Soil Dynamics and Earthquake Engineering 111 (2018) 1–7

Contents lists available at ScienceDirect

Soil Dynamics and Earthquake Engineering

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

Seismic indicators based earthquake predictor system using Genetic Programming and AdaBoost classification

Khawaja M. Asim[a,*], Adnan Idris[b], Talat Iqbal[a], Francisco Martínez-Álvarez[c]

[a] Centre for Earthquake Studies, National Centre for Physics, Pakistan
[b] Department of Computer Sciences and IT, The University of Poonch, Rawalakot, Pakistan
[c] Department of Computer Science, Pablo de Olavide University of Seville, Spain

ARTICLE INFO

Keywords:
Earthquake predictor system
Seismic indicators
Genetic Programming
AdaBoost
Earthquake prediction

ABSTRACT

In this study an earthquake predictor system is proposed by combining seismic indicators along with Genetic Programming (GP) and AdaBoost (GP-AdaBoost) based ensemble method. Seismic indicators are computed through a novel methodology in which, the indicators are computed to obtain maximum information regarding seismic state of the region. The computed seismic indicators are used with GP-AdaBoost algorithm to develop an Earthquake Predictor system (EP-GPBoost). The setup has been arranged to provide predictions for earthquakes of magnitude 5.0 and above, fifteen days prior to the earthquake. The regions of Hindukush, Chile and Southern California are considered for experimentation. The EP-GPBoost has produced noticeable improvement in earthquake prediction due to collaboration of strong searching and boosting capabilities of GP and AdaBoost, respectively. The earthquake predictor system shows enhanced results in terms of accuracy, precision and Matthews Correlation Coefficient for the three considered regions in comparison to contemporary results.

1. Introduction

Earthquakes are highly feared natural catastrophic events that pose threat to human lives and cause economic damages. The early predictions of such damaging events have a potential for saving human lives and diminish financial losses. Earthquake Predictor System (EPS) aims to generate an alarm about the occurrence of earthquakes [1]. The seismic indicators based EPS aims to predict earthquakes fifteen days prior to earthquake. The seismic indicators are computed using the temporal sequence of past earthquakes, recorded in earthquake catalog. These seismic indicators are provided to computationally intelligent algorithms to generate earthquake predictions, which eventually lead to creation of EPS regarding forthcoming earthquakes. Moreover, the contemporary literature sites numerous studies, which have employed seismic indicators in collaboration with machine learning based methods, to predict earthquake occurrences, thus findings of such studies are equally significant for the development of an earthquake predictor system.

Earthquake prediction is a challenging topic [2] and endeavors have been made to predict earthquakes for over a century [3]. Earthquake prediction approaches can be categorized into three types [4]: a) mathematical and statistical methods [5,6], b) precursor investigations [7–9], c) machine learning methodologies [10–12]. The recent encouraging results obtained in this field of research are the outcome of interdisciplinary interaction mainly, between seismology and Computational Intelligence (CI).

Machine learning based methodologies for earthquake prediction use seismic indicators in order to develop a correlation between the indicators and subsequent earthquakes. Thus, the study relies on the temporal seismic behavior of a region. The computation of seismic indicators is an effort to express the renowned principles of Gutenberg-Richter's law, seismic energy release, foreshock frequency and seismic rate changes, in numeric form. Machine learning based earthquake prediction methodologies can also be classified into two categories, depending upon the calculation approach of seismic indicators:

a) Computation of seismic indicators after a fixed duration of time [11–13].
b) Computation of seismic indicators after every earthquake, inclusive of the recent earthquake [10,14,15].

The former approach is designed to consider a fixed duration, such as 1 month, 2 weeks, so forth, for single prediction period. It does not address the issue, if multiple earthquakes strike the same region within the single prearranged prediction period. However, this issue of making multiple predictions for a single duration can be addressed through the

* Corresponding author.
E-mail addresses: asim.khawaja@ncp.edu.pk (K.M. Asim), adnanidris@upr.edu.pk (A. Idris), talat.iqbal@ncp.edu.pk (T. Iqbal), fmaralv@upo.es (F. Martínez-Álvarez).

https://doi.org/10.1016/j.soildyn.2018.04.020
Received 6 February 2018; Received in revised form 24 March 2018; Accepted 20 April 2018

latter approach. A seismic event occurrence may change the internal seismic state of the region. Therefore, fresh seismic indicators are essentially computed if an earthquake strikes the region during a prediction period. A new earthquake prediction is obtained based upon latest indicators, without waiting for the end of a prearranged prediction period. The latter approach is also advantageous in terms of number of feature instances to be used for developing prediction model. The greater is the number of feature instances, the better a model is trained. Since seismic indicators are computed for every recorded earthquake, therefore greater number of feature instances are available for training of model.

EPS is a field of science that has a potential for advancements. With the advent of computer based techniques, rapid progress has been observed in research and technology. CI and machine learning techniques have been used vastly for classification and, regression to obtain solution for many problems. For example, diagnosis through medical images [16], churn prediction through customer profiling [17], automatic surveillance in videos [18], differentiation between micro seismic events and quarry blasts [19], geological interpretation of structures [20] and so forth.

