IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 11, 2022, accepted March 24, 2022, date of publication March 30, 2022, date of current version April 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3163291

Anomalies Prediction in Radon Time Series for Earthquake Likelihood Using Machine Learning-Based Ensemble Model

ADIL ASLAM MIR[1,2], FATIH VEHBI ÇELEBI[1], HADEEL ALSOLAI[3], SHAHZAD AHMAD QURESHI[4], MUHAMMAD RAFIQUE[5], JABER S. ALZAHRANI[6], HANY MAHGOUB[7,8], AND MANAR AHMED HAMZA[9]

[1]Department of Computer Engineering, Ankara Yıldırım Beyazıt University, Ayvalı, Keçiören, 06010 Ankara, Turkey
[2]Department of Computer Science and Information Technology, King Abdullah Campus Chatter Kalas, The University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir 13100, Pakistan
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[4]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad 45650, Pakistan
[5]Department of Physics, King Abdullah Campus Chatter Kalas, The University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir 13100, Pakistan
[6]Department of Industrial Engineering, College of Engineering at Al-Qunfudhah, Umm Al-Qura University, Mecca 24382, Saudi Arabia
[7]Department of Computer Science, College of Science and Art at Mahayil, King Khalid University, Abha 62529, Saudi Arabia
[8]Faculty of Computers and Information, Department of Computer Science, Menoufia University, Menofia Governorate, Egypt
[9]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam Bin Abdulaziz University, AlKharj 16278, Saudi Arabia

Corresponding author: Adil Aslam Mir (adil.aslam@ajku.edu.pk)

This work was supported in part by the Deanship of Scientific Research at King Khalid University under Grant RGP 2/46/43; in part by the Princess Nourah Bint Abdulrahman University Researchers Supporting Project through Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, under Project PNURSP2022R303; and in part by the Deanship of Scientific Research at Umm Al-Qura University under Grant 22UQU4340237DSR04.

ABSTRACT The ability to predict the radioactive soil radon gas concentration is important for human beings because it serves as a precursor to earthquakes. Several studies have been conducted across the globe to confirm the correlation of radon emission dynamics and earthquakes, and concluded that the soil radon gas is the witness of anomalous behaviour before the occurrences of several earthquakes. This anomalous behavior can help to construct a better prediction model for earthquake forecasting. This paper aims at employing different ensemble and individual machine learning methods on real time radon time series data with different scenarios to predict anomalies in data caused by the seismic activities.The ensemble methods include boosted tree, bagged cart and boosted linear model while standalone machine learning methods include support vector machine with linear and radial kernels and k-nearest neighbors (*K*-NN). We tested the methods on a dataset recorded on the fault line located in Muzaffarabad. Time series data was collected over a period ranging from March 1, 2017 to May 11, 2018 including nine(09) earthquakes. The methods are tested in four different settings with 10 times 10 folds cross validation procedure over the time window of 1 to 4. The repeated 10 fold cross validation is performed to reduce the noise in the model performance estimation by replicating the 10 fold cross validation procedure 10 times. Statistical performance evaluation measures viz. root mean square error (RMSE), root mean squared log error (RMSLE), mean absolute percentage error (MAPE), percentage bias (PB), and mean squared error (MSE) have been calculated for the assessment of performance. In setting 1, the support vector machine with radial kernel performs better with the minimum RMSE score of 1381.023 when compared to other prediction models. In setting 3, it can be observed through different performance metrics such as RMSE, the value in the range [1262.864, 1409.616] which is minimum when other prediction models for predicting soil radon gas concentration dataset. For setting 4, the boosted tree model yielded the minimum RMSE and MAPE scores of 1573.174 and 0.056 respectively. Findings of the study shows that boosted tree and support vector machine with radial kernel proved to be better regression models for the prediction of anomalies in soil radon gas concentration during seismic activities. An important finding of this study suggests that by employing boosted tree ensemble method make us able to accurately predict soil radon gas concentration automatically from environmental parameters.

INDEX TERMS Automated system, earthquakes, ensemble methods, percentage bias, soil radon gas.

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

I. INTRODUCTION

The accuracy at which the decision support systems (DSSs) predict the samples, say for earthquake, medical diagnosis, etc. is of main concern in several domains especially where human lives are at stake. Earthquake is considered to be a major natural disaster and its unpredictability causes loss of human lives and infrastructure [1]. When talking about the earthquake prediction, there exist two different schools of thought. The first considers it to be a phenomenon which is impossible to predict in advance while others have spent a lot of resources and efforts to make it predictable. Various studies have been carried out in the past to tackle this challenging task through different angles [2]–[9]. The factor which makes it more challenging is the lack of technology to monitor the stress, pressure, changes occurring deep beneath the earth's crust using scientific instruments with more accuracy which may result in exploiting and extracting comprehensive seismic features for the purpose of analysis. During the earthquake preparation process beneath the surface, different geophysical and seismological processes occur's. Radon and one of its radioactive isotope thoron produced from uranium and thorium sources deep down the earth may potentially serve for the prediction of impending earthquakes. Radon has three naturally occurring isotopes viz. ^{222}Rn (usually called radon, stems its origin from radioactive ^{238}U series), ^{220}Rn (called as thoron, stems its origin into ^{232}Th radioactive series) and ^{219}Rn (called as action, stems its origin into ^{235}U radioactive series). Crustal abundance of ^{238}U (Uranium), ^{232}Th (Thorium) and ^{235}U (Actinum) isotopes are 2.7, 8.5 and 0.02 $\mu g\ kg^{-1}$ respectively. Though concentration of ^{232}Th is somewhat higher than ^{238}U in the earth crust but rate of production of ^{222}Rn and ^{220}Rn is about the same due to longer half life of ^{232}Th (14.1×10^9 years) as compared with ^{238}U(4.5×10^9 years). Out of three naturally occurring isotopes ^{222}Rn is more important due to its longer half life (3.825 days) as compared to ^{220}Rn (55.6 s) and ^{219}Rn (Actinon) [10]. Half-lives of later two isotopes restrict transport of these isotopes by diffusion method to short distances only. However thoron manages to reach earth surface but in lesser quantity than radon. In this article we shall focus on radon rather than other isotopes.

Several studies have been carried out across the globe focusing on earthquake prediction based upon anomalous behavior of radon gas in the atmosphere, soil, and water [3], [11]–[14]. The uneven behaviour of the radon in soil and water was correlated with the earthquake, first time, dated back in 1967 [15] and another study in 1976 also reported spikes in radon concentration before the occurrence of the earthquake [16]. Moreover, in 1978, another study reported unusual behaviour of radon concentration prior to earthquake [17], and resulted in extensive research activity to further explore the correlation between earthquakes and radon emission dynamics [18]–[25]. Moreover, the nature of carrier gases and other meteorological parameters definitely influence the radon emission underlying forces [25]–[28].

Consequently, with the recent advancements in computer science, different computational intelligence techniques have been successfully introduced to predict radon concentration from meteorological parameters [3]. Regression trees have been used to predict the radon soil gas concentration through environmental data such as pressure, rainfall, air temperature and soil temperature, and concluded that the prediction error increases a week before the earthquakes having magnitudes ranging from 0.8 to 3.3 [29], [30].A neural network system using radial basis function (RBF) has been tested that can be used as an alternative to traditional regression methods to isolate radon emission anomalies [31]. The proposed model was further tested and evaluated on future data set and the prediction accuracy 87.8% was acheived. Tareen *et al.* employed three different computational intelligence models to automatically detect anomalous behaviour in soil radon gas time series data by modelling the radon concentration with different statistical and meteorological parameters [11]. The findings of the study reveal that the irregular behaviour of radon concentration is caused by seismic activities. A study was conducted to optimize the machine learning model namely artificial neural network (ANN) for the accurate prediction of radon dispersion in Vietnam and concluded that ANN performed very well in order to predict radon dispersion with the lower values of performance metrics [32]. The soil radon gas concentration was estimated by employing a Deep Neural Network (DNN) using different environmental parameters and mapped the functional relationship between radon concentration and environmental parameters [33]. A new method was proposed which is based upon Adaptive Linear Neuron (Adaline) and estimated the soil radon gas concentration with associated environmental parameters [34]. The proposed methodology can efficiently differentiate the temporal variation of radon concentration related to environmental parameters. Sikder *et al.* employed the decision tree algorithm for the characterization of premonitory factors of low seismic activity that outperformed other regression-based techniques [35].

Machine learning explores the problem structure and construction of algorithms that can learn from and make predictions on data. It is a branch of artificial intelligence that deals with the development and study of algorithms that are capable of making models for predictions or decisions [36]. With the advent of technology, machine learning methods have shown significant results in various fields of studies such as medical diagnosis [37]–[42], banking [43], [44], market basket analysis [45], [46] and many more. A variety of methods are offered in this context such as Diagonal Linear Discriminant Analysis (DLDA) [47], k-Nearest Neighbors (k-NN) [48], Support Vector Machine (SVM) [49] and Random Forest (RF) [50] for classification and prediction purposes. Moreover, ensemble methods were also proposed such as bagging [51], boosting [52], [53] and stacking [54] where the final prediction is not made by a single model only but rather by aggregating the outcome of various weak learned models [55]. Ensemble methods

show significant improvement in performance than individual models in classification and prediction problems [56]–[59]. The improvement in the performance by employing ensemble methods is based upon the premise that prediction made by the ensemble is more accurate than relying on the individual classifier that constituted the ensemble [55].

The core idea of this research work is to investigate the ensemble methods and individual learning models for the accurate prediction of soil radon gas concentration time series data. Ensemble methods used in this paper are boosted tree, bagged cart and boosted linear model, and in the individual learning models' category, support vector machine with linear and radial kernels, and K-Nearest Neighbors (K-NN) are used. The testing of ensemble and individual learning methods are performed in different settings ranging from 1 to 4. Moreover, each setting consists of several windows from W_1 to W_4. The prediction of the soil radon gas concentration during the seismic activity or anomaly that captures the variations of the original concentration is of main interest in this study. The prediction model can better predict the soil radon concentration accurately that leads to the identification of anomalies in the time series. Instead of relying upon a single dataset for testing purposes, the testing phase is decomposed into different types of settings which are the incorporation of different seismic activities. Each of the settings leads to a different composition of training and testing sets. The time window scheme is employed to predict the radon concentration in different periods of time. The window comprises of the days before and after the occurrence of seismic events: a window size of 1 means 1 day before and after the seismic event. Likewise, window size of 3 and 4 means the samples which belong to 3 and 4 days before and after the occurrence of seismic activity. The impact of a seismic event ranged from its preparation phase (before the occurrence of a seismic event) to aftershocks. For experimentation, the dataset is recorded on the fault line present in Muzaffarabad; a city in Kashmir, administered by Pakistan over a period from 1^{st} of March 2017 to 11^{th} of May 2018 included 9 seismic events or earthquakes. The detailed description of seismic events along with their magnitude is presented below in Table 1. The cross validation procedure is applied which is 10 times 10 fold cross validation for training the models and tested using a test set provided by each setting.The original sample is randomly divided into 10 equal size subsamples in 10-fold cross-validation. One of the 10 subsamples is kept as validation data for testing the model, while the remaining 9 subsamples are used for training purposes. The cross-validation procedure is then repeated 10 times, with each of the 10 subsamples serving as validation data exactly once. To generate a single estimate, the 10 fold results are averaged. Further, this procedure is repeated 10 times and the model performance is estimated by averaging the performance across all folds and repeats. The basic idea behind the use of repeated cross validation is to incorporate all the samples in model training and validation as well as reduce the noise in the estimation of model performance. The experimentation is performed in the R language environment using the package CARET (Classification and Regression Training) [60].For performance evaluation, frequently used statistical metrics are computed such as RMSE, RMSLE, MAPE, PB and MSE. The ensemble and individual models are purely assessed upon the performance of the methods to efficiently capture temporal variations and functional relationships between radon concentration and environmental parameters.

