Hindawi
Journal of Sensors
Volume 2022, Article ID 5321530, 14 pages
https://doi.org/10.1155/2022/5321530

Research Article

A Novel Ensemble Earthquake Prediction Method (EEPM) by Combining Parameters and Precursors

Sumita Mukherjee,[1] **Prinima Gupta**,[1] **Pinki Sagar**,[2] **Neeraj Varshney**,[3] **and Manoj Chhetri**[4]

[1]*Manav Rachna University, Faridabad, India*
[2]*Manav Rachna International Institute of Research and Studies, Faridabad, India*
[3]*GLA University, Mathura, India*
[4]*Department of Information Technology College of Science and Technology, Royal University of Bhutan, Phuntsholing, Chukha, Bhutan*

Correspondence should be addressed to Manoj Chhetri; manoj_chhetri.cst@rub.edu.bt

Received 4 July 2022; Revised 18 August 2022; Accepted 19 September 2022; Published 25 October 2022

Academic Editor: Rajesh Kaluri

A leading cause of death from natural disasters over the last 50years is witnessed by none other than earthquake occurrences which have a negative economic impact on the world and claimed thousands of lives over the years, causing devastation to properties. In this paper, a novel Ensemble Earthquake Prediction Method (EEPM) is proposed and implemented to produce a strong learner (ensemble method) having better accuracy in prediction, less variance, and less errors. Data (parameters) which is continuous in nature is collected from two countries, India and Nepal, for five years, and surveyor's data (precursor) which is categorical in nature is collected from three countries India, Nepal, and Kenya for five years on the specific earthquake-prone regions. The preprocessed data is generated by combining parameters and precursor data. EEPM focuses on detecting the accurate and better early signs of an earthquake and finding the probability of occurrence of an earthquake in the specified region, i.e., better prediction and robustness. The results of EEPM produced better R^2 and less variance and less error in comparison to individual machine learning methods as well as better accuracy 87.8%, compared to state-of-the-art ensemble methods. The prediction of earthquake will alarm not only the people of the society but also the different organizations to explain the appropriate range of magnitude and dynamics of occurrence of earthquake.

1. Introduction

The meaning of prediction of an earthquake means concise forecasting of the time, size, and location of an impending earthquake in isotopic and geochemical precursors of earthquakes and volcanic eruptions. The researcher stated that earthquake forecasts in different regions considering factors such as local geological conditions, ground motion, and animal behaviour are taken into account for seismic casualties. For earthquake precursors, solutions have been searched in past studies, precisely in the USA, Japan, China, Israel, etc., by supervising regional studies based on parametric features. Researchers on the other hand in India, USA, Pakistan, Nepal, etc. have investigations on earthquake prediction using parameters which are based on the assumption on all regional factors, which can be filtered out, and general information about the earthquake patterns based on the parameters can be extracted. When the energy stored in elastically strained rocks is suddenly released, this release of energy causes intense ground shaking causing occurrence of earthquake in the area near the source of the earthquake, which generates energy in the form of waves throughout the Earth, elastic in nature called seismic waves. Earthquakes can be generated by many geographical factors like sudden volume changes in minerals, sudden slippage along faults, ground motion, bomb blasts, volcanic eruptions, heavy rainfall, rock bed material, regional tectonics, and altitude noticed by meteorologists, seismologists, and geologists. Based upon 80 systematically selected, high-quality peer-reviewed

research papers, using data mining methods, implementation on earthquake prediction is being thoroughly read and analysed. Different data mining methods are carried on the data sets to generate the accurate prediction on the probability of occurrence of earthquake.

In this research paper, the parameters and precursors of earthquake-related data are collected from different sources, then combined to get a unique preprocessor data having all the required features for prediction. The categorical data are converted to numerical so that a common numerical data set format having necessary attributes is considered. This unique preprocessed data can be utilized by different individual methods for analysis and logical results of regression. Pearson coefficient correlation through feature selection and proper data splitting ratio is selected, fitted with a contributing method, and then trained to minimize training errors by reducing dimensionality, increasing computational efficiency. This unique data set is used to apply on different data mining methods like KNN, SVM, XGBoost, decision tree, and random forest to generate the result of individual methods and then compared using performance measures like R^2, adjusted R^2, variance, Mean Square Error (MSE), and Root Mean Square Error (RMSE) and shows prediction with high variance and errors [1]. A novel ensemble method, Ensemble Earthquake Prediction Method (EEPM), using boosting over a single contributing method is framed for combining a set of weak learners into a strong learner in each iteration so that the ensemble method counteracts the behaviour to allow for generalization of the methods to new data sets. Boosting of the ensemble method is used to base learners which are generated sequentially by different individual methods in such a way so that the weak learners are eliminated after each iteration. The present base learner is always more effective than the previous one, as the decision tree is applied having a single feature to fit the training data set by choosing the right decision at each split of the tree. The predictions from the trees are combined using a random forest having overlapping results applying averaging for regression. The generated accuracy of the ensemble method generated is compared with previous developed ensemble methods for prediction of the probability of occurrence of earthquake.

2. Literature Survey

The proper selection of data sets plays an important role as few researchers used geological observations and historical data of a particular region or country so that attributes are justified and have a strong relationship with seismic activity. In the literature, the papers dealt with the selection of parameters and precursors with different data sets, to derive a powerful logical preprocessor data having independent and dependent attributes. The preprocessed data as a result becomes a strong and a new data set which is applied on different techniques like regression, KNN, SVM, decision tree, and random forest [2, 3].

The history of earthquake forecasting in a specific region is based on the identify location of fault characteristics or parameters such as depth, magnitude, length, latitude, longitude, and time and precursors such as movement of animals, trees, plant change in temperature, pressure, and radon gas to estimate the occurrence of earthquake. The identification of the magnitude of earthquakes is estimated simultaneously using all available fault parameters and precursors by excluding the chances of inconsistent estimations through data mining technique application for forecasting [4].

2.1. Comparison of Research Findings with Parameters, Precursors, and Previous Ensemble Methods. In the literature, it is concluded that to predict earthquakes, many researchers have asserted by observing multiple parameters based on observational data and developing patterns and relationships and some using multiple precursors and observing the change in seismicity patterns of the region. It has been observed that data mining techniques are capable of delivering better accuracy in terms of prediction for short-term and midterm earthquakes in comparison to large earthquakes. The analysis of the literature review using parametric data is shown in Table 1, and that using precursor data is shown in Table 2. Analyses of earlier ensemble methods are shown in Table 3.

Many researchers have worked on ensemble methods in different application areas. A system is proposed and designed by the ensemble method, combining different data mining techniques and network techniques which can detect an unexpected voltage leak from some electric devices to the house in order to save people's lives and resources [17]. In this paper, the proposed model achieves integrity by embedding a security feature Elliptic Curve Digital Signature Algorithm (ECDSA) for the predicted area of water bodies which helps to secure the key and the detected water bodies while transmitting in a channel using the ensemble method by combining data mining technique like networking XGBoost, and random forest [18]. In this paper, the effectiveness of the proposed model using various machine learning algorithms such as random forest,k-nearest neighbour, and decision tree s used and tested on using actual IoT-based data set which shows better accuracy using the ensemble method [19].

3. Sources of Data Collection

A uniform, nonredundant earthquake catalogue data have been compiled. The parameter data were collected from two countries, India and Nepal, for five years. The precursor data were collected from people who actually experienced earthquakes within the age range from 18 to 75 from three countries like India, Nepal, and Kenya, for five years. Some people have experienced the earthquake more than once. The parametric data has no null records and no outliers as the database is cleaned by the data source house. The earthquake catalogue has included four zones East, West, South, and North of two different countries referred in Table 4.

3.1. Data Source of Parameter Data

(1) The database was provided by the Meteorology Department of India and Disaster Research (India Today)

(2) The National Centre for Medium Weather Forecasting is located at Sector-50, Noida

Table 1: Research work using parameters.