In this study a novel idea of ensemble classification where Genetic Programming (GP) is evolved using boosting (GP-Adaboost), has been applied for earthquake prediction (EP-GPBoost). Seismically active regions of Hindukush, Chile and Southern California are considered in this study for modelling earthquakes and seismic indicators through GP-AdaBoost. GP-AdaBoost is a unique ensemble classifier, where searching capabilities of GP and boosting of AdaBoost are combined to develop a strong classifier. The GP's evolution is supported through boosting, where multiple GP strings are evolved per class, which act as single class classifier.

In rest of the manuscript, Section 2 contains details of the related literature. Section 3 encompasses the employed methodology, including the computation of seismic indicators along with details of GP and AdaBoost based methodology. Results and discussions can be found in Section 4.

2. Related work

This section provides the overview of the research methodologies offering the use of varied seismic indicators and precursors along with various machine learning techniques. Seismic precursory analysis has been carried out to develop earthquake prediction model through detecting anomalous patterns in these signals. It is presumed that unusual variations in seismic precursors are observed during earthquake preparation process caused by tectonic movements beneath earth surface [21]. The some of the studied seismic precursors are radon gas emission from soil, atmospheric vertical electric field and ionospheric variations. These earthquake precursory changes are bound together through a unified concept of Lithosphere-Atmosphere-Ionosphere Coupling (LAIC). The model signifies that anomalous activity caused due to earthquakes in lithosphere would be propagated into atmosphere and ionosphere [21]. Furthermore, animal behavior is also studied as an earthquake precursor due to their sharp receptors, capability of receiving very low frequencies of, acoustics and seismic waves [22]. In an experiment regarding animals' behavior prior to seismic activity, motion sensing cameras were installed at Yanachaga National Park, Peru, prior to 2011 Contamana earthquake of magnitude 7.0 [23]. A significant decrease in animal activity was observed three weeks period before the earthquake.

Probability based techniques have also been vastly applied for earthquake forecasting [24]. Recently, the application of different probability distributions, such as lognormal, gamma, Weibull distributions, are used for earthquake forecasting in Northeast India [5]. A mathematical technique based on Fibonacci, Dual and Lucas (FDL) numbers is also proposed for earthquake prediction [6]. The method is intended to predict local as well as global earthquakes, exploiting the planetary alignment in combination with FDL numbers.

CI based techniques have been increasingly used in earthquake prediction research in recent past. Thereby a new line of work is now introduced exploiting such CI techniques for earthquake prediction leading towards EPS. CI techniques can only be used, once a meaningful training dataset is provided. Therefore, it is considerably important to find multiple seismicity indicators without worrying about their highly non-linear relation with subsequent earthquakes. Eight seismic indicators have been proposed to use in combination with Back Propagation Neural Network (BPNN), Recurrent Neural Network (RNN) and Radial Basis Functions (RBF) [11]. These seismicity indicators [T, M_{mean}, $dE^{1/2}$, β, α, ΔM, μ, σ, η] hereafter called as Panakkat's indicators, are computed based upon concepts of Gutenberg-Richter's law, seismic energy release and foreshock frequency. The application of this methodology is observed for Southern California and San Francisco bay regions. RNN is reported to have performed better for these two regions. Later on, in a similar effort, Probabilistic Neural Network (PNN) is combined with Panakkat's set of seismicity indicators to predict earthquakes for the same regions [13].

The Panakkat's eight seismic parameters have also been exploited for generating prediction models for Hindukush and Northern Pakistan. Ensemble of tree based classifiers using Linear Programming Boost has shown encouraging results for Hindukush region [12], while feed forward neural network has performed better for Northern Pakistan [25]. Every region may possess different tectonic properties; therefore, it is logical to have the same CI based methodology manifesting dissimilar performances for different regions, even with the same seismicity indicators. In a recent effort, an earthquake early warning method has been generated for Southern California exploiting the Panakkat's eight seismic indicators in combination with four different classification algorithms. These algorithms are Neural Dynamic Classification (NDC), Support Vector Machine (SVM), PNN and Enhanced PNN (EPNN). NDC outperformed other applied techniques in generating earthquake early warnings for predefined thresholds [26].

A different set of seven seismic indicators [x_1, x_2, x_3, … x_7] is proposed based upon concepts of Gutenberg-Richter's law, Omori's law and Otsu law [10,14]. This set of seven seismic indicators is hereafter referred as Reyes's seismic indicators. These indicators are combined with different CI based techniques to generate earthquake prediction models. Artificial Neural Networks (ANN) demonstrated better prediction results for the regions of Chile and Iberian Peninsula as compared to K-Nearest Neighbor (KNN), SVM, Naive Bayes (NB) and so forth.

A study has been carried out to find the best performing set of seismicity indicators among Panakkat's and Reyes's seismic indicators. Every seismic indicator is evaluated using information gain and all the indicators possessing null information are excluded. The remaining indicators are together used in combination with ANN, KNN, NB and SVM to generate earthquake predictions for Chile and Iberian peninsula [27]. A setup has been arranged to analyze the sensitivity of both Panakkat's and Reyes's seismic indicators. This approach of sensitivity analysis is tested on four Chilean zones [28]. In this study, the seismic indicators are computed using various methods, which are onwards employed with various combination of training and test datasets to evaluate the variations in prediction performance. Another idea of obtaining maximum available seismic indicators was exercised by Asim et al. [15], in which all the seismic indicators are computed through multiple applicable approaches. This idea leads to the computation of more than 50 seismic indicators, which are further employed in combination with tree based ensemble learning methodologies, including decision tree J48, Random Forest, RotBoost and Rotation Forest. The Rotation Forest has exhibited better results amongst others, for earthquake prediction in Hindukush region.