II. MATERIAL AND METHODS

In this section, the statistical details of the soil radon gas time series dataset have been presented along with earthquake or seismic activities information. Moreover, a basic understanding of the ensemble and individual machine learning methods is also provided. The detailed information of the proposed simulation plan for prediction of soil radon gas concentration is also pictorially presented and discussed in details. Finally, the mathematical formulation of the performance metrics used for performance estimation of ensemble and individual machine learning methods for predicting soil radon gas concentration is also provided.

A. AREA OF STUDY

The Muzaffarabad city is the capital of state of azad Jammu and Kashmir, Pakistani administrated part of Jammu and Kashmir. It shares border with Pakistani provinces Khyber Pakhtunkhawa and Punjab towards west and south respectively. Eastern border is connected with the Indian administrated part of Kashmir. According to 2017 census, total population of city of Muzaffarabad was 149913. Muzaffarabad suffered from 2005 devastating earthquake with a magnitude 7.6M_w causing more than 80000 causalities in and around superbs of city. Muzaffarabad is a cup shaped valley. Air quality index (AQI) of Muzaffarabad is unhealthy for sentitive group of peoples. Particulate matter concentration ($PM_{2.5}$) in Muzaffarabad air is 6.6 times above the WHO air quality standards [61]. Since Muzaffarabad is seismically active area and has history of occurance of regular devastating earthquakes, so forecasting possible earthquake in future is a attractive field of study. We have installed RADON measuring station over a fault line passing beneath the Muzaffarabad.

B. DATA ACQUISITION

RTM-1688-2 SARAD nuclear Instrument was installed, for the continuous radiometric measurement of radon and meterological parameters, at Chehla location with latitude 34.39621 N and longitude 73.47347 E. Radioactive radon decays into its short living daughter products which are used to find radon concentration within the radon measurement chamber. Radon-222 decays into the Polonium-218 with the emission of alpha particle. Momentarily Polonium-218 becomes positively charged due to orbital electron scattering from emitted alpha particles. Positive ions of Polonium-218 is collected by working radon chamber and number of

TABLE 1. Earthquake details with date, magnitude and epicentre depth during the study period.

Earthquake #.	Earthquake Date	Earthquake Magnitude	Epicenter Depth (km)
E1	March 21, 2017	4.3	25
E2	March 23, 2017	2.5	156
E3	August 27, 2017	4.8	10
E4	September 23, 2017	4.6	61
E5	December 09, 2017	4.7	101
E6	February 03, 2018	0.8	157
E7	February 28, 2018	4.4	134
E8	March 14, 2018	4.9	10
E9	March 15, 2018	4.7	45

polonium-218 ions collected in chambers are proportional to the radon concentration. RTM 1688-2 works in slow and fast modes and stores the data on non-volatile memory using a circular architecture. The data acquired from the measurements is downloaded to a personal laptop using the seriel interface [62].

C. DATA DESCRIPTION

The dataset used for this work is ''soil radon gas time series data'', recorded on the fault line located at the Muzaffarabad city of Pakistan administered part of Kashmir as shown in Figure 1.The single reading was recorded after every 40 minutes, ensuing in 36 readings for the complete day. The concrete details of the radon measurement station and its instrumentation are reported elsewhere [3], [7], [11]. The dataset contains 15692 valid observations of radon concentration along with its environmental parameters such as thoron (Bq/m^3), temperature (^0C), relative humidity and pressure (mbar). During the data collection period, nine seismic activities were observed whose details with their magnitude are presented in Table 1. When considering the attribute of interest i.e. radon concentration (RN), the minimum and maximum observed radon concentration were 13743 Bq/m^3 and 28085 Bq/m^3 respectively. Moreover, the mean and median of the whole radon time series was found to be 21364 Bq/m^3 and 21569 Bq/m^3. During the seismic activity period, the minimum of radon concentration (RN) was observed with the concentration value of 16132 Bq/m^3 while the maximum was 26650 Bq/m^3. For thoron time series, the concentration of thoron,during seismic activities, varied from 2146 Bq/m^3 to 3734 Bq/m^3 respectively.

D. PROPOSED SIMULATION AND ANALYSIS PLAN

Figure 2 presents the complete experimental framework for this work. The simulation is executed for two different groups of machine learning methods presented as Group 1 and Group 2. Group 1 consists of ensemble methods for learning while Group 2 contains individual learning methods. The ensemble methods used in group 1 are boosted tree model, bagged cart model and boosted linear model while individual learning models are K-NN, SVMs with linear and radial kernels as presented in Group 2. The simulation is executed in 4 different settings ranging from setting 1 to 4 The basic purpose to introduce these settings is to investigate the prediction capability of the learned models on different test sets which included almost every seismic activity. Apart from the different distributions of training and testing data, the time window is also incorporated. The time window enables us to obtain data related to seismic activity along with all the samples of the days before and after the seismic activity as specified. Several investigations from the globe have confirmed the unusual behavior of soil radon gas concentrations prior to the occurrence of several earthquakes. This unusual behavior in soil radon before an earthquake could lead to the development of a better forecasting model that can lead to the prediction of soil radon gas concentration. The forecasting model can capture the temporal fluctuations in the soil radon time series by training and testing on multiple time windows. After every 40 minutes, a single reading is taken, totaling 36 readings for the complete day. The idea of the window is to extract the seismic activity along with the relevant time window (window of 1 means 36 readings before and after the seismic activity) which incorporates the non-seismic sample to seismic time series for better analyzing the variations.. The novelty of this work is to introduce settings that incorporate seismic activities in both training and testing sets to better analyze the forecasting models. The previously reported studies simply divide the dataset into seismic and non-seismic. The model was trained using non-seismic activity dataset. Further, the trained model is used to predict soil radon gas concentration in seismic activity dataset. For this work, the time window from 1 to 4 is used to extract the testing test. Each of the settings leads to a different distribution of training and testing data. Consider setting 1 presented in Figure 2, the training data consists of all non-seismic activity (NSA) samples along with the seismic activity (SA) data of 1,2,5,6,7,8,9 while testing set composed of samples belonging to seismic activity (SA) 3 and 4 with respect to time window ranged from 1 to 4. Thus, each setting along with a time window enables one to assess the performance of the models from group 1 and group 2 in a more efficient manner. The training set, splitted by each setting, is trained by ensemble methods as well as individual learning methods and results in their respective trained machine learning models. These ensemble

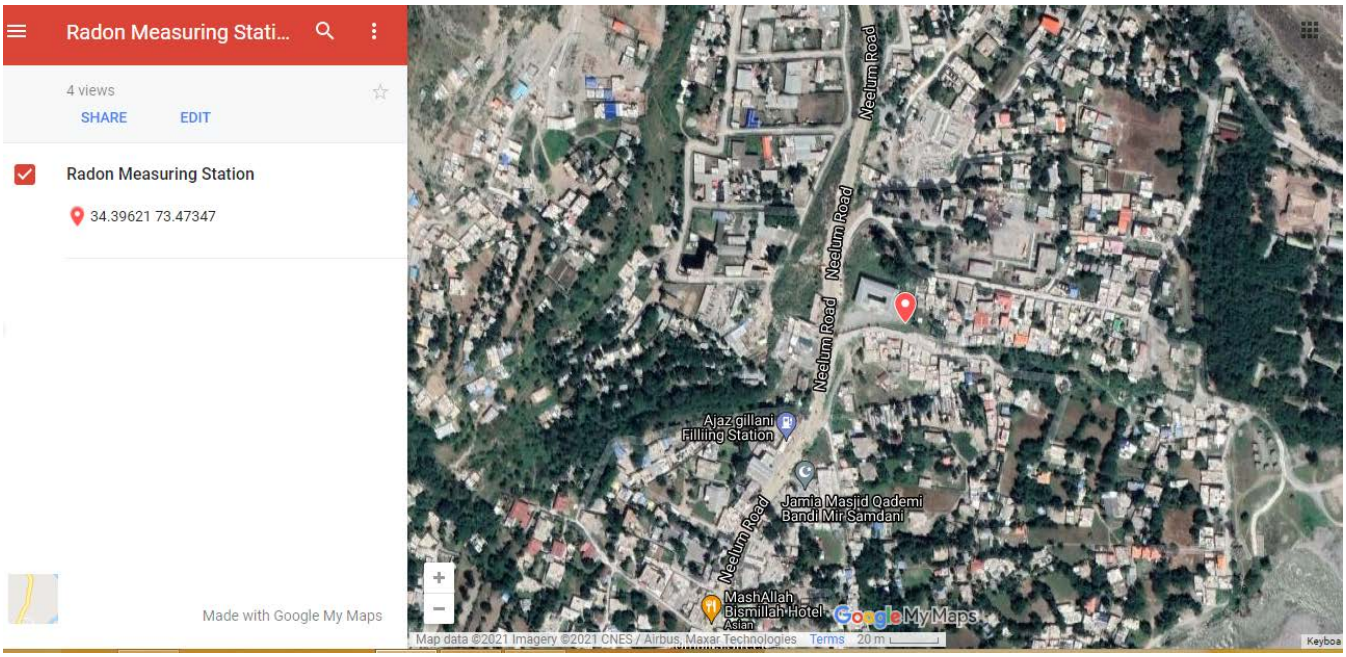

FIGURE 1. Soil radon gas concentration measuring station located at Muzaffarabad, Pakistan for data collection.

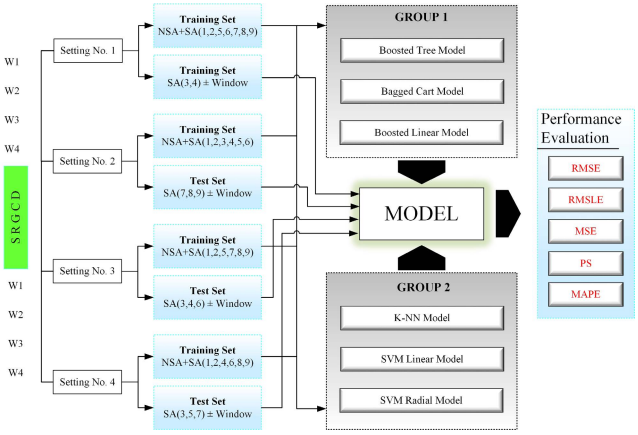

FIGURE 2. Proposed framework of this work.

and individual models are further tested by predicting the test set. The models are trained through a cross validation procedure which is 10 times 10 fold cross validation for this study. The predictions made by each model from groups 1 and 2 are assessed by calculating different statistical performance evaluation metrics. The performance metrics include RMSE, RMSLE, MAPE, PB and MSE.

