Engineered Science
DOI: https://dx.doi.org/10.30919/es1031

Predicting the Likelihood of an Earthquake by Leveraging Volumetric Statistical Data Through Machine Learning Techniques

Marat Nurtas,[1, 2,*] Zhumabek Zhantaev,[2] Aizhan Altaibek,[1, 2] Serik Nurakynov,[2] Nurbapa Mekebayev,[4] Kadrzhan Shiyapov,[3] Berik Iskakov[2] and Aizhan Ydyrys[1]

Abstract

This research paper presents an analysis of a dataset covering significant earthquakes over the past century, sourced from a publicly accessible seismic database. The dataset includes vital information such as the geographical coordinates, magnitudes, and depths of historical earthquake occurrences. The objective is to utilize machine learning techniques-specifically, k-nearest neighbors (KNN), support vector machines (SVM), random forests, and the XGBoost algorithm—to create predictive models that can anticipate future seismic events with magnitudes of 6 or higher. The models employ latitude, longitude, and depth as input parameters to define the spatial attributes of seismic activity, while the magnitude is used as the target output parameter, reflecting the event's strength and potential destructiveness. The research encompasses rigorous data preprocessing, including cleaning and feature scaling, followed by careful model training and validation through cross-validation methods to ensure the fidelity and robustness of the predictive models. Through iterative optimization, including hyperparameter tuning, feature selection, and performance assessment via suitable evaluation metrics, the models are continuously improved. The paper focuses on detailing these processes to demonstrate the methodology behind the development of machine learning models for earthquake prediction.

Keywords: Machine learning; Earthquake prediction; Seismic networks; Statistical analysis; Seismic activity.
Received: 09 October 2023; Revised: 16 November 2023; Accepted: 16 November 2023.
Article type: Research article.

1. Introduction

1.1 A brief review on earthquake and predictions

An earthquake[1] is a natural event where energy is released in the Earth's crust, causing the ground to shake. Predicting earthquakes accurately has been a long-standing goal in the field of seismology, and scientists have been working to develop effective forecasting methods to mitigate their impacts.

Throughout history, civilizations have tried to anticipate earthquakes by observing patterns or signs that could serve as precursors. However, these early attempts lacked scientific rigor and were often based on folklore or anecdotal evidence.

It wasn't until the 20th century that earthquake prediction began to emerge as a scientific discipline with the development of seismographs.[2] Seismographs, which can measure ground motion and record seismic waves, revolutionized the field of seismology. They allowed scientists to systematically monitor seismic activity and collect data for analysis. Various approaches have been explored to predict earthquakes, including monitoring changes in seismic waves, studying ground deformation, analyzing animal behavior, investigating electromagnetic signals and one of the least researched areas is forecasting based on historical statistical data. Hence, our research will concentrate on investigating this relatively unexplored approaches.

Advances in technology and data analysis techniques have

[1] *Department of mathematical and computer modeling, International IT University, 34/1 Manas Street, Almaty 050000, Kazakhstan.*
[2] *Institute of Ionosphere, Gardening community Ionosphere 117, Almaty 050020, Kazakhstan.*
[3] *Department of Mathematics and Mathematical Modeling, The National Pedagogical University named after Abai, 13 Dostyk Ave, Almaty 050010, Kazakhstan.*
[4] *Department of Computer Science, Kazakh National Women's Pedagogical University, 114 Gogol street, Almaty 050000, Kazakhstan.*
**Email:* m.nurtas@iitu.edu.kz; maratnurtas@gmail.com (M. Nurtas)

opened new possibilities for earthquake prediction based on statistical historical data, which we are going to explore in this article. Seismic networks,[3] with interconnected seismometers, can monitor seismic activity in real-time on a global scale. Combining this vast amount of data with sophisticated algorithms and machine learning techniques can help identify subtle patterns or precursors that may precede earthquakes.

In our research, our focal point lies in the realm of earthquake prediction within the highly seismic region where the Pacific plate converges with the Eurasian plate.[4] This geographical juncture has earned its reputation as one of the most earthquake-prone areas worldwide, making it an imperative area of study.

1.2 Literature review on seismicity detecting methodologies

In the following section, we will provide a concise introduction to traditional earthquake detection methods, and our primary emphasis will be on the utilization of machine learning and artificial intelligence in earthquake prediction. This discussion will center around the outcomes and insights derived from the integration of advanced technologies into seismology to enhance our understanding of seismic events and improve prediction accuracy.

1.2.1 Geodetic and Geophysical Monitoring

Geodetic and geophysical monitoring techniques[5] involve the continuous measurement of ground deformation, changes in gravity, and other physical parameters using instruments like GPS, strain meters, and tilt meters. These methods provide valuable information about the stress accumulation and release processes along fault lines. By analyzing these data, researchers have developed models that can estimate the likelihood of future earthquakes based on the greatest land movements, particularly in regions of high tectonic activity. Analysis presented in paper[5] shows that the greatest land movements in the region occur in spring, when average motions can be up to 1.5 m per month. It is demonstrated that integrated techniques provide a better means for monitoring landslide processes and gathering data for predictions of future movements.

Monitoring of the structures deformation and displacements of the earth's surface during landslides can be carried out using various types of systems and methods.[6] These methods and tools can be classified as remote sounding methods or satellite, photogrammetric and geodetic methods[7,8] The choice of instruments and measurement methods or the creation of a special monitoring system depends on various types of deformation, which will affect the method of stability analysis and, consequently, the entire deformation monitoring system.[9,10]

1.2.2 Seismicity-based models

Seismicity-based models[11] rely on historical earthquake data to identify patterns and trends that can be used to forecast future seismic activity. These models often consider factors such as earthquake magnitude, frequency, and spatial distribution.

While seismicity-based forecasts face constraints due to the limited time frame covered by instrumental data, notable progress has been achieved in their development and real-time implementation for earthquake forecasting. Some instances have seen the integration of seismicity-based forecasts into operational earthquake forecasting systems. It is worth noting that these methods have yielded probability improvements ranging from 2 to 5000 over time periods spanning from days to years.[12] Nevertheless, the absolute probabilities of seismic events remain relatively low.

Looking ahead, a promising trend emerges wherein time-dependent seismicity-based forecasts are anticipated to become standard practice in regions with moderate to high seismic hazard. These forecasts will offer continuous, real-time updates for local and regional seismic risk assessments.[13]

1.2.3 Precursor phenomena analysis

Researchers have explored the possibility of using precursor phenomena,[14] such as changes in electromagnetic fields, ground deformation, and foreshocks, as indicators of an impending earthquake. By monitoring these precursors, statistical analysis and machine learning techniques can be employed to detect patterns that precede seismic events. While these methods have shown promise, their practical application for accurate short-term earthquake prediction remains a topic of ongoing research.

Significant progress has been made in monitoring earthquake (EQ) anomalies[14] within the atmosphere and ionosphere, thanks to the Global Navigation Satellite System (GNSS), Detection of Electro-Magnetic Emissions Transmitted from Earthquake Regions (DEMETER), and other remote-sensing satellites. These advancements are grounded in the hypothesis of lithosphere-atmosphere-ionosphere coupling (LAIC). Traditional seismic monitoring tools, like seismometers and strong-motion accelerographs, complement these satellite observations by providing valuable data on the world's most active seismic regions. Both short-term and long-term patterns of EQ precursors, as detected by satellites, offer valuable insights into seismic activity.