Earthquake prediction studies have been carried out on a global scale, in which whole world is divided into four quadrants. The association rule mining and predicate logic has been applied on the earthquake data of all four quadrants. This predicate logic based prediction

system is trained to predict earthquake in whole quadrant of earth for forthcoming 12 h [29]. Fault lines are modelled in laboratory to study earthquakes. Acoustic signals are emitted during motion of faults. The physical properties of these acoustic signals are combined with machine learning techniques to predict the time before the upcoming lab-earthquake [30]. Recently, a machine learning technique, called Long-Short Term Memory (LSTM) networks has been used to develop the earthquake prediction model [4]. This model exploits the idea that earthquakes occur in a region due to spatial and temporal interaction of multiple regions. Therefore, this relationship has been learned using LSTM.

EPS based on seismic indicators and machine learning has been a subject undergoing intense study according to contemporary literature. Quest for finding new meaningful seismic indicators is still underway. Evolutionary algorithms, in particular GP and AdaBoost based machine learning methods have not been explored so for in this field of research. Thereby, the application of GP-AdaBoost based ensemble classification in combination with innovatively computed seismic indicators is a unique contribution for the field of earthquake prediction.

3. Data and methods

The regions considered for performing research on seismic indicators based EPS are Hindukush, Chile and Southern California. A large number of earthquakes have occurred in aforementioned regions which make them interesting for earthquake prediction. In seismic indicators based EPS, the required raw dataset is temporal sequence of past seismicity for the selected regions. The past seismicity is available in the form of a catalog, which is publicly available from the United States Geological Survey (USGS) [31]. The catalog considered for this study is taken from period of January 1980 to December 2016. The coordinate boundaries of the regions are kept same as taken in the previous earthquake prediction studies for the respective regions [10–12]. The catalogs of the said regions are evaluated for cut-off magnitude. Cut-off magnitude refers to the minimum magnitude, below which earthquake catalog is deemed incomplete. There are numerous techniques to assess the cut-off magnitude for a catalog, whereas in this study it is obtained using the analysis of Gutenberg-Richter curve [32]. Table 1 summarizes the range of coordinates and cut-off magnitude for each region.

3.1. Seismic indicators computation

In this study, seismic indicators are considered as base line for the development of EP-GPBoost. These indicators are grounded on the well-known seismic facts of Gutenberg-Richter's law, seismic energy release, seismic rate changes, foreshock frequency and recurrence time of earthquakes. The Panakkat's and Reye's indicators are considered, in addition to some other indicators of seismic rate changes, standard deviation of b-value, as considered in Zamani et al. [33]. Table 2 introduces the seismic parameters introduced in this research. These indicators are explained in detail in previous research studies carried out in this field [10–12,33]. The state of the art in this research regarding indicators is their computation strategy, which is based on idea of retaining maximum information regarding seismic state of the region. Some of the specific seismic indicators can be computed via multiple approaches, while there are few other indicators which are dependent upon a variable parameter.

Table 1
Range of coordinate boundaries considered for respective regions and cut-off magnitude of earthquake catalogs.

Regions	Latitude range	Longitude range	Cut-off magnitude
Southern California	32–36.5 N°	114.75–121 W°	2.6
Chile	32.5–36 S°	70–72.5 W°	3.4
Hindukush	35–39 N°	69–74.6 E°	4.0

Table 2
Seismic indicators employed in combination with GP-Adaboost in this study.

Symbolic representation	Description
b	Slope of Gutenberg-Richter curve
a	y-intercept of Gutenberg-Richter curve
σb	Standard deviation of b-value
$T_{recurrence}$	Total recurrence time between earthquake magnitudes
β	Seismic rate change proposed by Matthews and Reasenberg [34]
z	Seismic rate change proposed by Habermann and Wyss [35]
M_{mean}	Mean magnitude
$dE^{1/2}$	Rate of square root of energy
ΔM	Magnitude deficit
η	Mean square deviation
x_6	Maximum magnitude earthquake recorded during last week
x_7	Probability of occurring a magnitude larger than or equal to 6.0
T	Time elapsed for last "n" seismic events

Every region may possess different geological properties and different nature of relation with earthquakes, which is intended to be modelled in this study by the combination of seismic indicators and CI. A set of seismic indicators showing best performance for one region may not show the similar performance for other regions. But it would be impractical and undesired to discover the best suited combination of indicators for every different region in a real time EP-GPBoost. Therefore, keeping this in view, the seismic indicators are computed through multiple possibilities. All the computed indicators are simultaneously employed for developing EP-GPBoost. Seismic indicators are classified into two categories based upon their computing strategy as described below.

3.1.1. Non-parametric seismic indicators

The seismic indicators are computed mathematically using temporal sequence of past seismicity. The indicators which are mathematically independent of any other variable factor apart from past seismicity, are directed as non-parametric seismic indicators. Table 3 briefs about all the non-parametric seismic indicators and mathematical expressions.

For example, b-value is the slope of Gutenberg-Richter's law curve and corresponds to seismic rate of a region [36]. It is based upon the past seismicity only, without involvement of any other variable parameters, therefore, it is considered as non-parametric seismic indicator. It can be computed using two different approaches, namely, least square regression analysis (*lsq*) and maximum likelihood method (*mlk*). Similarly, a-value is y-intercept of Gutenberg-Richter's law curve. Thus, two b-values lead to the calculation of two a-values. The rest of seismic indicators provide single value each, which lead to the total of ten non-parametric seismic indicators.