E. ENSEMBLE METHODS

In machine learning and statistics, the ensemble is the collection of multiple models and is one of the self-efficient methods as compared to other basic models [55]. Supervised learning algorithms are extremely useful in searching through different solution spaces to predict suitable hypothesis space for certain problems.The ensemble technique combines different hypotheses to provide the best hypothesis. Basically, ensemble technique is used for obtaining a strong learner with the help of a combination of weak learners. While performing classification using ensemble methods, more computations are performed as compared to making predictions with a particular model so multiple models can be a way to help poor algorithms for performing well after doing extra computations. The ensemble method is also an example of supervised learning as firstly it is trained and then it makes predictions and represents a single hypothesis space. Experimentally, ensemble methods provide more accurate results provided that there is considerable diversity between the models.

1) BOOSTING AND BAGGING

In order to generate the different base learners in ensemble methods, sequential and parallel ensemble methods are used, such as boosting and bagging [63]. Sequential ensemble methods, such as boosting, are employed to exploit the dependence between the different base learners generated whereas in parallel ensemble method, bagging as a representative, is to exploit the independence between the base learners generated. Boosting ensemble method boosts the overall performance of a base learning algorithm in a residual-decreasing way [63]. On the other hand, the bagging ensemble method combines the independent base learner to reduce the error. The word bagging is the abbreviation of Bootstrap AGGregatING [51]. Bagging is designed to improve the accuracy of predictions in decision support systems by model averaging that helps to reduce the variance and minimizes the overfitting problem. In order to

perform bagging, m different bootstraps are created from the original training data. The base learning algorithm either for classification or regression is trained upon each bootstrap and this result in m individual base learners. In the areas where classification is of main concern, the final classifications are made by combining the base learners' classifications by plurality voting or averaging the probabilities of the estimated class.For regression problems, the new predictions are made by averaging the predictions of the individual models generated using different bootstraps. Consider X is a sample for which the prediction needs to be made, $BL_1(x), BL_2(X), \ldots\ldots, BL_m(X)$ are the predictions generated from individual base learners. The bagged prediction P_bag is the aggregation of the predictions from individual base learners formulated as:

$$P_{bag} = BL_1(x), BL_2(X), \ldots\ldots, BL_m(X). \quad (1)$$

This aggregation results in the reduction of the variance of an individual base learner and minimizes the overfitting problem as discussed above. For the base learning algorithms having larger variance (decision trees) than others, bagging works very well and improves its performance whereas the algorithm having higher bias (linear regression), the bagging results in less improvement of performance in classification and regression problems [55], [64]. The higher variance base learners are those learners for which a small change in the training data can make a major change in response values.

Boosting works by finding many rules of thumb using a subset of the training examples simply by sampling repeatedly from the distribution [65]. In subsequent iterations a new rule is generated using the subset of training examples. To make the boosting approach workable one of the methods is to focus on the difficult to predict/classify examples and to increase the weights of the examples that are misclassified. Therefore, the hardest examples would be included in the next iteration during sampling, enabling it to be predictable in the next rule of thumb. The accuracy of each weak rule is measured by how much it accurately classifies the examples. Finally, the predictions about unseen samples are made by aggregating the predictions of all the weak rules to make a single prediction rule with the hope that the aggregate is better than using a single prediction rule. A general boosting procedure is given below in Figure 3.

F. K-NEAREST NEIGHBOR

K-NN technique is a non-parametric method first developed in 1951 [66] and further expanded by Thomas Cover [48]. The algorithms work by finding the feature similarity to predict the values for test samples. The feature similarity is calculated in such a way that the distance is computed for new data samples from all the training sets. For distance calculation, there exists a variety of methods such as Euclidian and Manhattan distances. The Euclidean distance is computed by the sum of the squared difference between the existing (y) and new point (x).

Input:
Samples distribution for trainingTD;
The base learning algorithm BL;
Number of learning iterations T.
Process:
1. $TD_1 = TD$
2. *for m in 1 to T:*
   a. $h_m = BL(TD_m)$; % *training of a weak learner from the distribution* TD_m
   b. $Error_m = P_{x\sim TD_m}(h_m(x) \neq f(x))$ % *error evaluation of model* h_m
   c. $TD_{m+1} = Adjust\ Distribution\ (TD_m, Error_m)$

Output:
$H(x) = combine\ outputs(h_1(x), \ldots\ldots h_m(x))$

FIGURE 3. A general boosting framework to boost the performance of the base learner.

$$\text{Euclidean distance} = \sqrt{\sum_{n=1}^{N}(x_n - y_n)^2} \quad (2)$$

Moreover, the Manhattan distance is the sum of the absolute difference between existing and a new point formulated as:

$$\text{Manhattan distance} = \sum_{n=1}^{N}|x_n - y_n| \quad (3)$$

After calculating the distance of a new sample from each sample in the training set, the K number of neighbors needs to be selected to find the classification or prediction for the new sample. The step-by-step working of the K-NN algorithm is given below.

1) Read the training and test dataset.
2) Initialize the value of K to the optimum number of neighbors
3) For every sample in test data.
   a) Compute the distance between the test sample and the training set.
   b) Distance and index of the sample is added to the ordered collection.
   c) The ordered collection is sorted in ascending order by their distances computed in step 3a.
   d) Choose the first K entries from the sorted collection.

Return the mean of the K response values to serve as the predicted value for the current testing sample.

G. SUPPORT VECTOR MACHINE

The support vector machine (SVM) [67] is a deterministic technique and considered to be the most useful machine learning tool where classification and regression tasks are of concern. It was originally designed for a binary classification task that separates the samples of different classes with hyperplanes having maximum margin [68]. However, the minimum distance of instances of different classes from the classification hyperplane is called the margin. The SVM with some modifications can be used for regression tasks where the output is a real value known as support vector regression (SVR). For regression, the epsilon-insensitive regression ($\epsilon - SVM$), the data for

training the algorithm consists of predictor variables and associated observed response values. Here, the goal is to find a function $g(x)$ that does not deviate more than epsilon (ϵ) for each training point x. In the case of linear SVM regression, let us consider a training data where x_n a multivariate set of M samples with associated response values y_n. In order to find the linear function $g(x)$ that is as flat as possible, the task is to find the function $g(x)$ with norm having minimum value ($\beta\beta'$) [69].

$$\mathrm{g(x)} = x'\beta + a \tag{4}$$

To do so, the formulation results as a convex optimization problem to minimize the function put through all residuals with the value less than epsilon (ϵ) as given by:

$$J(\beta) = \frac{1}{2}\beta'\beta, \forall n : |y_n - (x_n'\beta + a) \leq \epsilon| \tag{5}$$

For the points when there is no such function g(x) to satisfy all the constraints above, the slack variables are introduced for each point to deal with this situation as given by:

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{n=1}^{M}(\xi_n + \xi_n^*) \tag{6}$$

The C is known as a box constant that helps to get rid of overfitting. It is a positive numeric value that controls the penalty imposed on samples lying outside the epsilon margin epsilon margin (ϵ) and tolerates the trade-off between the flatness of $g(x)$ and the extent to which the deviations are larger than ϵ. Moreover, the loss is measured from the distance between the epsilon boundary and observed value y as given by:

$$L_\epsilon = \begin{cases} 0, & \text{if } |y - g(x)| \leq \epsilon \\ |y - g(x)| - \epsilon, & \text{otherwise} \end{cases} \tag{7}$$

For linear SVM regression, the Lagrange dual (L_D) can be obtained by introducing different non-negative multipliers α_n and α_n^* for each of the instances x_n. The L_D for the Lagrange primal function L_p is given below where we minimize the function as given by:

$$\mathrm{L}(\alpha) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)x_i'x_j + \epsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} y_i\left(\alpha_i^* - \alpha_i\right) \tag{8}$$

Finally, the function that is used to predict the test set or new values are given by:

$$\mathrm{g(x)} = \sum_{n=1}^{M}\left(\alpha_n + \alpha_n^*\right)\left(x_n'x\right) + a \tag{9}$$

III. PERFORMANCE MEASURE

In order to assess the accuracy of the predictions of radon concentration (RN) from other attributes such as thoron, temperature, relative humidity and pressure, different frequently used performance metrics are computed. RMSE is considered to be a frequently used performance evaluation measure that has been applied to various fields of studies where prediction models are of concern. It is more sensitive to outliers because a large difference between actual and predicted values results in a markedly larger effect on its value. RMSE can be computed from:

$$\mathrm{RMSE} = \sqrt{\frac{1}{V}\sum_{n=1}^{V}(Actual_n - Predicted_n)^2} \quad \text{where } V \text{ represents total number of samples} \tag{10}$$

The presence of outliers when calculating RMSE can explode the error term but RMSLE can scale down the outliers and result in nullification of their effect. The RMSLE can be calculated from the equation given below:

$$\mathrm{RMSLE} = \sqrt{\frac{1}{V}\sum_{n=1}^{V}(log(Actual_n + 1) - log(Predicted_n + 1))^2} \quad \text{where } V \text{ represents total number of samples} \tag{11}$$

It is used mostly to avoid the excessive effect of huge differences in the predicted and actual values in the case when these values are higher in number. Moreover, the MAPE is also frequently used performance metric which is used to assess the accurateness of prediction model, computed from:

$$\mathrm{MAPE} = \frac{1}{V}\sum_{n=1}^{V}\left|\frac{Actual_n - Predicted_n}{Actual_n}\right| \tag{12}$$

MAPE is the average of absolute percentage error. The features that make MAPE popular and useful are its scale independency and easy interpretation [70]. Apart from its advantages, it has certain disadvantages such as resultant undefined or infinite values when the actual values are zero or close to zero. The actual values with a magnitude less than 1 yielded the MAPE to a higher percentage value whilst the actual zero values resulted in infinite MAPE values [71].

Moreover, Mean Squared Error (MSE) is a performance metric that estimates how much the actual and predicted values are closer to each other, computed with V number of samples from the equation given below:

$$\mathrm{MSE} = \frac{1}{V}\sum_{n=1}^{V}(Predicted_n - Actual_n)^2 \tag{13}$$

More simply put, it is the average square difference between the actual and predicted value. The lower the value of MSE indicates the better fit of the prediction model. The tendency of predicted value to be smaller or larger in average to its

TABLE 2. RMSE and MAPE statistics of ensemble and individual leaning methods for predicting radon concentration from other environmental attributes keeping setting 1and window from 1 to 4.

