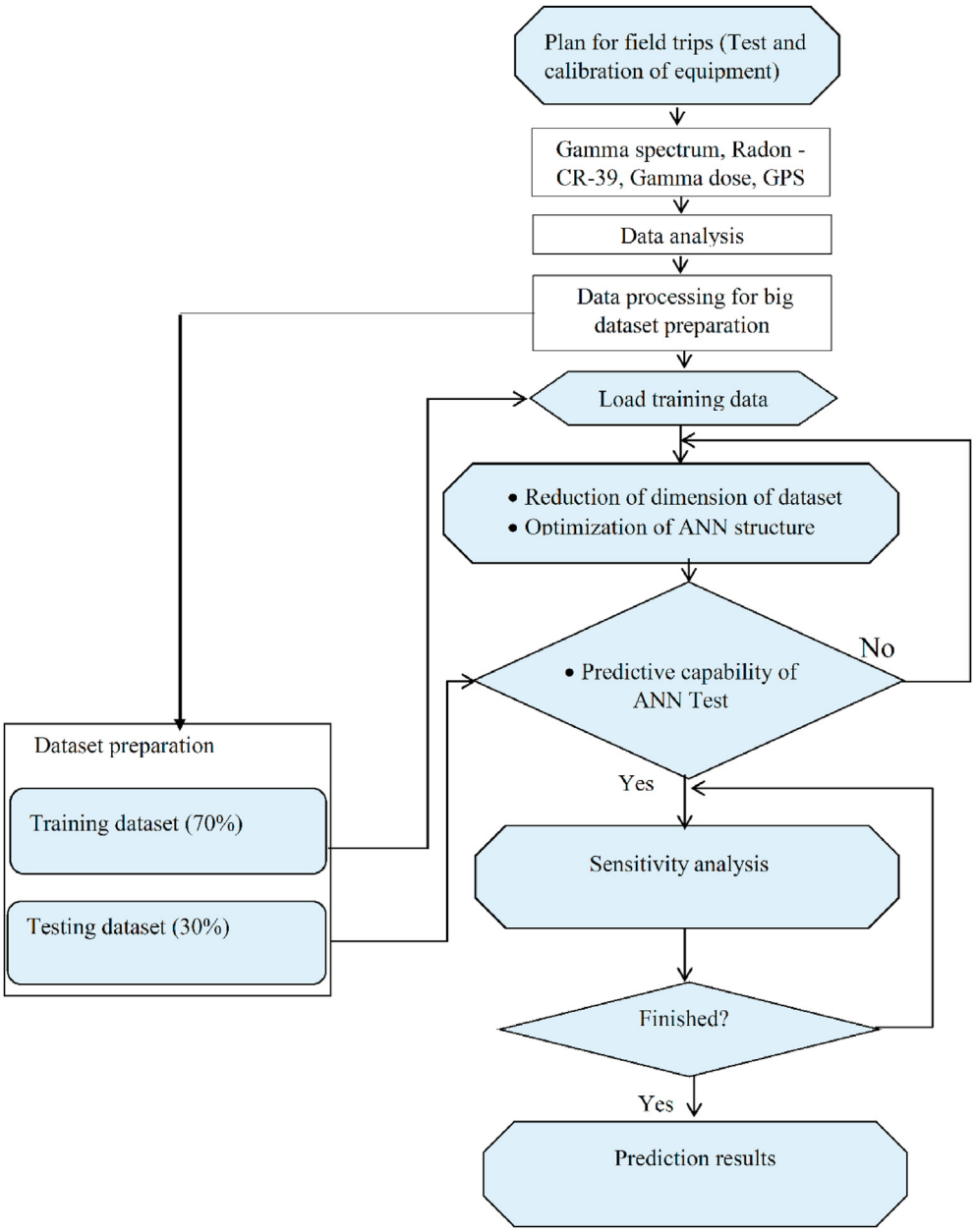


Fig. 3. Methodological flowchart of this study.

the effect of each input variable on the predictive results of output variable. Matlab packages and software were used for data processing and modeling in this study.

2. Materials and methods

2.1. Data used

The uranium concentration (ppm) as the first important input data, the data was recorded from the field radiometric measurements which were performed using the potable spectrometer GF5 with NaI(Tl) scintillation crystal of the Czech Gamma Surveyor Company™ and 3 × 3m grid net of about 350 × 1250 m of area inside red line of Sinquen deposit (Fig. 1). In the field, each measurement point was performed in 3 min and the detector was placed at 1 m from the Earth surface with the display on screen of potassium, uranium and thorium concentrations, and gamma absorbed dose rate D [nGy/h]. The radionuclide concentrations were represented and contributed from surface near soil and rocks with thick about 30 cm from ground surface. The final results and their adequate uncertainty were estimated using the data obtained. The uranium concentration is the main source to release radon (^{222}Rn) from this mine to surrounding environment.

The gamma dose (μSv/h) at 1 m from ground was used as the second input data, the data was measured together with radon test points at the same position (inside 21 dwellings). The gamma doses have original from all of three natural decay chains (^{238}U, ^{232}Th, and ^{4} K). The data will provide information the rest contribution of radon at the local of radon test data (the radon is from ^{238}U at situ). The gamma dose were recorded by potable gamma survey model DKS-96 with NaI(Tl) scintillation crystal of Russian. Both of the potable Gamma and Gamma Surveyor were calibrated at the

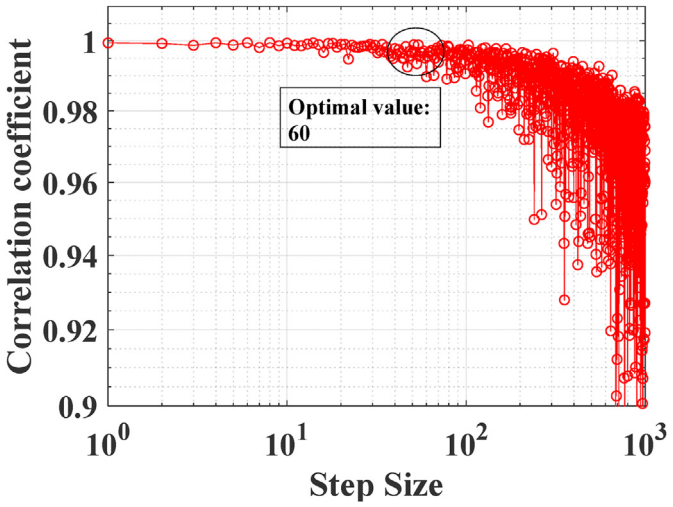


Fig. 4. The correlation coefficient between the raw and reduced data in function of Step Size.

National Atomic Energy Agency of Vietnam. The third and fourth input data are distance (m) and strike (degree) parameters which are defined distance from Gamma Surveyor points to radon test and strike follow trending direction from Gamma Surveyor to radon test points respectively. The distance parameter was inputted related to dispersion by time accumulation of movement, which were different by each of data points. It is similar with strike related to trending direction from each of uranium points (the point has uranium concentration values which were measured) and each of radon test points. Although parameters such as topography, wind direction, wind velocity, climate, temperature are also affecting the radon dispersion but when using the radon concentration by accumulated record of measurement of CR-39 detectors during 3 mouths, those parameters could be neglected.

For the radon test data at 21 dwellings surrounding mine, the detector CR-39 was used for this survey, which were placed inside RADUET chambers made by Radosys Ltd-Hungary. The CR-39 track etch detector with two diffusion chambers are dedicated to detect the radon and thoron activity (Tokonami et al., 2005). The first chamber is used for detecting only radon and the secondary chamber is sensitive for both radon and thoron. A simple calculation factors separate radon and thoron activity data results follow (Dung et al., 2014). At every dwelling the CR-39 detector was hung at a distance of 2 m from door and walls and at a height of the 1.5–1.8 m from the Earth surface. The CR-39 detectors were exposed for a period of three months. The average radon and thoron concentrations was estimated by averaging measured concentrations in this time period. After the exposure time the track detector pairs were recovered and transferred to the National Atomic Energy Agency of Vietnam for processing where the detectors were chemically treated and determined of the radon and thoron activities. All the coordinates of the measurement of gamma, spectrum gamma and radon, thoron points were recorded by Garmin GPS, version 60CSx with high accuracy ± 2–3 m.

The number of 1,000,000 experiments was then used to construct the database to perform the simulation. It included 4 variables as inputs, namely the uranium concentration (ppm), gamma dose (μSv/h), distance (m) and strike (degree). The dispersion radon activity (Bq/m^3) was considered as the output variable (Table 1). The dataset was scaled into the range of [0; 1] to minimize the bias between variables. For illustration purpose, Fig. 2 displays the correlation graphs and histograms of all input variables versus the dispersion radon activity in the present database.

2.2. Methods used

2.2.1. Artificial Neural Network (ANN)

ANN is a well-known machine learning algorithm which aims at mimicking the behavior of the human brain by means of connection. ANN can reconstruct a number of functions of human behavior, performed by a finite number of layers with different computing elements called neurons (Nathan et al., 2016). The ANN structure has three layers: input, hidden and output. Hidden layers provide a relationship between the input and output layers. ANN has been generally used for prediction, estimation, forecasting, pattern recognition, optimization, and to establish relationships between complex featured variables. The advantage of ANN is that no prior knowledge of object attributes is required, because even if the exact relationship between inputs and outputs is not known. ANN has the ability to learn the exact behavior between the inputs and outputs from the examples without any kind of physical involvement, can be trained to learn relationship, and able to extract the exact pattern between the input and output variables without any additional explanation (Mustafa et al., 2012). ANN can be considered black box models because they do not learn based on assumptions regarding input-output transmission function as well as physical interaction of parameters. In this study, ANN was used to predict the radon dispersion.