3.1.2. Parametric seismic indicators

The seismic indicators, which are mathematically dependent upon any other variable in addition to seismicity, are called as parametric seismic indicators. In the previous research studies, a single value of such parameters was considered only. In order to obtain the maximum internal seismic information of a region, seismic indicators are computed for multiple values of variable parameter. For example, standard deviation of b-values (Σb) is dependent upon b-value itself. Since we already computed two b-values, so employing both values separately for computing Σb. This leads to the availability of two seismic indicators for Σb. Similarly, total recurrence time ($T_{Recurrence}$) is dependent upon b, a-values along with a varying threshold magnitude (M_j). It is computed for different combination of variable parameters, thereby

Table 3
Details of non-parametric seismic indicators including frequency and mathematical expresson.

Indicator name	No. of features	Mathematical expression
b	2	Least square regression analysis (*lsq*), $\frac{(n(\sum M_i \log N_i) - \sum M_i \sum \log N_i)}{((\sum M_i)^2 - n\sum M_i^2)}$ Maximum likelihood (*mlk*), $\frac{\log_{10} e}{\text{mean}(M) - \min(M)}$
a	2	Least square regression analysis (*lsq*), $\sum (\log_{10} N_i + b_{lsq} M_i)/n$ Maximum likelihood (*mlk*), $\log_{10} N + b_{mlk} \min(M)$
$dE^{1/2}$	1	$\frac{\sum (10^{(11.8+1.5M)})^{\frac{1}{2}}}{T}$,
T	1	$t_n - t_1$, t= time in days
M_{mean}	1	$\frac{\sum_i M}{n}$, where n = total no. of earthquakes used for computation of seismic indicators, which in this study is taken as 50
z	1	$\frac{R_1 - R_2}{\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}}$ where R_1 and R_2 correspond the seismic rate for two different intervals. S_1 and S_2 represent the standard deviation of rate. n_1 and n_2 show the number of seismic event in both intervals
β	1	$\frac{M(t,\delta) - n\delta}{\sqrt{n\delta(1-\delta)}}$ where n represents total events in the whole earthquake dataset, t is total time duration and δ is the normalized duration of interest. $M(t, \delta)$ shows the number of events observed, defined using end time t and interval of interest δ
x_6	1	$\max\{M_i\}$, when $t \in [-7, 0)$
Total	10	

Table 4
Details of non-parametric seismic indicators including frequency and mathematical expresson.

Indicator name	No. of features	Mathematical expression
x_7	2	$e^{\frac{-3b_i}{\log e}}$, for i = [*lsq, mlk*]
η	2	$\frac{\sum(\log N - a_i - b_i M)^2}{n-1}$, for i = [*lsq,mlk*]
σb	2	$2.3b_j^2\sqrt{\frac{\sum_{i=1}^{n}(M_i - \text{mean}(M))^2}{n(n-1)}}$, for j = [*lsq, mlk*]
ΔM	2	$M_{max,actual} - a_i/b_i$, for i = [*lsq, mlk*]
$T_{recurrence}$	42	$T_r = \frac{T}{10^{a_i - b_i M_j}}$, for i = [*lsq, mlk*] and j = [4.0, 4.1,4.2,…, 6.0]
Total	50	

providing multiple seismic features for $T_{Recurrence}$. Overall fifty seismic features have been obtained for parametric seismic indicators. The names of parametric seismic indicators and their computation strategy is summarized in Table 4.

3.2. Genetic Programming and AdaBoost

A classification algorithm establishes criteria for deciding a target label for the test instance on the basis of values of features. A classifier is induced with labeled instances and then a learned classifier differentiates between test instances of binary classes, in the case of earthquake prediction system. GP has been applied in number of problems, for attaining an optimal solution due to its searching capabilities. In this work, GP has been ensembled with AdaBoost to construct a strong classifier, with enhanced classification capabilities. AdaBoost algorithm offers weight updating for hard instances in an iterative process. The weight updating through boosting helps in dealing with hard instances, which ultimately improves the classification performance. In GP-AdaBoost ensemble, *N* GP strings are evolved per class through boosting. AdaBoost contributes in evolving multiple GP strings through an iterative process in such a way that each new GP string recognizes the incorrectly classified instances of previous iteration. The boosting performs weight update for instances in a bid to tackle hard instances. Area Under Curve (AUC) is considered as a fitness function for the evolution of GP strings per class. Moreover, each of the evolved GP strings act as a single class classifier. The decision of a class label for a test instance is made on the basis of higher value from a weighted sum of the outputs of GP strings evolved for each class.

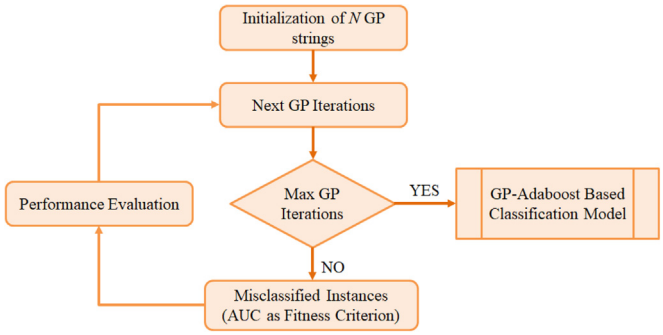

Fig. 1. Flow chart of GP-AdaBoost based earthquake predictor system.