ML Methods	RMSE				MAPE			
	W_1	W_2	W_3	$W4$	W_1	W_2	W_3	W_4
BstTree	1399.237	1267.969	1199.362	1191.145	0.046	0.042	0.04	0.04
BagCrt	1828.311	1646.186	1519.141	1532.077	0.062	0.056	0.052	0.052
BstLm	2466.482	2333.622	2175.823	2179.625	0.082	0.078	0.074	0.074
K-NN	1829.133	1660.596	1569.133	1568.406	0.062	0.057	0.053	0.054
SVML	1859.246	1685.892	1547.506	1543.875	0.063	0.056	0.051	0.052
SVMR	1381.023	1258.624	1176.552	1166.37	0.045	0.041	0.039	0.039

real or actual value can be described by percentage bias (PB), formulated with V number of samples as:

$$\mathrm{PB} = 100 \times \frac{\sum_{n=1}^{V} (Predicted_n - Actual_n)}{\sum_{n=1}^{V} Actual_n} \tag{14}$$

The larger positive values of PB indicate overestimation bias whilst larger negative values indicate model underestimation bias. On the other hand, PB of value 0 is considered to be an optimal value representing accurate model simulation.

IV. RESULT AND DISCUSSION

The RMSE and MAPE statistics for ensemble and individual learning methods are presented in Table 2. The ensemble methods include boosted tree method, bagged cart and boosted linear model and individual learning models are K-NN, support vector machine (SVM) with the linear and radial kernel. The statistics presented in Table 2 are calculated by employing all the methods from groups 1 and 2 on the soil radon gas concentration dataset in setting 1. The setting 1, as shown in Figure 2, is the distribution of training and testing samples in such a way that training data is composed of the non-seismic activity data (NSA) and seismic activities (E1, E2, E5, E6, E7, E8 and E9) while testing data is constituted by E3 and E4 with respect to time window from 1 to 4. The statistics calculated in Table 2 reveal that when predicting radon concentration as a function of environmental parameters, the minimum RMSE is achieved by a support vector machine with a radial kernel across all the time windows. For time window 1, the minimum RMSE is 1381.023 yielded by a support vector machine with a radial kernel. A similar trend can be observed across all the windows achieving a minimum of RMSE by a support vector machine (SVM) with a radial kernel when predicting radon gas concentration. Considering MAPE, like RMSE, the minimum value of MAPE is observed for SVM with radial kernel across all the time windows ranging from 1 to 4. The minimum MAPE value of SVM with radial kernel is 0.045 for time window 1 when compared to the maximum value of the ensemble method of 0.082 by the boosted linear model. The statistics presented above in Table 2 reveal that the individual learning model, SVM with radial kernel, performs better than all the other methods especially from ensemble methods in terms of RMSE and MAPE. Although, the RMSE and MAPE statistics for SVM with radial kernel is smaller than all the other methods but boosted tree method performs as a next rival to SVM with radial and achieves an approximately similar type of results when compared to SVM with radial kernel. The minimum difference of RMSE can be observed in time window 2 with the value of 9.345 when compared to SVM with radial kernel. Likewise, only a difference of 0.001 for the value of MAPE is observed when comparing SVM with the radial and boosted tree model. The similar type of results discussed above can be seen for SVM with radial kernel and boosted tree model when compared to other ensemble and individual learning methods in Figure 4-7 (a-f), presenting actual and predicted soil radon gas concentration when splitting data according to setting 1 and time window of 1 to 4. The actual radon concentration is presented by a red curve while the predicted radon concentration is presented in a black color curve. It can be seen that boosted tree and SVM with radial kernel are the two competent models from the rest because both models perform very close to each other and overlapping most of the original radon time series. The boosted linear model performs worst in setting 1 (window from 1 to 4) and does not capture the temporal variations in the time series. However, bagged cart and K-NN perform nearly equivalent to each other and perform better and result in capturing some temporal variations efficiently when compared to boosted linear model and SVM with linear kernel. Table 3 presented the different statistics when comparing actual and predicted radon concentration by the different ensemble and individual learning methods keeping setting 2 (see Figure 2) by time window of 1 to 4. It can be seen from Table 3; although, the boosted linear model performs better than other machine learning models specified with the value of RMSE in the range [1082.2, 1173.95] for windows from 1 to 4 but the boosted linear model did not capture the temporal variation as per original radon concentration (see Figure 8 (c)). This can be easily observed through percentage bias (PB) value of -0.002, -0.004 and -0.004 for time windows of 2, 3 and 4 respectively, showing negative bias which is the clear indication of model underestimation bias. A similar type of patterns can be observed in Figure 8(a-f) presenting actual and predicted radon concentrations for setting 2 and a time window of 3 days. Refer to Figure 8 (a-f), the actual and predicted radon concentration showing in red and black color, apart from lowest RMSE value of the boosted linear model, the predicted values did not follow the variations in the radon time series data. Hence, this results in negative

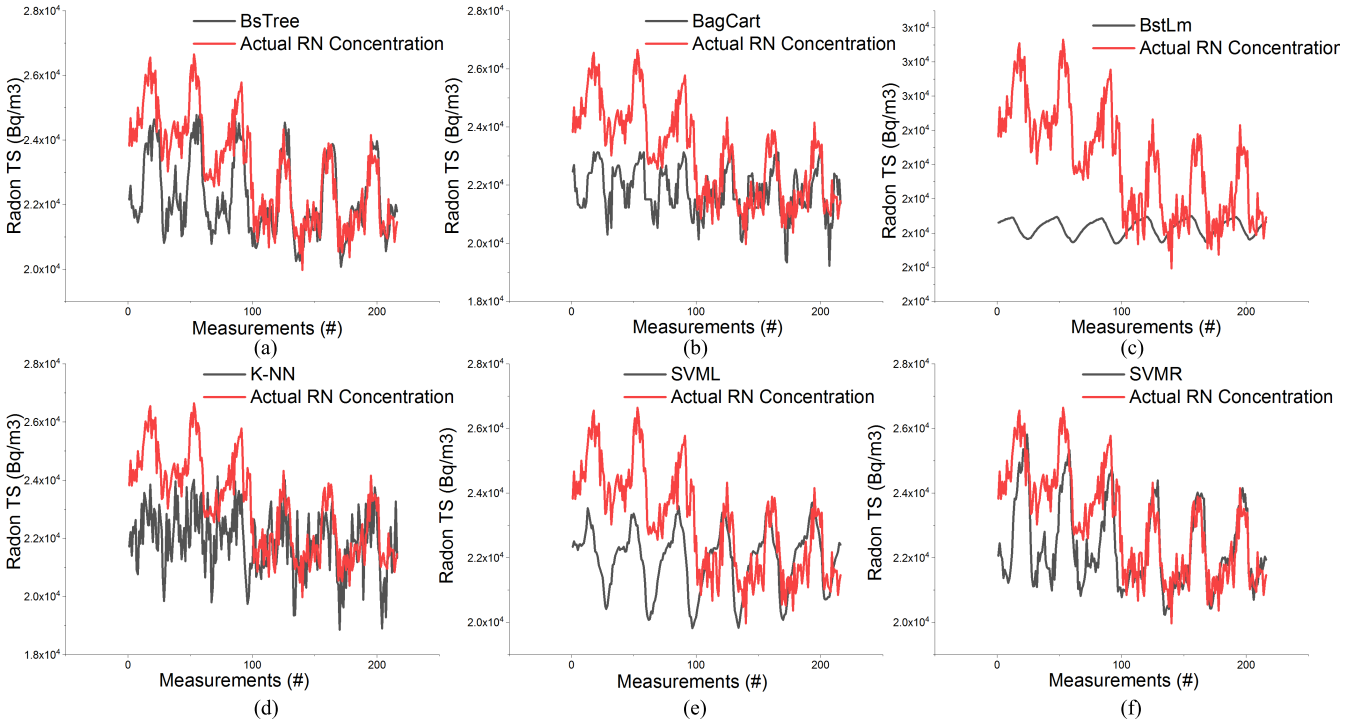

FIGURE 4. (a-f) Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 1 and time window of 1.

TABLE 3. RMSE, RMSLE, MAPE, PB and MSE statistics for ensemble and individual leaning methods for predicting radon concentration from other environmental attributes keeping setting 2 and time window from 1 to 4.

ML Methods	W_1				W_2					
	RMSE	RMSLE	MAPE	PB	MSE	RMSE	RMSLE	MAPE	PB	MSE
BstTree	1509.981	0.071	0.056	0.037	2280042	1485.69	0.07	0.057	0.021	2207275
BagCrt	1341.842	0.062	0.05	0.029	1800540	1375.277	0.065	0.053	0.012	1891387
BstLm	1173.948	0.053	0.044	0.014	1378155	1162.958	0.054	0.044	-0.002	1352472
K-NN	1607.295	0.075	0.06	0.036	2583396	1547.189	0.073	0.059	0.021	2393795
SVML	1218.65	0.056	0.046	0.02	1485107	1289.89	0.06	0.051	0.004	1663815
SVMR	1598.121	0.075	0.058	0.043	2553992	1602.216	0.076	0.06	0.025	2567096
	W_3					W_4				
BstTree	1431.614	0.068	0.054	0.017	2049519	1443.34	0.069	0.053	0.016	2083231
BagCrt	1333.089	0.063	0.051	0.01	1777127	1343.384	0.063	0.05	0.009	1804682
BstLm	1097.825	0.051	0.041	-0.004	1205221	1082.203	0.05	0.04	-0.004	1171163
K-NN	1488.037	0.071	0.056	0.018	2214253	1481.303	0.07	0.055	0.016	2194258
SVML	1203.55	0.056	0.046	0.002	1448532	1188.107	0.055	0.046	0.001	1411599
SVMR	1550.929	0.074	0.058	0.02	2405379	1572.864	0.075	0.058	0.019	2473901

percentage bias. However, support vector machine (SVM) with a linear kernel is the better option to be considered because it overlaps the original radon time series by capturing temporal variations throughout the tested time series. It can also be seen from Table 3, after a boosted linear model, the support vector machine (SVM) with a linear kernel has the lowest RMSE value as well as a percentage bias closer to "0". These statistics leads to a conclusion that SVM with linear kernel performs better in setting 2 with the time window of 1 to 4. From Table 4, by experimenting with setting 3 (see Figure 2), the boosted tree model results in value of RMSE in the range [1262.864, 1409.616] for windows from 1 to 4 which is minimum when compared to other prediction methods. The support vector machine with radial kernel performs closer to boosted tree model having RMSE with the difference of 93, 99.258, 88.17 and 81.359 for the time window of 1, 2, 3 and 4 respectively. For other performance metrics, the average RMSLE value across the entire time window for the boosted tree model is 0.0595. On the other hand, the average RMSLE values for bagged cart model, boosted linear model, K-NN, SVM with linear and radial kernel are 0.069, 0.089, 0.072, 0.066 and 0.064 respectively. Similarly, the average MAPE and MSE values for the boosted tree model are 0.047 and 1745326 respectively which is