Recent developments in space-based atmospheric and ionospheric measurements[15] have provided compelling evidence that supports the identification of anomalous patterns over seismic hotspots. These anomalies are believed to originate from stress-induced changes within the Earth's crust, propagating upward through the lithosphere and atmosphere to the ionosphere due to interactions between ions and molecules. This holistic approach to EQ monitoring, encompassing both ground-based and space-based data, enhances our understanding of seismic processes and enables more effective earthquake prediction and preparedness.

1.2.4 Machine learning approaches

Machine learning algorithms,[16,17] including K-Nearest Neighbors,[18] Support vector machines,[19] Random forests,[20] XGBossting[21] and *etc*.[22] have been applied to earthquake forecasting. These methods learn from historical seismic data, incorporating a wide range of features such as seismic waveforms, earthquake catalogs[23-24], and other geophysical parameters. By training on known earthquake occurrences, these models can make predictions about future seismic events with varying degrees of accuracy.

The inherent complexities of earthquake forecasting, including determining their size and precise location, have posed formidable challenges. However, recent years have witnessed significant progress in this field, with machine learning (ML)-based approaches emerging as promising tools for earthquake prediction. These studies primarily focus on predicting earthquake magnitude, trends, and occurrences. The analysis encompasses various ML algorithms, with a particular emphasis on assessing their performance against different seismic indicators. The findings offer valuable insights into the efficacy of these ML algorithms, shedding light on the seismic indicators that exhibit the best performance. Notably, this ML studies highlights the highest-performing ML algorithm for earthquake magnitude prediction, potentially paving the way for future research in this domain.

By monitoring these precursors, conducting statistical analysis, and utilizing machine learning techniques[25,26], scientists can identify patterns that precede seismic events. Although these methods show promise, accurately predicting earthquakes in the short term is still an ongoing research topic.

1.3 Challenges and Problem statments

One of the main challenges associated with earthquake prediction is timing. Predicting exactly when an earthquake will occur is extremely difficult, even with our best theories and models. Another challenge is the complexity of the factors that contribute to earthquakes. Earthquakes are caused by a combination of crustal movements, plate movements, and collisions, and understanding the underlying patterns of these phenomena is a challenging multidisciplinary field of investigation.[27]

We have detailed in the above sections that data-driven solution have the potential to help overcome these challenges by providing new insights into the factors that contribute to earthquakes. By analyzing large amounts of data, it is possible to identify patterns and correlations that may take time to be apparent through traditional methods. This can help us better understand the underlying causes of earthquakes and improve our ability to predict them.

2. Data preparation

2.1 Data availability and data quality

We analyze extensive seismological data, including information about past earthquakes, historical trends and places of increased seismic activity. Earthquake data have been collected over a hundred years, from 1923 to 2023, allowing for the investigation of trends and changes over time. The dataset includes earthquakes of diverse magnitudes and types, enabling the analysis of various earthquake scenarios and their consequences and *etc*. The data's availability and reliability are ensured by the U.S. Geological Survey (USGS), an organization specialized in earthquake monitoring and the "significant-earthquake-dataset-1900-2023" data, which guarantees the quality and reliability of the data. The datasets provided by our research contain a variety of earthquake-related parameters, including magnitude, depth, epicenter coordinates, time of occurrence, and other metadata. These parameters allow us to study various aspects of earthquakes and explore the connections between them.

2.2 Feature selection

We aim to identify patterns and connections between various factors that may portend earthquakes and choosing a subset of relevant features or variables from a larger set of available features to build a predictive model. In this study include longitude, latitude, magnitude, depth and time attributes.
The importance of using longitude, latitude, magnitude, depth, and time attributes:

The longitude and latitude traits are the geographic coordinates of an earthquake's epicenter. They play a key role in predicting earthquakes because location is one of the main factors affecting seismic activity. By analyzing these attributes, the model can detect spatial patterns and dependencies between earthquakes of different magnitudes.

The magnitude trait is a numerical value of an earthquake's magnitude that measures its energy and destructiveness. This trait is an important factor for modeling and predicting earthquakes because it allows us to estimate the potential level of hazard and loss associated with earthquakes.

The depth trait in earthquake prediction is a critical factor in seismic research. Shallow earthquakes tend to cause more surface damage than deeper ones. Additionally, understanding the depth assists seismologists in uncovering the earthquake's tectonic origin and associated geological processes, enhancing prediction models and deepening our understanding of seismic activity's root causes.

The time trait reflects the date and time of the earthquake. This attribute can be useful for analyzing time patterns and trends in seismic activity. It can also be used to predict future earthquakes based on historical data. Our primary research goal is to forecast future earthquakes using historical data (In our case, the historical data is latitude, longitude, magnitude, depth and time), as previously indicated.

2.3 Dataset preparation and model selection

Dataset contains information about earthquakes that occurred during a specific time period. Each sample includes the following attributes:
"id": the unique identifier of the earthquake.

"time": date and time of the earthquake in UTC format.
"latitude": the geographic latitude of the earthquake's epicenter.
"longitude": the geographic longitude of the earthquake's epicenter.
"depth": the depth of the earthquake in kilometres.
"mag": the magnitude of the earthquake.
"place": the location of the earthquake (the name of the region or country).
"type": type of earthquake (*e.g.* "earthquake", "explosion", "quarry blast", *etc*.).
The list continues in this way and consists of 22 basic parameters. We have divided the calculation process into two large groups, in the initial case we only need the very necessary different 5 parameters. The remaining parameters depend on our further work.

Data from USGS Earthquake Search API were pre-processed and reduced to a usable format. Some additional transformations, such as date and time format conversion, were applied to ensure compatibility (Table 1).

The "Significant Earthquake Dataset from 1900-2023" is an extensive compilation of data regarding significant earthquakes that have taken place globally over the last 123 years. The dataset is continually revised and managed by the National Earthquake Information Center (NEIC), part of the U.S. Geological Survey (USGS). Its purpose is to offer precise and current information on earthquake events. With over 37,000 entries, this dataset also provides comprehensive information on each earthquake, including the date, time, location, magnitude, and depth of the seismic activity.

We have collected all recorded earthquakes of the Eurasian continent from 1900 to May 3, 2023. As a result, there were 6997 recorded cases with a magnitude from 2.9 to 8.02.

However, there is a different time interval between each data point, since only those days when earthquakes occurred are recorded in our dataset. To set an equal time interval between each record, we need to resample the data to have a record of each day, even if no earth activity occurred on that day. Resampling is the process of changing the frequency or time intervals of a time series. This can be useful when working with time series data that has irregular time intervals or when the frequency of the data needs to be changed to suit the analysis. Our research employs two primary resampling techniques commonly used in time series analysis: upsampling and downsampling.[28]

Our dataset, edited and reprocessed from the two main data sources mentioned above, covers earthquakes greater than magnitude 6 between 1923 and 2023 in the following area. Our research focus and the historical database we aim to compile encompass the regions highlighted in the Fig. 1. Specifically, we are concentrating on earthquake prediction in the seismically active zone where the Pacific Plate intersects with the Eurasian Plate. This geographic location is renowned as one of the most earthquake-prone areas globally.