In GP-AdaBoost methodology, the GP's evolution is conducted through adopting boosting for weight update. The boosting is employed to generate P number of GP programs for each class. The obtained results are added and the decision of a class label for a test instance is made on the basis of higher value from a weighted sum of the outputs of GP strings evolved for each class. This GP-AdaBoost methodology developed for earthquake prediction is inspired from the work presented by Idris et al. [17]. Fig. 1 shows the sequence of the processes involved in GP-AdaBoost algorithm, whereas Fig. 2 lists the steps involved in the algorithm.

In GP-AdaBoost algorithm, multiple parameters are involved which need to be adjusted before training the methodology. The parameters along with their values are shown in Table 5. These values are empirically chosen after exhaustive experimentation during the training phase. The overall flowchart of the proposed research methodology for EPS is shown Fig. 3.

In GP-AdaBoost algorithm, multiple parameters are involved which needs to be adjusted before training the methodology. The parameters along with their values are shown in Table 5. These values are empirically chosen after exhaustive experimentation during the training phase. The Elite Size is the parameter which identifies the number of GP program evolved through boosting, for each class involved in earthquake prediction. Similarly, cross over, mutation and reproduction rates are set to 0.07, 0.90 and 0.03, respectively. The higher value of mutation is used to ensure diversity for each of the next generation. Moreover, Ramped half and half method is applied for initializing the population, whereas AUC is applied for evolution of earthquake predictor system

4. Results and discussion

In this research EPS is modelled as a binary classification task with aim to generate prediction for earthquakes of magnitude 5.0 and greater 15 days prior to an earthquake. The results of the proposed methodology are evaluated through parameters given below.

4.1. Evaluation parameters

The outcome of a binary classification model is either called as True Positive (TP), False Positive (FP), True Negative (TN) or False Negative (FN) when compared with original data. These terms are defined as:

Step by step procedure of GP-AdaBoost

Inputs: C, P, T, N
[C = Number of target classes (in our case 2)
P = Number of GP strings required in boosting
T = Training dataset
N = Number of instances in T]
For j = 1 to C (GP strings are obtained for each class)
GP_Init(POP) (Population Initialization)
Init_Weights(W) (Weight Initialization as $W_i = 1/N$ for each instance i)
For k=1 to P (GP strings evolved using Ada boosting)
Fill POP. New set of GP strings
Evolving GP string capable of recognizing Class C_j, keeping AUC as fitness function
end For
end For
(P * C) GP strings are acquired for every class Cj
Accumulative classification result of each class Cj is calculated by weighted sum of α_{jk} and $output_{jk}$ of the GP strings.

$$max_j\left(\sum_{k=1}^{P}\left(a_{jk} * output_{jk}\right)\right)$$

The j which, scores maximum indicates the class of a test instance

Fig. 2. List of steps involved in GP-AdaBoost algorithm.

Table 5
List of varying parameters for GP-AdaBoost and their selected values.

Parameter values used in GP-AdaBoost algorithm	
Parameter Name	Value
Number of generation	20
Number of GP strings per class	5
Fitness function	AUC
Functions	+, /, *, If, > , < , Pow, &, |, Max, Min, Exp, Log
Max depth of trees	20
Mutation	0.90
Cross over	0.07
Reproduction of new programs	0.03
Population size	100
Population initializer	Ramped Half and Half method

- True Positive (TP): An earthquake actually occurred and also predicted by EPS.
- False Positive (FP): No earthquake occurred but falsely predicted by EPS.
- True Negative (TN): No earthquake occurred and no alarm generated by EPS.
- False Negative (FN): An earthquake occurred but EPS was unable to predict.

The performance evaluation metrics are computed based upon aforementioned quantities. These metrics and their mathematical formulae are given in Table 6.

The reason of observing performance through multiple criteria is to examine the different aspects of EP-GPBoost. Sensitivity signifies the capability of EP-GPBoost to sense the earthquakes while specificity does the same for non-earthquakes instances. P_1 represents the ratio of actual true predictions out of all the generated earthquake predictions. In other words, it can be related to false alarms. Higher P_1 refers to the lower false alarm ratio, which is of utmost importance in EP-GPBoost. The total accurate predictions, whether positive or negative, made by EP-GPBoost are expressed by accuracy. However, accuracy is generally not considered as a measure which truly reflects the competence of a classifier [37]. Therefore, Matthews Correlation Coefficient (MCC) and R Score (R) have also been introduced for performance evaluation. MCC and R incorporate all four types of predictions (TP, TN, FN, TN)