2.2.2. Validation indicators

To validate the performance of ANN, the coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and slope are used. Precisely, the R^2 values allow to identify the statistical relationship between the measurement and predicted dispersion radon activity. Its value ranging from 0 to 1, where 0 means no correlation and 1 is perfect correlation. Two criteria such as RMSE and MAE have the same units as the radon activity and lower values of RMSE, MAE indicate good accuracy of prediction of dispersion radon activity using ANN algorithm. Slope indicates the slope of regression fit compared with the perfect linear fit. The values of R^2, RMSE and MAE are estimated using the following equations:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \widehat{u}_i\right) \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(u_i - \widehat{u}_i\right)^2} \tag{2}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(u_i - \widehat{u}_i\right)^2}{\sum_{i=1}^{n}\left(u_i - \widehat{u}\right)^2} \tag{3}$$

where u_i is defined as the measurement values of dispersion radon activity, $\widehat{u}_i$is the predicted dispersion radon activity given by ANN algorithm, $\widehat{u}$ is defined as the mean value of the u_i, and n is the number of the considered samples.

2.2.3. Methodological flow chart of this study

Methodology of this study is presented in Fig. 3 which can be carried out by several main steps including (1) data preparation: data used in this study was collected from Singquen mine, Vietnam with 1,000,000 experiments which included the values of 4 input

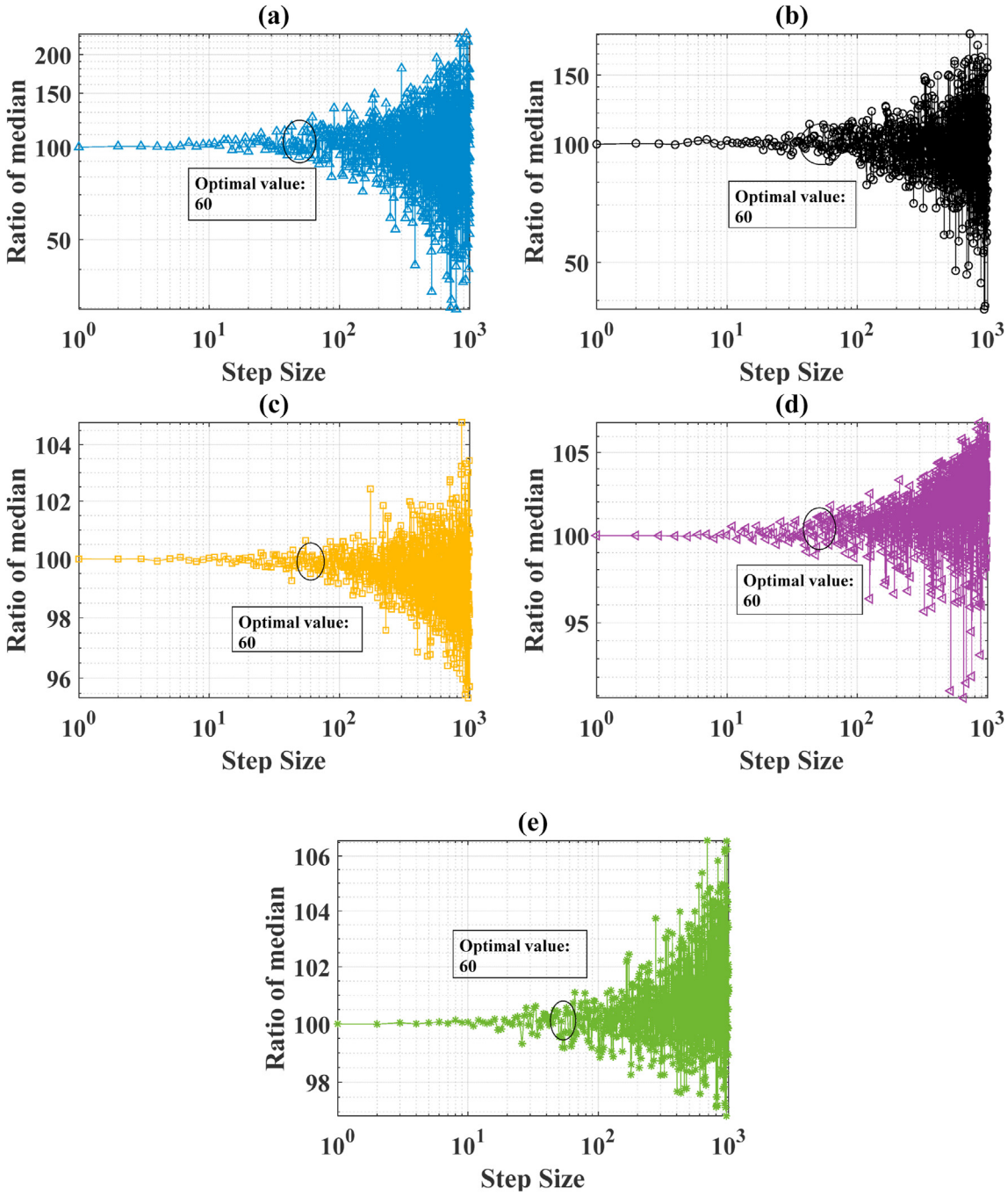


Fig. 5. The ratio of median evaluation between the raw and reduced data in function of Step Size for (a) uranium concentration; (b) gamma dose, (c) distance, (d) strike and (e) dispersion radon activity.

variables namely the uranium concentration (ppm), gamma dose (μSv/h), distance (m) and strike (degree) and one output variable namely dispersion radon activity (Bq/m^3), (2) dataset splitting: In this work, we have used the holdout sample method for model validation (Keane and Wolpin, 2007), the data collected was split into two parts to construct training (70%) and testing (30%) datasets, out of these, training dataset was used to train the ANN model and testing dataset was used to validate the predictive capability of the ANN model, (3) Training and optimizing the structure of ANN to find the best architecture (4) evaluation of the ANN predictive capability: the ANN predictive capability was validated using testing datasets and quantitative indicators (RMSE, MAE, and R^2), and its predictive capability was compared with several benchmark ML models such as Deep Neural Networks (DNN), Random Forest (RF), and Gausian Process Regression (GPR), and (5) sensitivity analysis is performed to evaluate the influence of input variables on the output.

3. Results and discussion

3.1. Reduction of dimension of dataset

As the amount of measurements in the dataset is considered as very high volume (i.e. over a million of data), the reduction of such

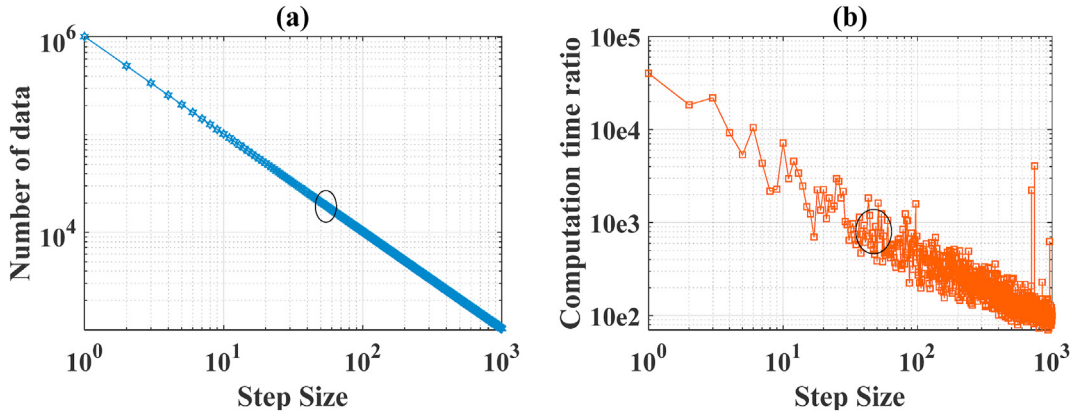


Fig. 6. The number of data and computation time ratio in function of "Step Size".

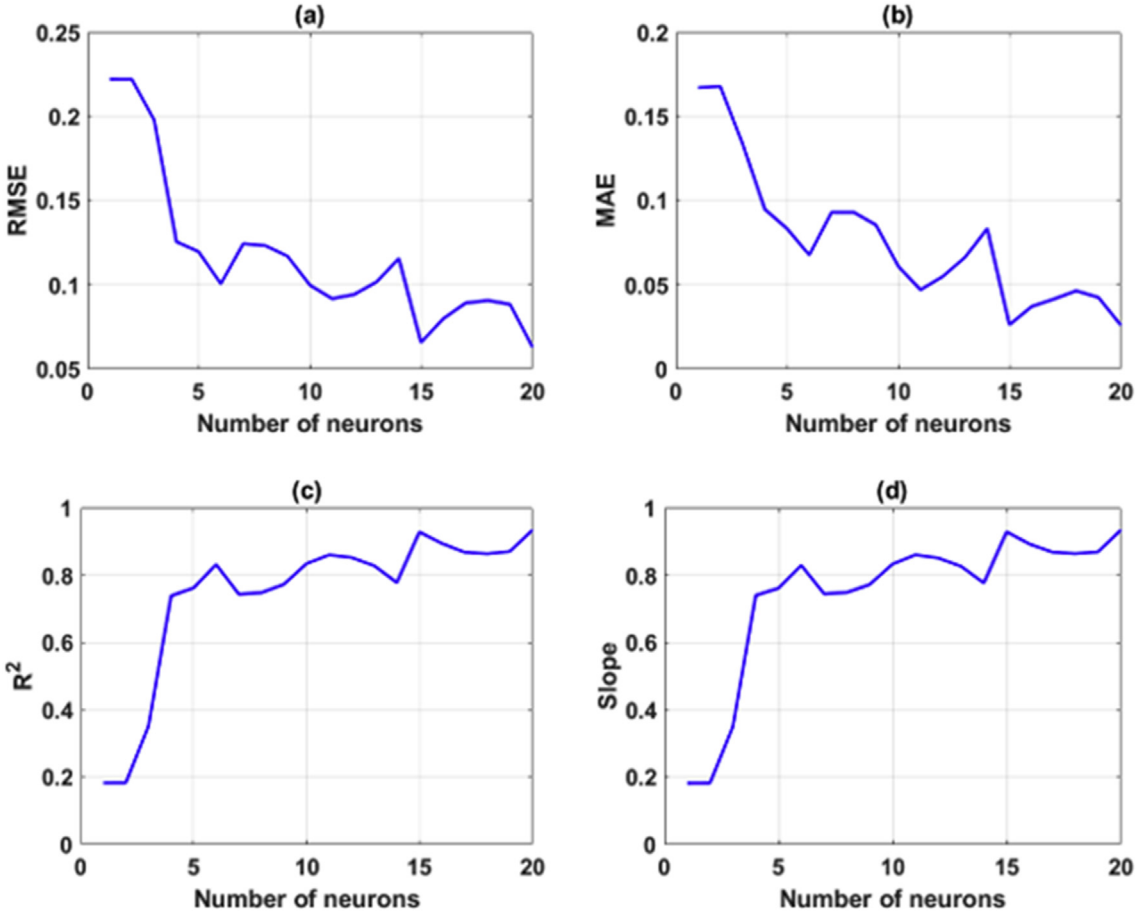


Fig. 7. Results of simulations with 1 hidden layer and the number of neurons from 1 to 20 neurons for (a) RMSE; (b) MAE; (c) R^2; and (d) Slope.