Before starting to build a machine learning model, we need to prepare the dataset for further training of the model. Fig. 2 illustrates the historical earthquake activity in the specified region from 1900 to the present, focusing on events with magnitudes exceeding 1. This graphic(data) aids our aim to forecast future earthquakes in this area, specifically those reaching or surpassing a magnitude of 6, by analyzing this extensive historical data.

The dataset sourced from the "USGS database" and the "significant-earthquake-dataset-1900-2023" contains certain parameters that are non-essential for model training and can be omitted, if desired. Moreover, some parameters within the dataset exhibit a high percentage of missing values, exceeding 50%. Additionally, earthquakes resulting from human activities, such as nuclear weapons testing and mining, were included, but we have opted to exclude such data as our study exclusively centers on natural earthquakes.

Figure 3 features two graphs: the left graph displays all 22 main parameters considered in our study, while the right graph focuses on the 5 primary parameters chosen for detailed analysis. The right graph also showcases the cleaned and

Table 1. A real-time view of the USGS Database.

	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	...	updated	place	type	horizontalError
6993	1907-09-15 17:45:41.560000+00:00	39.310	71.614	15.0	6.27	mw	NaN	NaN	NaN	NaN	...	2022-04-25T20:37:03.805Z	11 km NE of Karakenja, Tajikistan	earthquake	NaN
6994	1907-08-21 05:11:17.580000+00:00	42.775	48.965	15.0	5.51	mw	NaN	NaN	NaN	NaN	...	2022-04-25T20:41:31.577Z	90 km ENE of Novokayakent, Russia	earthquake	NaN
6995	1906-12-22 18:21:12.800000+00:00	43.988	84.930	15.0	7.95	mw	NaN	NaN	NaN	NaN	...	2022-04-25T20:28:00.767Z	95 km WSW of Shihezi, China	earthquake	NaN
6996	1906-11-12 17:32:23.250000+00:00	41.635	81.979	15.0	6.48	mw	NaN	NaN	NaN	NaN	...	2022-04-25T20:40:56.031Z	79 km W of Kuqa, China	earthquake	NaN
6997	1906-03-02 06:14:55.730000+00:00	43.219	84.179	20.0	6.64	mw	NaN	NaN	NaN	NaN	...	2022-04-25T20:24:45.290Z	78 km ESE of Xinyuan, China	earthquake	NaN

5 rows × 22 columns

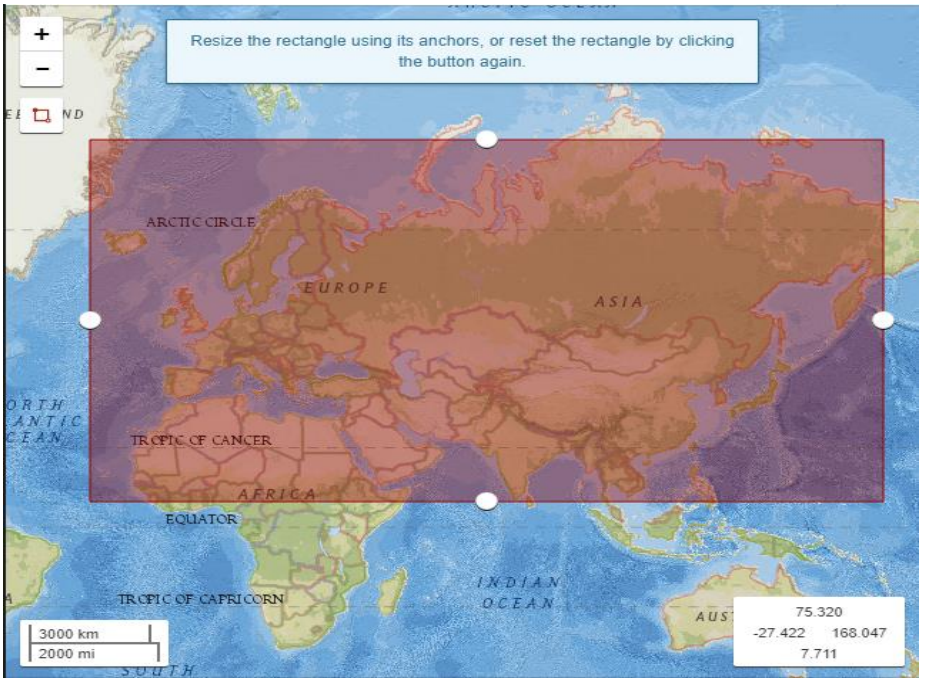

Fig. 1 The area covered by our study data is a zone of high seismicity where the Pacific plate meets the Eurasian plate.

filtered dataset, represented uniformly in a single color for clarity.

3. Methods and analysis

3.1 Evaluation of methods and analysis

Firstly, we use K-Nearest Neighbor (KNN) in our prediction model.[29] KNN is a non-parametric machine learning algorithm used for classification and regression tasks. Earthquake prediction would start with selecting the right set of features (variables). These could include previous seismic activity, geological data, and other relevant measurements. These features are critical as they will serve as the basis for determining "distance" between data points in the KNN algorithm. KNN algorithm does not create a model immediately from the training data. KNN just memorizes the data, but when it is the turn of prediction, it runs through all the data finding the nearest neighbors and based on these neighbors it makes a prediction. KNN finds the K (number of nearest points) labeled samples in the closest proximity to the point that is to be predicted. This K can be defined by the user. For a new observation (a particular situation where you want to predict the likelihood of an earthquake), KNN will look for the 'k' training examples that are closest to this observation. And the closest proximity, hence the distance measure, in our

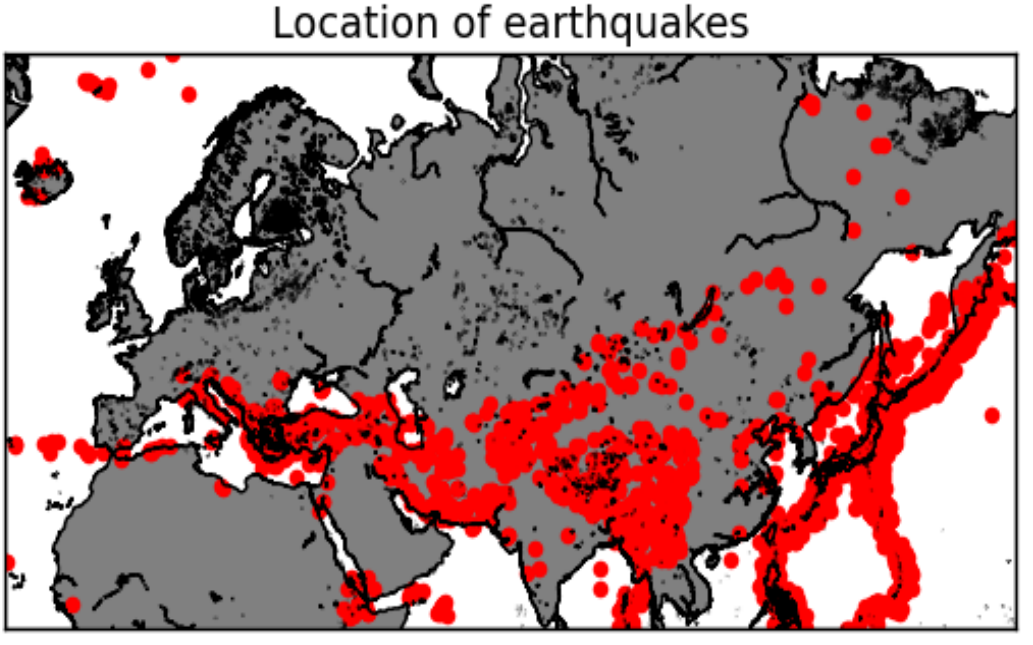

Fig. 2 Dataset for further training of the model.

