A

COLLABORATIVE RESEARCH PROPOSAL

ON

"Predicting Large Earthquakes from small Earthquake in Nepal Himalaya using Machine Learning Techniques "

Submitted By:

Himal Bhandari

Madan Bhandari College of Engineering

Submitted To:

Pokhara University Research Center

Pokhara-30, Kaski, Nepal

April 25, 2025

ABSTRACT

Earthquake magnitude prediction for Nepal has been carried out in this research using the temporal sequence of historic seismic activities in combination with the machine learning classifiers. Prediction has been made on the basis of mathematically calculated seismic indicators using the earthquake catalog of the region. This work evaluates artificial neural networks' accuracy when used to predict earthquakes magnitude in Nepal. Several seismicity indicators have been retrieved from the literature and used as input for the networks. Some of them have been improved and parameterized in order to extract more valuable knowledge from datasets. These parameters are based on the well-known geophysical facts of Gutenberg–Richter's inverse law, distribution of characteristic earthquake magnitudes and seismic quiescence. Machine learning methods, such as random forest, artificial neural network and long short-term memory (LSTM) neural networks, excel at identifying patterns in large-scale databases and offer a potential means to improve earthquake prediction performance. Differing from physical and statistical approaches to earthquake prediction, we explore whether small earthquakes can be used to predict large earthquakes within the framework of machine learning. Specifically, we attempt to answer two questions for a given region: (1) Is there a likelihood of a large earthquake (e.g., $M \geq 6.0$) occurring within the next year? (2) What is the maximum magnitude of an earthquake expected to occur within the next year?

Keywords: earthquake prediction; machine learning; random forest; long short-term memory neural network; declustering; statical seismology.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS/LIST OF ABBREVIATIONS

LSTM	:	Long Short-Term Memory
ML	:	Machine Learning
ANN	:	Artificial Neural Network
RF	:	Random Forest
MHT	:	Main Himalaya Thrust
USGS	:	United States Geological Survey
NCS	:	National Seismological Centre, Nepal
ISC	:	International Seismological Centre
PSHA	:	Probabilistic Seismic Hazard Analysis
GMPE	:	Ground Motion Prediction Equation
SDG	:	Sustainable Development Goals
NDRRMA	:	National Disaster Risk Reduction and Management Authority, Nepal
AUC	:	Area Under Curve

Table of Contents

1. Introduction

1.1 Background of the Study

The conventional perception of earthquakes has been changing in recent decades; cascading hazards and their effects along with damage to structures and infrastructure, casualties, socioeconomic and environmental losses are nowadays considered under multidisciplinary aspects of earthquake impact[1].They cause damage to infrastructure leading to significant loss and catastrophic damage in the affected areas, both ecologically and economically[2], [3]. Firstly, seismic events are the result of intricate interactions between tectonic plates, faults, and other geological factors[4]. Large earthquakes often re-occur at lengthy intervals (hundreds to thousands of years), making it arduous to identify trends and patterns over a prolonged time frame[4].

Nepal is a seismically active due to the continuous convergence of the Indian plate beneath the Eurasian plate, the entire area of Hindu-Kush-Himalaya (HKH) is hit by strong to major earthquakes frequently[5], [6]. As Nepal is centrally located in the HKH, hundreds of earthquakes of magnitude greater than 4 occur every year[1]. The Subduction of the Indian Plate beneath the Eurasian Plate, resulting an uplift at the rate 21 ± 3 mm /year in southern Nepal during the past 10,000 years, makes Nepal highly susceptible to earthquakes[7]. The notable earthquake of 2015 was recorded by instruments installed in the Kathmandu valley and the aftershock activities of the 2015 Gorkha earthquake are also well documented by the National Seismological Center, Nepal (for details see http://www.seismonepal.gov.np/)[1]. The main shock of the 2015 Gorkha seismic sequence was followed by 480 aftershocks of local magnitude equal to or greater than 4; and some of the aftershocks, like those of April 25, 2015 (MW 6.7), April 26, 2015 (MW 6.9), and May 12, 2015 (MW 7.3), April 25, 2015 (MW 7.8), aggravated the damage throughout the affected areas. Seismotectonic and Engineering Seismological Aspects of the MW 7.8 Gorkha, Nepal, Earthquake, by Rajesh Rupakhety[1].

During the past several years a number of earthquake prediction research were successfully implemented in the way of Theoretical, Mathematical, Computational and Statistical techniques[8].The Gutenberg-Richter inverse power-law establishes an inverse linear relationship between the magnitude of seismic events and the logarithm of frequency of occurrence of events of magnitude equal or lower than that magnitude[9], [10], [11].Another temporal earthquake distribution model that is being given increasing scientific attention in recent years is the characteristic earthquake distribution model originally proposed by Kagan and Jackson[9].Some scholars proposed the use of machine learning in the area of research in seismology[12]. The research of time series and magnitude prediction, neural networks, including LSTM and convolution neural networks, have been widely used in recent years[9], [13]. Machine Learning (ML) and Artificial Neural Networks (ANN) have been used in a variety of fields for prediction and classification purposes, like computer vision, object recognition, genetics, weather forecasting[14].

1.2 Literature Review

B. Gutenberg and C.F. Richter[11], [15]**,** in this study, pointed out that earlier estimates of how often strong earthquakes happen in California were mostly based on old and unreliable historical records. In this note, they try to improve those estimates by comparing earthquake activity in California to earthquake activity around the world using statistics. To tell the difference between big and small earthquakes, we can't just rely on damage reports, especially since some quakes happen under the ocean or in places where no one lives. Instead, we need to use data from instruments. They use the instrumental magnitude scale, which was first made for California and later used globally.

The earthquake magnitude was originally defined based on how big the wave looks on a seismograph (a special machine that records earthquakes), taken 100 kilometers away from the earthquake's center. This method works for shallow earthquakes, which are the kind that usually happen in California. Deep earthquakes haven't been recorded in that region.

HemChandra Chaulagain, Dipendra Gautam and Hugo Rodrigues[1], In Nepal, earthquake impacts vary across communities. For example, after the 1260 earthquake, more people died from famine than the quake itself. While studies on earthquakes began mainly after the 1988 quake, there’s still a lack of research that looks at earthquakes from different perspectives. Local-level models for predicting damage, deaths, and losses are limited, and models from other countries may not work well for Nepal.