Paper Ref.	Year	Number of records	Accuracy	Magnitude range	Data mining methods
[2]	2015	Average	79	≤5.2	Classification, KNN, SVM, regression
[5]	2016	Medium	81	≤4.4	Classification, KNN, SVM, regression
[6]	2017	Average	81	≤5.1	Regression, SVM, classification, XGBoost
[7]	2018	Large	82	≤4.1 to 5.0	SVM, XGBoost, neural network
[8]	2020	Average	84	≤4.1	Regression, classification, SVM, random forest

Table 2: Research work using Precursors.

Paper Ref.	Year	Number of records	Accuracy	Magnitude range	Data mining methods
[9]	2019	Medium	83.2	≤4.9	Statistical features-based approach (SFBA), multi-layer perceptron, regression
[3]	2020	Medium	79to 82	≤5.1	Regression, multi precursor system, XGBoost, SVM, clustering
[10]	2021	Large	79.41	≤5.1	SVM, KNN, regression
[11]	2021	Medium	80.7	≤3.9	Time series, classification, magnitude and inter-event time (IET), through visibility graph, classification, random forest
[12]	2021	Average	81.4	≤4.8	Decision tree, Radom Forest, XGBoost

Table 3: Research based on the ensemble method.

Paper Ref.	Year	Method	Type of data	Accuracy	Combination of methods
[13]	2019	Ensemble boosting	Precursor based	81.9	Support vector machines (SVMs), decision trees (DTs), artificial neural networks (ANNs), extreme learning machines (ELMs), and regression
[14]	2020	Ensemble stacking	Parameter based	82.8	Neural network, support vector machine, decision tree, k -nearest neighbours, classification
[15]	2021	Ensemble stacking	Precursor based	84.3	Models are autoregressive conditional heteroscedasticity (GARCH), autoregressive integrated moving average (ARIMA), time series
[16]	2021	Ensemble stacking	Parameter precursor based	85.3	Different swarm intelligence algorithms, KNN, classification, regression

Table 4: Samples of parameter data collected from Nepal and India.

Date	Time (UTC)	Time (IST)	Latitude	Longitude	Magnitude	Depth	Location
31/01/15	13:59:43	19:29:43	28.37° N	84.07° E	5.0	10	Pokhara, Nepal
09/05/15	1:01:41	6:31:41	23.3° N	70.4° E	10	3.4	Kutch, Gujarat
01/01/16	21:30:45	3:00:45	27.75° N	86.12° E	4.2	10	Xizang-Nepal border region
26/12/16	8:45:48	14:15:48	30.8° N	77.9° E	10	3.5	District. Dehradun, Uttarakhand
22/08/17	0:50:26	6:20:26	29.3931° N	81.1053° E	4.6	10	NE of Dipayal, Nepal
07/09/17	14:32:00	20:02:00	25.2° N	90.0° E	10	4.4	India-Bangladesh, Border Region
10/03/18	3:21:28	8:51:28	35.7° N	76.1° E	10	4.5	Jammu & Kashmir
19/10/18	19:57:23	1:27:23	28.57° N	82.88° E	4	10	Janakpur, Nepal
26/09/19	13:04:25	18:34:25	28.1° N	84.6° E	4.3	10	Prayagraj, Bharatpur, Nepal

Table 5: Samples of precursor data collected from Nepal, India, and Kenya.

Date	Time (UTC)	Magnitude	Depth	Temperature	Pressure	Animal behaviour	Falling of leaves	Location
31/01/15	13:59:43	5.0	10	Low	Medium	Overactive	More	Nepal
22/01/16	3:42:00	4.0	45	Very low	High	Overactive	More	Kenya
05/01/17	19:41:00	4.6	10	Very low	Medium	Active	Normal	Nepal
30/04/18	0:37:01	4.7	10	Low	Low	Overactive	More	India
29/04/19	11:27:05	4.2	14	Medium	Medium	Abnormal	More	India

Table 6: Attributes of parameter data set.

Attribute name	Attribute description	Type
X1 = magnitude	Y1 = predicted R^2 from linear regression based on independent variable magnitude	Dependent
X2 = depth	Y2 = predicted R^2 from linear regression based on independent variable magnitude	Independent
X3 = location	Y3 = predicted earthquake magnitude using regression	Independent
X4 = date	Y4 = predicted earthquake magnitude from decision tree using regression	Independent
X5 = time (UTC)	Y5 = predicted earthquake magnitude from random forest using regression	Independent
X6 = time (IST)	Y6 = predicted earthquake magnitude from KNN using regression	Independent
X7 = longitude	Y7 = predicted earthquake magnitude from SVM using regression	Independent
X8 = latitude	Y8 = predicted earthquake magnitude from XGBoost using regression	Independent

(3) India Meteorology Department located at Block M, Lodhi Road, Delhi, has extended their support to avail the historical data. There are many sources used as the historical data were collected from two countries. Few are mentioned below (Source Link file:https://en.wikipedia.org/wiki/Geology_of_India,https://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=1&d=1,https://en.wikipedia.org/wiki/National_Geophysical_Data_Center, andhttps://www.indiatoday.in/diu/story/300-disasters-80-000-deaths-100-crore-affected-india-s-two-decade-tryst-with-natural-calamities-1767202-2021-02-08)

3.2. Data Source of Precursor Data. Through Google Form survey analysis, some survey shows that people having the experience of earthquake more than once in a lifetime are also considered. Repetitive data from survey data is removed, and an accurate database is used for relevance and real-time earthquake events (refer to Table 5.

3.3. Data Set Description. For the parameter data, the input values and attributes are from X1 to X8 and the target values are from Y1 to Y8. The types of target values are numeric in nature. The critical parameter magnitude is the time interval between the arrival of the P-waves and S-waves, and dependent attribute and other parameters are in the dependent attribute. The details of the attributes for parameter data set are tabulated below in Table 6.

For the precursors, the input values and attributes are from P1 to P10 and the target values are from Q1 to Q10 as tabulated below in Table 7. The target values are converted to numeric values. The value of the categorical data of the surveys is denoted by a number like very high temperature is represented by 1—high, 2—low, 3—very low, 4—not sure, and 5—magnitude which is the dependent attribute, and others are independent attributes.

3.4. Preprocessor Data. A preprocessed data is generated by combining both parameters and precursor data sets. The parameters are collected from two countries India and Nepal, and the precursors are collected from three countries India, Nepal, and Kenya. A graphical representation which is easy to identify provides the different elements of a process and understands the interrelationships among the various steps. The sequential steps for developing a preprocessors data are put in a logical order so that data mining techniques can be applied in the future to predict the occurrence of earthquakes as shown in Figure 1.

3.5. Data Integration. Data preprocessing involves different operations which can organise the data in rule-based applications and database driven into a proper format for better interpretation in data mining process. As data integration is a part of data preprocessing which is derived by combining parameters and precursors, standardizing is applied to normalizing routines to transform the data into its preferred and consistent format. Reduced representation in volume is obtained from the number of attributes, number of attribute values, and the number of tuples which produces the same or similar analytical results.

Table 7: Attributes of precursor data set.