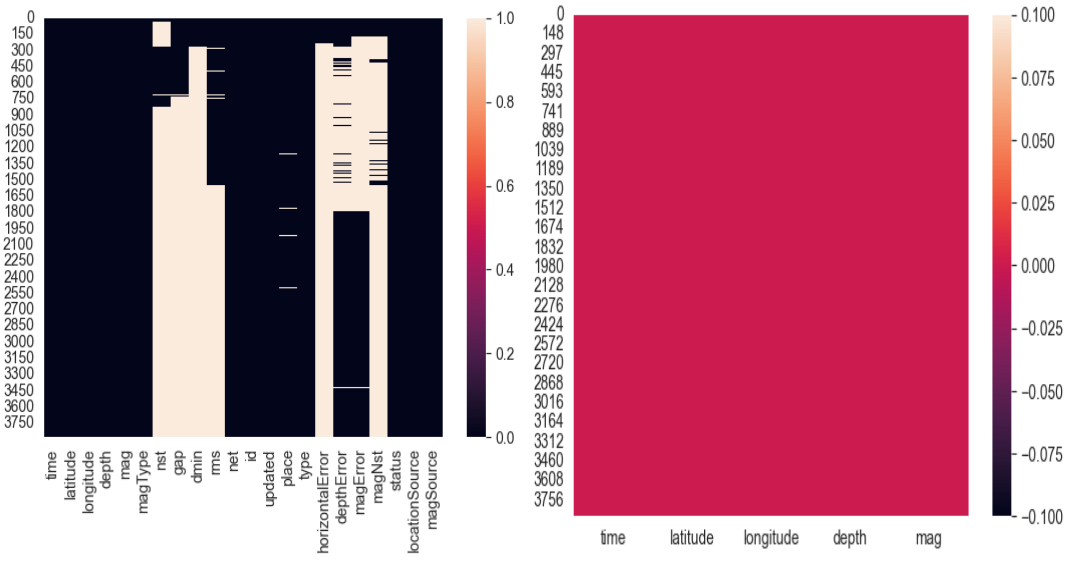

Fig. 3 Dataset before changes and after removing unnecessary parameters.

study can be calculated by Euclidean distance ($EuclideanDistance => sqrt\left(sum\left((x - y)\text{^}2\right)\right)$). To find these nearest points, indexing algorithms like 'kd-tree' and 'ball-tree' are used. A 'kd-tree' separates data in Cartesian axes and a 'ball-tree' in a nested hypersphere. When the number of dimensions is larger, the 'ball-tree' is more efficient than the 'kd-tree'. A simple approach to select k is set square root of number of total data points (length of dataset). For instance, in the "Significant Earthquake Dataset from 1900-2023" that we are examining, there are 294,846 recorded earthquakes. Based on our calculations, the optimal number of K is approximately 543.

The Support Vector Machines (SVM)[30] for earthquake prediction involves harnessing their ability to handle nonlinear patterns and classify complex datasets. We shortly discuss some practical facts when applying SVMs to earthquake prediction. Earthquake prediction models based on SVM require the selection of appropriate features that could be indicative of seismic activity. These features can include historical seismic data, such as magnitude, depth, location of past earthquakes. Earthquake data is inherently nonlinear and complex. SVMs can handle this nonlinearity effectively, especially when using the kernel trick. By choosing a suitable kernel function (like the Radial Basis Function or RBF), SVM can transform the input space into a higher-dimensional space where a linear separation is possible.

The research on earthquake prediction using deep learning[31-35] is one of the most comprehensive and growing areas in the field. However, we will not be exploring it in this discussion. A thorough investigation into deep learning, as a subject, necessitates a more expansive study on its own.

The Random Forests algorithm,[36] represent a class of ensemble machine learning techniques suitable for both classification and regression applications. When applied to the domain of seismology, the algorithm can be trained using historical seismic datasets with the objective of predicting future seismic occurrences or associated metrics such as earthquake magnitude. For regression-based applications, evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are commonly employed; lower values of these metrics denote superior model performance. Conversely, in classification scenarios, a model's efficacy is often gauged by its accuracy, with higher values being preferable. Additionally, a robust R-squared value is indicative of an effective model representation relative to the underlying data.

The XGBoost algorithm,[37] known for its high efficiency and capability to handle structured data, is particularly suited for both classification and regression problems. Its application in earthquake prediction involves processing extensive seismic data to identify potential precursors to tremors. By training on historical data, including quake magnitudes, locations, and geological features, XGBoost can discern intricate patterns that may indicate imminent seismic activity. Its robustness and scalability make it a valuable tool for our study seeking to enhance forecasting models and potentially extend early warning times. During the experiment we also analyze the importance of different features in the XGBoost model to gain insights into which factors are most predictive of earthquakes.

3.2 Model training

We possess a labeled dataset containing records of past seismic activities, detailing the times, locations, and magnitudes of these earthquakes. Our aim is to leverage this historical information to forecast future events. The KNN algorithm was trained on the two aforementioned combined and sorted datasets. Within our study's region, the valid dataset encompassed historical earthquake data, such as location (latitude, longitude), magnitude, depth, and time. The dataset was partitioned randomly into training and testing subsets, allocating 80% for training and the remaining 20% for

testing. Subsequently, the performance of the KNN algorithm was assessed using the Mean Squared Error (MSE) metric.

Figure 4 displays the performance of the K-Nearest Neighbors (KNN) algorithm, measured by the Train Score and Test Score for different K values up to 500. Both graphs show a decline in score as K increases, suggesting lower prediction accuracy with a higher number of neighbors.

The best K is: 15

Accuracy with this K: 0.7989561829588675

As discussed in Section 3.1, the outcomes from various algorithms used to determine the optimal K are as follows:

{'Default-Train': 0.7802614101842166,

'K-Best-Train': 0.7919845368046313,

'GridSearch-Best-Train': 0.8002999271639116,

'RandomSearch-Best-Train': 0.804140220252088}

The results from our experiments indicated that the optimal number of K-neighbors for accurate earthquake prediction using all tested algorithms was 15. This configuration achieved a Mean Squared Error (MSE) of approximately 80%. These findings underscore the efficacy of the KNN method in earthquake prediction and offer significant insights into fine-tuning parameters for more accurate forecasts.