Only Kathmandu Valley has a detailed loss estimation, and other areas are still not studied properly. Historical data is especially valuable because it's more reliable than predictions. However, detailed records of Nepal's major earthquakes including effects by gender, urban vs rural damage, and economic impact are still missing.

This chapter aims to fill that gap by analyzing major earthquakes (above magnitude 6.5) that hit Nepal in 1833, 1934, 1980, 1988, 2011, and 2015 to understand the pattern of damage and help prepare for future quakes.

Roger Bilham, Vinod K. Gaur and Peter Molnar[7], since 1800, major earthquakes in the Himalayas, like those in 1905, 1934, and 1950, have caused widespread destruction, but less than half of the region has experienced such events in that time. No surface ruptures have been found, making it hard to predict future quakes. Estimates suggest these earthquakes involved about 4 to 8 meters of ground slip. With the Himalayas moving at 20 mm per year, some areas may now have enough built-up strain for a quake similar to the 1934 event, with potential slips of 4 to 6 meters or more. Parts of the region may not have ruptured in 500–700 years, possibly leading to slips over 10 meters.

The population in the Ganges Plain has grown significantly, with 50 million people now at risk. Weak building codes and rapid urbanization increase the danger. A repeat of a 1905-like quake could cause 200,000 deaths, and a major quake near a megacity could be far worse.

Xi Wang, Zeyuan Zhong, Yuechen Yao, Zexu Li, Shiyong Zhou, Changsheng Jiang and Ke Jia[16], Earthquake prediction is a persistent challenge in seismology, with no reliable physical or statistical models for forecasting large earthquakes. This study investigates the potential of machine learning, specifically random forest and long short-term memory (LSTM) neural networks, to predict large earthquakes ($M \geq 6.0$) and their maximum magnitudes in the Sichuan–Yunnan region using data from small earthquakes. By extracting seismicity parameters from earthquake catalogs, we addressed two questions: the likelihood of a large earthquake occurring within a year and the expected maximum magnitude. Results indicate that random forest outperforms other methods in classifying large earthquake occurrences,[17]. while LSTM provides reasonable magnitude estimations. These findings suggest that small earthquakes contain valuable information for predicting future large events, highlighting machine learning as a promising tool for improving earthquake prediction. The study also identified key seismicity parameters consistent with physical interpretations, offering insights into critical features for prediction models. However, limitations exist for earthquake swarms, which are statistically rare, are challenging to predict due to minimal magnitude differences. Additionally, longer-term seismic monitoring is needed to support advanced models like transformers for better performance. The study also did not account for spatial earthquake locations, an important factor for future models. Addressing these limitations represents the next steps for enhancing prediction accuracy. Overall, this work underscores the potential of machine learning to advance earthquake forecasting, providing a foundation for future research to refine models and incorporate spatial data for more robust predictions.

Bertrand Rouet-Leduc, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin Humphreys, Paul A. Johnson[17]**,** Machine learning (ML) applied to acoustic signals from laboratory shear experiments reveal hidden patterns that accurately predict fault failure times, offering new insights into earthquake forecasting. By analyzing instantaneous acoustic emissions from a lab fault, ML identifies a previously unrecognized signal, once considered low-amplitude noise enabling precise failure predictions throughout the fault's slip cycle without relying on historical data. This signal likely stems from continuous grain movements in the fault gouge as blocks shift. These findings suggest that ML could uncover similar signals in Earth's seismic data, advancing our understanding of fault physics and constraining failure times. Unlike earlier studies limited to earthquake catalogs, ML mitigates human bias by autonomously detecting patterns across vast datasets. Scaling this approach to natural faults, such as those with repeating earthquakes on the San Andreas Fault near Parkfield, could reveal analogous signals detectable by borehole or surface instruments. Ongoing research explores this potential. Beyond earthquakes, this ML-driven method may predict failures in various materials, leveraging advances in instrumentation, computing power, and data handling. This pioneering application of ML to continuous acoustic/seismic data sets the stage for transformative progress in earthquake science and material failure prediction.

Zefeng Li, Men-Andrin Meier, Egill Hauksson, Zhongwen Zhan, and Jennifer Andrews[18], Earthquake Early Warning (EEW) systems often face false alerts due to local impulsive noise from natural or human sources. To address this, we trained a Generative Adversarial Network (GAN) using 300,000 P-wave waveforms from southern California and Japan to learn key characteristics of earthquake P waves. The GAN's critic serves as an automatic feature extractor, feeding into a Random Forest classifier trained on approximately 700,000 earthquake and noise waveforms. This approach achieves exceptional accuracy, identifying 99.2% of earthquake P waves and 98.4% of noise signals, significantly reducing false EEW triggers. By combining GANs and Random Forests, we eliminate the need for manual feature selection and leverage large seismic datasets to deliver state-of-the-art performance. Our findings highlight GAN’s ability to uncover compact seismic wave representations, offering broad potential for seismology applications, including improved EEW systems and noise discrimination.

Table 1: Review of some of existing literature on Landslide hazard across Nepal

S.N.	Reference Literature	Study Region	Key Findings
1	B. Gutenberg and C.F. Richter	California	Earthquake frequency estimates in California by comparing global instrumental magnitude data, revealing inaccuracies in earlier estimates based on unreliable historical records.
2	HemChandra Chaulagain, Dipendra Gautam	Nepal	The study analyzes major Nepalese earthquakes (1833–2015) to identify damage patterns and improve local-level preparedness, highlighting the inadequacy of existing models and the need for detailed, region-specific historical data.
3	Roger Bilham, Vinod K. Gaur and Peter Molnar	Ganges plain	Major Himalayan earthquakes caused significant destruction, but less than half the region has ruptured, with no surface ruptures detected, complicating future quake predictions.

			With 20 mm/year tectonic movement, some areas may have accumulated strain for quakes with 4–10 meters of slip, threatening 50 million people in the Ganges Plain, where weak building codes and urbanization amplify risks.
4	Xi Wang, Zeyuan Zhong, Yuechen Yao	China	Predicting large Earthquake with the help of small Earthquake using ML Models
5	Bertrand Rouet-Leduc, Claudia Hulbert	Laboratory Experiment	Machine learning predicts slip failure time in shear experiments using statistical features of seismic signals. This approach could identify new geophysical signals and enhance earthquake forecasting. Applicable to various failure scenarios, including brittle failure, avalanches, landslides, and volcanic eruptions.
6	Zefeng Li, Men-Andrin Meier,	Southern California and Japan	Machine learning discriminator, trained on a large dataset, achieves 99.2% accuracy for earthquake P waves and 98.4% for impulsive noise. Significantly reduces false alerts in earthquake early warning systems.