dimension is important to reduce efforts in computational time and memory. The analysis of representative subset of data is presented in this section. In accordance with statistical point of view, the reduction of number of data should ensure the correlation relation between the raw and the reduced data. Moreover, the reduced subset of data must be reliable enough to not distort the tendency of the raw data. To this aim, the correlation coefficient between the raw and the reduced data was analyzed as plotted in Fig. 4. In this study, the "Step Size" is defined as the dividend with respect to the total number of the raw data, the divisor. It means that if "Step Size" = 10, the reduced database contains the initial samples (i.e. 1,000,000 data) divided by "Step Size", making the total number of samples in the reduced dataset is equal to 100,000. It is seen that the value of 60 Step Size could be reasonably accepted as the optimal value to reduce the input space (Fig. 4).

Another statistical analysis showing the influence of "Step Size" such as ratio of median value (in %) between the raw and reduced data is presented in Fig. 5a, b, c, d and e for uranium concentration, gamma dose, distance, strike and predicted dispersion radon activity, respectively. The number of 60 step size is also confirmed as the optimal value.

In addition, the number of data and computation time presented

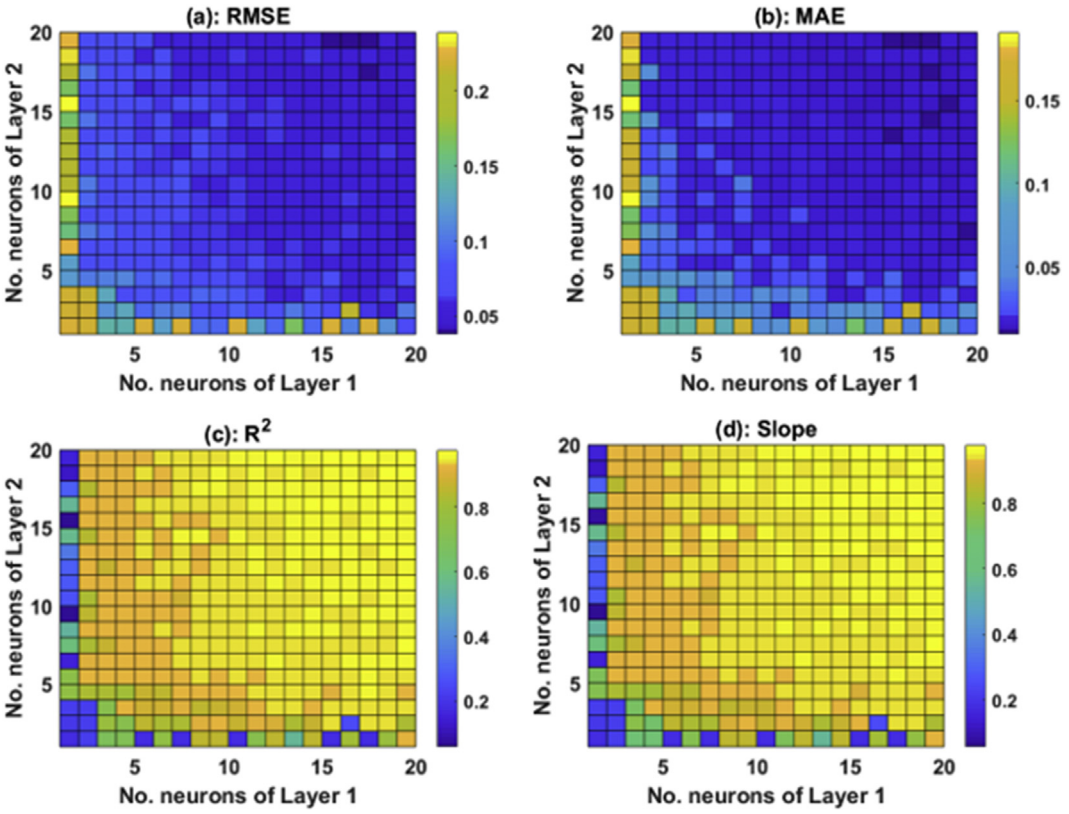


Fig. 8. Results of simulations with 2 hidden layer and the number of neurons from 1 to 20 neurons for (a) RMSE; (b) MAE; (c) R^2; and (d) Slope.

in ratio, with respect to the raw dataset, are also shown in function of Step Size in Fig. 6. As expected, it is seen that the relationship between the number of data and computation time was highly linear with "Step Size". Finally, the "Step Size" of 60 was chosen to perform the ANN simulations. It exhibited a gain of 40 times in function of computation time than using the raw dataset (Fig. 6b). Moreover, the reduced dataset contained only 17,000 data instead of 1,000,000 of data in the raw dataset, while keeping the same statistical information.

3.2. Validation of Artificial Neural Networks

3.2.1. Optimization of ANN structure

Performance of ANN model depends significantly on the selection of structure of ANN. In this section, the structure of ANN was analyzed and optimized to get the best performance of the model for prediction. Firstly, the parametric study was performed with ANN using 1 hidden layer. The number of neurons in the hidden layer is varied from 1 to 20. Fig. 7 shows the results of error criteria (RMSE, MAE, R^2 and Slope) in function of different number of hidden neurons. The Slope is defined as the gradient between the linear fit line and the perfect linear fit. A similar trend is observed for all error criteria, as the performance of the ANN model is increased with higher number of neurons in the hidden layer. It is noticed that at least 5 neurons in the hidden layer are required to achieve reasonable prediction results.

Fig. 8 compares the prediction performance of ANN with 2 hidden layers, using RMSE, MAE, R^2 and Slope as validation criteria. The number of neurons in each hidden layer is varied from 1 to 20. Again, it is easily observed that the best prediction results are observed when using the maximum number of neurons in both hidden layers. The values of RMSE and MAE have a strong tendency toward lower values, whereas those of R^2 and Slope tends to increase with higher number of neurons.

The highest value of R^2 with respect to the case of ANN using 1 hidden layer is $R^2 = 0.934$, whereas that of the case using 2 hidden layers is $R^2 = 0.9766$. As a conclusion for the parametric study, the higher the number of neurons, the better the predictive results achieved. In short, the ANN structure containing 2 hidden layers were chosen for further investigations as the prediction results exhibited higher precisions.

3.2.2. Predictive capability of ANN

As a consequence of the parametric study in the previous section, the structure ANN [4-20-20-1] was chosen to perform the predictive capability of ANN. Fig. 9 shows the prediction results in a probability distribution form for the training, testing and overall dataset. Besides, Fig. 10 explores the cumulative distribution for the three datasets. All the values with respect to the performance indicators are indicated in Table 2. It is observed that a highly statistical correlation was achieved between actual values of dispersion radon activity and its corresponding predicted values, for the training, testing as well as the whole dataset. In terms of the quality assessment indicators, for the training part, $R = 0.974$, $R^2 = 0.948$, $RMSE = 0.056$, $MAE = 0.020$, whereas for the testing part, those values were $R = 0.970$, $R^2 = 0.942$, $RMSE = 0.059$ and $MAE = 0.020$. It was also noticed that the regression line between the actual and predicted data exhibited a slope of 43.45°, 43.38° and 43.43° for training, testing and all data, respectively. The slope was very close to 45°, showing a great performance in regression capability of the ANN prediction model. Graphs of cumulative distribution between the actual and predicted data for the training, testing and all dataset are also given (Fig. 10). Other performance indicators such as the mean of error (ErrorMean) and the standard