Training an SVM model for earthquake prediction involves several steps, including the selection of relevant features and the tuning of model parameters. Once these features and parameters are determined, the model is trained on a dataset by finding the optimal hyperplane that minimizes predictive errors, based on the chosen loss function. The model's performance is evaluated using metrics such as R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The model training performance for our data is as follows:

R-squared (R2) Score: 0.5391776104068513

MAE Score: 0.5368172071764645

MSE Score: 0.5153235425121637

In interpreting these metrics for earthquake prediction using SVM, which is a complex problem with high inherent variability, an R^2 score of 0.54 might be reasonably good. This score indicates that the model can explain around 54% of the variability in the data, which, given the unpredictability of earthquakes, could be seen as a positive outcome. However, it's important to compare this R^2 value, along with the MAE and MSE scores, against other models or benchmarks in the same domain to gauge the relative performance of the SVM accurately. Contextualizing these metrics is crucial for a meaningful evaluation of the model's predictive capabilities.

To train a Random Forest model, we utilize the training dataset. It's essential to fine-tune hyperparameters, including the number of trees, maximum depth of the trees, and minimum samples per leaf, among others. One advantage of Random Forest is its ability to rank features based on their importance. This can provide insights into which features are most influential in predicting earthquakes. The values provided are metrics used to evaluate the performance of a regression model implemented with the Random Forest algorithm, with each metric representing a different aspect of model accuracy:

Mean Squared Error (MSE): 0.18452689889821955

Mean Absolute Error (MAE): 0.3112474504734689

Root Mean Squared Error (RMSE): 0.42956594243284646

R-squared (R^2): 0.8349888575244284

In Random Forest algorithm results, these metrics collectively suggest that our Random Forest model has a relatively good fit to the seismic data, explaining a significant portion of the variance in the target variable and with errors that are reasonable given the context (although without context, it's hard to assess the absolute goodness of these values).

Training a predictive model using XGBoost necessitates a careful specification of the objective function. In the study of earthquake prediction using XGBoost, which is inherently a regression task due to the continuous nature of the seismic magnitude scale, the mean squared error (MSE) serves as an apt evaluation metric. MSE quantifies the variance between the predicted seismic activities and the recorded data, thus offering a clear measure of the model's predictive accuracy.

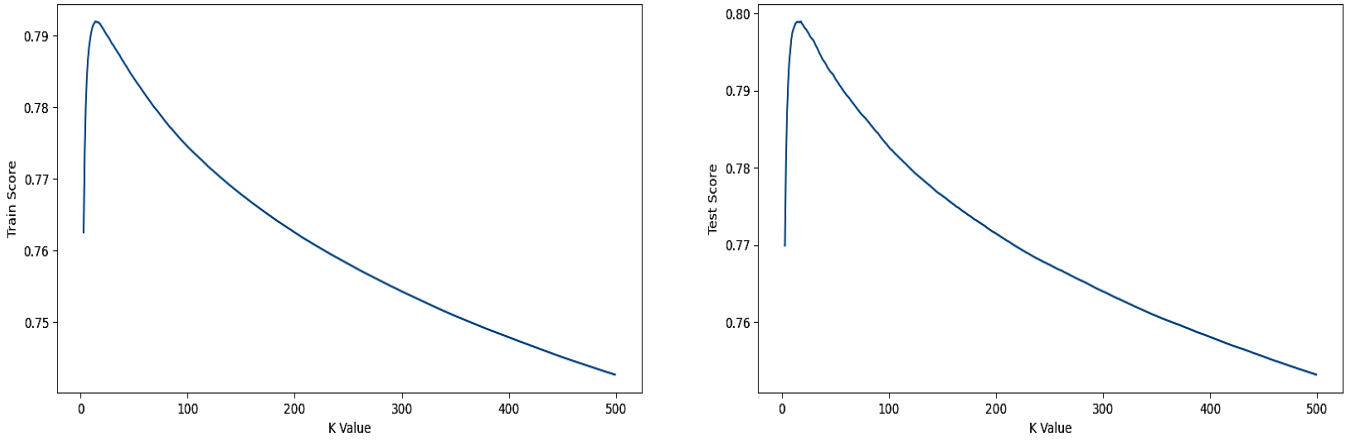

Fig. 4 The results from the KNN algorithm for determining the optimal number of neighbors were obtained for all K values up to 500.

Also using the test set and metrics suitable for regression tasks, such as R-squared, Mean Absolute Error (MAE), or Root Mean Squared Error (RMSE). The result of this algorithm based on our data is as follows:

GS.best_params_

{'gamma': 0.1, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 300}

GS.best_score_

0.8190620737633614

RS.best_params_

{'learning_rate': 0.31499832889131046, 'max_depth': 3, 'n_estimators': 44}

RS.best_score_

0.7851752427523718

The results we have provided to come from hyperparameter tuning of an XGBoost model, using two different search strategies: Grid Search (GS) and Random Search (RS). In our study, GS.best_params_ and RS.best_params_ show the best combination of hyperparameters found by Grid Search and Random Search, respectively, for an XGBoost model.

These results suggest that the XGBoost model with the hyperparameters found by Grid Search (higher best_score_) is expected to perform better than the model with hyperparameters found by Random Search on the given dataset for earthquake prediction.

4. Results and discussion

4.1 Indirect results

Our first side indirect result is to process and sort the two large datasets used to create a new dataset and graphically analyze it.

Despite the complexity of earthquake processes, the absence of consistent precursor events, variability in fault behavior, limitations in historical data, inadequate monitoring infrastructure, and the intricacies of stress accumulation and release, alongside human and economic considerations, predicting earthquakes remains a formidable challenge. However, overcoming these hurdles, we managed to derive significant insights from our analysis using appropriate methodologies.[38-41] One aspect of our research centers on the hypothesis that earthquakes may display periodic behavior, suggesting that seasonality could be discerned from existing data. To test this, we initially generated a new dataset and plotted a graph with the aim of detecting any discernible patterns or recurrent trends that might reveal seasonal characteristics.

Our approach is reasonable for exploring the potential periodicity of earthquakes in a specific geographical area. By dividing our collected dataset into equal segments of 20 years, we can create a structured framework for analyzing historical earthquake data.

We analyzed earthquake occurrences in 20-year segments over a total span of 100 years to identify consistent patterns or indications of periodicity. Observing a recurrence of earthquakes in a given area over this century-long period suggests the potential for this pattern to persist into the next 100 years. Such an observation supports the notion that seismic activity may exhibit cyclical patterns.

Mean Magnitude: 6.40

Median Magnitude: 6.29

Standard Deviation of Magnitude: 0.42

Figure 5 visually analyzes historical earthquakes with magnitudes above 6 in the designated area, covering a specific each 20-year intervals. This analysis aims to investigate the presence of any periodic patterns in earthquake occurrences within these years. In the graph, clear blue lines depict the average magnitudes, while the shaded areas illustrate the frequency range, encompassing all maximum and minimum magnitude events.

Our analysis suggests a tentative pattern of 100-year periodicity in earthquake occurrences, based on the current data and study methods. It's crucial to note, however, that seismic processes are complex and influenced by numerous geological and tectonic factors. Although our data may show signs of periodicity, we must approach the interpretation of these findings with caution. It is essential to consider that overarching geological and tectonic dynamics could significantly impact earthquake patterns over the long term.