1.3 Research Gap Analysis/ Problem Statements

It has been a huge leap in seismology, and even with the high-resolution seismic data, predicting large earthquakes has still remained a very difficult science, especially in tectonically complicated places like the Nepal Himalaya. Small-magnitude earthquakes occur frequently, and it has been understood that these earthquakes carry precursory messages. Yet, for prediction modeling, they have been underutilized. Empirical or deterministic approaches characterize most of the existing models; they cannot capture the process of earthquake preparation, which is nonlinear and multivariate within the range of common patterns. The Gutenberg-Richter law and their variation of β-value, energy release, or micro seismic clustering offer some early promise, yet they are seldom used in common with machine-learning approaches. Moreover, machine learning applications on seismology have hence mostly left Nepal with no region-specific, data-driven models as regard its very specific geophysical characteristics.

It justifies the need for interdisciplinary approaches bringing together seismological theories with machine learning to extract some meaningful patterns from small earthquakes and improve the forecasting of large ones in the Nepal Himalaya.

1.4 Research Objectives/Questions

The main objective of the research is: To predict and forecast the Large Earthquakes (magnitude ≥6.0) from small Earthquake in Nepal Himalaya using Machine Learning Techniques. The following are the secondary objectives to achieve the main objective.

- **To explore the use of small earthquakes for predicting large earthquakes**
  To investigate whether small seismic events contain patterns or information that can help predict the occurrence of large earthquakes.
- **To apply machine learning techniques to earthquake prediction**
  Assess the performance of machine learning models, such as Random Forest and Long Short -Term Memory (LSTM), in predicting large earthquakes based on historical seismic data.
- **To determine the likelihood of large earthquake**
  Develop a predictive model to classify the probability of a significant earthquake. (e.g, magnitude> 6.0) occurring in a given region within the next year.
- **To estimate the maximum magnitude of future earthquake**
  Utilize machine learning approaches to provide a rough estimation of the maximum magnitude of an earthquake expected within a specific timeframe.
- **To compare the performance of machine learning models**
  Evaluate and compare the effectiveness of Random Forest and LSTM models in classifying earthquake occurrences and estimating magnitudes.
- **To contribute to earthquake Risk Mitigation**

Provides insights that can improve the predictive accuracy of earthquake forecasts, contributing to disaster preparedness and risk reduction efforts.

The following Questions will be addressed in this research:

- Predict the likelihood of a large earthquake in a specific region within the next year?
- The maximum magnitude of an earthquake is expected within the same timeframe.
- Determination of best machine learning models for prediction of large earthquake.

These objectives and questions are academically interesting, adequate, and achievable, aligning with the significant impacts of earthquake and the need for prediction and effective disaster management strategies in Nepal.

1.5 Hypothesis

Severe earthquakes tend to be rarer here, but the region isn't devoid of minor-magnitude earthquakes either-small in scale, such minor earthquakes are believed to reflect the changing stress in the crust. These small earthquakes, while relatively low in individual energy release, provide important information on the lithosphere stress state. The entire energy released by clusters of small seismic events and their temporal-spatial development often precedes a heavy seismic rupture. In other words, as suggested by the Gutenberg-Richter relation, there exists logarithmic relationship between magnitude and frequency of earthquakes of a particular region, however, it is usually expressed as:

$$\log N = a - bM \text{ ---(i)}$$

where N=number of earthquakes having magnitude ≥M, and α and β are constant. Such alterations of the β-value through time and space are considered, among other factors, as diagnostic of stress regime; that is, decreasing β-values point into an increasing concentration of stress that raises the probabilities of a large event.

The study seeks to understand if small events could be strong precursors of large earthquakes in the Nepal Himalaya by such an investigation that compares frequency, energy release, and β-value evolution of small-magnitude earthquakes. This will entail unsupervised clustering, supervised classification, and machine learning approaches less supervised, such as temporal sequence models (for example, LSTM networks), to derive the features from seismic catalogs to pick out subtle, nonlinear patterns that may exist before large earthquakes. The integration of seismological theory and data-driven models aims to predict large earthquakes from small earthquakes, however at framing a predictive capability that could predict the probability of a large-magnitude earthquake on micro seismic activity and the associated energy dynamics preceding it.

This hypothesis addresses the main objective of the research, which is to develop a predictive model of large earthquakes from small earthquakes using machine learning approaches.

1.6 Rationale of the Study

The Central Himalayan region, especially Nepal, lies at the collision boundary of the Indian plate and the Eurasian tectonic plate-an active thrust zone that causes some of the most devastating earthquakes in the world, such as the 2015 Gorkha earthquake. There is an urgent need for better methods of earthquake forecasting, going beyond traditional empirical models and static hazard maps, given the highly seismic region. Even when traditional seismology is vital in understanding the long-term seismic potential, the traditional way has not been effective for short-term to medium-term prediction because the phenomena that produce earthquakes are inherently nonlinear, high-dimensional, and partially observable. Preceding a mega earthquake is a complex preparation phase, usually identified by an increase in micro seismicity, localized energy release, changes in β-value, as well as evolving stress fields, which are too fine to be detected with conventional statistical techniques alone.

The major idea underlying the study is that small-magnitude earthquakes are not independent random background events, but rather random events resulting from slow strain accumulation and consequent fault loading. Their frequency, patterns of distribution, and long-term seismic energy released are indirect yet quantifiable indicators of stress conditions in the lithosphere. The Gutenberg-Richter law, especially by virtue of, its β-value, is a statistical lens to observe changes in seismic regimes; for instance, a declining β-value is often linked with increasing differential stress and greater potential for a larger rupture. With machine learning entering the scene and capable of handling vast amounts of noisy and non-linear datasets, we now have tools that can hide in seismic sequences as they were never able to be seen before. When trained on the historical seismic data, models like recurrent neural networks (e.g., LSTM), convolutional architectures, and clustering algorithms 'learn' to associate certain micro seismic behaviors with the likelihood of future large events.

However, the application of machine learning to Himalayan conditions has not yet fully matured. Most such models are designed for tectonically very different, highly instrumented areas like California or Japan, and therefore cannot be directly adapted to the peculiar geodynamic context of the Nepal Himalaya. Thus, there is an imperative need to develop regional-specific, seismologically informed machine learning models that are responsive to fault mechanics, data paucity, and energy release dynamics of the region.