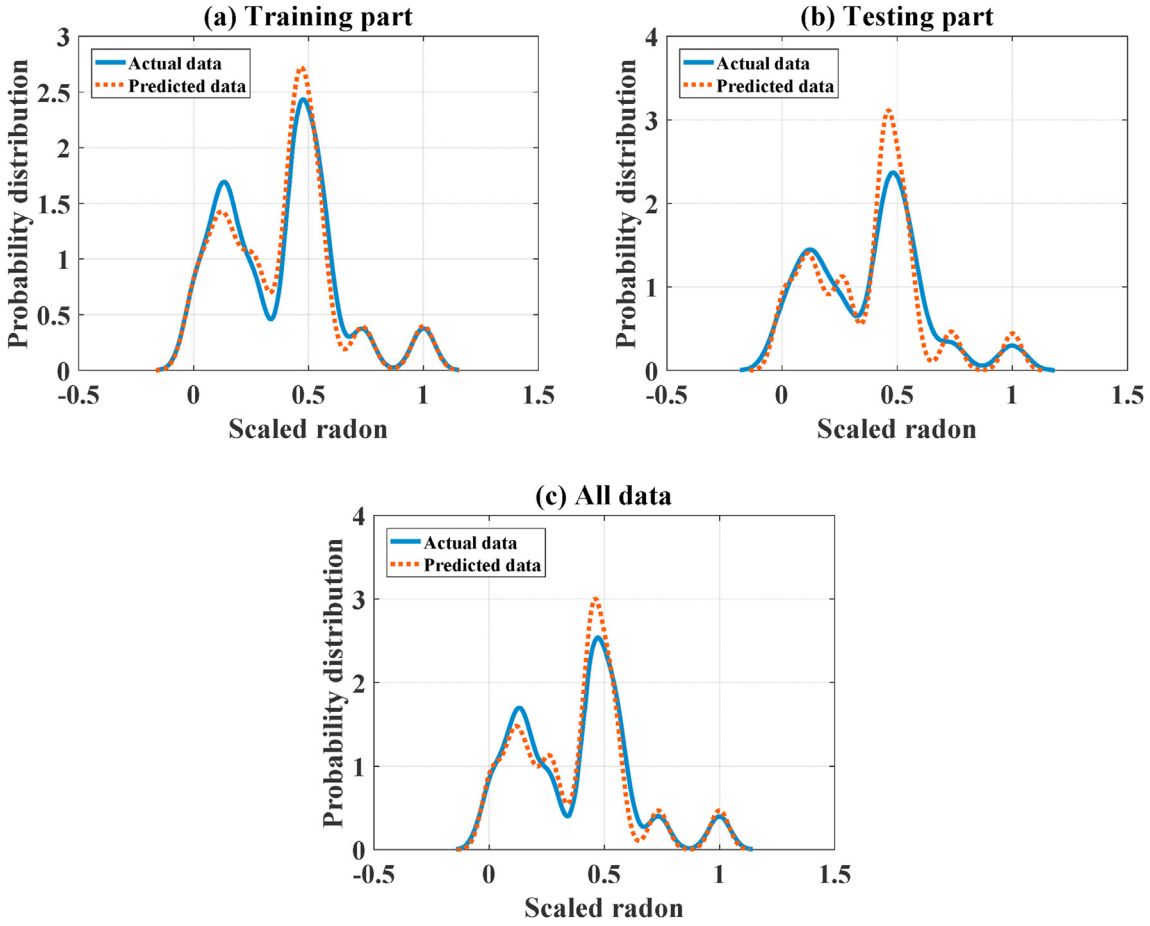


Fig. 9. Probability distribution between actual and predicted data for (a) training part, (b) testing part and (c) all data.

Table 2
Value of performance indicators.

Indicator	Training part	Testing part	All data
R	0.9737	0.9703	0.9728
R^2	0.9485	0.9415	0.9460
RMSE	0.0561	0.0589	0.0572
MAE	0.0202	0.0203	0.0202
ErrorMean	0.0003	0.0004	0.0003
ErrorStD	0.0561	0.0589	0.0570
Slope	0.9473	0.9471	0.9473
SlopeAngle (°)	43.45	43.45	43.43
Intercept	0.0200	0.0200	0.0200

deviation of error (ErrorStD) were also given in Table 2 for comparison propose.

In order to characterize locally the performance of the prediction model, comparison between actual and predicted data was analyzed at different quantile levels. To this aim, quantiles from 10 to 90% with a step of 10% were employed for tracking locally the behavior of the probability distribution of actual and predicted data. It is worth noticed that only quantiles from 10% to 90% were used, focusing on the most important statistical distribution. The results of such analysis are presented in Fig. 11a, b, c for training, testing and all data, respectively, whereas precise values of the quantiles are indicated in Table 3. The ratio (%) between the predicted and actual dispersion radon activity in each quantile level was also calculated and presented in Table 3. It is seen that for the testing part at Q_{30}, a high ratio of 141.8% was observed, similarly at Q_{40}, a ratio of 126.1% was obtained. However, it is deduced that the proposed model exhibited locally a strong efficiency in prediction the dispersion radon activity, as the average value for training, testing and all data are 97.4%, 102.3% and 100.2%, respectively. In general, the performance of ANN is good for prediction of radon dispersion at the study area.

3.3. Comparison of the optimized ANN with Deep Neural Networks, Random Forest, and Gausian Process Regression

In this section, we compared our optimized ANN with several models namely Deep Neural Networks (DNN), Random Forest (RF), and Gausian Process Regression (GPR). Out of these models, DNN is known as advanced ANN models with more than 2 hidden layers in the network structure (Cichy and Kaiser, 2019), RF is well-known as popular decision tree based machine learning model (Svetnik et al., 2003), and GPR is a nonparametric approach based on Bayesian theory (Schulz et al., 2018). In this study, the DNN was trained using different numbers of hidden layers (3−12) (Tables 4 and 5, Fig. 12 and Annex), the RF was trained using a minimum leaf size of 5 and 800 trees, and the GPR was trained with two different kernels such as Matern32 (GPR-32) and Matern52 (GPR-52). The comparison results show that ANN exhibits an excellent prediction capability, with the values of R2 higher than RF, and GRP. Besides, the values of errors such as RMSE and MAE of ANN are lower than those of GPR and RF. Similarly, the results of ANN compared with DNN using more than 3 hidden layers are presented in Table 5. It could be observed that an increase in the hidden layer number reduces the

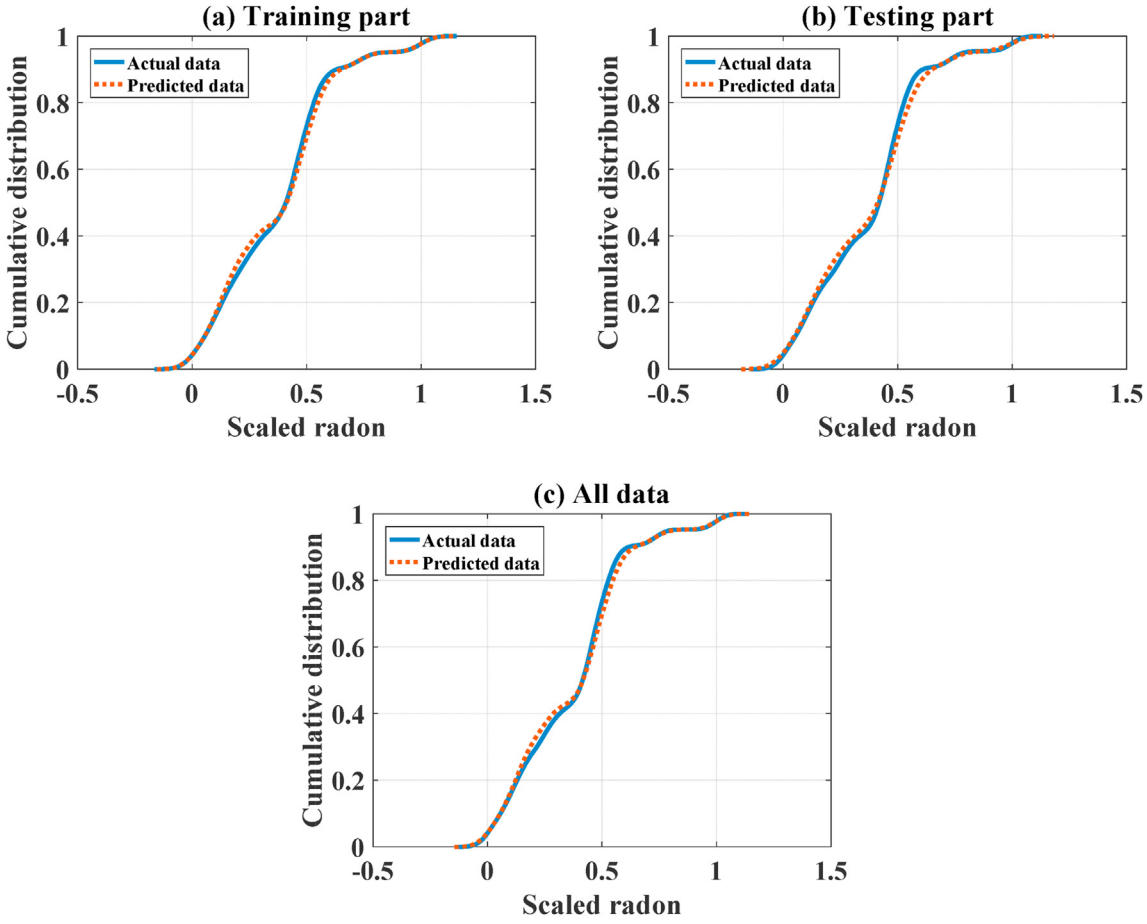


Fig. 10. Cumulative distribution between actual and predicted data for (a) training part, (b) testing part and (c) all data.

prediction accuracy of neural network. Therefore, it could be concluded that the ANN [4-20-20-1] is the best predictor of the problem.

3.4. Sensitivity analysis

In this section, a sensitivity analysis of input variables based on PDP characterization is presented. PDP characterization technique allow tracking the influence of input variables through the numerical prediction model such as ANN. In this study, the local effect of each input variable using the PDP technique is presented in Fig. 13a, b, c and d for uranium concentration, gamma dose, distance and strike, respectively. It is seen that all input variables govern the dispersion radon activity, but at different amplitudes. Moreover, several appropriate fits were applied for the PDP curves in order to quantify the variation. For instance, a quadratic fit was used for tracking information for uranium concentration and distance, a linear fit was used for strike and a Fourier equation was applied for gamma dose. It should be noticed that those equations are also indicated in Table 6. The fit equations could be useful as a quick empirical formulation to estimate the effect of the variation of input variables.

In order to analyze the amplitude of the PDP curves determined previously, two criteria were proposed based on (i) the area of the PDP curve with respect to its min value and (ii) the fluctuation level of each curve. Trapezoidal numerical integration was applied for calculating the area level of each PDP curve, whereas the fluctuation level was quantified through its standard deviation. The results are presented in bar graph mode in Fig. 14a for area level and 14 b for fluctuation level. It is seen that the gamma dose was the most influenced variable on the dispersion radon activity. The gamma dose exhibited the highest value of area level as well as the highest level of fluctuation. Same level of influence is considered for all other variables.