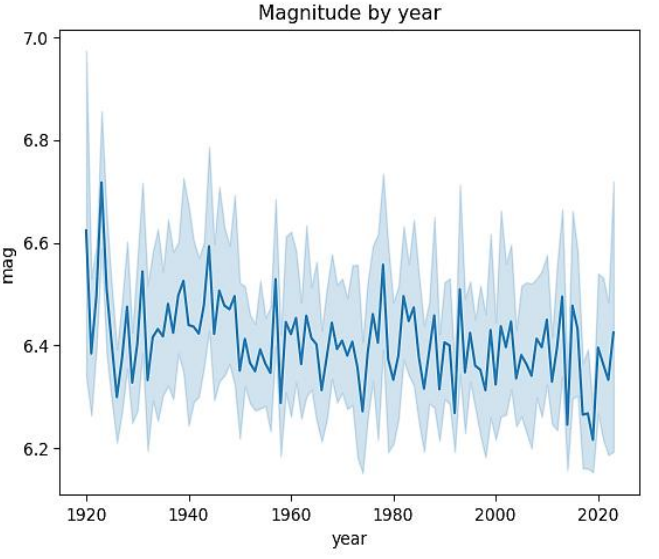

Fig. 5 Index of 100-year earthquakes in the study area over a 20-year interval.

4.2 Main results

One of the main results of our study is that to apply the machine learning models mentioned in the above sections to different regions and time intervals to discern potential long-term earthquake trends and enhance predictive. It may reveal patterns or trends in different regions and time periods that contribute to a better understanding of earthquake behavior on a larger scale. These insights can be valuable for seismic hazard assessment and preparedness efforts.

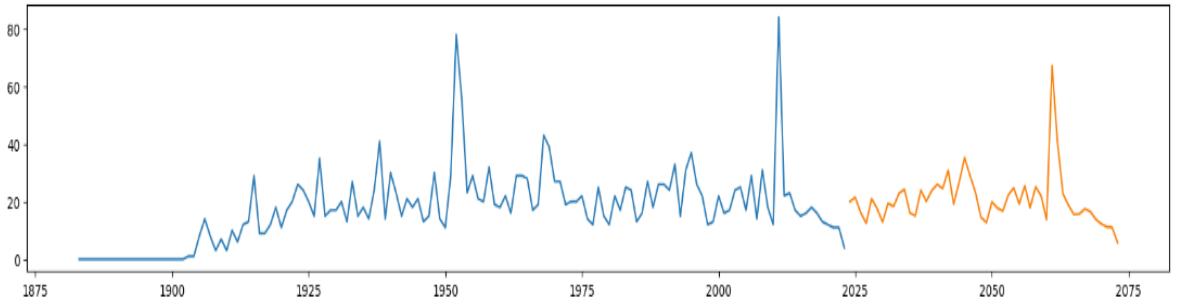

Fig. 6 6-point earthquake prediction by years and frequency of shaking.

Figure 6 displays two distinct lines: the blue line represents the historical data of earthquakes with a magnitude of exactly 6, while the red line illustrates the predicted occurrences of earthquakes potentially reaching a magnitude of 6. In this figure, the y-axis indicates the recurrence frequency of magnitude 6 earthquakes, and the x-axis denotes the corresponding years.

Based on our research using appropriate machine learning models and indirect result analysis for earthquake prediction, we found that although the results were not as accurate as we had hoped, we are confident that with a larger and more diverse dataset, it will be possible to achieve more precise predictions. This is because the graphical representation of the result obtained above shows that there is at least seasonality between the predicted future earthquakes and the past historical earthquakes.

We processed the data qualitatively and applied it to various machine learning techniques, which demonstrated adaptability to earthquake data through the accuracy metrics achieved. Analyzing our most accurate model, we identified areas with a probable future earthquake exceeding magnitude 6 in the region under study. The predictions are illustrated in the subsequent graph.

Figure 7 depicts the forecasted probability of earthquakes occurring in the next 20 years within our research area, specifically focusing on events predicted to have magnitudes of 6 or higher. In this figure, the red crosses mark the locations where the likelihood of such earthquakes is highest.

To verify the accuracy of our results, we cross-referenced them with the outcomes presented in Graph 6. This comparison was also conducted in the section on indirect results, where we juxtaposed our findings with the centennial seasonality theory.

As previously stated, accurately predicting earthquakes remains an elusive challenge in science. Hence, the results we present do not guarantee that earthquakes will recur in the same areas. What we have accomplished is formulating predictions by effectively employing models and appropriately analyzing data.

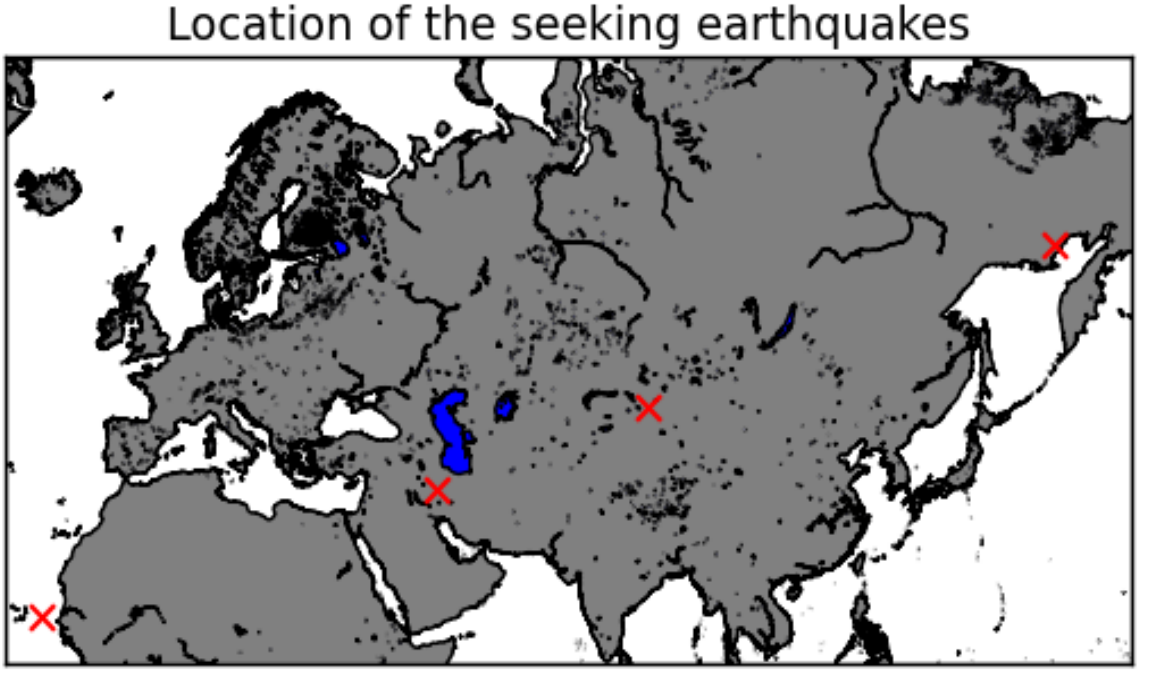

Fig. 7 Prediction of earthquakes for the next 20 years with a magnitude of at least 6. Red crosses - actual forecast focuses.

5. Conclusion

Our study has conducted an extensive analysis of over a century's worth of historical statistical data from a targeted region to forecast potential future earthquakes using various machine learning methods. However, our findings reveal that no single model distinctly outperforms others due to the complex and unpredictable nature of seismic activity; none of the models we used achieved an accuracy greater than 80%.