This study is therefore motivated to fill in this disciplinary and methodological gap-seismological theory, statistical insight, and machine learning, exploring whether the minor ones might inform us better about larger ones. The goal is not only to understand scientifically what causes earthquake precursors but also to provide practical outputs for disaster risk reduction and resilience.

2. Methodology

This study attempts to examine whether small earthquakes are precursors to large earthquake prediction in the Nepal Himalaya region. Earthquake catalog data will be obtained systemically from two major sources: the United States Geological Survey (USGS) and the National Seismological Centre (NSC), Nepal, over the period of this study. The raw data will be subjected to rigorous preprocessing, including but not restricted to duplicate removal, correction of outliers, and interpolation of the missing values to ensure temporal series continuity and reliability.

The ZMAP toolbox commonly used for establishing the distinction between mainshocks (large earthquakes) and aftershocks (smaller events) will then be introduced through the very wide range of accurate research done with MATLAB. Using spatiotemporal algorithms for declustering and homogeneity analysis, the values β and α will then be computed using the Gutenberg-Richter frequency-magnitude relationship. The completeness analysis will be done on the available earthquake dataset. At the end of this seismological modeling, several engineering attributes will be derived by mathematical formulations that describe the spatially and temporally evolving seismicity, such as cumulated seismic moment, event rate, inter-event time, and energy release indexes.

These features will serve as training inputs for high-class machine learning models like Random Forests (RF), Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) networks. The main aim is to test how well a model can predict the occurrence of large earthquakes in earlier patterns of small magnitude seismicity with model quality assessed primarily on predictive accuracy, precision, and temporal reliability.

Together, these engineering features and machine learning models provide the necessary infrastructure for conducting high-quality research, integrating USGS and NCS data with advanced training of ML models and extensive literature reviews to achieve the study's objectives.

2.1 Study Area

This research is particularly concerned with the Nepal Himalaya, which has been one of the most seismically active segments of the entire Himalayan arc located on the convergent boundary of the Indian Plate and Eurasian Plate. The region falls within a longitude range of approximately 80°E to 89°E and a latitude range of 26°N to 31°N, hosting MHT, or the Main Himalayan Thrust, the primary fault responsible for large earthquakes.

It is one of the regions that has witnessed disastrous events, such as the 1934 Bihar–Nepal Earthquake (Mw 8.0) and the 2015 Gorkha Earthquake (Mw 7.8), proving that it is at a very high seismic hazard. The study uses an extensively dense catalog of earthquakes with small-magnitude events (Mw < 4.5) from the last 10 decades, coming from the National Seismological Centre (NSC), USGS.

On the whole, the natural laboratory of Nepal Himalaya proves ideal for active tectonics, fault systems well documented, and with recent advances in seismic monitoring. This study proposes the development of a machine learning-based predictive model tailored very specifically to the unique seismicity of the region for better hazard assessment and early warning systems.

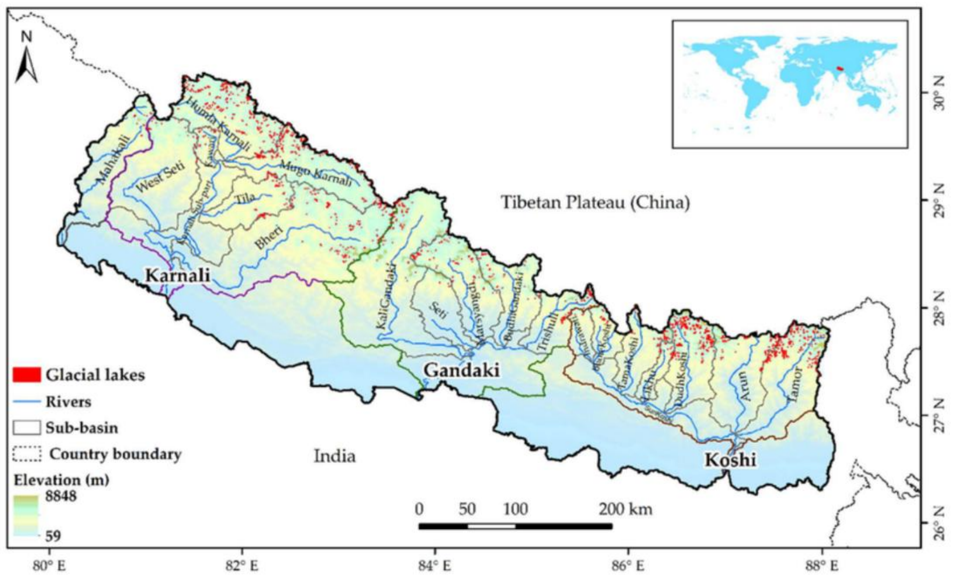

Figure 1: Location map of study area with topographic features

2.2 Data Collection and Analysis

The seismic event data required for the proposed research was obtained from the following authoritative sources:

- United States Geological Survey (USGS)
- National Seismological Centre (NSC), Nepal

- International Seismological Centre (ISC)

The earthquake catalog includes the following features:

- Event time (UTC)
- Epicentral coordinates (latitude, longitude)
- Hypo central depth (km)
- Magnitude (ML, Mw)
- Focal mechanism parameters (when available)
- Seismic moment and energy release

Events with a magnitude ≥ 2.5 will be considered to ensure consistency and reliability in instrumental records. Data from different sources were cross-validated and harmonized to maintain consistency in format and units.

The raw data will be subjected to rigorous preprocessing and analysis, including but not restricted to duplicate removal, correction of outliers, and interpolation of the missing values to ensure temporal series continuity and reliability. The analysis of the data set is carried out in the following methods:

- **De-duplication**: Redundant events reported across multiple catalogs were merged using spatial and temporal tolerance windows.
- **Outlier Removal**: Events with implausible coordinates, depths, or magnitudes were filtered using statistical thresholds.
- **Magnitude Homogenization**: All magnitudes were converted to moment magnitude (Mw) using empirical relationships, enabling consistent modeling.
- **Spatial Binning**: The study region was partitioned into uniform spatial grids (e.g., 0.1° x 0.1°) to support spatial-temporal analysis.

2.3 Research Approach

This research is directed towards creating a predictive strategy that could ably predict earthquake occurrence using patterns of small earthquakes and seismological and modern machine learning approaches.