In preliminary, it could notice here that the fit equations with quick empirical formulation to understand and estimate the effect of the variation of input variables follow Strike < Distance < Uranium concentration < Gamma dose in order. So, it could say that the gamma dose is the most important variable in comparison with others. It is worthy when gamma dose is the most important variable, the data variable concerning to close and direct to local radon concentration at situ of radon monitoring measured point. Because of the radon resource is at monitoring measured place, the places are dwellings surrounding mine. Therein, the radon has original from local ^{226}Ra which one of the main contribute to local monitoring measured gamma dose. This radon resource at situ, inside closed dwellings, be accumulated by time, do not affect by wind or not much effect by distance and strike, and easy to reach to CR-39 detectors. The reasons setup for gamma dose (related to local radon concentration) became the most important variable. The rest of radon contribute to CR-39 was reached to CR-39 have original from uranium concentration which distribute inside Sinquyen Mine. The radon needs time and effected dispersion by distance and strike from mine to measurement points (dwellings). In general, the uranium concentration, distance and strike could be the same role of important variables. But in different case

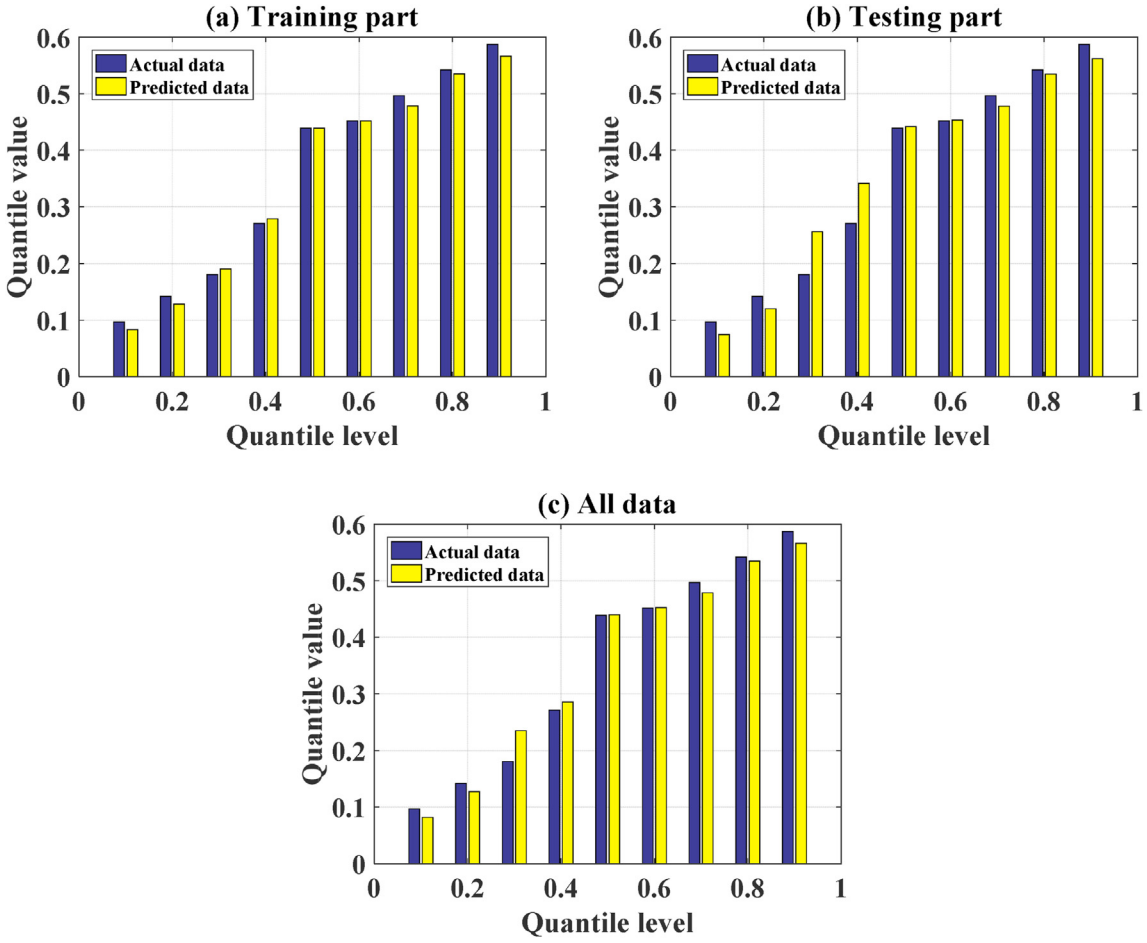


Fig. 11. Quantile values between actual and predicted data for (a) training part, (b) testing part and (c) all data.

Table 3
Value of quantile statistical analysis.

Quantile level	Training part			Testing part			All data		
	Actual data	Predicted data	Ratio (%)	Actual data	Predicted data	Ratio (%)	Actual data	Predicted data	Ratio (%)
Q_{10}	0.097	0.083	85.7	0.097	0.074	76.8	0.097	0.082	84.6
Q_{20}	0.142	0.128	90.5	0.142	0.120	84.5	0.142	0.127	89.6
Q_{30}	0.181	0.191	105.6	0.181	0.256	141.8	0.181	0.235	130.1
Q_{40}	0.271	0.279	102.9	0.271	0.342	126.1	0.271	0.285	105.3
Q_{50}	0.439	0.439	100.1	0.439	0.442	100.7	0.439	0.440	100.2
Q_{60}	0.452	0.452	100.1	0.452	0.453	100.3	0.452	0.452	100.2
Q_{70}	0.497	0.479	96.3	0.497	0.478	96.	0.497	0.478	96.3
Q_{80}	0.542	0.535	98.7	0.542	0.534	98.6	0.542	0.535	98.6
Q_{90}	0.587	0.566	96.4	0.587	0.562	95.7	0.587	0.566	96.4
		Average	97.4		Average	102.3		Average	100.2

Table 4
Comparison of the optimized ANN and other benchmark ML models.

Indicator	ANN		GPR-52		GPR-32		RF	
	Train	Test	Train	Test	Train	Test	Train	Test
R	0.9739	0.9703	0.8570	0.8790	0.8619	0.8813	0.9867	0.9698
R^2	0.9485	0.9415	0.7344	0.7726	0.7429	0.7767	0.9736	0.9405
RMSE	0.0561	0.0589	0.1272	0.1163	0.1252	0.1152	0.0429	0.0620
MAE	0.0202	0.0203	0.0716	0.0629	0.0699	0.0596	0.0221	0.0322
Err.Mean	0.0003	0.0004	0.0011	0.0008	0.0008	0.0002	0.0004	0.0006
Err.StD	0.0561	0.0589	0.1273	0.1164	0.1252	0.1152	0.0429	0.0620

Table 5
Performance of DNN with different number of hidden layers (Appendix).

Datasets	Number of hidden layers									
	3	4	5	6	7	8	9	10	11	12
Train	0.963	0.956	0.955	0.948	0.942	0.933	0.928	0.919	0.91	0.906
Test	0.942	0.933	0.935	0.922	0.914	0.919	0.909	0.901	0.881	0.886

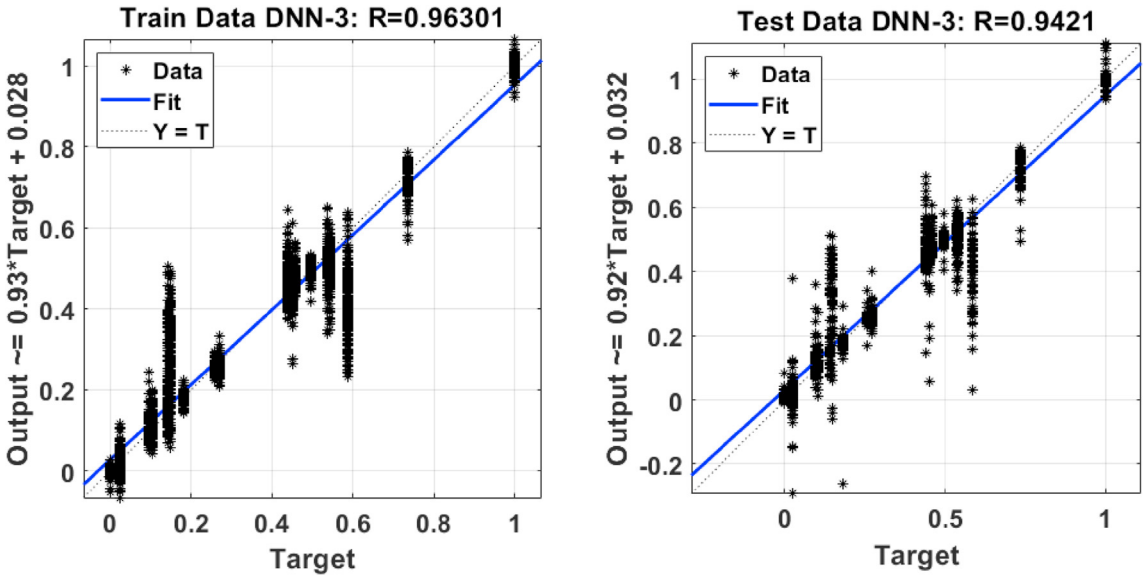


Fig. 12. Typical plot of DNN performance with 3 hidden layers.