Attributes name	Attribute description	Type
P1 = magnitude	Q1 = predicted R^2 from linear regression based on magnitude	Dependent
P2 = depth	Q2 = predicted R^2 from linear regression based on magnitude	Independent
P3 = location	Q3 = predicted R^2 from linear regression based on magnitude	Independent
P4 = date	Q4 = predicted earthquake magnitude from decision tree using regression	Independent
P5 = time	Q5 = predicted earthquake magnitude from random forest using regression	Independent
P6 = temperature	Q6 = predicted earthquake magnitude from KNN using regression	Independent
P7 = pressure	Q7 = predicted earthquake magnitude from SVM using regression	Independent
P8 = animal behaviour	Q8 = predicted earthquake magnitude from XGBoost using regression	Independent
P9 = water movement	Q9 = predicted earthquake magnitude from KNN using regression	Independent
P10 = falling of leaves	Q10 = predicted earthquake magnitude from SVM using regression	Independent

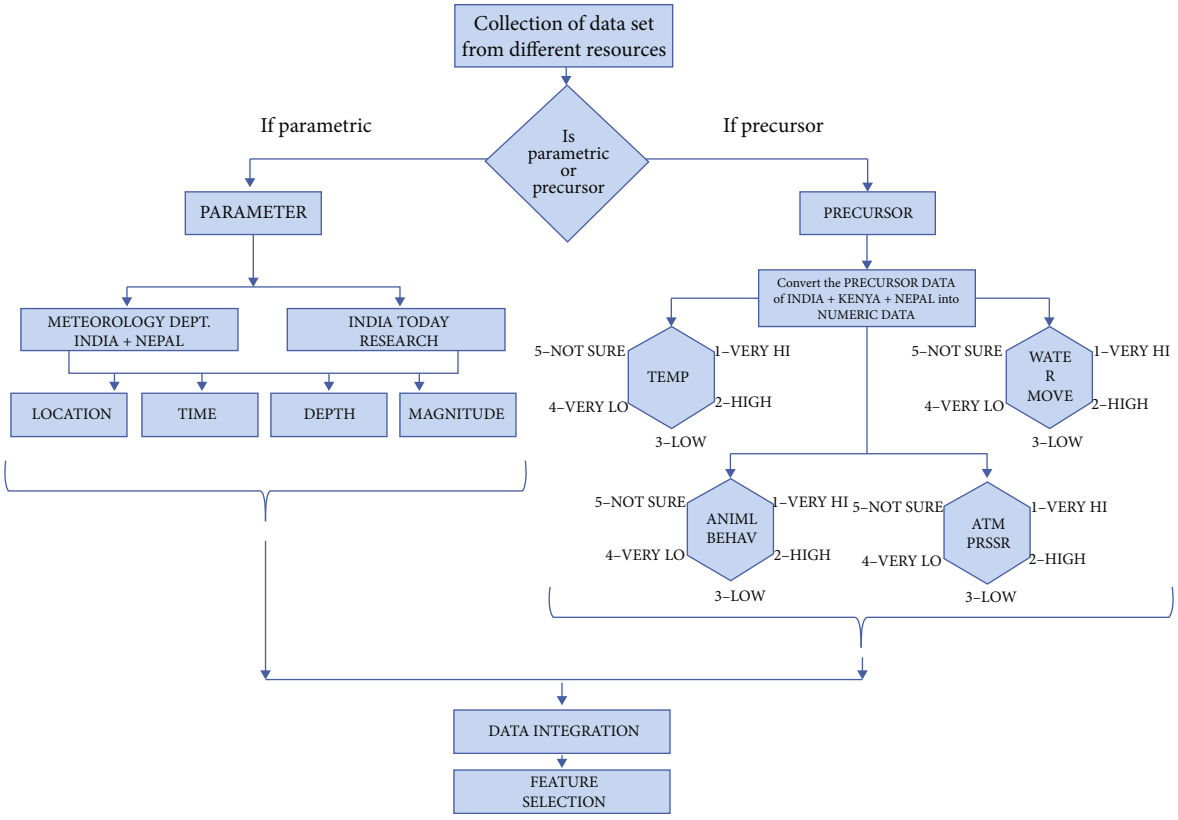


Figure 1: Flowchart of preprocessor data.

3.6. Feature Selection. After data integration, the attributes are selected and extracted for a new preprocessor data to develop, which is required for linear association strength measurement between two variables by applying Pearson coefficient. The feature selection of the preprocessor data associates independent and dependent attributes.

Some attribute features are not very significant for model construction and prediction, whereas they raise the dimensionality of the feature sets, which can be hard to analyse and require more time for training the data, thus making the decision complex. Lesser attributes give better accuracy; thus, less significant attributes must not be considered in the data set. The attributes which were selected are logically justified and explained in Table 8.

Pearson's correlation coefficient: here, the correlation is obtained by Pearson's correlation coefficient by measuring statistical and surveyor data for continuous variables using the method of covariance. The correlation relationships are defined below:

(i) It has built a significant relationship with numeric data and also encoded the character data to numeric for stronger relationship and correlation

Table 8: List of significant attributes.

Attributes	Significant attribute description	Type
Country	The data of three countries India, Nepal, and Kenya are combined derived from precursor data	Independent
Year	We have considered the data for the span of 5 years (2015-2019), and it has been derived from the period at which the earthquake occurred above mentioned countries	Independent
New_time_ist_ hour	The time at which earthquake occurred is obtained by converting the local time of Nepal and Kenya to Indian standard time and included data which already have Indian time	Independent
Day_period	The day was bucketed into 6 slices like the readings of the day are grouped into six categories, early morning (2 am to 5 am), morning (6 am to 9 am), afternoon (10 am to 12 pm) late afternoon (1 pm to 4 pm) evening (5 pm to 9 pm), and late evening (10 pm to 1 am)	Independent
Weekdays	The name of the specified week of the day is represented by the name like 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} week	Independent
Timestamps	It is considered the date and time of country I; here, it is Indian standard time and date	Independent
Local_time	Nepal is 15 minutes ahead, thus the conversion of local time of Nepal to India; Kenya local time is 2.5 hours behind India, thus conversion of Kenya time to India	Independent
Day	Whether the time of occurrences is during day or night time	Independent
Weekend	Whether the day of occurrence is weekend or not	Independent
Country_India	Out of three countries which country India's time is considered by conversion of other countries	Independent
Year_2015	Out of five different years, individual year 2015 based on the mean data of that specified year	Independent
Year_2016	Out of five different years, individual year 2016 based on the mean data of that specified year	Independent
Year_2017	Out of five different years, individual year 2017 based on the mean data of that specified year	Independent
Year_2018	Out of five different years, individual year 2018 based on the mean data of that specified year	Independent
Q1	Quarter is divided into four categories like summer, winter, autumn, or spring; it is Q1 of the year summer having high temp	Independent
Q2	Quarter is divided into four categories like summer, winter, autumn, or spring; it is Q2 of the year winter having low temp	Independent
Q3	Quarter is divided into four categories like summer, winter, autumn, or spring; it is Q3 of the year autumn having moderate temp	Independent

(ii) Higher correlation coefficient and attributes are strongly correlated, and one of them can be discarded

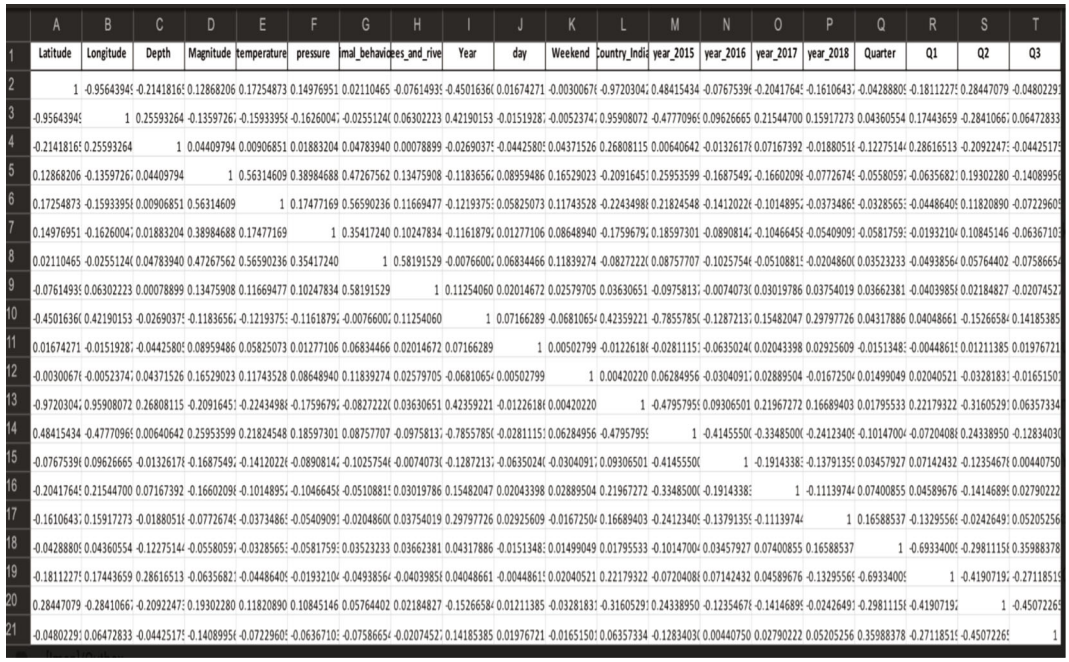
(iii) If the correlation constant is 0, then the attributes are independent, and if it is negative, then one attribute discourages the other; i.e., if the value of one attribute increases, then the value of the other decreases