The earthquake catalog data will be used from the United States Geological Survey (USGS) and finally from National Seismological Centre (NSC), Nepal. This catalog will be processed using ZMAP toolbox in MATLAB to allow for accurate separation of the two events: mainshock events from aftershock events with spatial-temporal declustering algorithms. The main reason for this classification is to ensure isolation of large independent seismic events (mainshocks) from dependent aftershock sequences in preparation for precise labeling for the machine learning models. The b-value will be computed in ZMAP toolbox using Gutenberg-Richter inverse power law model analysis. The model will be analyzed using the following equation:

$$log_{(10)}(N) = a - b(M) \text{---------------------(ii)}$$

Where;

- N = number of earthquakes with magnitude ≥ M
- M = magnitude of the earthquake
- a-value = represents the seismicity rate of the region (i.e., the overall level of earthquake activity)
- b-value = describes the relative proportion of small to large earthquakes
    - A higher b-value (>1**)** means many small earthquakes relative to large ones
    - A lower b-value (<1) suggests more large earthquakes, indicating higher stress

The earthquake dataset will be declustred in ZMAP toolbox using either Gardner and Knopoff or Reasenberg's algorithm. After declustering of dataset, A certain quantitative amount of seismological and engineering characteristics will go through derivation and these will include:

- Seismicity rate,
- Inter-event times and spatial distances,

- Cumulative energy release,
- Gutenberg-Richter β-value and α -value and
- Temporal clustering metrics.

All these features comprise the basis of selection as parts to be significant reflectors of crustal stress evolution and fault dynamics preceding large seismic events. The extracted feature dataset will be worked out in testing three different machine learning models.

- Random Forest (RF) is well-suited to discovering nonlinear relationships and ranking feature importance.
- Artificial Neural Networks (ANN) learn from complex high-dimensional relations among features.
- Long Short-Term Memory (LSTM) networks that are effective for learning from long-range temporal dependencies in seismic sequences.

The model will be built by randomly splitting available data into training and validation samples. Both samples will make 80% and 20%, respectively, to be used for training and validation, calibration. This controls overfitting where the model learns on the historical patterns but would have little generalization on unseen data. Model performance will also be robustly evaluated using metrics like precision, recall, F1 score, ROC-AUC to assess accuracy and reliability.

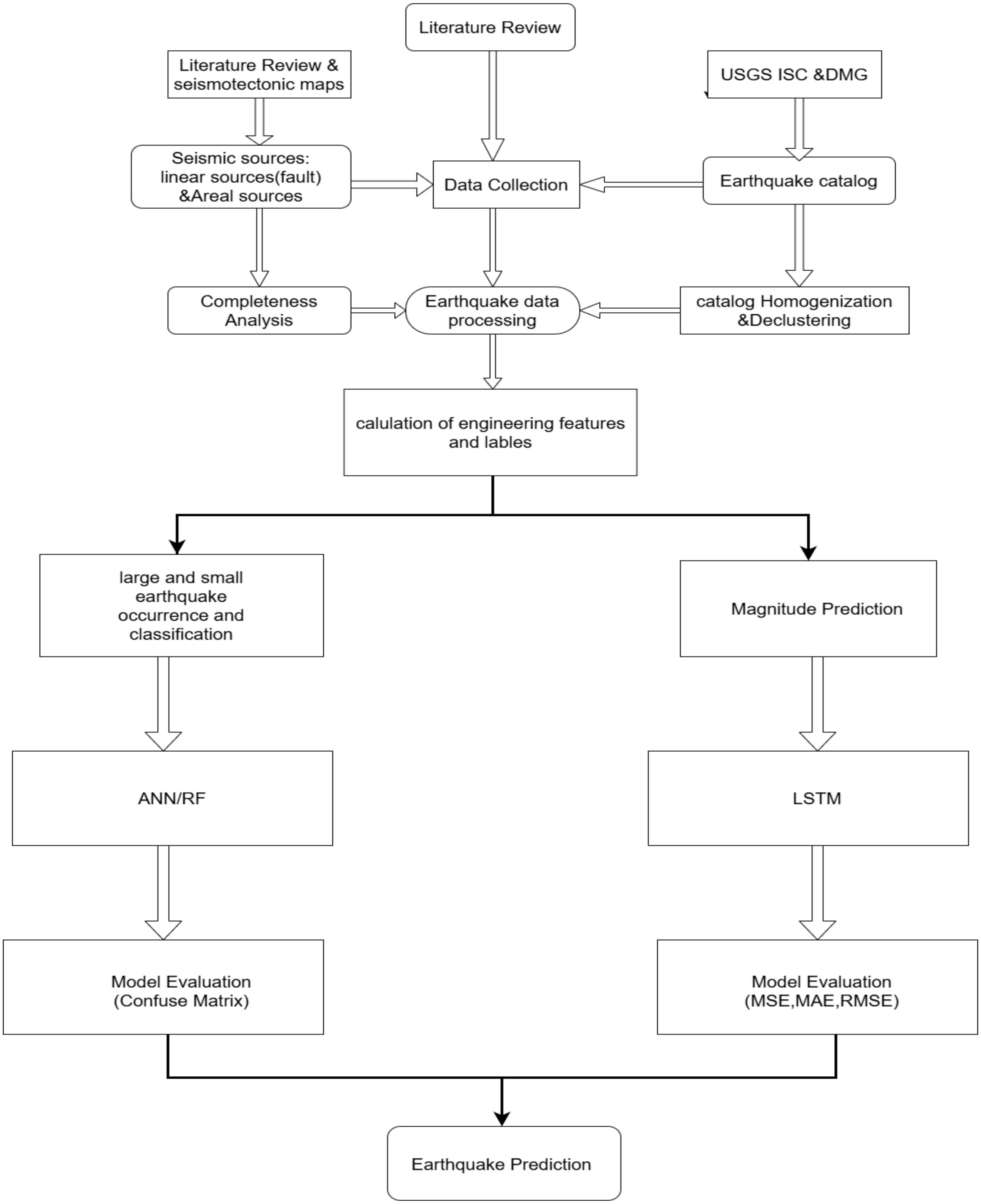
Literature Review
Literature Review & seismotectonic maps
USGS ISC &DMG
Seismic sources: linear sources(fault) &Areal sources
Data Collection
Earthquake catalog
Completeness Analysis
Earthquake data processing
catalog Homogenization &Declustering
calulation of engineering features and lables
large and small earthquake occurrence and classification
Magnitude Prediction
ANN/RF
LSTM
Model Evaluation (Confuse Matrix)
Model Evaluation (MSE,MAE,RMSE)
Earthquake Prediction