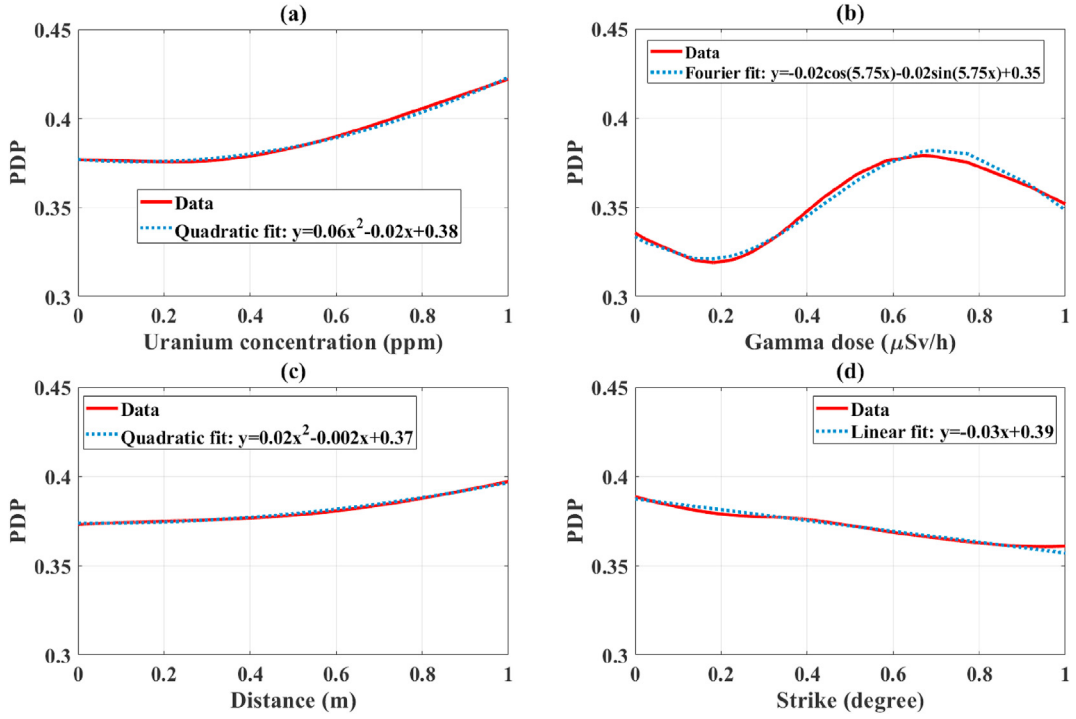


Fig. 13. PDP curves for (a) uranium concentration, (b) gamma dose, (c) distance and (d) strike as well as the most appropriate fits.

Table 6
Most appropriate fits for PDP characterization.

Variable	Most appropriate fit	Equation
Uranium concentration	Quadratic	$y = 0.06x^2 - 0.02x + 0.38$
Gamma dose	Fourier	$y = -0.02\cos(5.75x) - 0.02\sin(5.75x) + 0.35$
Distance	Quadratic	$y = 0.02x^2 - 0.002x + 0.37$
Strike	Linear	$y = -0.03x + 0.39$

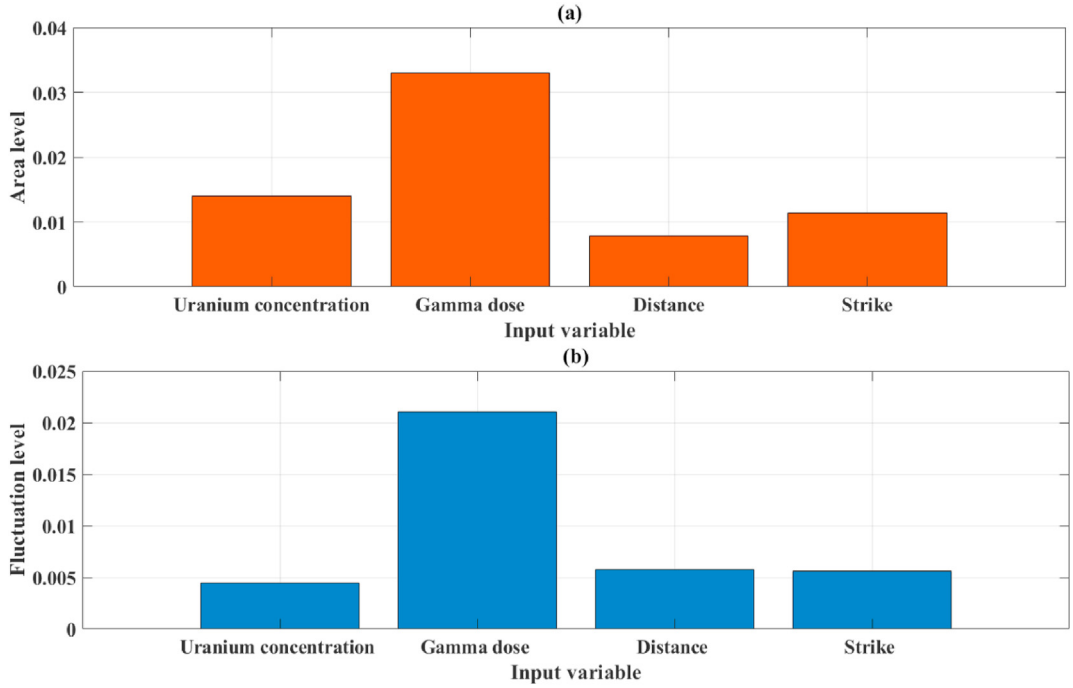


Fig. 14. Bar graphs for analysis of (a) area level and (b) fluctuation level from PDP curves.

they can replace of position such uranium concentration for the second important variable for assessment of area level but the last one when mention of fluctuation level.

4. Conclusions

In this study, ANN machine learning method was developed for prediction of radon dispersion based on four input data variables including the distance, strike, Uranium concentration, gamma dose at the case study of the Sinquyen mine. The results show that ANN performed with for prediction of radon dispersion in this study ($R^2 = 0.9342$ with 1 hidden layer and $R^2 = 0.9766$ for two hidden layers). The sensitivity analysis results showed all input variables govern the dispersion radon activity with different amplitudes and fitted with different equations of linear, quadratic and Fourier equations, but the gamma dose is the most influenced and important variable in comparison with strike, distance and uranium concentration variables for prediction of radon dispersion.

With the achievements of developing of machine learning method for prediction of radon dispersion will promise useful and effective for social and reduce cost of radon monitoring and measurement for the similar structure mines as well as areas have high natural radioactivity at Vietnam in particular and in the world. It means when we want to predict or mapping of radon concentration, this study might help to reduce the cost, time, procedure of radon measurement with only low cost of gamma spectrometry and gamma dose surveys. The great machine learning method application will open new method to predict and mapping radon in future which is one of the important parameters to assess of high quality and safe of fresh air. To extend this study, we would like to deal with real-time monitoring of input mutil-data collection of different parameters using the proposed model. This study can also be applied for the data collected by CR-39 measurement.

Credit author statement

Hao Duong Van, Hai-Bang Ly, Binh Thai Pham, Conceptualization, Methodology, Software. **Trinh Dinh Huan, Son Nguyen Thai**, Data curation. All authors, Writing – original draft. **Hao Duong Van, Trinh Dinh Huan**, Visualization, Investigation. **Hai-Bang Ly, Binh Thai Pham**, Supervision. **Hai-Bang Ly, Binh Thai Pham**, Software, Validation. **Hao Duong Van, Hai-Bang Ly, Binh Thai Pham**, Writing- Reviewing and Editing.