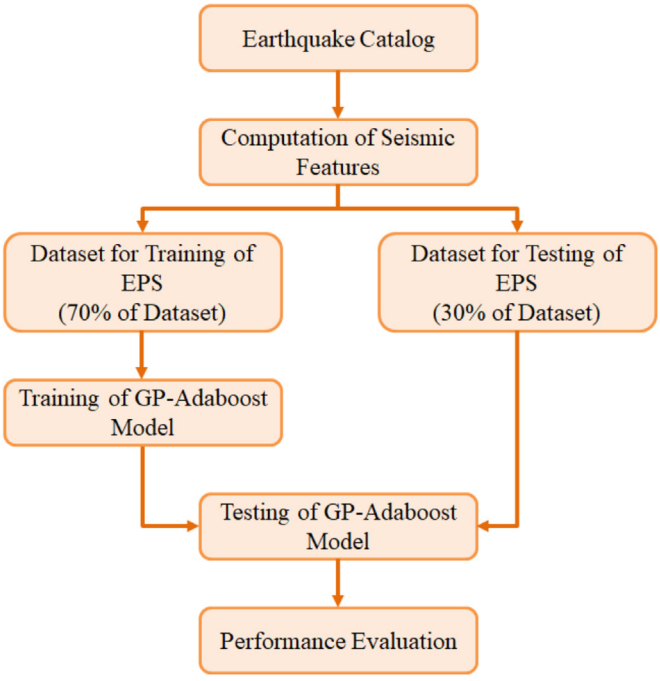

Fig. 3. Flowchart of overall research methodology.

simultaneously to give an overall balanced view of performance. Their values vary between − 1 and + 1, where 1 signifies the perfect prediction model while − 1 shows the opposite behavior of model. However, 0 depicts the total random behavior of algorithm.

4.2. Performance of the EP-GPBoost

The datasets of all three regions contain different number of instances, depending upon the quantity of seismic events recorded in the catalogs of the respective regions. The quality of the earthquake catalog depends upon the density of instrumentation in a certain region. In this regard, the Southern California takes lead with the least cut-off earthquake magnitude, followed by Chile and then Hindukush region with the highest cut-off magnitude. It must be noted that before processing

Table 6
Performance evaluation measures and the respective mathematical equations.

Performance metrics	Mathematical equation
Sensitivity (S_n)	$S_n = \frac{TP}{TP + FN}$
Specificity (S_p)	$S_p = \frac{TN}{TN + FP}$
Positive Predictive Value/Precision (P_1)	$P_1 = \frac{TP}{TP + FP}$
Negative Predictive Value (P_0)	$P_0 = \frac{TN}{TN + FN}$
Matthews Correlation Coefficient (MCC)	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
R Score	$R = \frac{(TP \times TN) - (FP \times FN)}{(TP + FN)(FP + TN)}$
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

the earthquake catalog for seismic indicators' computation, all the seismic events below cut-off magnitude are removed. This ensures that misleading and incomplete information is not considered to determine the trends in seismic indicators.

In this study, due to the approach used for computation of seismic indicators, the number of instances are dependent upon the quantity of seismic events available in catalogs. Thus, the number of instances available for Southern California are 33,543, while the dataset of Chile contains 7656 instances. The Hindukush contains the least number of instances, i.e. 4350. Every seismic region is different from other region and may show varied behavior for different seismic indicators. Therefore, separate training of EP-GPBoost is carried for every region using 70% of the available seismic instances in the respective datasets. Once the trained EP-GPBoost is obtained, the results are evaluated over the rest of unseen 30% of datasets.

The performance of EP-GPBoost for all three regions is summarized in Table 7. The EP-GPBoost shows remarkable performance for all three regions, particularly in terms of low false alarms generation. The precision value of 74%, 80% and 84% for Hindukush, Chile and Southern California, respectively implies that the ratio of false alarms is considerably low. The model also shows remarkable performance in terms of MCC and R score. MCC of 0.47 for Hindukush represents a noteworthy positive relation between seismic indicators and subsequent earthquakes, which improves for Chile to 0.60 and further strengthens for Southern California to 0.67.

Inter-regional comparison of results obtained through GP-AdaBoost expresses that Southern California takes lead followed by Chile and then Hindukush. The reason for such trend is evident from the quality of relevant earthquake catalog. Therefore, it is needed to improve the quality of earthquake catalogs in terms of improved cut-off magnitude that can be achieved by increasing the density of instrumentation for recording earthquakes.

4.3. Comparison to existing works

The reason for selecting three aforementioned regions for developing seismic indicators based EP-GPBoost is to draw a comparison with previous results obtained for these regions. Different seismic indicators based predictions have already been carried out for these regions using various machine learning techniques. Thus, in order to prove the superiority of proposed methodology based on seismic indicators computation, the comparison has been drawn with previously proposed methodologies through said evaluation measures. It is evident from Table 7 that the predictions obtained by the proposed methodology have outperformed the previously obtained prediction results for the respective regions.

Table 7
Prediction results for binary classification problem of GP AdaBoost based EPS.

Performance metrics	Hindukush	Chile	Southern California
S_n (%)	57.7	61.2	68.7
S_p (%)	89.2	93.9	94.4
P_1 (%)	74.3	80.2	84.2
P_0 (%)	79.6	85.7	87.4
Accuracy (%)	78.7	84.5	86.6
MCC	0.50	0.60	0.67
R Score	0.47	0.55	0.63

The MCC recorded for Hindukush region by [12] is 0.33 which has been improved to 0.50 in this study, thus showing notable improvement in the prediction results. Furthermore, the current methodology has also shown improved results for other criteria as well. Precision has improved from 61% to 74% and accuracy from 65% to 78%. The only decreased performance is S_n, which is affordable, given the improvement recorded in all other performance measures. The prediction results for Chile has improved using the proposed methodology in terms of all the used performance measures. MCC has increased from 0.39 to 0.6 while noticeable decrease in false alarms can also be witnessed through increased precision from 61% to 80%. The similar trend of improvement is also observed for Southern California region in which MCC is improved from 0.51 to 0.66. Precision has improved from 71% to 84%, while noteworthy improvement has also been observed in all the performance measures except for S_n. The trade-off between small decrease in S_n and rest of performance measures is acceptable. Therefore, the obtained results show that the proposed approach, EP-GPBoost, has outperformed the previously obtained results for the considered regions.