Figure 2: Research approach and methodological framework of study

2.3.1 Earthquake and Causative Factors

The sudden release of tectonic stress accumulated along faults within the Earth's crust generates earthquakes. The seismicity in the Nepal Himalaya is dependent on the convergence of the Indian and Eurasian plates at a rate of ~18 mm/year and concentrates stress along the Main Himalayan Thrust (MHT), the main seismogenic structure of this region. The rupture characteristics are determined by the fault geometry, slip rate, and locking depth. The distribution of foreshocks, aftershocks, and background seismicity constrains understanding of stress transfer and strain accumulation behaviors. Additional influences on seismicity might include crustal heterogeneity and fluid pressures. Ultimately, a comprehensive understanding of both large and small earthquakes is necessary for seismic hazard assessment. This research combines seismological theory and data-driven modeling for exploration into predictive indicators.

2.3.2 Data-Driven and Machine Learning Approaches

Machine learning is among the very powerful tools in predicting seismic forecasting to model nonlinear, dynamic, and immensely complicated patterns subsumed in the earthquake catalogs. This study concentrated on the three forms of machine learning which have been widely popular: Random Forest (RF), Artificial Neural Networks (ANN), and the Long Short-Term Memory (LSTM) networks, each quite different in terms of their advantage for predicting large earthquakes on the basis of smaller preceding earthquake events.

- **Random Forest (RF)**

Random Forest is an ensemble learning technique that simply builds a huge number of decisions trees in the training phase while aggregating their outputs to improve prediction capabilities and generalizability. While in regression tasks, such as in determining the probability or magnitude of occurrence of large earthquakes, the response to prediction is the average response of the individual tree predictions. The mathematical formulation is expressed as follows:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T\, i^{(x)}\text{-----------------------(iii)}$$

where $\hat{y}$ denotes the final prediction, N is an integer denoting the number of decision trees, and $Ti^{(x)}$ denotes the output from i^{th} the tree for the input feature vector x. The RF is particularly suited for handling noisy and high-dimensional datasets, as are common with seismic records; it is robust against overfitting and affords interpretability through feature importance metrics.[19], [20]

- **Artificial Neural Networks (ANNs)**

Artificial Neural Networks (ANNs) are biologically inspired computational frameworks composed of interlinked processing units (neurons) established on layers. These networks have complex functions by transforming the input through a cascade of weighted summations and nonlinear activation functions. A feedforward ANN with a single hidden layer can be mathematically formulated as:

$$\hat{y} = f\left(\sum_{J=1}^{n} \omega_J^{(2)} \cdot \sigma\left(\sum_{i=1}^{m} \omega_{iJ}^{(2)} x_i + b_J^{(1)}\right) + b^{(2)}\right) \text{----------------(iv)}$$

where x_i represents the input features, ω and b denote the weights and biases, σ is an activation function (e.g., ReLU or sigmoid), and f represents the output layer function. ANNs are particularly advantageous in identifying nonlinear relationships among seismic features such as inter-event time, depth, magnitude, and energy release, making them suitable for event classification and magnitude estimation tasks[21].

- **Long Short-Term Memory (LSTM)**

Networks with LSTM architecture belong to the family of Recurrent Neural Networks (RNNs) that can model sequential dependencies present in temporal data. LSTMs maintain a cell state and gating mechanism to control the flow of information, allowing the network to maintain long-term dependencies while avoiding the vanishing gradient problem. The fundamentals of LSTM unit operations consist of:

- Forget gate:

$$f_t = \sigma\left(w_f\left[h_{t-1,x_t}\right] + b_f\right)\text{-----------(v)}$$

- Input gate and candidate update:

$$i_t = \sigma\left(w_i\left[h_{t-1,x_t}\right] + b_i\right)\text{-----------(vi),}$$

$$c_t = tanh\left(w_c\left[h_{t-1,x_t}\right] + b_c\right)\text{-----------(vii)}$$

- Cell state update:

$$Ct = ft \odot C_{t-1} + it \odot \tilde{C}_t\text{------------(viii)}$$

- Output gate and hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), h_t = o_t \odot \tanh(C_t)\text{--------(ix)}$$

Here, x_t is the input at time t, h_t is the hidden state, C_t is the cell state, and σ denotes the sigmoid activation function. LSTM is particularly suitable for capturing temporal dependencies in earthquake sequences, such as identifying patterns in foreshocks, aftershocks, and cumulative seismic release over time[22].

2.3.3 Model evaluation and Validation

The machine learning models are validated and evaluated to ensure they generalize well to unseen data and are not overfitted to the training set. It helps assess model reliability, stability, and real-world predictive performance. The models will be calibrated as follows:

1. Train-Validation-Test Split

The data set is divided into three subsets, which are training (70-80%), validation (10-15%), and test (10-20%). The training set fits the model, the validation set tunes the hyperparameters, and the test set evaluates final performance. This avoids overfitting, thus ensuring some degree of generalization for unseen data.

2. K-Fold Cross-Validation

The data is split into k equal parts, each fold being used for validation once while the remaining k-1 folds are used for training. Very useful for Random Forest and ANN models when the data arrives in a non-sequential manner. This gives a solid estimate for performance and reduces the variance from the random splits.

3. Time Series Validation (Rolling Forecast Origin)

Retaining time series order, training on historical data, and validating upon future sequences. It is integral for models like LSTMs, which are reliant on sequential input. Avoids data leakage and mimics real-world forecasting conditions.

4. Hyperparameter Tuning (Grid Search/Random Search)

Goes through the parameter combinations to optimize validation model performance. It's possible to combine hyperparameter tuning with cross-validation to allow for a more systematic approach.

5. Evaluation Metrics

- Regression: RMSE, MAE, R^2 Score, either for magnitude of prediction or continuous.
- Classification (if framed): Accuracy, Precision, Recall, F1-Score, and area under the curve-AUC- for event classification or threshold.
- Specialized Seismology: The residual error analysis, prediction intervals, and space-time match rates.

3. Expected outcome, Impact and Application of the research

From the research title: " Predicting Large Earthquakes from small Earthquake in Nepal Himalaya using Machine Learning Techniques ", the following results and outputs are expected.