The success of these computational models is highly dependent on the availability of comprehensive, high-quality datasets that include a wide range of geological settings and cover substantial time periods. The models are designed more to calculate probabilities and identify patterns that may indicate risk, rather than to make exact predictions.

Our research highlights the importance of tailoring models to specific tasks, such as predicting the location, magnitude, or timing of an earthquake, as well as the need for precise performance metrics. A key observation is that machine learning in seismology is most effective when it involves a combination of different algorithmic strategies, supported by extensive geological knowledge.

The study emphasizes the importance of understanding earthquake patterns and trends to enhance seismic hazard assessment and preparedness. We advise readers and fellow researchers to consider our findings as predictive insights that require cautious interpretation and continued refinement through ongoing research.

Looking forward, our research will advocate for a synergistic approach that merges machine learning algorithms with geological analysis.[42-45] This collaborative effort aims to improve the accuracy of earthquake risk assessments and mitigate the impacts of potential mispredictions.

Acknowledgements

This work was performed under the grant №AP09058367 of the Committee of Science of the Ministry of Science and Higher Education of Kazakhstan (2021-2023). Data Availability: This study is based upon earthquake catalogs. The authors downloaded catalogs from United States Geological Survey (USGS, https://earthquake.usgs.gov/earthquakes/search/).

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable.

References

[1] F. Magrini, D. Jozinovi, F. Cammarano, A. Michelini, L. Boschi, Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale, *Artificial Intelligence in Geosciences*, 2020, **1**, 1-10, doi: 10.1016/j.aiig.2020.04.001.

[2] K. Li, S. Chen, G. Hu, Seismic labeled data expansion using variational autoencoders, *Artificial Intelligence in Geosciences*, 2020, **1**, 24-30, doi: 10.1016/j.aiig.2020.12.002.

[3] T. A. Stabile, V. Serlenga, C. Satriano, M. Romanelli, E. Gueguen, M. R. Gallipoli, E. Ripepi, J. Saurel, S. Panebianco, J. Bellanova, The INSIEME seismic network: a research infrastructure for studying induced seismicity in the High Agri Valley (southern Italy), *Earth System Science Data*, 2020, **12**, 519-538, doi: 10.5194/essd-12-519-2020.

[4] K. W. Wegmann, S. F. Gallen, Tectonic Geomorphology Above Mediterranean Subduction Zones, Treatise on Geomorphology (Second Edition), 2022, **2**, 87-119, doi: 10.1016/B978-0-12-818234-5.00223-6.

[5] M. Zeybek, İ. Şanlıoğlu, A. Özdemir, Monitoring landslides with geophysical and geodetic observations, *Environmental Earth Sciences*, 2015, **74**, 6247-6263, doi: 10.1007/s12665-015-4650-x.

[6] V. Zaalishvili, K. Chotchaev, D. Melkov, O. Burdzieva, B. Dzeranov, A. Kanukov, I. Archireeva, A. Gabaraev, L. Dzobelova, Geodetic, geophysical and geographical methods in landslide investigation: Luar case study, *E3S Web of Conferences*, 2020, **164**, 01014, doi: 10.1051/e3sconf/202016401014.

[7] A. Bichler, P. Bobrowsky, M. Best, M. Douma, J. Hunter, T. Calvert, R. Burns, Three-dimensional mapping of a landslide using a multi-geophysical approach: the Quesnel Forks landslide, *Landslides*, 2004, **1**, 29-40, doi: 10.1007/s10346-003-0008-7.

[8] E. Brückl, F. K. Brunner, K. Kraus, Kinematics of a deep-seated landslide derived from photogrammetric, GPS and geophysical data, *Engineering Geology*, 2006, **88**, 149-159, doi: 10.1016/j.enggeo.2006.09.004.

[9] F. V. De Blasio, A. Elverhøi, D. Issler, C. B. Harbitz, P. Bryn, R. Lien, On the dynamics of subaqueous clay rich gravity mass flows—the giant Storegga slide, Norway, *Marine and Petroleum Geology*, 2005, **22**, 179-186, doi: 10.1016/j.marpetgeo.2004.10.014.

[10] J.-P. Malet, O. Maquaire, E. Calais, The use of Global Positioning System techniques for the continuous monitoring of landslides: application to the Super-Sauze earthflow (Alpes-de-Haute-Provence, France), *Geomorphology*, 2002, **43**, 33-54, doi: 10.1016/s0169-555x(01)00098-8.

[11] K. F. Tiampo, R. Shcherbakov, P. Kovacs, Probability Gain From Seismicity-Based Earthquake Models, Risk Modeling for Hazards and Disasters. Elsevier, 2018, **7**, 175-192, doi: 10.1016/B978-0-12-804071-3.00007-0.

[12] C. Beauval, O. Scotti, F. Bonilla, The role of seismicity models in probabilistic seismic hazard estimation: comparison of a zoning and a smoothing approach, *Geophysical Journal International*, 2006, **165**, 584-595, doi: 10.1111/j.1365-246x.2006.02945.x.

[13] Y. Rong, Y. Bai, M. Ren, M. Liang, Z.Wang, Seismicity-based 3D model of ruptured seismogenic faults in the North-South Seismic Belt, China, *Frontiers in Earth Science*, 2023, **10**, 1023106, doi: 10.3389/feart.2022.1023106.

[14] M. Shah, Earthquake ionospheric and atmospheric anomalies from GNSS TEC and other satellites, Computers in