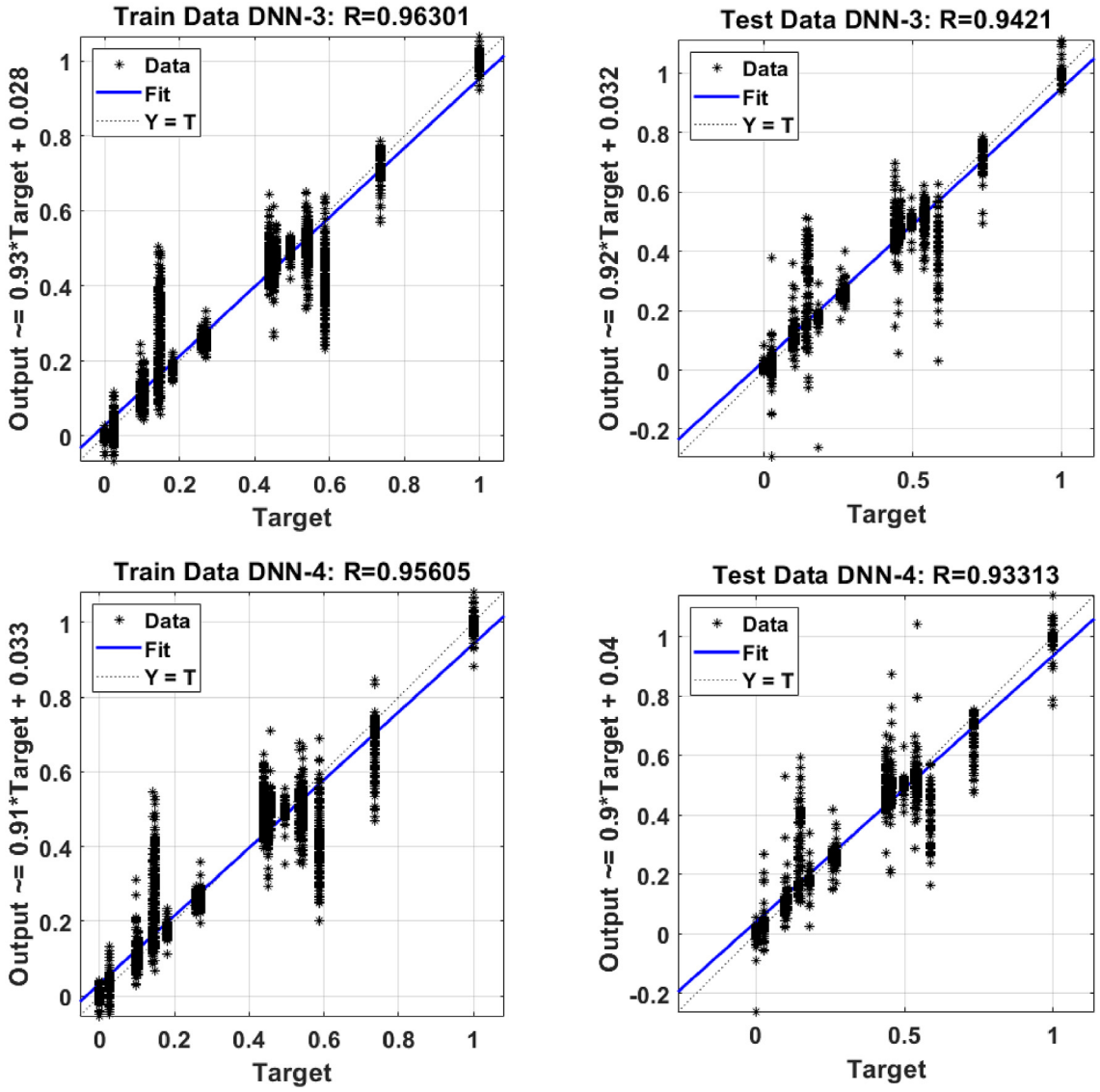
Declaration of competing interest

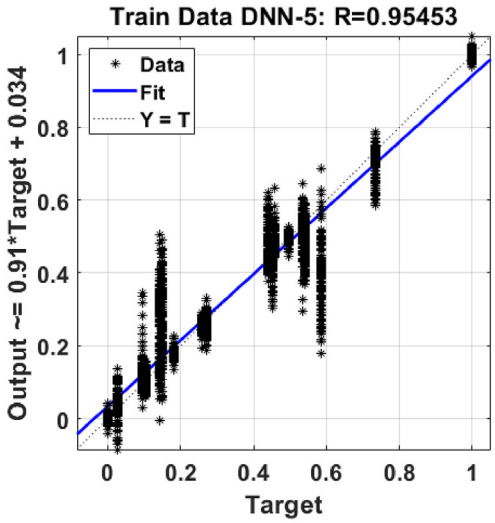
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

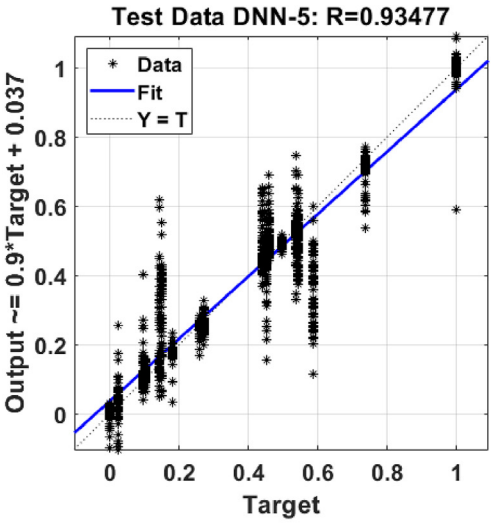
Appendix

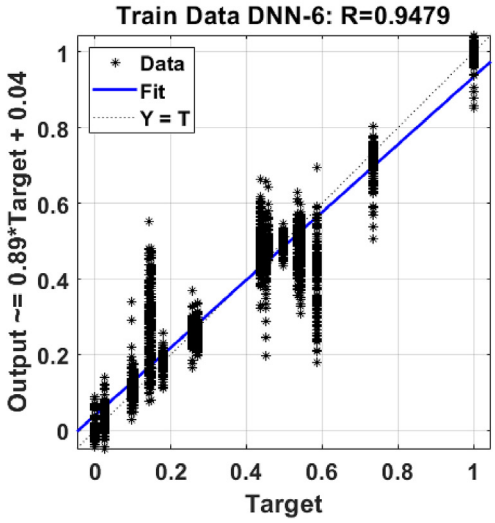
Performance of DNN with different number of hidden layers.

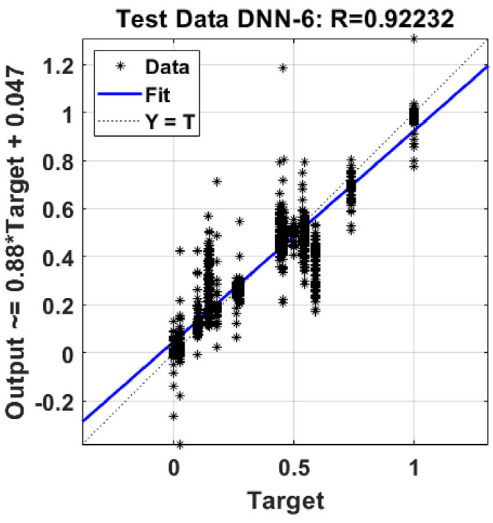
Performance of DNN with different number of hidden layers

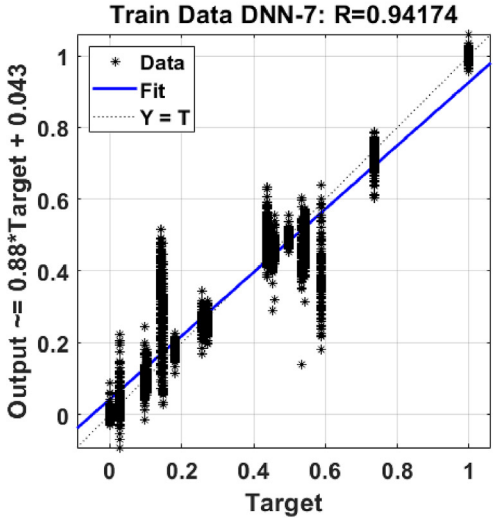


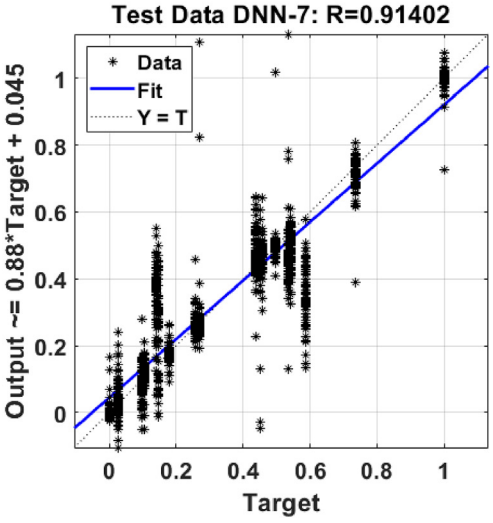

Train Data DNN-5: R=0.95453
Data
Fit
Y = T
Output ~= 0.91*Target + 0.034
Target


Test Data DNN-5: R=0.93477
Data
Fit
Y = T
Output ~= 0.9*Target + 0.037
Target

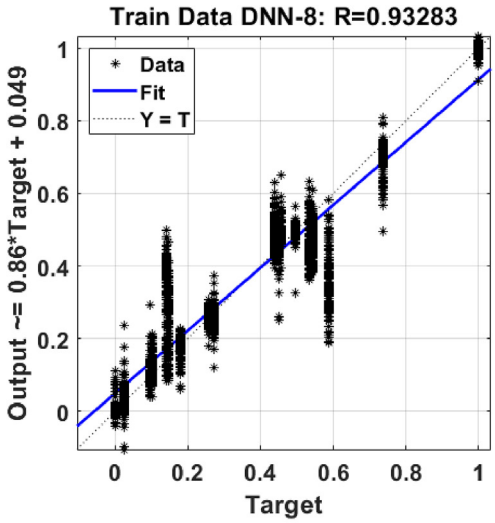

Train Data DNN-6: R=0.9479
Data
Fit
Y = T
Output ~= 0.89*Target + 0.04
Target

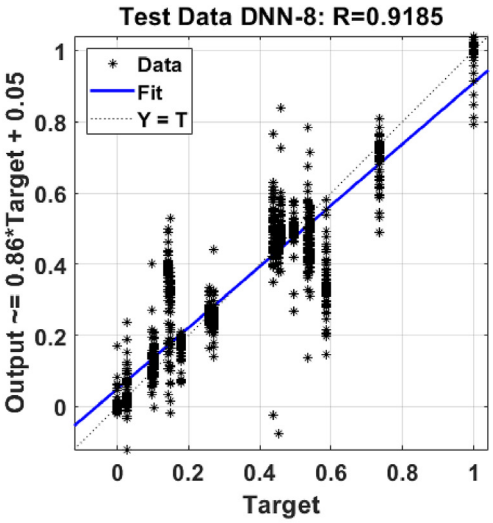

Test Data DNN-6: R=0.92232
Data
Fit
Y = T
Output ~= 0.88*Target + 0.047
Target

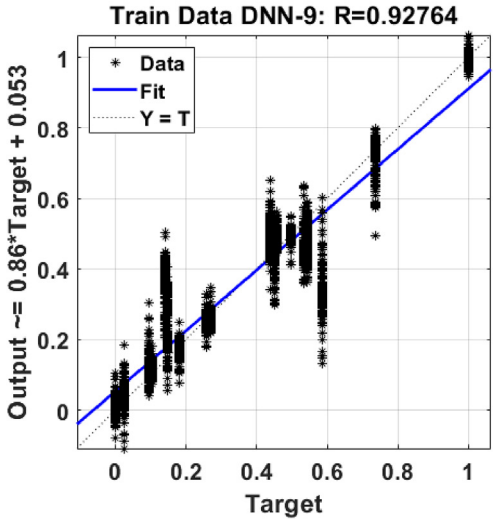

Train Data DNN-7: R=0.94174
Data
Fit
Y = T
Output ~= 0.88*Target + 0.043
Target

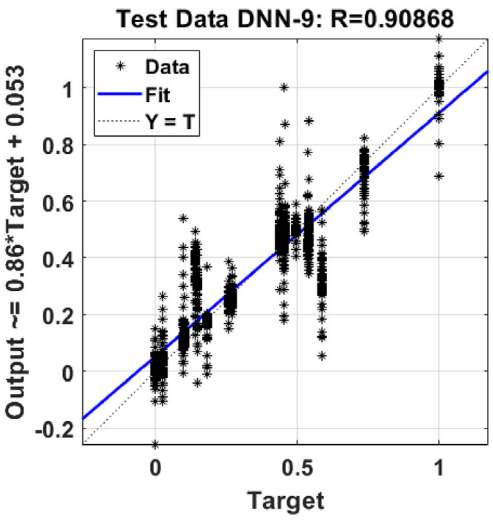

Test Data DNN-7: R=0.91402
Data
Fit
Y = T
Output ~= 0.88*Target + 0.045
Target

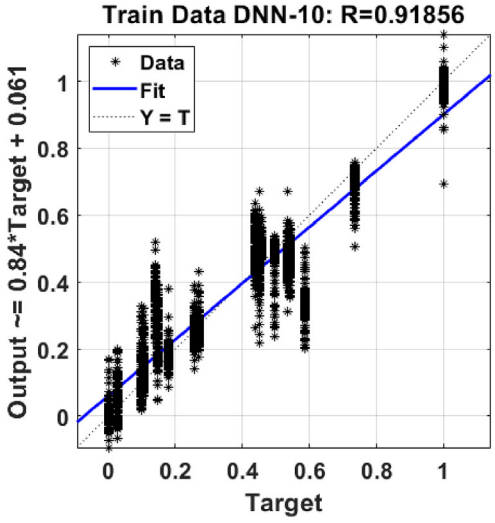
(***continued***).