(iv) The correlation coefficient is measured on a scale that varies between "+1" and "-1"

(v) When one variable increases and the other variable also increases, the correlation is positive; when one decreases and the other increases, it is negative. Complete absence of correlation is represented by 0. The values of all correlation attributes are shown in Figure 2

4. Implementing Existing Techniques for Earthquake Prediction Using Regression

Techniques like SVM, KNN, XGBoost, decision tree, and random forest based on supervised machine learning are applied on the data set to generate the results. Mathematical equations for supervised learning techniques are shown in Table 9. The most satisfactory result is generated by random forest technique as the random forest splits the nodes of preprocessed data and then selects the split which results in homogeneous subnodes. In KNN, a relationship of pressure with animal behaviour is observed; in most of the cases, time has more or less range. In SVM depth with cross-validation, although the time monitored is different, i.e., morning and evening, but there is a relation between falling of leaves and temperature. Applying XGBoost gradient and values shows that the relationship between the atmospheric pressure and temperature is within a range of closer depth of the occurrence magnitude and the different ranges of depth. The findings of decision tree are constructed suggesting that there is a strong relationship with the possibility of occurrence of earthquake having ranges of magnitude with temperature and animal behaviour. The creation of subnodes like temperature, atmospheric pressure, location, and magnitude in random forest shows the relationship with falling of leaves of the tree, water movement in water bodies, pressure, and temperature having specific ranges of magnitude. The graphs generated by all five techniques showed more and less similar results, i.e., attributes among all five techniques which are compared and the factors which are resulting a closer proximity also reflecting the accuracy of findings of the individual existing technique. Once the mapping is done with regression, using linearly on preprocessor data generating a relationship and correlations on attributes



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Latitude	Longitude	Depth	Magnitude	temperature	pressure	imal_behavi	ees_and_rive	Year	day	Weekend	Country_Indi	year_2015	year_2016	year_2017	year_2018	Quarter	Q1	Q2	Q3
2	1	-0.9564394	-0.2141816	0.12868206	0.17254873	0.14976951	0.02110465	-0.0761493	-0.4501636	0.01674271	-0.0030067	-0.9720304	0.48415434	-0.0767539	-0.2041764	-0.1610643	-0.0428880	-0.1811227	0.28447079	-0.0480229
3	-0.9564394	1	0.25593264	-0.1359726	-0.1593395	-0.1626004	-0.0255124	0.06302223	0.42190153	-0.0151928	-0.0052374	0.95908072	-0.4777096	0.09626665	0.21544700	0.15917273	0.04360554	0.17443659	-0.2841066	0.06472833
4	-0.2141816	0.25593264	1	0.04409794	0.00906851	0.01883204	0.04783940	0.00078899	-0.0269037	-0.0442580	0.04371526	0.26808115	0.00640642	-0.0132617	0.07167392	-0.0188051	-0.1227514	0.28616513	-0.2092247	-0.0442517
5	0.12868206	-0.1359726	0.04409794	1	0.56314609	0.38984688	0.47267562	0.13475908	-0.1183656	0.08959486	0.16529023	-0.2091645	0.25953599	-0.1687549	-0.1660209	-0.0772674	-0.0558059	-0.0635682	0.19302280	-0.1408995
6	0.17254873	-0.1593395	0.00906851	0.56314609	1	0.17477169	0.56590236	0.11669477	-0.1219375	0.05825073	0.11743528	-0.2243498	0.21824548	-0.1412022	-0.1014895	-0.0373486	-0.0328565	-0.0448640	0.11820890	-0.0722960
7	0.14976951	-0.1626004	0.01883204	0.38984688	0.17477169	1	0.35417240	0.10247834	-0.1161879	0.01277106	0.08648940	-0.1759679	0.18597301	-0.0890814	-0.1046645	-0.0540909	-0.0581759	-0.0193210	0.10845146	-0.0636710
8	0.02110465	-0.0255124	0.04783940	0.47267562	0.56590236	0.35417240	1	0.58191529	-0.0076600	0.06834466	0.11839274	-0.0827222	0.08757707	-0.1025754	-0.0510881	-0.0204860	0.03523233	-0.0493856	0.05764402	-0.0758665
9	-0.0761493	0.06302223	0.00078899	0.13475908	0.11669477	0.10247834	0.58191529	1	0.11254060	0.02014672	0.02579705	0.03630651	-0.0975813	-0.0074073	0.03019786	0.03754019	0.03662381	-0.0403985	0.02184827	-0.0207452
10	-0.4501636	0.42190153	-0.0269037	-0.1183656	-0.1219375	-0.1161879	-0.0076600	0.11254060	1	0.07166289	-0.0681065	0.42359221	-0.7855785	-0.1287213	0.15482047	0.29797726	0.04317886	0.04048661	-0.1526658	0.14185385
11	0.01674271	-0.0151928	-0.0442580	0.08959486	0.05825073	0.01277106	0.06834466	0.02014672	0.07166289	1	0.00502799	-0.0122618	-0.0281115	-0.0635024	0.02043398	0.02925609	-0.0151348	-0.0044861	0.01211385	0.01976721
12	-0.0030067	-0.0052374	0.04371526	0.16529023	0.11743528	0.08648940	0.11839274	0.02579705	-0.0681065	0.00502799	1	0.00420220	0.06284956	-0.0304091	0.02889504	-0.0167250	0.01499049	0.02040521	-0.0328183	-0.0165150
13	-0.9720304	0.95908072	0.26808115	-0.2091645	-0.2243498	-0.1759679	-0.0827222	0.03630651	0.42359221	-0.0122618	0.00420220	1	-0.4795795	0.09306501	0.21967272	0.16689403	0.01795533	0.22179322	-0.3160529	0.06357334
14	0.48415434	-0.4777096	0.00640642	0.25953599	0.21824548	0.18597301	0.08757707	-0.0975813	-0.7855785	-0.0281115	0.06284956	-0.4795795	1	-0.4145550	-0.3348500	-0.2412340	-0.1014700	-0.0720408	0.24338950	-0.1283403
15	-0.0767539	0.09626665	-0.0132617	-0.1687549	-0.1412022	-0.0890814	-0.1025754	-0.0074073	-0.1287213	-0.0635024	-0.0304091	0.09306501	-0.4145550	1	-0.1914338	-0.1379135	0.03457927	0.07142432	-0.1235467	0.00440750
16	-0.2041764	0.21544700	0.07167392	-0.1660209	-0.1014895	-0.1046645	-0.0510881	0.03019786	0.15482047	0.02043398	0.02889504	0.21967272	-0.3348500	-0.1914338	1	-0.1113974	0.07400855	0.04589676	-0.1414689	0.02790222
17	-0.1610643	0.15917273	-0.0188051	-0.0772674	-0.0373486	-0.0540909	-0.0204860	0.03754019	0.29797726	0.02925609	-0.0167250	0.16689403	-0.2412340	-0.1379135	-0.1113974	1	0.16588537	-0.1329556	-0.0242649	0.05205256
18	-0.0428880	0.04360554	-0.1227514	-0.0558059	-0.0328565	-0.0581759	0.03523233	0.03662381	0.04317886	-0.0151348	0.01499049	0.01795533	-0.1014700	0.03457927	0.07400855	0.16588537	1	-0.6933400	-0.2981115	0.35988378
19	-0.1811227	0.17443659	0.28616513	-0.0635682	-0.0448640	-0.0193210	-0.0493856	-0.0403985	0.04048661	-0.0044861	0.02040521	0.22179322	-0.0720408	0.07142432	0.04589676	-0.1329556	-0.6933400	1	-0.4190719	-0.2711851
20	0.28447079	-0.2841066	-0.2092247	0.19302280	0.11820890	0.10845146	0.05764402	0.02184827	-0.1526658	0.01211385	-0.0328183	-0.3160529	0.24338950	-0.1235467	-0.1414689	-0.0242649	-0.2981115	-0.4190719	1	-0.4507226
21	-0.0480229	0.06472833	-0.0442517	-0.1408995	-0.0722960	-0.0636710	-0.0758665	-0.0207452	0.14185385	0.01976721	-0.0165150	0.06357334	-0.1283403	0.00440750	0.02790222	0.05205256	0.35988378	-0.2711851	-0.4507226	1

Figure 2: Generated preprocessed data. In this we have provided snapshot of sample data table.