Earth and environmental Sciences. Elsevier, 2022, **28,** 387-399, doi: 10.1016/B978-0-323-89861-4.00009-9.
[15] M. S. Satti, M. Ehsan, A. Abbas, M. Shah, J. F. de Oliveira-Júnior, N. A. Naqvi, Atmospheric and ionospheric precursors associated with M ≥ 6.5 earthquakes from multiple satellites, *Journal of Atmospheric and Solar-Terrestrial Physics*, 2022, **227**, 105802, doi: 10.1016/j.jastp.2021.105802.
[16] D. Douglas Atsa'am, T. Gbaden, R. Wario, A machine learning approach to formation of earthquake categories using hierarchies of magnitude and consequence to guide emergency management, *Data Science and Management*, 2023, **6**, 208-213, doi: 10.1016/j.dsm.2023.06.005.
[17] N. S. M. Ridzwan, S. H. M. Yusoff, Machine learning for earthquake prediction: a review (2017-2021), *Earth Science Informatics*, 2023, **16**, 1133-1149, doi: 10.1007/s12145-023-00991-z.
[18] M. H. Al Banna, K. Abu Taher, M. S. Kaiser, M. Mahmud, M. S. Rahman, A. S. M. S. Hosen, G. H. Cho, Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges, *IEEE Access*, 2020, **8**, 192880-192923, doi: 10.1109/access.2020.3029859.
[19] K. M. Asim, A. Idris, T. Iqbal, F. Martínez-Álvarez, Earthquake prediction model using support vector regressor and hybrid neural networks, *PLOS ONE*, 2018, **13**, e0199004, doi:/10.1371/journal.pone.0199004.
[20] K. M. Asim, F. Martínez-Álvarez, A. Basit, T. Iqbal, Earthquake magnitude prediction in Hindukush region using machine learning techniques, *Natural Hazards*, 2017, **85**, 471-486, doi: 10.1007/s11069-016-2579-3.
[21] T. Wang, Y. Bian, Y. Zhang, X. Hou, Classification of earthquakes, explosions and mining-induced earthquakes based on XGBoost algorithm, *Computers & Geosciences*, 2023, **170**, 105242, doi: 10.1016/j.cageo.2022.105242.
[22] Z. Bao, J. Zhao, P. Huang, S. Yong, X. Wang, A Deep learning-based electromagnetic signal for earthquake magnitude prediction, *Sensors*, 2021, **21**, 4434, doi: 10.3390/s21134434.
[23] C. Cao, X. Wu, L. Yang, Q. Zhang, X. Wang, D. Yuen, Long short-term memory networks for pattern recognition of synthetical complete earthquake catalog, *Sustainability*, 2021, **13**, 9-14, doi: 10.3390/su13094905.
[24] P. Xiong, L. Tong, K. Zhang, X. Shen, R. Battiston, D. Ouzounov, R. Iuppa, D. Crookes, C. Long, H. Zhou, Towards advancing the earthquake forecasting by machine learning of satellite data, *IEEE Access*, 2021, **771**, 145256, doi: 10.1016/j.scitotenv.2021.145256.
[25] M. Yousefzadeh, S. Ahmad Hosseini, M. Farnaghi, Spatiotemporally explicit earthquake prediction using deep neural network, *Soil Dynamics and Earthquake Engineering*, 2021, **144**, 106663, doi: 10.1016/j.soildyn.2021.106663.
[26] M. H. Al Banna, K. A. Taher, M. S. Kaiser, M. Mahmud, M. S. Rahman, A. S. Hosen, G. H. Cho, Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges, *IEEE Access*, 2020, **8**, 192880-192923, doi: 10.1109/ACCESS.2020.3029859.
[27] G. Asencio-Cortés, A. Morales-Esteban, X. Shang, F. Martínez-Álvarez, Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure, *Computers & Geosciences*, 2018, **115**, 198-210, doi: 10.1016/j.cageo.2017.10.011.
[28] M. R. Vimal, S. M. Naseem, Time series analysis: forecasting with sarimax model and stationarity concept, *Journal of Emerging Technologies and Innovative Research (JETIR)*, 2020, **7**, 156-161.
[29] K. M. Asim, F. Martínez-Álvarez, A. Basit, T. Iqbal, Earthquake magnitude prediction in Hindukush region using machine learning techniques, *Natural Hazards*, 2017, **85**, 471-486, doi: 10.1007/s11069-016-2579-3.
[30] K. M. Asim, A. Idris, T. Iqbal, F. Martínez-Álvarez, Earthquake prediction model using support vector regressor and hybrid neural networks, *PLoS One*, 2018, **13**, e0199004, doi: 10.1371/journal.pone.0199004.
[31] Z. Bao, J. Zhao, P. Huang, S. Yong, X.-A. Wang, A deep learning-based electromagnetic signal for earthquake magnitude prediction, *Sensors*, 2021, **21**, 4434, doi: 10.3390/s21134434.
[32] C. Cao, X. Wu, L. Yang, Q. Zhang, X. Wang, D. Yuen, Long short-term memory networks for pattern recognition of synthetical complete earthquake catalog, *Sustainability*, 2021, **13**, 9-14, doi: 10.3390/su13094905.
[33] S. Chanda, M. C. Raghucharan, K. S. K. Karthik Reddy, V. Chaudhari, S. N. Somala, Duration prediction of Chilean strong motion data using machine learning, *Journal of South American Earth Sciences*, 2021, **109**, 103253, doi: 10.1016/j.jsames.2021.103253.
[34] F. Corbi, L. Sandri, J. Bedford, F. Funiciello, S. Brizzi, M. Rosenau, S. Lallemand, Machine learning can predict the timing and size of analog earthquakes, *Geophysical Research Letters*, 2019, **46**, 1303-1311, doi: 10.1029/2018gl081251.
[35] P. Debnath, P. Chittora, T. Chakrabarti, P. Chakrabarti, Z. Leonowicz, M. Jasinski, R. Gono, E. Jasińska, Analysis of earthquake forecasting in India using supervised machine learning classifiers, *Sustainability*, 2021, **13**, 971, doi: 10.3390/su13020971.
[36] Y. Essam, P. Kumar, A. N. Ahmed, M. A. Murti, A. El-Shafie, Exploring the reliability of different artificial intelligence techniques in predicting earthquake for Malaysia, *Soil Dynamics and Earthquake Engineering*, 2021, **147**, 106826, doi: 10.1016/j.soildyn.2021.106826.
[37] M. Fernández-Gómez, G. Asencio-Cortés, A. Troncoso, F. Martínez-Álvarez, Large earthquake magnitude prediction in Chile with imbalanced classifiers and ensemble learning, *Applied Sciences*, 2017, **7**, 625, doi: 10.3390/app7060625.
[38] E. Florido, G. Asencio-Cortés, J. L. Aznarte, C. Rubio-Escudero, F. Martínez-Álvarez, A novel tree-based algorithm to discover seismic patterns in earthquake catalogs, *Computers & Geosciences*, 2018, **115**, 96-104, doi: 10.1016/j.cageo.2018.03.005.
[39] J. Huang, X.-A. Wang, Y. Zhao, C. Xin, H. Xiang, Large earthquake magnitude prediction in Taiwan based on deep

learning neural network, *Neural Network World*, 2018, **28**, 149-160, doi: 10.14311/nnw.2018.28.009.
[40] F. Khosravikia, P. Clayton, Machine learning in ground motion prediction, *Computers & Geosciences*, 2021, **148**, 104700, doi: 10.1016/j.cageo.2021.104700.
[41] J.-W. Lin, C.-T. Chao, J.-S. Chiou, Determining neuronal number in each hidden layer using earthquake catalogues as training data in training an embedded back propagation neural network for predicting earthquake magnitude, *IEEE Access*, 2018, **6**, 52582-52597, doi: 10.1109/access.2018.2870189.
[42] A. Meirmanov, M. Nurtas, Mathematical models of seismics in composite media: elastic and poro-elastic components, *Journal of Differential Equations*, 2016, **2016**, 1-22.
[43] M. Nurtas, Zh. Baishemirov, S. Alpar, F. Tokmukhamedova, Numerical simulation of wave propagation in mixed porous media using finite element method, *Journal of Theoretical and Applied Information Technology*, 2021, **99**, 4163-4172.
[44] M. Nurtas, Zh. Baishemirov, A. Ydyrys, A. Altaibek, 2-D Finite Element method using "eScript" for acoustic wave propagation, *Proceedings of the 6th International Conference on Engineering & MIS*, 2020, **2020**, 3410774, doi: 10.1145/3410352.3410774.
[45] M. Nurtas, Zh. Baishemirov, Zh.Zhanabekov, Convalutional Neural Networks as a method to solve estimation problem of acoustic wave propogation in poroelastic media, *News of the National Academy of sciences of the Republic of Kazakhstan*, 2020, **4**, 52-60, doi: 10.32014/2020.2518-1726.65.