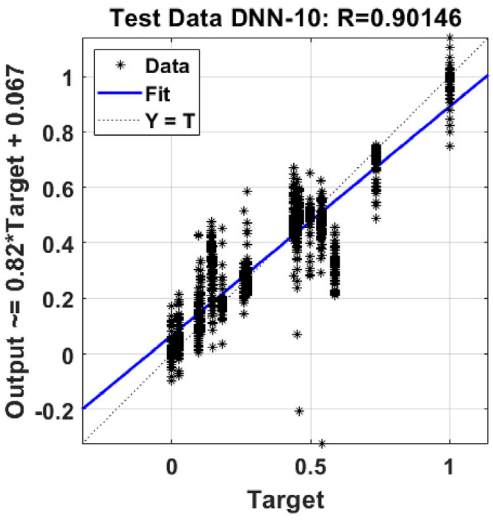

Train Data DNN-8: R=0.93283
Data
Fit
Y = T
Output ~= 0.86*Target + 0.049
Target


Test Data DNN-8: R=0.9185
Data
Fit
Y = T
Output ~= 0.86*Target + 0.05
Target


Train Data DNN-9: R=0.92764
Data
Fit
Y = T
Output ~= 0.86*Target + 0.053
Target


Test Data DNN-9: R=0.90868
Data
Fit
Y = T
Output ~= 0.86*Target + 0.053
Target


Train Data DNN-10: R=0.91856
Data
Fit
Y = T
Output ~= 0.84*Target + 0.061
Target


Test Data DNN-10: R=0.90146
Data
Fit
Y = T
Output ~= 0.82*Target + 0.067
Target

(***continued***).

References

Ahmad, N., Nasir, T., Rehman, J., Ullah, H., Uddin, Z., 2019. Risk assessment of radon in soil collected from chromite mines of Khanozai and Muslim Bagh, Balochistan, Pakistan. Environ. Technol. Innov. 16, 100476.

Banzi, F., Msaki, P., Mohammed, N., 2017. Assessment of radioactivity of 226Ra, 232Th and 40K in soil and plants for estimation of transfer factors and effective dose around Mkuju River Project, Tanzania. Min. Miner. Depos. 11, 93–100.

Carvalho, F.P., Madruga, M.J., Reis, M.C., Alves, J.G., Oliveira, J.M., Gouveia, J., Silva, L., 2007. Radioactivity in the environment around past radium and uranium mining sites of Portugal. J. Environ. Radioact. 96, 39–46.

Carvalho, F.P., Reis, M.C., 2006. Radon in Portuguese houses and workplaces. Proceed. International Conference Healthy Buildings HB2006, Held in Lisbon. 4–8 June 2006, pp. 507–511.

Chalupnik, S., Wysocka, M., 2008. Changes of radium concentration in discharge waters from coal mines in Poland as a result of mitigation. In: Uranium, Mining and Hydrogeology. Springer, Berlin, Heidelberg, pp. 839–850.

Cichy, R.M., Kaiser, D., 2019. Deep neural networks as scientific models. Trends Cognit. Sci. 23, 305–317.

Doering, C., McMaster, S.A., Johansen, M.P., 2018. Modelling the dispersion of radon-222 from a landform covered by low uranium grade waste rock. J. Environ. Radioact. 192, 498–504.

Dung, B.D., Van Giap, T., Kovacs, T., Toan, T.N., Minh, T.K., Quyet, N.H., Van Khanh, N., 2014. Estimation of radon and thoron caused dose at extraction and processing sites of mineral sand mining area in Vietnam (HA TINH province). J. Radioanal. Nucl. Chem. 299, 1943–1948.

Grant, C.N., Lalor, G.C., Balcázar, M., 2012. Radon monitoring in sites of economical importance in Jamaica. Appl. Radiat. Isot. 71, 96–101.

Hadad, K., Doulatdar, R., Mehdizadeh, S., 2007. Indoor radon monitoring in Northern Iran using passive and active measurements. J. Environ. Radioact. 95, 39–52.

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.

Hilton, J., 2008. Towards a management and regulatory strategy for phosphoric acid and phosphogypsum as co-products. In: Naturally Occurring Radioactive Material (NORM V)(Proc. Int. Symp. Seville, 2007), IAEA, Vienna, pp. 281–295.

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.

Keane, M.P., Wolpin, K.I., 2007. Exploring the usefulness of a nonrandom holdout sample for model validation: welfare effects on female behavior. Int. Econ. Rev. 48, 1351–1378.

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.

Laiolo, M., Cigolini, C., Coppola, D., Piscopo, D., 2012. Developments in real-time radon monitoring at Stromboli volcano. J. Environ. Radioact. 105, 21–29.

Le Khanh, P., Bui, D.D., Nguyen, D.C., Tibor, K., Nguyen, V.N., Duong, V.H., Nguyen, T.S., Vu, T.M.L., 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.

Ly, H.B., Pham, B.T., Dao, D.V., Le, V.M., Le, L.M., Le, T.T., 2019. Improvement of ANFIS model for prediction of compressive strength of manufactured sand concrete. Appl. Sci. 9, 3841.

Mustafa, M.R., Rezaur, R.B., Saiedi, S., Isa, M.H., 2012. River suspended sediment prediction using various multilayer perceptron neural network training algorithms—a case study in Malaysia. Water Resour. Manag. 26, 1879–1897.

Nathan, D., Thanigaiyarasu, G., Vani, K., India, C., 2016. Compasison of artificial neural network approach and data mining technique for the prediction of surface roughness in end milled components with texture images. Int. J. Adv. Eng. Technol. Vol, 592. VII/Issue I/Jan.-March, 587.

Nguyen, D.C., Le Khanh, P., Jodlowski, P., Pieczonka, J., Piestrzyński, A., Van, H.D., Nowak, J., 2016. Natural radioactivity at the sin quyen iron-oxide-copper-gold deposit in North Vietnam. Acta Geophys. 64, 2305–2321.

Oyedele, J.A., 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. Accel. Spectrometers Detect. Assoc. Equip. 619, 467–469.

Pham, B.T., Nguyen, M.D., Van Dao, D., Prakash, I., Ly, H.B., Le, T.T., Ho, L.S., Nguyen, K.T., Ngo, T.Q., Hoang, V., 2019. Development of artificial intelligence models for the prediction of Compression Coefficient of soil: an application of Monte Carlo sensitivity analysis. Sci. Total Environ. 679, 172–184.

Pourghasemi, H.R., Kariminejad, N., Amiri, M., Edalat, M., Zarafshar, M., Blaschke, T., Cerda, A., 2020. Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. Sci. Rep. 10, 1–11.

Ramola, R.C., Negi, M.S., Choubey, V.M., 2005. Radon and thoron monitoring in the environment of Kumaun Himalayas: survey and outcomes. J. Environ. Radioact. 79, 85–92.

Rashidi, A., Sigari, M.H., Maghiar, M., Citrin, D., 2016. An analogy between various machine-learning techniques for detecting construction materials in digital images. KSCE J. Civ. Eng. 20, 1178–1188.

Rivera, J.I., Bonilla, C.A., 2020. Predicting soil aggregate stability using readily available soil properties and machine learning techniques. Catena 187, 104408.

Schulz, E., Speekenbrink, M., Krause, A., 2018. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. J. Math. Psychol. 85, 1–16.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.

Thanh Duong, N., Van Hao, D., Duong, D.T., Phan, T.T., Le Xuan, H., 2020. Natural Radionuclides and Assessment of Radiological Hazards in MuongHum, Lao Cai, Vietnam. Chemosphere, p. 128671.

Tokonami, S., Takahashi, H., Kobayashi, Y., Zhuo, W., Hulber, E., 2005. Up-to-date radon-thoron discriminative detector for a large scale survey. Rev. Sci. Instrum. 76, 113505.

Van Dao, D., Jaafari, A., Bayat, M., Mafi-Gholami, D., Qi, C., Moayedi, H., Van Phong, T., Ly, H.B., Le, T.T., Trinh, P.T., 2020. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. Catena 188, 104451.

Van Hao, D., Dinh, C.N., Jodlowski, P., Kovacs, T., 2019. High-level natural radionuclides from the Mandena deposit, South Madagascar. J. Radioanal. Nucl. Chem. 319, 1331–1338.

Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., Lei, M., 2019. Machine learning in materials science. InfoMat 1, 338–358.

Who, 2009. World Health Organization Handbook on Indoor Radon: a Public Health Perspective. World Health Organization.

Wu, H.X., Wei, Q.L., Yang, B., Liu, Q.C., 2014. Fast prediction method of radon concentration in environment air. In: Applied Mechanics and Materials. Trans Tech Publ, pp. 819–822.

Xie, D., Wang, H., Kearfott, K.J., 2012. Modeling and experimental validation of the dispersion of 222Rn released from a uranium mine ventilation shaft. Atmos. Environ. 60, 453–459.

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.