Table 9: Equations used in different techniques.

SVM	$F(x) = \sum n = 1\, N(a\, n - a\, n^*)\, \left(x\, n^{'}\, x\right)_+ b$
KNN	$\sqrt{\sum (x_i - y_i)^2}$
XGBoost	$F_0\,(X) = \text{argmin}\mu \sum (i-1)\text{^}n\{L(y\_(i))\mu\}$
Decision tree	$E(S) = \sum_{i-1}^{c} - p_i \log_2 p_i$
Random forest	$F(x) = \frac{1}{j} \sum_{j=1}^{j} c_{j\text{full}} + \sum_{k=1}^{k} \left(\frac{1}{j} \sum_{j=1}^{j} \text{contribution}_j(x,k) \right)$

in most cases, a consolidated magnitude range is found between 4.1 and 5.14, and the findings of implementation of existing techniques are shown in Table 10.

5. Novel Ensemble Earthquake Prediction Method

Many of the earlier research works on earthquake prediction are carried out on parameters only, i.e., on historical data, as well as many research works are done only on precursors, i.e., based on survey data only by implementing one or more data mining methods. Only a few research works are carried on combining both parameters and precursors as well as using multiple data mining techniques. The facts of limited research on combined data are motivated to generate a preprocessor data by combining parameters and precursors. The principle of combining data sets has been of interest for predicting the possibility of occurrence of earthquake using regression which is sure to produce better results. Although the result generated by individual technique has high variance and errors and restricted to only linear data set as discussed in Section 5, thus, it is motivated to develop an ensemble method by combining the individual method effectively so that the result on prediction not only has low variance and low errors but also has a better accuracy than earlier developed ensemble methods. This new ensemble method will be a precise, robust method by having less training errors and dimensionality as well as can be used on nonlinear data sets in the future. It will also operate on two visible components, one finds the average squared error of the individual models, and the other quantifies the prediction by interacting with individual predictions generated by individual techniques.

5.1. Work Flow of Novel Ensemble Earthquake Prediction Model (EEPM). The novel Ensemble Earthquake Prediction Method (EEPM) is discussed in a stepwise logical manner, which has been described in different steps like data collection, preprocessing of data, data splitting, implementation, and performance measure evaluation. The work flow of EEPM is shown in Figure 8.

5.2. Working of EEPM

5.2.1. Step 1: Data Collection. The parameters like location, date, time, magnitude, depth, temperature, longitude, and latitude are collected for two countries India and Nepal for five years (refer to Table 1). The precursors like location, date, time, temperature, atmospheric pressure, water movement, animal behaviour, falling of leaves, and rainfall are collected from people who actually experienced earthquakes within the age range from 18 to 75 from three countries like India, Nepal, and Kenya for five years shown in Table 2. The details of data collection are explained in Section 3.

Table 10: Implementation findings.

Technique	Findings
KNN	The feature similarity is generated by KNN algorithm application used to predict the values of any new data points, assigning a value that closely resembles the points in the training set. Thus, for low and damp pressure, animal behaviour is overactive, temperature is cold, falling of leaves is high, time is early morning, and depth can be different, having a magnitude between 4.1 and 5.14 shows a high possibility of occurrence of earthquake (refer to Figure 3).
SVM	Once the mapping is done by SVM with regression using linear magnitude and depth with cross-validation, although the time monitored is different, i.e., time is both morning and evening, for low pressure, animal behaviour is overactive, temperature is very cold, falling of leaves shows high, but the variation in depth is negligible, location is negligible, and falling of leaves having magnitude between 4.1 and 4.9 shows high possibility of occurrences of earthquake (refer to Figure 4).
XGBoost	Applying XGBoost gradient on preprocessor data generates relationship and correlation on attributes, and values show that even though the atmospheric pressure is damp and dry, temperature is very cold, time is morning, the falling of leaves is medium and high, and both dry and damp pressures show cold or very cold temperature within a range of closer depth and location having a magnitude between 4.1 and 5.1 which shows high possibility of occurrences of earthquake (refer to Figure 5).
Decision tree	The findings of the graphs by decision tree are constructed suggesting that there is a strong relationship on possibility of occurrences of earthquake having the ranges of magnitude ranging from 4.0 to 5.04, animal behaviour is more active than normal, atmospheric pressure is dry, temperature is either very cold or cooler than normal, falling of leaves is high, and depth variation is minimum, showing high possibility of occurrences of earthquake (refer to Figure 6).
Random Forest	The random forest splits the nodes of preprocessed data and then selects the split which results in homogeneous subnodes. The creation of subnodes like temperature, atmospheric pressure, and location magnitude shows increases in the homogeneity of resultant subnodes like longitude, latitude, and depth showing that leaves of the tree are falling, there is more water movement, more water bodies are high recorded on the earthquake prone locations, the range of magnitude is from 4.1 to 5.14, animal behaviour is more active, temperature is cold, time is early morning or late night, and variation in depth is negligible which shows high possibility of occurrences of earthquake (refer to Figure 7).

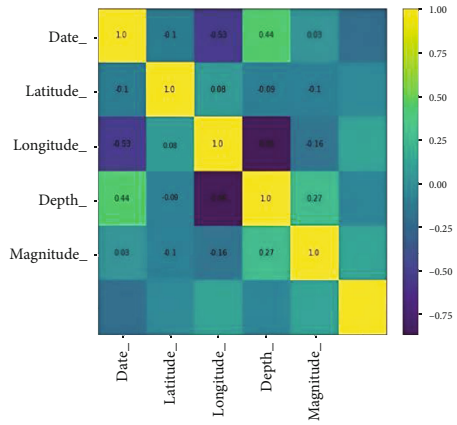


Figure 3: Result generated by KNN.

5.2.2. Step 2: Preprocessor Data. A preprocessor data is generated for accurate, consistent, and having completeness by combining both the parameters collected from two countries India and Nepal and the precursors of three countries India, Nepal, and Kenya which are used by different techniques shown in Figure 2 explained above in Section 4. In this step, the process of combining parameters and precursors will be elaborated and shown in Figure 1. Here, parameters X1 and Y1 and precursors P1 and Q1 are dependent data which are combined to get a preprocessed data set used by individual techniques.

5.2.3. Step 3: Splitting Data. Data set is split into two parts; training data and test data are the substrate for estimating parameters, comparing models, and all other activities required to reach a final algorithm. The splitting is performed on 70% of the training data, and 30% of the test set is used for estimating a final, unbiased assessment of the algorithm's performance.

5.2.4. Step 4: Framing of Ensemble Earthquake Prediction Model (EEPM). Boosting is defined by a combination of algorithms where weak learners are converted to a strong

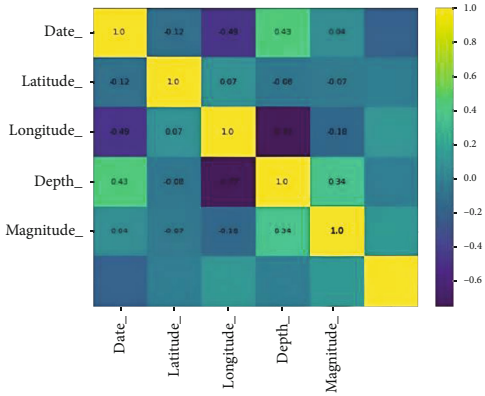


Figure 4: Result generated by SVM.