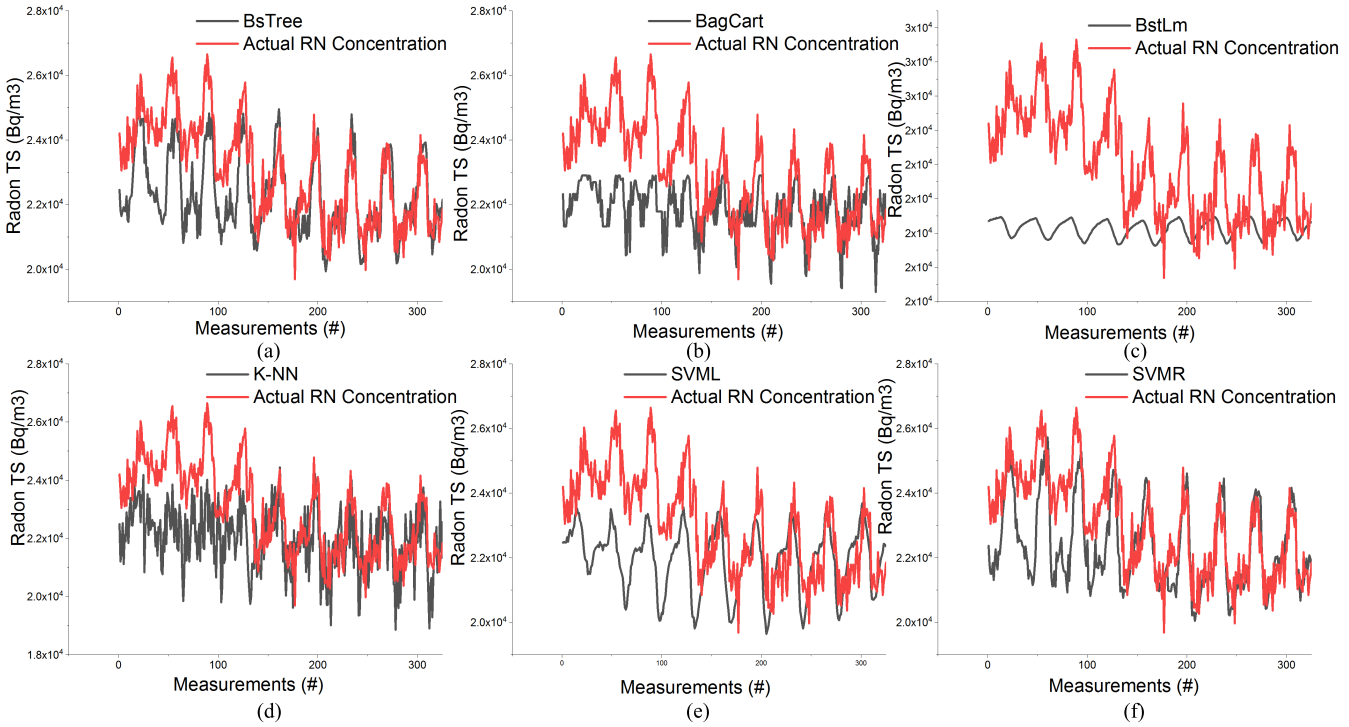

FIGURE 5. (a-f) Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 1 and time window of 2.

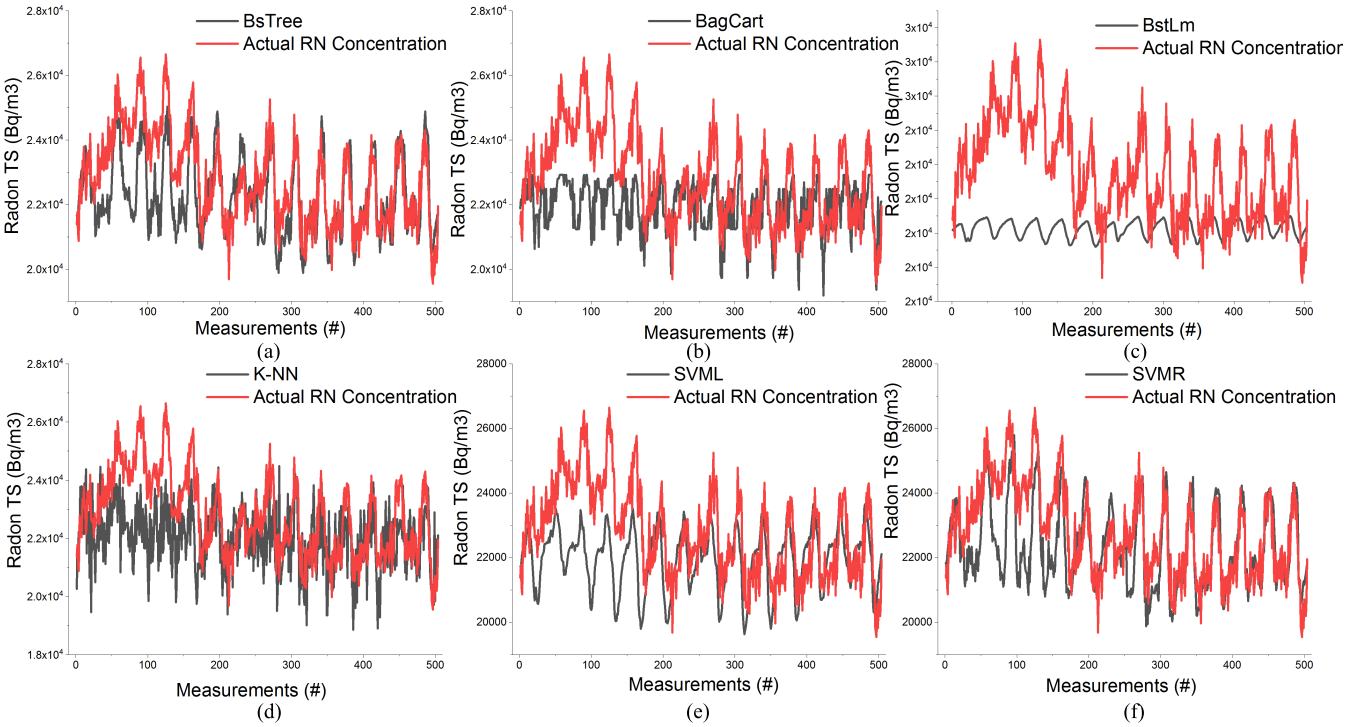

FIGURE 6. (a-f) Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 1 and time window of 3.

relatively promising when compared with the boosted linear model with highest value and SVM with radial kernel with closer average MAPE, MSE statistics of 0.067, 3976024 and 0.05, 1992770 respectively. Refer to Figure 9 (a-f), the actual and predicted radon concentration for the ensemble and individual learning methods are shown in red and black curves. It can be seen from Figure 9a; the predicted radon gas concentration during the seismic activities overlaps the

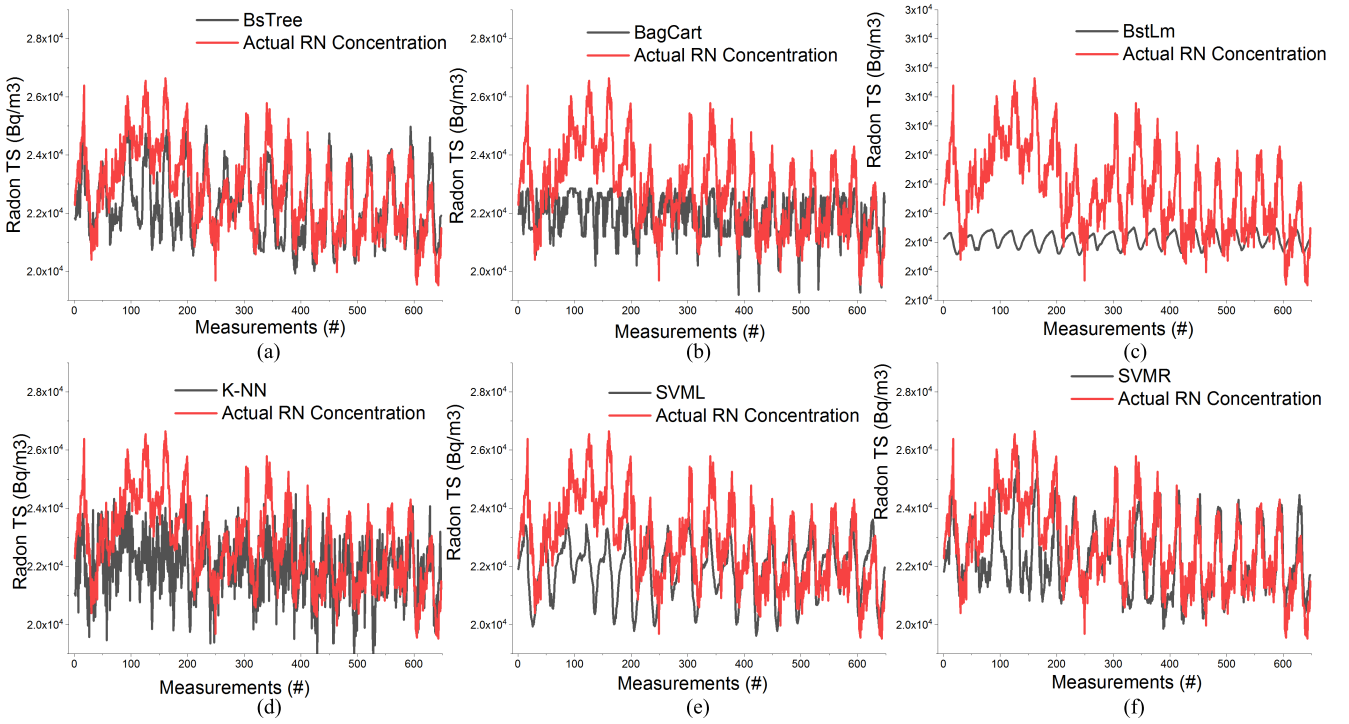

FIGURE 7. (a-f) Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 1 and time window of 4.