GP-AdaBoost and inclusion of new seismic indicators through the principle of retaining maximum information, both have contributed towards the improvement of results. Newly introduced seismic indicators covering different aspects of the already proposed indicators have provided the detailed information of the seismic region. This detailed information regarding the seismic region has been exploited by the searching capabilities of GP. As GP searches for the best performing features, is further coupled with ensemble methodology of AdaBoost, thereby providing a robust model for earthquake prediction. Table 8 details the comparison of proposed methodology with previously proposed methodologies for three regions, whereas Fig. 4 visualizes the performance in terms of MCC. In previous research studies, such a variety of seismic indicators have not been exploited simultaneously. Furthermore, combined usage multiple machine learning techniques for earthquake prediction is the novelty of this research.

5. Conclusion

In this research, seismic indicators based EP-GPBoost has been proposed. A unique methodology is devised, which encompasses the maximum information of a region through the computation of available seismic indicators. These indicators are fed to a Genetic Programming (GP) and AdaBoost (GP-AdaBoost) based ensemble classification methodology. GP-AdaBoost is a unique combination of strong searching and boosting capabilities of GP and AdaBoost, respectively. The GP-AdaBoost based model has been trained and tested for the Hindukush, Chile and Southern California regions. The obtained prediction results for these regions exhibit improvement when compared with already available studies. Inclusion of maximum available seismic indicators and application of GP-AdaBoost, has resulted to enhance earthquake prediction performance 15 days prior to an earthquake. Thus, the computation of maximum seismic parameters and employing of GP-AdaBoost has developed a new and robust EPS, called as EP-GPBoost. Future efforts are aimed towards finding more suitable seismic indicators and application of deep learning methodologies for earthquake predictor system.

Table 8
Comparison of EP-GPBoost with previous methodologies applied for Hindukush, Chile and Southern California.

	Hindukush		Chile		Southern California	
Performance evaluation	Asim et al. [12]	Proposed methodology	Reyes et al. [10]	Proposed methodology	Panakkat and Adeli [11]	Proposed methodology
S_n (%)	91	57.7	43.1	61.2	80	68.7
S_p (%)	36	89.2	91.3	93.9	71	94.4
P_1 (%)	61	74.3	61.1	80.2	71	84.2
P_0 (%)	79	79.6	83.5	85.7	86	87.4
Acc. (%)	65	78.7	79.7	84.5	75.2	86.6
MCC	0.33	0.50	0.39	0.60	0.51	0.67
R Score	0.27	0.47	0.34	0.55	0.51	0.63

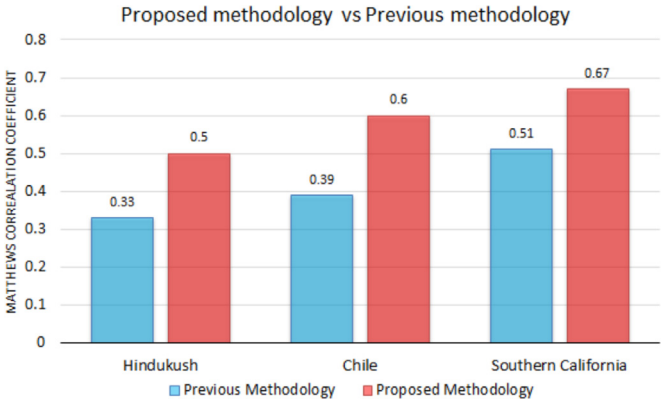

Fig. 4. Performance comparison between proposed methodology and previous methodology for respective regions in terms of MCC.