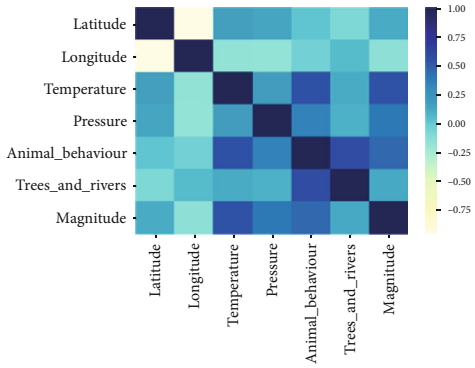


Figure 5: Results generated by XGBoost.

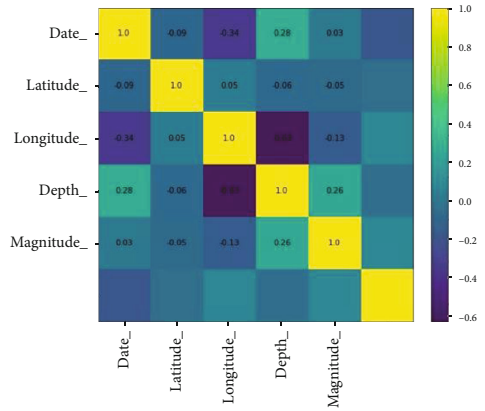


Figure 6: Results generated by decision tree.

learner after multiple iterations. The individual new model generates a strong learner with lower bias at the end of each process to focus its efforts on the most difficult observations to fit up for lower variance. The ensemble model is a weighted sum of L weak learners. The preprocessor data is initialized, and an equal weight to each of the data points is assigned. Boosting will keep on performing until the correct result is obtained. The steps are as follows (refer to Figure 9).

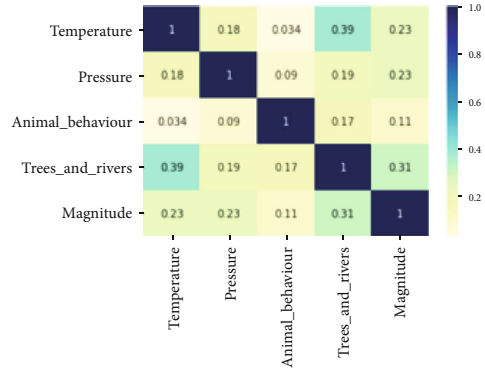


Figure 7: Result generated by random forest.

(1) (a) In the training data set, a base model is built to predict the observations in the record training set (P_i, Q_i), a differential loss function "L" which is used to identify the model with a constant value, and the number of iterations is used to find a "$\eth$" (predicted value) for which the loss function is minimum

$$F_n(P_i) = \operatorname{argmin} \sum_{i=1}^{n} L(Q_i, \eth_i) \tag{1}$$

(b) As the target column is continuous, the loss function follows the number of iterations with the loss function

$$\mathrm{L} = \frac{1}{n} \sum_{i=0}^{n} (Q_{i-}\eth_i)^2 \tag{2}$$

whereas "P_i" is the vector of input variable which is defined as $X_1 \cdots X_N$ and $P_1 \cdots P_N$; "Q_i" output variable or observed variable from $Y_1 \cdots Y_N$ and $Q_1 \cdots Q_N$; "L" loss function; "$\eth$ " predicted value; "argmin" argument of the minimum; "n" number of iterations; "m" number of DT (decision tree) made on iterations ($m = 1$ first DT $m = n$ means last DT); "i" number of records; "$F(P_i)$" the previous model; "r_{im}" pseudoresidual generated on DT; and "$h_m(P_i)$" DT made of residuals.

Initialize the probabilities of the distribution as$1/n$, where$n$is the number of data points, when$n$data points$F(P_i)$are looking for the function which produces the output almost equal toQ_i. But in real case scenarios, there is a difference between predicted output and actual output Q_i. This difference is called a residual $Q_{i-}\, Q_i\eth_i$. Now, in gradient boosting, another model is trained on the data points Q_i. The target variable is$Q_{i=1}$to$n$ or$Q_i - F_n(P_i)$, on training models$F_n$up to the final model. And the target variable for model is represented as $Q - [F_0 + F_1 + F_2 + \cdots \cdots \cdots \cdots + F_n + F_{n-1}]$.

An algorithm is fitted on the training data using the respective probabilities. The pseudoresidual is found r_{im} to fit a new model on the residual. For making a change in the model, the loss function is calculated.

The new model is added to the older model, and the next iteration is continued.

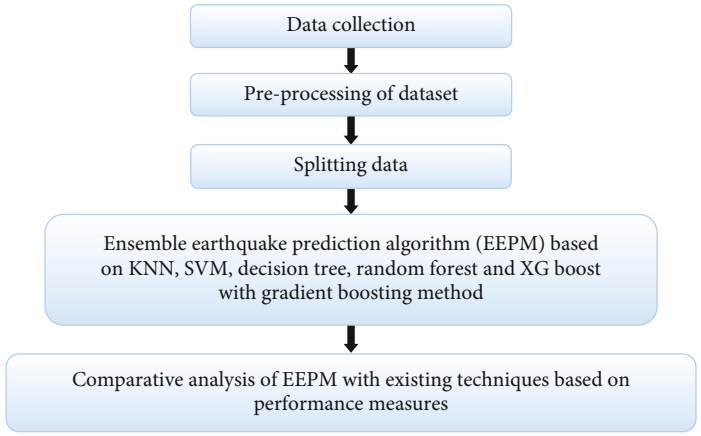


Figure 8: Work flow of novel Ensemble Earthquake Prediction Model.

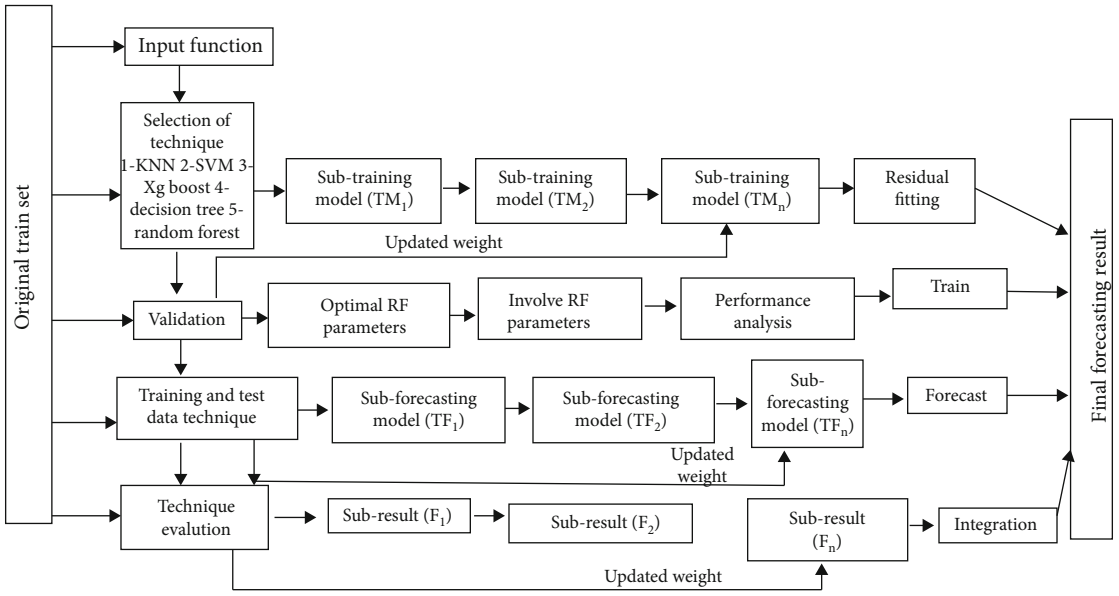


Figure 9: Working of the ensemble model.