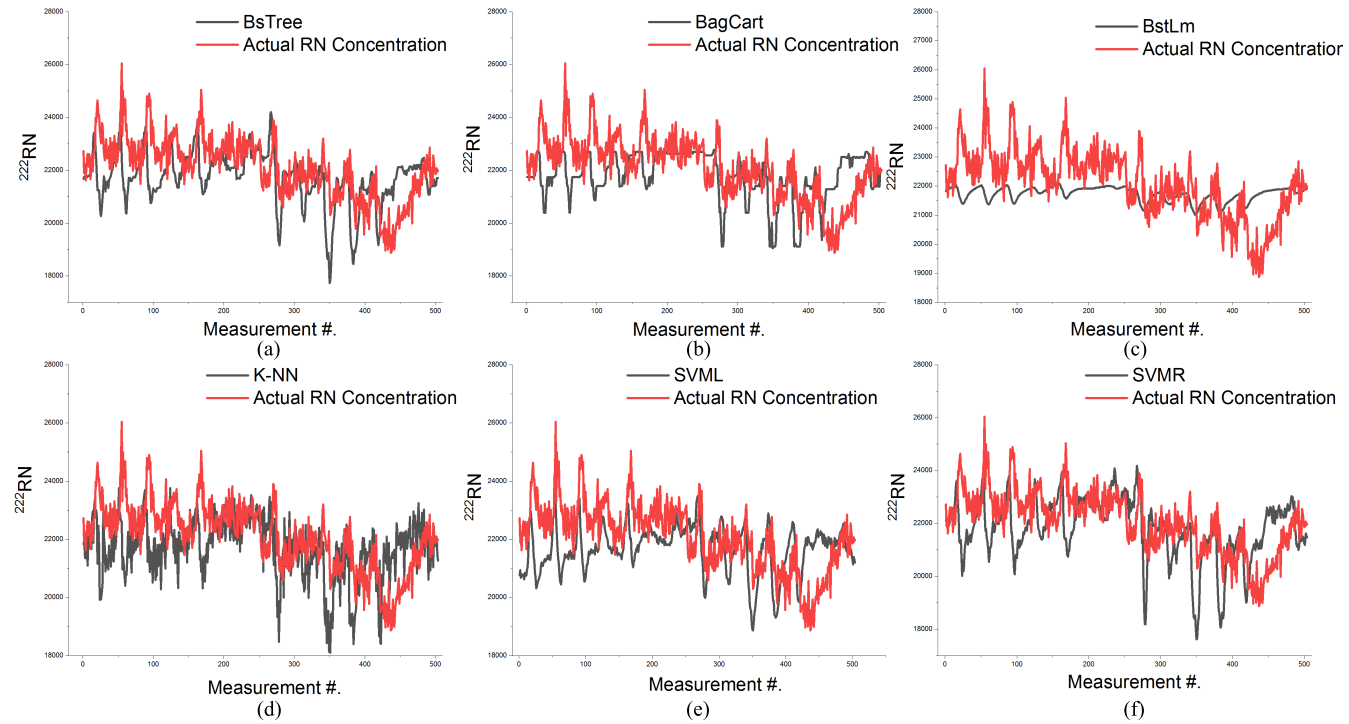

FIGURE 8. (a-f) Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 2 and time window of 3.

original radon concentration and captures temporal variations in the time series more effectively than other methods. Apart from the boosted tree model shown in Figure 8d, the support vector machine (SVM) with radial kernel performs closer to the boosted tree model by overlapping original radon concentration when compared to others. The boosted linear model (see Figure 8c), the boosted linear model did not capture variations in the original radon time series

TABLE 4. RMSE and MAPE statistics for ensemble and individual leaning methods for predicting radon concentration from other environmental attributes keeping setting 3 and time window from 1 to 4.

ML Methods	W_1				W_2					
	RMSE	**RMSLE**	**MAPE**	**PB**	**MSE**	**RMSE**	**RMSLE**	**MAPE**	PB	MSE
BstTree	1409.616	0.063	0.05	0.016	1987018	1328.651	0.06	0.047	0.013	1765313
BagCrt	1698.05	0.075	0.06	0.027	2883375	1566.813	0.07	0.055	0.025	2454903
BstLm	2139.817	0.095	0.072	0.039	4578815	2026.835	0.09	0.068	0.04	4108058
K-NN	1742.754	0.078	0.063	0.026	3037193	1608.145	0.072	0.058	0.022	2586131
SVML	1640.299	0.073	0.055	0.03	2690582	1505.165	0.067	0.051	0.027	2265521
SVMR	1502.624	0.068	0.053	0.01	2257879	1427.909	0.065	0.05	0.006	2038924
	W_3				W_4					
BstTree	1278.338	0.058	0.045	0.009	1634148	1262.864	0.057	0.045	0.009	1594826
BagCrt	1459.378	0.065	0.052	0.02	2129785	1459.905	0.065	0.052	0.021	2131322
BstLm	1894.414	0.084	0.064	0.038	3588803	1904.841	0.085	0.065	0.039	3628418
K-NN	1529.142	0.069	0.055	0.018	2338276	1534.578	0.069	0.055	0.018	2354931
SVML	1390.41	0.062	0.047	0.023	1933239	1403.878	0.063	0.048	0.025	1970873
SVMR	1366.508	0.062	0.048	0.003	1867343	1344.223	0.061	0.047	0.004	1806934

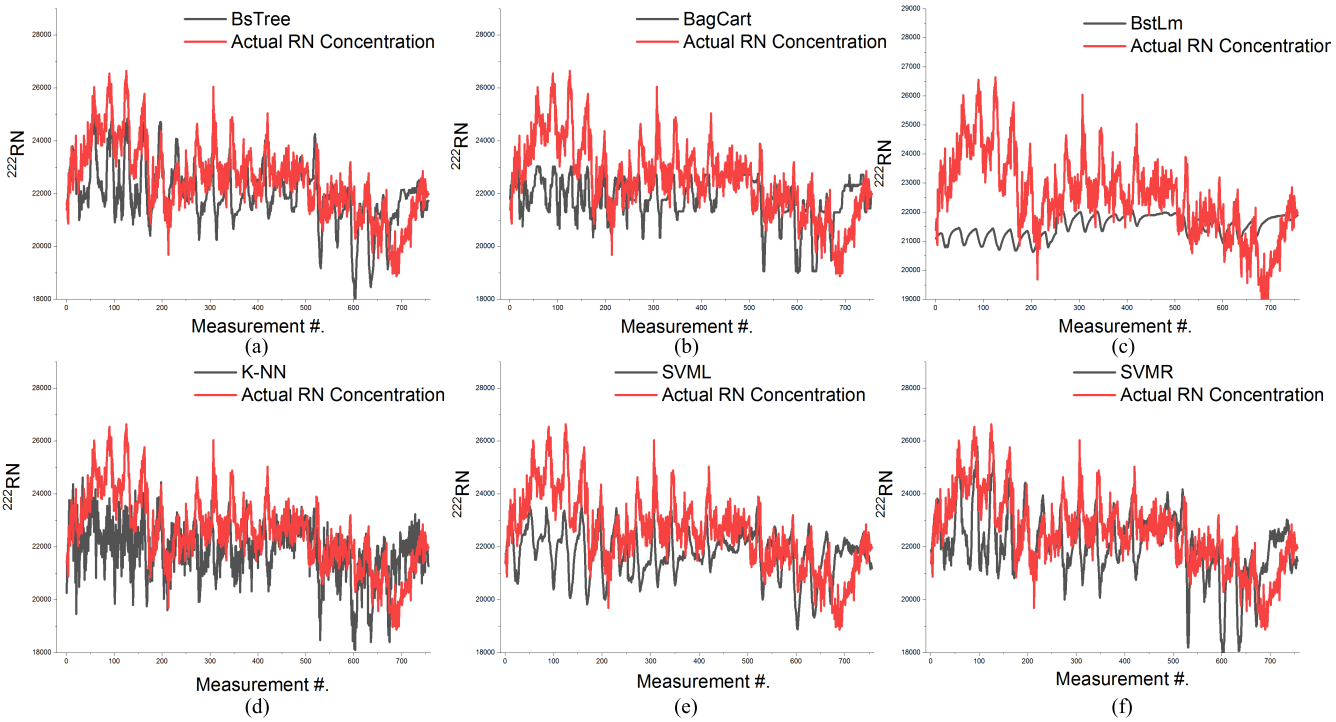

FIGURE 9. (a-f) Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 3 and time window of 3.

and resulted in larger values of different error metrics presented in Table 4. Table 5 presents the RMSE and MAPE statistics for ensemble and individual learning methods keeping setting 4 and the time window of 1 to 4. The minimum RMSE and MAPE score of 1573.174 and 0.056 is achieved by boosted tree model and it can be easily seen from Figure 10 (a-f) that the predictions made by boosted tree model overlap with the original radon concentration presented in red color. Similarly, from Figure 10f, the support vector machine (SVM) with radial kernel performs similar to other experimentation results calculated above in different settings, the performance of SVM with radial kernel is similar to boosted tree model by overlapping most of the variations in original radon concentration time series. The statistics computed above in different settings from 1 to 4 across all the time windows, it is concluded that boosted tree based ensemble method performs better than the individual models when predicting soil gas radon time series data during the seismic activities. It is also observed that a support vector machine with a radial kernel is the second choice after boosted tree method for this task because of its performance is slightly better than boosted tree method in setting 1 while in setting 3 and 4, its performance is closer to boosted tree method.

TABLE 5. RMSE and MAPE statistics for ensemble and individual leaning methods for predicting radon concentration from other environmental attributes keeping setting 4 and time window from 1 to 4.

ML Methods	RMSE				MAPE			
	W_1	W_2	W_3	W_4	W_1	W_2	W_3	W_4
BstTree	1573.174	1478.825	1379.781	1314.423	0.056	0.053	0.048	0.045
BagCrt	1765.261	1653.585	1500.508	1450.629	0.062	0.058	0.051	0.049
BstLm	2116.983	2014.034	1830.583	1793.024	0.069	0.067	0.06	0.059
K-NN	1811.603	1669.05	1538.659	1516.08	0.064	0.059	0.053	0.051
SVML	1773.471	1688.794	1549.615	1507.272	0.063	0.06	0.054	0.053
SVMR	1608.636	1502.272	1405.168	1352.815	0.056	0.052	0.048	0.046

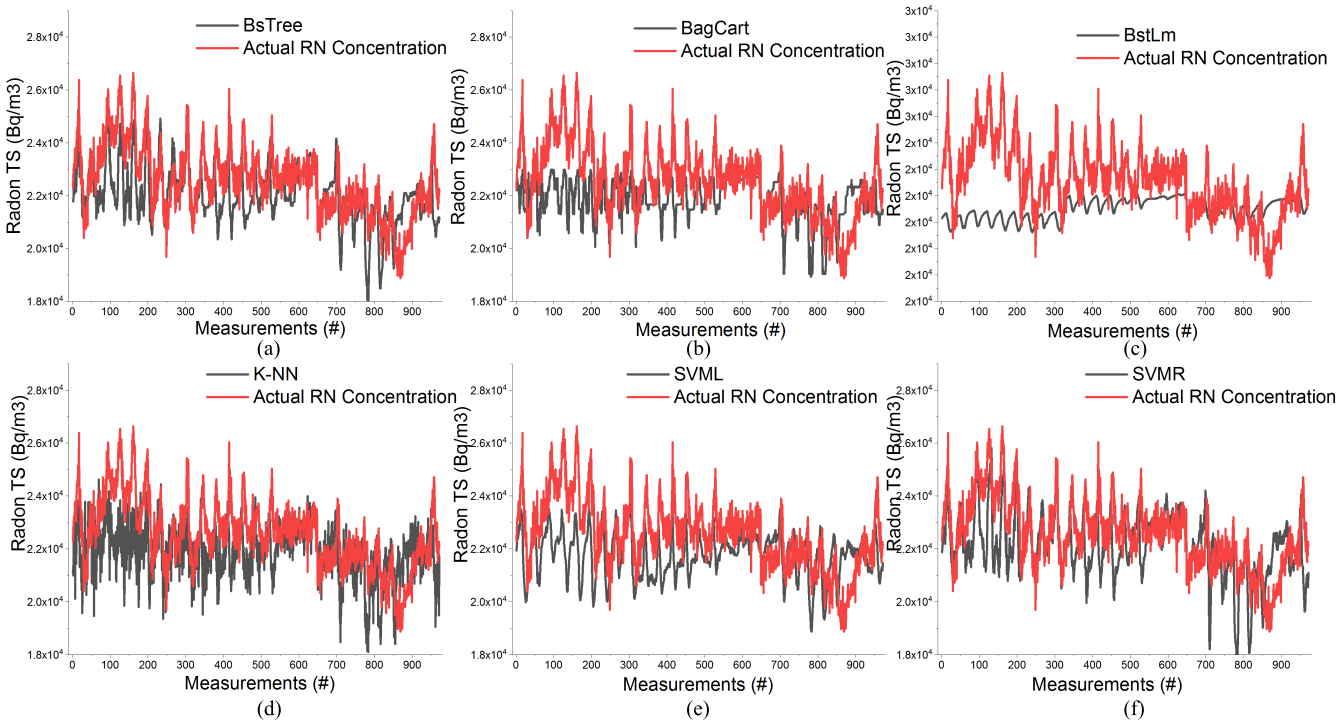

FIGURE 10. Represents actual and predicted radon concentration for ensemble and individual learning methods keeping setting 4 and time window of 4.