- **Establishment of a Full-Fledged Predictive Framework**

This research work would result in a scientifically validated, machine learning-based framework capable of providing an outlook on the occurrence of large earthquakes based on the statistical and physical behavior of small magnitude seismicity.

It will convert the existing seismological approaches from passive observations to proactive forecasting; thus, contributing to the new generation of earthquake prediction models established on data science and geophysics.

- **Advancing Seismology by Interdisciplinary Innovation**

By integrating physical seismological features with AI models, this work introduces an exceptionally strong hybrid methodology.

Not only does it advance knowledge in the area relating to the initiation processes of earthquakes, but it also goes well beyond such achievements by setting the new standard for interdisciplinary research in earthquake prediction. It is a great effort bringing together seismology, applied mathematics, and machine learning.

- **Contributing to the Global Scientific Community**

The model, method, and findings will provide a replicable and open framework to the international research community that can be used in other active seismic tectonic regions. It supports the global efforts in more reliable forecasting of earthquakes and will add to high-impact academic journals, seismic workshops, and collaborative research across discipline

- **Enhanced Capabilities for Real-Time Monitoring and Alerts**

The model could be integrated into national seismic networks for real-time monitoring and early warning systems.

This continuous analysis of incoming micro seismic data would facilitate the early detection of interludes of increased seismic risk, providing the necessary lead time for governments and emergency services.

- **Strategic Alignment to Nepal’s Disaster Risk Reduction Agenda**

It supports national resilience efforts directly through earthquake forecasting for evidence-based planning, thereby improving disaster planning and optimization of resources and community engagement by agencies like NDRRMA.

Under the Sendai Framework, this contributes to the Nepal development goal as well as to the realization of long-term infrastructure safety and risk management in the country.

- **Construction Resilience Improvement and Urban Safety**

Forecast data could help home-engineer design practices for critical infrastructures, seismic retrofitted structures and land use planning.

Outputs from model simulations could be invaluable to urban planners and engineers in selecting high-risk areas for consideration in the contribution toward building more earthquake-resilient cities and communities in the Himalayan region.

- **Empowerment through Community Preparedness and Awareness**

The research output will also support the public awareness campaigns informing communities how very small earthquakes can be indicators of much larger events.

At the grassroots levels, it will prepare citizens to be active participants in local disaster response systems.

- **Real World Application in Emergency Response Systems**

By forecasting large seismic events, the model enables governments and humanitarian organizations to pre-position resources, plan evacuations, and deploy emergency services efficiently.

By achieving these outputs, the research will provide significant contributions to earthquake prediction and earthquake disaster management strategies in Nepal, supporting both practical applications and academic advancements. This holistic approach will ensure that the findings are not only theoretically robust but also practically implementable, leading to tangible benefits for the local population and infrastructure.

4. Work Plan and Schedule

Table 2. Work Plan

S.N.	Tasks	Sub Tasks	Months (Duration: Jan. 2026 – Dec. 2026)											
			1	2	3	4	5	6	7	8	9	10	11	12
1	Planning and Preparation	1. Finalize Research Team and Roles	✓											
		2. Develop Detailed Project Plan and Timeline		✓										
		3. Conduct Preliminary Literature Review		✓										
		4. Identify data sources and collect initial data			✓									
2	Data Collection and Analysis	1. Determine Seismic sources				✓								
		2. Data Collection (NCS, Seismotectonic maps)				✓	✓							
		3. Preprocessing of Data					✓	✓						
		4. Begin Preliminary Data Analysis						✓						
		5. Perform Detail Data Analysis							✓					
3	Machine Learning Model Development	1. Random Forest and Artificial Neural Network Model								✓				
		2. Long Short-Term Memory Model								✓				
		3. Validate & refine Models								✓				
4	Document Preparation and expert insights	1. Present document finding and interim results									✓			
		2. Expert discussion for qualitative insights										✓		
		3. Refining Machine Learning model										✓		
5	Recommendation and Submission	1. Submit Final Report											✓	
		2. Submit article peer reviewed journal											✓	
		3. Present final report and recommend policy												✓

Table 3: Expected schedule of the planned study

S.N.	**Tasks**	**Months**											
		1	2	3	4	5	6	7	8	9	10	11	12
1	Grant Agreement	■											
2	Literature Review and Data Collection		■	■	■	■	■						
3	Data Analysis							■					
4	Model development								■	■			
5	Result and Report Writing										■		
6	First Draft Submission & Presentation											■	
7	Final Submission												■

5. Budget Distribution Plan

Table 4: Expected Budget of the planned study

Expense Category	Amount (NRs.)	Justification
Data Acquisition (Earthquake Catalog data)	35,000 /-	High resolution Seismotectonic maps, NCS and USGS earthquake, fault line data.
Travel and Transportation	40,000 /-	During research and data collection
Supportive Staff	15,000 /-	Supportive Staff worker's wages
Software and Modeling tools	50,000 /-	MATLAB, and Python expert charge
Miscellaneous	10,000 /-	Report printing charge, communication
Total	**1,50,000 /-**	**Including 15% Tax**

Total Budget According to Work Plan, Contribution Requested to PURC

Competitive Components	NRs	
Data Acquisition and Transportation	45,000 /-	
Compensation for Key Experts, Supporting Staffs	95,000/-	
Miscellaneous Expenses	10,000 /-	
Total	1,50,000/-	

Breakdown of Compensation for Key Experts

S.N.	Position	Person-per-days Compensation rate	Time Input in Person/Days	Total Amount (Rs.)	Remarks
A	**Key Experts**				
1.	Principal Investigator	5000.00	6	30,000/-	
2.	Supportive Staff	1500.00	10	15,000 /-	
3.	MATLAB & Python Expert	5000.00	10	50,000 /-	
	Total Cost			95,000 /-	

Breakdown of Miscellaneous Expenses

S.N.	Type of Reimbursable Expenses	Unit	Quantity	Unit rate	Amount (Rs.)
1.	Communication Cost	LS			2,000/-
2.	Stationery and Printing Cost	LS			8,000/-
	Total Miscellaneous Expenses				10,000/-

6. Relevance to the UN Sustainable Development Goals

The proposed research on " Predicting Large Earthquakes from small Earthquake in Nepal Himalaya using Machine Learning Techniques " aligns with UN Sustainable Development Goal 11: Sustainable Cities and Communities.