(2) (a) The pseudoresiduals are calculated as the number of iteration records; all data points are on the same preprocessor data, and data are converted to numeric

$$r_{\text{im}} = -\left[\frac{\partial L(Q_i, F(P_i))}{\partial F(P_i)}\right] F_{(P)=}F_{(m-1)}P_i \quad \text{for } i = 1\text{--------}n \quad (3)$$

(b) The output value for each leaf of DT are calculated in terms of residuals by taking average numbers in a leaf, refer to the above table for variables

$$\eth_m = \text{argmin} \sum_{i=1}^{n} L(Q_i, F_{m-1}(P_i) + \eth h_m(P_i)) \text{ for } i = 1\text{--------}n \quad (4)$$

(3) Multiple "$\eth_m$" is computed by solving the following optimization problem

$$\eth_{m=}\text{argmin} \sum_{i=1}^{n} \text{L}(\, Q_i, F_N - 1 + Q_i P_i \eth \text{h}_m(\text{P}_i)) \text{ (refer to the above table for variables)} \quad (5)$$

(4) The model is framed

$$F_m\, P_i = F_{m-1} P_i + \eth_m, h_m(Pi) \text{ (refer to the above table for variables)} \quad (6)$$

(5) Output $F_m(P_i)$

The EEPM has resulted in the following findings that probability of occurrences of earthquake is more during morning and abnormal behaviour of animal and falling of

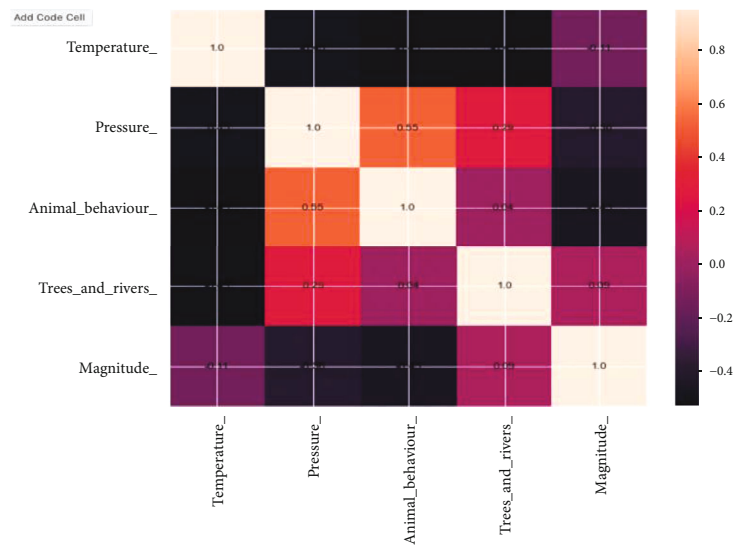


Figure 10: Generated by EEPM.

Table 11: Comparative analysis results.

Methods	R^2	Adjusted R^2	Variance	MSE	RMSE
SVM	0.64	0.58	0.24	0.36	0.60
KNN	0.76	0.62	0.26	0.39	0.62
XGBoost	0.74	0.59	0.31	0.33	0.55
Decision tree	0.80	0.75	0.28	0.32	0.56
Random Forest	0.82	0.76	0.27	0.26	0.50
EEPM	0.88	0.85	0.19	0.20	0.44

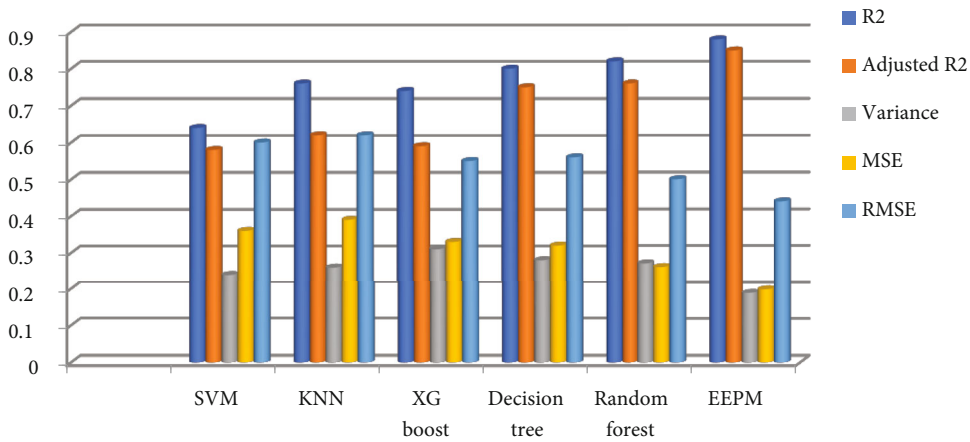


Figure 11: Different performance measures.

leaves are predominant with rise in water movement and mostly in cold temperature. oosting uses gradient algorithm by fitting many models on samples of preprocessed data set and involves fitting of many different techniques using another method and learns to best combine the predictions by different decision trees and random forest, if any extra tree or trees to be included. The training models can be represented by TM_1, TM_2 up to TM_n model which are constructed here by applying data mining techniques like linear regression mode, which is a comprehensive technique, by combining different algorithms, generating graphs, and establishing a relationship having forecasting model like TF_1, TF_2 up toTF_n (refer to Figure 10). The generalization ability of an ensemble is usually significantly better than that of a single learner, so

Table 12: Comparative analysis of EEPM with earlier ensemble methods.

Paper reference	Year	Method	Types of data set	Accuracy	Precision	MSE	Quality of prediction
[13]	2019	Ensemble boosting	Precursor based	81.9	0.72	0.36	Satisfactory
[14]	2020	Ensemble stacking	Parameter based	82.8	0.79	0.34	Satisfactory
[15]	2021	Ensemble stacking	Precursor based	84.3	0.81	0.30	Quite good
[16]	2021	Ensemble stacking	Parameter precursor based	85.3	0.83	0.28	Good
EEPM	2022	Ensemble boosting	Parameter precursor based	87.8	0.85	0.20	Best

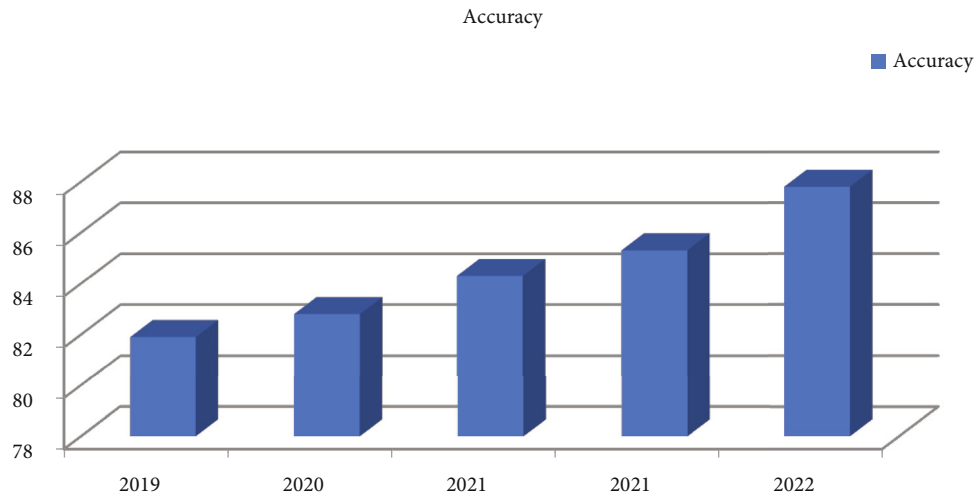


Figure 12: Accuracy analysis of EEPM with earlier ensemble methods.

ensemble methods are very attractive and definitely, the ensemble model is considered as shown in Figure 9.

5.2.5. Step 5: Comparative Analysis of EEPM with Different Individual Methods Based on Linear Regression. The comparative analysis of EEPM with existing techniques will be done based on performance measures like R^2, adjusted R^2, Mean Square Error (MSE), Root Mean Square Error (RMSE), and variance. In this paper, these above important measures on different individual methods which are derived using regression are compared to determine the correctness of the results as well as compared with the EEPM to understand the better result of the ensemble method in comparison to individual method. The value of performance measures R^2 in SVM is 0.64, KNN is 0.76, XGBoost is 0.74, decision tree is 0.80, random forest is 0.82, and EEPM which is maximum is 0.88. The different values of other measurements like adjusted R^2, variance, MSE, and RMSE are also calculated, respectively (refer to Table 11). The graphical analysis of the results is shown in Figure 11.