A. COMPARISON WITH EXISTING LITERATURE

In this section, the methodology and experimental results, obtained by testing soil radon gas concentration data have been compared with most recent studies. Mir *et al.* [72] proposed a methodology that categorizes soil radon gas concentration data into seismically active and non-active using stacking and automatic anomaly indication function. The radon concentration along with the labeled anomaly data was trained by a meta-learner that classifies it into seismic and non-semic ones. Further, these classifications are passed to an automatic anomaly indication function that labels the time series by calculating the indication percentage. The points where indication percentage gets higher or equal to indication factor were considered to be an anomaly. Tareen *et al.* [7] proposed an earthquake prediction model based on boxplot interpretation using soil radon gas concentration data. The specific patterns were observed in soil radon gas concentration by analyzing boxplots. This is due to the different geological and seismic activities before the occurrence of the earthquake. Tareen *et al.* [11] also employed computational intelligence techniques for the detection of anomalous behavior in soil radon gas before seismic activities. The authors reported that the seismic activity or noise could be responsible for the abnormality in soil radon time-series data. Rafique *et al.* [3] proposed a methodology based on delegation for the accurate prediction of soil radon gas concentration data. The methodology is tested by splitting the data into seismically active and non-active time series. The delegated regressor and other methods were trained using non-seismic time series. The trained models were used to predict the seismically active time series data. Further, the root mean squared error for actual and predicted soil radon concentration was calculated for each model. The delegated regressor model outperforms when compared to other machine learning models. This research study provides more exhaustive experimentation by introducing settings that lead to different compositions of training and testing sets. Instead of training by using

non-seismic data only, as performed in delegated regressor model, each setting incorporates different seismic activities for training and testing purposes. These settings along with time windows lead us to choose a better prediction model for the prediction of soil radon gas concentration at radon measuring stations. This is the novel methodology to gauge the importance of machine learning based ensemble and individual learning methods to forecast the radon concentration efficiently.

V. CONCLUSION

In order to predict radon concentration, a precursor for an earthquake, this study has employed different ensemble and individual machine learning methods for the prediction of soil radon gas concentration using different environmental attributes. The performance of the methods is assessed more vividly by incorporating different training and test set distributions through settings from 1 to 4. The training set is composed of different seismic activities and normal data while testing data is based upon seismic activities with its associated time window from 1 to 4. In setting 1, boosted tree and support vector machine (SVM) with radial kernel performed alike and captured temporal variations in the time series more effectively. For setting 2, boosted linear model has the least RMSE and other performance metrics did not capture temporal variations in the time series. Moreover, support vector machine with linear kernel and boosted tree performed relatively better than other models. In setting 3 and 4, the boosted tree model outperformed when compared to other ensemble and individual models by predicting soil gas radon concentration more accurately. This study concludes that ensemble methods results in relatively better regressed models, and support vector machine with radial kernel performs closer to boosted tree model in setting 3 and 4. This study suggests a boosted tree method to automatically predict soil radon gas radon concentration from environmental parameters in the soil radon time series. The prime focus of this study is to predict the soil radon gas concentration during the anomalies. However, this study can be extended to classify the anomalies in predicted radon concentration. Moreover, the post-processing methods such as automatic anomaly indication function may also be applied.

REFERENCES

[1] B. Adhikari, S. R. Mishra, S. B. Marahatta, N. Kaehler, K. Paudel, J. Adhikari, and S. Raut, "Earthquakes, fuel crisis, power outages, and health care in nepal: Implications for the future," *Disaster Med. Public Health Preparedness*, vol. 11, no. 5, pp. 625–632, Oct. 2017.

[2] V. G. Gitis and A. B. Derendyaev, "Machine learning methods for seismic hazards forecast," *Geosciences*, vol. 9, no. 7, p. 308, Jul. 2019.

[3] M. Rafique, A. D. K. Tareen, A. A. Mir, M. S. A. Nadeem, K. M. Asim, and K. J. Kearfott, "Delegated regressor, a robust approach for automated anomaly detection in the soil radon time series data," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 3004.

[4] M. Awais, A. Barkat, A. Ali, K. Rehman, W. Ali Zafar, and T. Iqbal, "Satellite thermal IR and atmospheric radon anomalies associated with the Haripur earthquake (Oct 2010; Mw 5.2), Pakistan," *Adv. Space Res.*, vol. 60, no. 11, pp. 2333–2344, Dec. 2017.

[5] Z. Jilani, T. Mehmood, A. Alam, M. Awais, and T. Iqbal, "Monitoring and descriptive analysis of radon in relation to seismic activity of northern Pakistan," *J. Environ. Radioactivity*, vol. 172, pp. 43–51, Jun. 2017.

[6] B. Aslam, A. Zafar, U. A. Qureshi, and U. Khalil, "Seismic investigation of the northern part of Pakistan using the statistical and neural network algorithms," *Environ. Earth Sci.*, vol. 80, no. 2, pp. 1–18, Jan. 2021.

[7] A. D. K. Tareen, M. S. A. Nadeem, K. J. Kearfott, K. Abbas, M. A. Khawaja, and M. Rafique, "Descriptive analysis and earthquake prediction using boxplot interpretation of soil radon time series data," *Appl. Radiat. Isot.*, vol. 154, Dec. 2019, Art. no. 108861.

[8] A. A. Mir, K. J. Kearfott, F. V. Çelebi, and M. Rafique, "Imputation by feature importance (IBFI): A methodology to envelop machine learning method for imputing missing patterns in time series data," *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0262131.

[9] A. A. Mir, F. V. CÇelebi, M. Rafique, L. Hussain, A. S. Almasoud, M. Alajmi, F. N. Al-Wesaabi, and A. M. Hilal, "An improved imputation method for accurate prediction of imputed dataset based radon time series," *IEEE Access*, vol. 10, pp. 20590–20601, 2022.

[10] L. Tommasino, "Radiochemical methodsl radon," pp. 32–44, 2005.

[11] A. D. K. Tareen, K. M. Asim, K. J. Kearfott, M. Rafique, M. S. A. Nadeem, T. Iqbal, and S. U. Rahman, "Automated anomalous behaviour detection in soil radon gas prior to earthquakes using computational intelligence techniques," *J. Environ. Radioactivity*, vol. 203, pp. 48–54, Jul. 2019.

[12] M. Janik, P. Bossew, and O. Kurihara, "Machine learning methods as a tool to analyse incomplete or irregularly sampled radon time series data," *Sci. The Total Environ.*, vol. 630, pp. 1155–1167, 2018.

[13] R. Törnqvist, J. Jarsjö, J. Pietroá, A. Bring, P. Rogberg, S. M. Asokan, and G. Destouni, "Evolution of the hydro-climate system in the lake Baikal basin," *J. Hydrol.*, vol. 519, pp. 1953–1962, Nov. 2014.

[14] S. K. Sahoo, M. Katlamudi, C. Barman, and G. U. Lakshmi, "Identification of earthquake precursors in soil radon-222 data of kutch, gujarat, India using empirical mode decomposition based Hilbert Huang transform," *J. Environ. Radioactivity*, vol. 222, Oct. 2020, Art. no. 106353.

[15] V. I. Ulomov and B. Mavashev, "A precursor of a strong tectonic earthquake," *Doklady Akademii Nauk*, vol. 176, no. 2, pp. 319–321, 1967.

[16] A. Sultankhodzhayev, I. Chernov, and T. Zakirov, "Hydroseismological premonitors of the gazli earthquake," *Proc. Uzb. SSR Acad. Sci.*, vol. 7, pp. 51–53, Dec. 1976.

[17] H. Wakita, "Earthquake prediction and geochemical studies in China," *Chin. Geophys.*, vol. 1, no. 2, pp. 443–457, 1978.

[18] C.-Y. King, "Radon monitoring for earthquake prediction in China," *EPR. Earthq. Predict. Res.*, vol. 3, no. 1, pp. 47–68, 1985.

[19] C.-Y. King, "Episodic radon changes in subsurface soil gas along active faults and possible relation to earthquakes," *J. Geophys. Res., Solid Earth*, vol. 85, no. 6, pp. 3065–3078, 1980.

[20] C.-Y. King, "Radon emanation on San Andreas fault," *Nature*, vol. 271, no. 5645, pp. 516–519, Feb. 1978.

[21] A. Mogro-Campero, R. Fleischer, and R. Likes, "Changes in subsurface radon concentration associated with earthquakes," *J. Geophys. Res., Solid Earth*, vol. 85, no. B6, pp. 3053–3057, 1980.

[22] S. A. Pulinets, V. A. Alekseev, A. D. Legen'ka, and V. V. Khegai, "Radon and metallic aerosols emanation before strong earthquakes and their role in atmosphere and ionosphere modification," *Adv. Space Res.*, vol. 20, no. 11, pp. 2173–2176, Jan. 1997.

[23] R. C. Ramola, "Relation between spring water radon anomalies and seismic activity in Garhwal Himalaya," *Acta Geophys.*, vol. 58, no. 5, pp. 814–827, Oct. 2010.

[24] J. Vaupotič, A. Riggio, M. Santulin, B. Zmazek, and I. Kobal, "A radon anomaly in soil gas at Cazzaso, NE Italy, as a precursor of an ML=5.1 earthquake," *Nukleonika*, vol. 55, pp. 507–511, 2010.

[25] V. Walia, H. S. Virk, T. F. Yang, S. Mahajan, M. Walia, and B. S. Bajwa, "Earthquake prediction studies using radon as a precursor in N-W Himalayas, India: A case study," *Terr., Atmos. Ocean. Sci.*, vol. 16, no. 4, p. 775, 2005.

[26] H. Virk, "Radon monitoring of microseismicity in the Kangra and Chamba Valleys of Himachal Pradesh, India," *Nucl. Geophys.*, vol. 9, no. 2, pp. 141–146, 1995.