References

[1] Allen CR. Responsibilities in earthquake prediction: to the seismological Society of America, delivered in Edmonton, Alberta, May 12, 1976. Bull Seismol Soc Am 1976;66:2069–74.
[2] Geller RJ, Jackson DD, Kagan YY, Mulargia F. Enhanced: earthquakes cannot be predicted. Science 1997;275:1616–20.
[3] Geller RJ. Earthquake prediction: a critical review. Geophys J Int 1997;131:425–50.
[4] Wang Q, Guo Y, Yu L, Li P. Earthquake prediction based on spatio-temporal data mining: an LSTM network approach. IEEE Trans Emerg Top Comput 2017. http://dx.doi.org/10.1109/TETC.2017.2699169.
[5] Sil A, Sitharam T, Haider ST. Probabilistic models for forecasting earthquakes in the northeast region of India. Bull Seismol Soc Am 2015.
[6] Boucouvalas A, Gkasios M, Tselikas N, Drakatos G. Modified-Fibonacci-dual-Lucas method for earthquake prediction. Proc SPIE Vol 2015. [95351A-95351A].
[7] Jilani Z, Mehmood T, Alam A, Awais M, Iqbal T. Monitoring and descriptive analysis of radon in relation to seismic activity of Northern Pakistan. J Environ Radioact 2017;172:43–51.
[8] Barkat A, Ali A, Siddique N, Alam A, Wasim M, Iqbal T. Radon as an earthquake precursor in and around northern Pakistan: a case study. Geochem J 2017;51:337–46.
[9] Awais M, Barkat A, Ali A, Rehman K, Zafar WA, Iqbal T. Satellite thermal IR and atmospheric radon anomalies associated with the Haripur earthquake (Oct 2010; Mw 5.2), Pakistan. Adv Space Res 2017;60:2333–44.
[10] Reyes J, Morales-Esteban A, Martínez-Álvarez F. Neural networks to predict earthquakes in Chile. Appl Soft Comput 2013;13:1314–28.
[11] Panakkat A, Adeli H. Neural network models for earthquake magnitude prediction using multiple seismicity indicators. Int J Neural Syst 2007;17:13–33.
[12] Asim K, Martínez-Álvarez F, Basit A, Iqbal T. Earthquake magnitude prediction in Hindukush region using machine learning techniques. Nat Hazards 2017;85:471–86.
[13] Adeli H, Panakkat A. A probabilistic neural network for earthquake magnitude prediction. Neural Netw 2009;22:1018–24.
[14] Morales-Esteban A, Martínez-Álvarez F, Reyes J. Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence. Tectonophysics 2013;593:121–34.
[15] Asim KM, Idris A, Martínez-Álvarez F, Iqbal T. Short term earthquake prediction in Hindukush region using tree based ensemble learning. In: Proceedings of the 2016 international conference on frontiers of information technology (FIT); 2016, p. 365–70.
[16] He C, Micallef L, Tanoli Z-u-R, Kaski S, Aittokallio T, Jacucci G. MediSyn: uncertainty-aware visualization of multiple biomedical datasets to support drug treatment selection. BMC Bioinforma 2017;18:393.
[17] Idris A, Iftikhar A, ur Rehman Z. Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. Clust Comput 2017:1–15. http://dx.doi.org/10.1007/s10586-017-1154-3.
[18] Asim KM, Murtza I, Khan A, Akhtar N. Efficient and supervised anomalous event detection in videos for surveillance purposes. In: Proceedings of the frontiers of information technology (FIT), 2014 12th international conference on; 2014, p. 298–302.
[19] Shang X, Li X, Morales-Esteban A, Chen G. Improving microseismic event and quarry blast classification using artificial neural networks based on principal component analysis. Soil Dyn Earthq Eng 2017;99:142–9.
[20] Shang X, Li X, Morales-Esteban A, Dong L, Peng K. K-means cluster for seismicity partitioning and geological structure interpretation, with application to the Yongshaba Mine (China). Shock Vib 2017;2017.
[21] Pulinets S, Ouzounov D. Lithosphere–atmosphere–ionosphere coupling (LAIC) model – an unified concept for earthquake precursors validation. J Asian Earth Sci 2011;41:371–82.
[22] Buskirk RE, Frohlich C, Latham GV. Unusual animal behavior before earthquakes: a review of possible sensory mechanisms. Rev Geophys 1981;19:247–70.
[23] Grant RA, Raulin JP, Freund FT. Changes in animal activity prior to a major (M = 7) earthquake in the Peruvian Andes. Phys Chem Earth Parts A/B/C 2015;85:69–77.
[24] Kagan YY, Jackson DD. Probabilistic forecasting of earthquakes. Geophys J Int 2000;143:438–53.
[25] Asim KM, Awais M, Martínez–Álvarez F, Iqbal T. Seismic activity prediction using computational intelligence techniques in northern Pakistan. Acta Geophys 2017;65:919–30.
[26] Rafiei MH, Adeli H. NEEWS: a novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. Soil Dyn Earthq Eng 2017;100:417–27.
[27] Martínez-Álvarez F, Reyes J, Morales-Esteban A, Rubio-Escudero C. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. Knowl-Based Syst 2013;50:198–210.
[28] Asencio-Cortés G, Martínez-Álvarez F, Morales-Esteban A, Reyes J. A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction. Knowl-Based Syst 2016;101:15–30.
[29] Ikram A, Qamar U. Developing an expert system based on association rules and predicate logic for earthquake prediction. Knowl-Based Syst 2015;75:87–103.
[30] Rouet-Leduc B, Hulbert C, Lubbers N, Barros K, Humphreys CJ, Johnson PA. Machine learning predicts laboratory earthquakes. Geophys Res Lett 2017;44:9276–82.
[31] Survey USG. Quaternary fault and fold database for the United States [Online]. Available: ⟨http//earthquake.usgs.gov/hazards/⟩.
[32] Wiemer S, Wyss M. Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the western United States, and Japan. Bull Seismol Soc Am 2000;90:859–69.
[33] Zamani A, Sorbi M, Safavi A. Application of neural network and ANFIS model for earthquake occurrence in Iran. Earth Sci Inform 2013;6:71–85.
[34] Matthews MV, Reasenberg PA. Statistical methods for investigating quiescence and other temporal seismicity patterns. Pure Appl Geophys 1988;126:357–72.
[35] Habermann RE, Wyss M. Reply [to "Comment on Habermann's method for detecting seismicity rate changes"]. J Geophys Res: Solid Earth 1987;92:9446–50.
[36] Gutenberg B, Richter CF. Seismicity ofthe Earth and Associated Phenomena. Princeton, New Jersey: Princeton University Press; 1954. Second Ed.
[37] Idris A, Rizwan M, Khan A. Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies. Comput Electr Eng 2012;38:1808–19.