5.3. Comparative Analysis of EEPM with Earlier Ensemble Methods. Most of the previous ensemble methods are carried out using either regression or classification and performed either by stacking which ensembles with the best predictions from multiple well-performing machine learning methods. All previous ensemble methods are compared in Table 12 based on different measures. Here, EEPM has used boosting which transforms weak learning methods into strong ones and the accuracy of prediction is an important measure which is used to compare with other previous ensemble methods, and a conclusion is drawn (refer to Figure 12).

The individual technique does not show accuracy in greater extent in the pattern of data and relationship having high bias and high variance using training the data set and test data set, whereas the result in the ensemble method has considered patterns of data and relationship using training data set which has low bias errors and low variance. EEPM by integrating different methods through unique processed data (parameter and precursor) and by combining many weak learners generated one strong learner. The relationship of pressure and temperature with animal behaviour is more prominent, and occurrences of possibility of earthquakes is more in the morning at cold regions having specific a magnitude range between 4.1 and 5.8. The generated R^2 is 0.88, adjusted R^2 is 0.85, variance is 0.19, MSE is 0.20, and RMSE is 0.44 and having an accuracy of 87.8 by EEPM so it is concluded that EEPM can predict a better forecasting on earthquake-prone areas having better accuracy. It is also concluded that there must be a regular checking by seismologist, metrologies, or different institutions related to earthquake studies during morning, during fall in temperature, unusual animal movement, fall in pressure, and

unusual behaviour of trees and water bodies at earthquake-prone areas, which can give higher rate of possibility of forecasting earthquake.

6. Contribution of Work

The great threat of earthquakes in the area, earthquake-prone location, obviously is very important to develop an effective system of risk assessment, and prevention of negative effects of earthquakes is a necessity. In this respect, it should be said that forecasting of the natural disaster is really quite problematic, but still, the application of the ML methods with proper data set and then through ensemble gives an opportunity to predict the possibility of location of an earthquake to occur, having a rage of magnitude and observing some external features like unusual animal movement, falling of leaves, and rapid rise and fall in temperature and pressure though without a definition of the precise date and time of a disaster. Nevertheless, it is possible to forecast an earthquake within a few months, for instance, making people more prepared for the natural disaster taking into account all above mentioned; it is possible to conclude that earthquakes occur regularly in the IUB area and, therefore, they represent a serious threat to people living in the area. The EEPM can induce interest among seismologists and researchers to apply new technologies as well as other ML techniques and different ensemble methods using this as a base and get more accurate result.

The novelty of this research is the genuine data set recorded from the location of actual occurrences of earthquake. The data from people experiencing earthquake is also recorded, and then, a unique data set is prepared based on both data sets. This novel data set is applied on ML for individual prediction, and then, ensemble technique is applied on individual techniques to get a better and accurate prediction.

The limitations of the study are claims of breakthroughs have failed to withstand scrutiny on finding reliable precursors. Occurrences of earthquake are highly sensitive to unmeasurable fine details of the state of the earth in a large volume not just in the immediate vicinity of the hypocentre.

7. Conclusion and Future Scope

The conclusion of the research work states that there must be a regular checking by seismologists, metrologies, or different institutions related to earthquake studies during the morning, during fall in temperature, unusual animal movement, fall in pressure, and unusual behaviour of trees and water bodies at earthquake-prone areas, which can give a higher rate of possibility of forecasting earthquake.

The unique preprocessor data generated can be used by other data mining methods for regression as well as classification. This preprocessor data can be amended by adding one or more attributes for better results and prediction, which can be utilized by different data mining methods.

This novel EEPM is certainly going to enrich the researchers, seismologist, and metrological department to understand the application of different data mining methods individually as well as the power of the ensemble methods for better and accurate prediction. The importance of EEPM can be made stronger applying regression as well as classification by combining more individual data mining methods, more iteration, and multiple decision trees and can also be used by using nonlinear data. EEPM supports to use stacking and applying more individual methods in neural networks and other systems from its decentralized origin.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

I with gratitude like to pass my appreciation to my supervisor Dr. Prinima Gupta and cosupervisor Prof. (Dr) Felix Musua for their expert advice and encouragement. I would also like to thank Mr. Vobbani Venkateswarlu, Mr. Atharva Kulkarni, Mr. Nikhil Sahu, and Mr. Saikat Das for their honest support and cooperation.

References

[1] R. Agarwal, K. V. Arya, S. Shekhar, and R. Kumar, "An efficient weighted algorithm for web information retrieval system," in *In 2011 International Conference on Computational Intelligence and Communication Networks*, pp. 126–131, IEEE, 2011.

[2] K. Goda, T. Kiyota, R. M. Pokhrel et al., "The 2015 Gorkha Nepal earthquake: insights from earthquake damage survey," *Frontiers in Built Environment*, vol. 1, 2015.

[3] G. Martinelli, "Previous, current, and future trends in research into earthquake precursors in geofluids," *Geosciences*, vol. 10, pp. 189–3574, 2020.

[4] R. Kumar and A. Chaturvedi, "Improved cuckoo search with artificial bee colony for efficient load balancing in cloud computing environment," in *Smart Innovations in Communication and Computational Sciences*, pp. 123–131, Springer, Singapore, 2021.

[5] A. Morales-Esteban, G. Asecio-Cortes, F. Martínez-Álvarez, and J. Reyes, "A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction," *Knowledge Based Systems*, vol. 101, pp. 15–31, 2016.

[6] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2017.

[7] K. M. Asim, A. Idris, T. Iqbal, and F. Martínez-Álvarez, "Earthquake prediction model using support vector regressor and hybrid neural networks," *Plos One*, vol. 13, no. 7, p. e0199004, 2018.

[8] P. Bangar, D. Gupta, S. Gaikwad, B. Marekar, and J. Pati, "Earthquake prediction using machine learning algorithm," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 4684–4688, 2020.

[9] M. Kachakhidze, N. K. Murphy, and B. Khvitia, "Large earthquake prediction methods," *Open Journal of Earthquake Research*, vol. 8, no. 4, 2019.

[10] M. Kulkarni, C. Mulay, and S. Marathe, "Earthquake prediction using machine learning," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 7, no. 3, 2021.

[11] S. Bellamkonda, L. Settipalli, R. Vedantham, and M. K. Vemulad, "An enhanced earthquake prediction model using long short-term memory," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 4, pp. 2397–2403, 2021.

[12] K. Budiman and Y. N. Ifriza, "Analysis of earthquake forecasting using random forest," *Journal of soft computing exploration*, vol. 2, no. 2, pp. 153–162, 2021.

[13] M. H. Alobaidi, M. A. Meguid, and F. Chebana, "Predicting seismic-induced liquefaction through ensemble learning frameworks," *Scientific Reports*, vol. 9, article 11786, 2019.

[14] S. Cui, Y. Yin, D. Wang, Li, and Wang, "A stacking-based ensemble learning method for earthquake casualty prediction," *Applied Soft Computing*, vol. 101, article 107038, 2021.

[15] S. Arshia, H. Taghavizade, A. Bigdeli, and A. Shishegaran, "Predicting the earthquake magnitude along Zagros fault using time series and ensemble model," *Journal of Soft Computing in Civil Engineering*, vol. 3, p. 67, 2021.

[16] J. Zhang and J. Zhang, "An ensemble method to improve prediction of earthquake-induced soil liquefaction: a multi-dataset study," *Neural Computing and Applications issue*, vol. 33, no. 5, pp. 1533–1546, 2021.

[17] M. Hussein, "An automatic system for detecting voltage leaks in houses to save people's lives," *Science*, vol. 21, no. 3, p. 1485, 2021.

[18] S. G. Rupa and C. Iwendi, "ECDSA-based water bodies prediction from satellite images with UNet," *MDPI Journal*, vol. 14, no. 14, p. 2234, 2022.

[19] P. Kumar, "Design of anomaly-based intrusion detection system using fog computing for IoT network," *Automatic Control and Computer Sciences*, vol. 55, no. 2, pp. 137–147, 2021.