[27] H. Virk, A. K. Sharma, and V. Walia, "Correlation of alpha-logger radon data with microseismicity in NW Himalaya," *Current Sci.*, vol. 4, pp. 656–663, Oct. 1997.

[28] R. C. Ramola, Y. Prasad, G. Prasad, S. Kumar, and V. M. Choubey, "Soil-gas radon as seismotectonic indicator in Garhwal Himalaya," *Appl. Radiat. Isot.*, vol. 66, no. 10, pp. 1523–1530, Oct. 2008.

[29] S. Džeroski, L. Todorovski, B. Zmazek, J. Vaupotic, and I. Kobal, "Modelling soil radon concentration for earthquake prediction," in *Proc. Int. Conf. Discovery Sci.* Berlin, Germany: Springer, Oct. 2003, pp. 87–99.
[30] B. Zmazek, L. Todorovski, S. Džeroski, J. Vaupotiá, and I. Kobal, "Application of decision trees to the analysis of soil radon data for earthquake prediction," *Appl. Radiat. Isot.*, vol. 58, no. 6, pp. 697–706, Jun. 2003.
[31] D. Gupta and D. Shahani, "Estimation of radon as an earthquake precursor: A neural network approach," *J. Geol. Soc. India*, vol. 78, no. 3, p. 243, 2011.
[32] V.-H. Duong, H.-B. Ly, D. H. Trinh, T. S. Nguyen, and B. T. Pham, "Development of artificial neural network for prediction of radon dispersion released from Sinquyen mine, Vietnam," *Environ. Pollut.*, vol. 282, 2021, Art. no. 116973.
[33] A. Negarestani, S. Setayeshi, M. Ghannadi-Maragheh, and B. Akashe, "Layered neural networks based analysis of radon concentration and environmental parameters in earthquake prediction," *J. Environ. Radioactivity*, vol. 62, no. 3, pp. 225–233, Jan. 2002.
[34] A. Negarestani, S. Setayeshi, M. Ghannadi-Maragheh, and B. Akashe, "Estimation of the radon concentration in soil related to the environmental parameters by a modified adaline neural network," *Appl. Radiat. Isot.*, vol. 58, no. 2, pp. 269–273, Feb. 2003.
[35] I. U. Sikder and T. Munakata, "Application of rough set and decision tree for characterization of premonitory factors of low seismic activity," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 102–110, Jan. 2009.
[36] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2020.
[37] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, pp. 89–109, Aug. 2001.
[38] I. Kononenko, I. Bratko, and M. Kukar, "Application of machine learning to medical diagnosis," *Mach. Learn. Data Mining, Methods Appl.*, vol. 389, p. 408, Jun. 1997.
[39] B. Erickson, P. Korfiatis, Z. Akkus, and T. Kline, "Machine learning for medical imaging," *RadioGraphics*, vol. 37, no. 2, pp. 505–515, 2017.
[40] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Commun.*, vol. 11, no. 1, pp. 1–9, Dec. 2020.
[41] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *Proc. Electric Electron., Comput. Sci., Biomed. Engineerings' Meeting (EBBT)*, Apr. 2018, pp. 1–4.
[42] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," 2003.
[43] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, p. 29, Mar. 2019.
[44] P. S. Patil and N. V. Dharwadkar, "Analysis of banking data using machine learning," in *Proc. Int. Conf. I-SMAC*, Feb. 2017, pp. 876–881.
[45] Y.-L. Chen, K. Tang, R.-J. Shen, and Y.-H. Hu, "Market basket analysis in a multiple store environment," *Decis. Support Syst.*, vol. 40, no. 2, pp. 339–354, Aug. 2005.
[46] R. Gangurde, B. Kumar, and S. Gore, "Building prediction model using market basket analysis," *Int. J. Innov. Res. Comput. Commun. Eng*, vol. 5, no. 2, pp. 1302–1309, 2017.
[47] C. H. Park and H. Park, "A relationship between linear discriminant analysis and the generalized minimum squared error solution," *SIAM J. Matrix Anal. Appl.*, vol. 27, no. 2, pp. 474–492, Jun. 2005.
[48] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[49] V. N. Vapnik, "Introduction to statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, pp. 988–999, 1979.
[50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[51] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[52] P. Bartlett, Y. Freund, W. S. Lee, and R. E. Schapire, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.
[53] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.
[54] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.
[55] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.* Berlin, Germany: Springer, 2000, pp. 1–15.
[56] S.-T. Luo and B.-W. Cheng, "Diagnosing breast masses in digital mammography using feature selection and ensemble methods," *J. Med. Syst.*, vol. 36, no. 2, pp. 569–577, Apr. 2012.
[57] D. Lavanya, "Ensemble decision tree classifier for breast cancer data," *Int. J. Inf. Technol. Converg. Services*, vol. 2, no. 1, pp. 17–24, Feb. 2012.
[58] T. Gneiting and A. E. Raftery, "Weather forecasting with ensemble methods," *Science*, vol. 310, no. 5746, pp. 248–249, 2005.
[59] I. Maqsood, M. R. Khan, and A. Abraham, "An ensemble of neural networks for weather forecasting," *Neural Comput. Appl.*, vol. 13, no. 2, pp. 112–122, 2017.
[60] M. Kuhn. (2020). *Caret: Classification and Regression Training. R Package Version 6.0-86*. Accessed: Mar. 20, 2020. [Online]. Available: https://cran.r-project.org/web/packages/caret/caret.pdf
[61] IQAir. *Air Quality in Muzaffarabad*. Accessed: Dec. 2, 2021. [Online]. Available: https://www.iqair.com/pakistan/azad-kashmir/muzaffarabad
[62] SARAD. *SARAD Closer to Your Application*. Accessed: Dec. 2, 2021. [Online]. [Online]. Available: https://www.sarad.de/
[63] Z.-H. Zhou, *Ensemble Methods: Foundation Algorithms*. Boca Raton, FL, USA: CRC Press, 2019.
[64] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.
[65] R. E. Schapire, "A brief introduction to boosting," in *Proc. IJCAI*, vol. 99, 1999, pp. 1401–1406.
[66] E. Fix and J. L. Hodges, "Nonparametric discrimination: Consistency properties," Project, Randolph Field, TX, USA, 1951, pp. 21–49.
[67] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2013.
[68] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[69] Accessed: Jan. 4, 2022. [Online]. Available: https://au.mathworks.com/help/stats/understanding-support-vector-machine-regression.html
[70] R. F. Byrne, "Beyond traditional time-series: Using demand sensing to improve forecasts in volatile times," *J. Bus. Forecasting*, vol. 31, no. 2, pp. 13–19, 2012.
[71] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecasting*, vol. 32, no. 3, pp. 669–679, 2016.
[72] A. A. Mir, F. V. Çelebi, M. Rafique, M. R. I. Faruque, M. U. Khandaker, K. J. Kearfott, and P. Ahmad, "Anomaly classification for earthquake prediction in radon time series data using stacking and automatic anomaly indication function," *Pure Appl. Geophys.*, vol. 4, pp. 1–15, May 2021.

ADIL ASLAM MIR received the B.S. and M.Phil. degrees in computer sciences from The University of Azad Jammu and Kashmir, Muzaffarabad, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with Ankara Yıldırım Beyazıt University, Turkey. He is also a Research Associate with the Department of Computer Science and Information Technology, The University of Azad Jammu and Kashmir. His research interests include machine learning-based decision support systems (DSSs), data mining, and artificial intelligence.

FATIH VEHBI ÇELEBI received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, in 1988, the M.Sc. degree in electrical and electronics engineering from Gaziantep University, in 1996, and the Ph.D. degree in electrical and electronics engineering from Erciyes University, in 2002. He is currently a full-time Professor and the Dean of the Faculty of Engineering and Applied Sciences, Ankara Yıldırım Beyazıt University (AYBU). He has published so many scientific papers. His current research interests include cyber security, artificial intelligence, machine learning, and optoelectronics.

HADEEL ALSOLAI is currently an Academic Teacher at the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University. She has expertise in conducting empirical studies of software engineering techniques (particularly software maintainability, software quality, and open-source systems), along with machine learning techniques (particularly ensemble techniques, data pre-processing, and parameter tuning). Her teaching and research interests include artificial intelligence and software engineering.

SHAHZAD AHMAD QURESHI received the Ph.D. degree from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Pakistan, in 2008. He has been a Research Associate at the University of Warwick, U.K., during his Ph.D. study. He is currently an Associate Professor with the Department of Computer and Information Sciences, PIEAS. His research interests include medical image analysis, deep learning, biophotonics, bioinformatics, evolutionary computing, and computer vision on applications for solving problems related to mammography, mitosis detection in breast cancer, retinopathy, He-La cervical cancer cells investigation, 3D-DSA, acoustic signal processing, satellite imaging, and brain tumor segmentation.

MUHAMMAD RAFIQUE received the Ph.D. degree in computational physics from the Department of Physics and Applied Mathematics, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan, in 2005. He worked as a Visiting Faculty Member and a Postdoctoral Fellow at the Nuclear Engineering and Radiological Science Department, University of Michigan, Ann Arbor, MI, USA, in 2014. He was also the Chairman at the Department of Physics, and the Director of quality enhancement and ORIC. He is currently working as a Professor of physics with the Department of Physics, The University of Azad Jammu and Kashmir, Muzaffarabad. He is also the Director of Advanced Studies, as an additional charge, at The University of Azad Jammu and Kashmir. He has published more than 100 research articles in international and national journals of repute. He has also produced more than 35 M.Phil./M.S. and five Ph.D. students as a Supervisor and a Co-Supervisor. His research interests include reactor physics, radiation physics, computational physics and mathematics, geophysics, and medical physics.

JABER S. ALZAHRANI received the Ph.D. degree from Lamar University, in 2015. He is currently an Associate Professor at the Department of Industrial Engineering, College of Engineering at Al-Qunfudhah, Umm Al-Qura University, Saudi Arabia. He has many peer-reviewed articles. His research interests include optimization, supply chain, scheduling, and AI.

HANY MAHGOUB received the Ph.D. degree in computer science from the Faculty of Computers and Information, Menoufia University, Egypt. In 2010, he was an Assistant Professor of computer science at the Department of Computer Science, Faculty of Computers and Information, Menoufia University. Since February 2017, he has been an Assistant Professor of computer science at the Department of Computer Science, Faculty of Science and Arts, King Khalid University, Saudi Arabia. He is the author of more than 20 articles, and many funded research projects. His research interests include AI, intelligent systems and bioinformatics, the IoT, smart cities, human computation, software testing, machine learning, data mining, text mining, web mining, information retrieval, information extraction, big data, semantic web, and distributed systems.

MANAR AHMED HAMZA received the Ph.D. degree from Omdurman Islamic University, Omdurman, Sudan, in March 2021. She is currently a Lecturer with the Department of Computer and Self Development, Prince Sattam Bin Abdulaziz University, and the Faculty of Computer Science and Information Technology, Omdurman Islamic University. Her research interests include data mining, text mining, and machine learning.