Justification:

SDG 11: Sustainable Cities and Communities aims to make cities and human settlements inclusive, safe, resilient, and sustainable. One of the key targets under this goal is to reduce the number of people affected by disasters, including natural hazards like earthquakes, by enhancing disaster risk reduction and resilience at local and national levels.

Earthquake, particularly in Nepal Himalaya, pose significant threats to infrastructure, livelihoods, and human safety. Earthquakes have a huge impact on human existence, taking lives, injuring human beings, and causing mental trauma. On the other hand, they cause the destruction of houses, structures, and critical facilities with long-lasting social and economic consequences. The research addresses this critical issue by developing a predictive model using a Data-Driven Machine Learning Approach that integrates seismic energy release, main shocks, aftershocks and foreshocks factors. By predicting and assessing earthquake in Nepal Himalaya, the research directly contributes to enhancing the resilience of infrastructure and the surrounding communities, aligning with the objectives of SDG 11.

How the Research Contributes to SDG 11?

1. Enhancement of Earthquake Early Warning and Risk Preparedness

Correct prediction of large earthquakes using patterns from smaller seismic events would substantially improve the early warning systems. If the authorities can rely on these accurate predictions, they can provide timely warnings and activate emergency procedures to minimize human and economic losses. This would further increase preparedness in and around earthquake regions, especially in the Nepal Himalaya, where premature actions can save the lives of many and mitigate structural damage.

2. In Pursuit of Resilient Urban Planning and Infrastructure Design

Knowing where and how big earthquakes are likely to occur is important for designers and civil engineers in establishing cities to be safer. Thus, predictive models based on seismic records enable more informed decisions about land use, zoning, and the building of critical infrastructure. By using GMPE and PSHA this research presents numerical estimations of ground shaking levels that are expected in a wide variety of areas.

This practice equips designers with instrumentations for hazard-specific design guidelines that ensure buildings, bridges, and lifeline infrastructures are prepared for future seismic occurrences. Such scientific inputs are important for the updating of construction codes as well as the promotion of sustainable construction practices in vulnerable urban regions. Communities are better poised to suffer with minimal impact from future earthquakes with more resilient infrastructure, making way for further sustainable urban growth.

3. Assist Data-Driven Policy and Governance for Disaster Risk Reduction

Scientific investigation that propounds a valid tool for seismic forecasting allows governments and decision-makers to plan disaster risk management strategies. With robust information, authorities can ensure investment in resilient infrastructure, enforce safety regulations, and raise community awareness. Such policies will assure safety and sustainability in the long term, especially in developing mountain regions with high exposure to seismic threats.

4. Strengthening Science, Technology, and Innovation for Safer Communities

Seismic data applications combined with machine learning provide one of the means of furthering the science of earthquake phenomena and creating technological advancements in predicting disasters. It would also provide a scientific basis for engineers and policymakers alike. An investment in knowledge such as this promotes the development of adaptive communities with a contemporary toolset of risk reduction and long-term resilience.

In conclusion, the proposed research on predicting large earthquake from small earthquake on Nepal Himalaya directly supports SDG 11: Sustainable Cities and Communities by contributing to the development of disaster-resilient infrastructure, promoting safer communities, and enhancing disaster risk management strategies in Nepal.

References

[1] H. Chaulagain, D. Gautam, and H. Rodrigues, “Revisiting major historical earthquakes in Nepal: Overview of 1833, 1934, 1980, 1988, 2011, and 2015 seismic events,” in *Impacts and Insights of the Gorkha Earthquake*, Elsevier, 2018, pp. 1–17. doi: 10.1016/B978-0-12-812808-4.00001-8.

[2] R. Bilham, V. K. Gaur, and P. Molnar, “Earthquakes: Himalayan seismic hazard,” Aug. 24, 2001. doi: 10.1126/science.1062584.

[3] H. Chaulagain, H. Rodrigues, V. Silva, E. Spacone, and H. Varum, “Seismic risk assessment and hazard mapping in Nepal,” *Natural Hazards*, vol. 78, no. 1, pp. 583–602, Apr. 2015, doi: 10.1007/s11069-015-1734-6.

[4] M. Liu and S. Stein, “Mid-continental earthquakes: Spatiotemporal occurrences, causes, and hazards,” Nov. 01, 2016, *Elsevier B.V.* doi: 10.1016/j.earscirev.2016.09.016.

[5] M. Liu and S. Stein, “Mid-continental earthquakes: Spatiotemporal occurrences, causes, and hazards,” Nov. 01, 2016, *Elsevier B.V.* doi: 10.1016/j.earscirev.2016.09.016.

[6] K. M. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, “Earthquake magnitude prediction in Hindukush region using machine learning techniques,” *Natural Hazards*, vol. 85, no. 1, pp. 471–486, Jan. 2017, doi: 10.1007/s11069-016-2579-3.

[7] R. Bilham, V. K. Gaur, and P. Molnar, “Earthquakes: Himalayan seismic hazard,” Aug. 24, 2001. doi: 10.1126/science.1062584.

[8] S. Narayanakumar and K. Raja, “A BP Artificial Neural Network Model for Earthquake Magnitude Prediction in Himalayas, India,” *Circuits and Systems*, vol. 07, no. 11, pp. 3456–3468, 2016, doi: 10.4236/cs.2016.711294.

[9] A. Panakkat and H. Adeli, “NEURAL NETWORK MODELS FOR EARTHQUAKE MAGNITUDE PREDICTION USING MULTIPLE SEISMICITY INDICATORS,” 2007. [Online]. Available: www.worldscientific.com

[10] “gutenberg1936”.

[11] A. de Santis, G. Cianchini, P. Favali, L. Beranzoli, and E. Boschi, “The Gutenberg-Richter law and entropy of earthquakes: Two case studies in central Italy,” *Bulletin of the Seismological Society of America*, vol. 101, no. 3, pp. 1386–1395, Jun. 2011, doi: 10.1785/0120090390.

[12] Y. Shimshoni and N. Intrator, “Classification of Seismic Signals by Integrating Ensembles of Neural Networks,” 1998.

[13] M. Moustra, M. Avraamides, and C. Christodoulou, “Artificial neural networks for earthquake prediction using time series magnitude data or Seismic Electric Signals,” *Expert Syst Appl*, vol. 38, no. 12, pp. 15032–15039, Nov. 2011, doi: 10.1016/j.eswa.2011.05.043.

[14] K. M. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, “Earthquake magnitude prediction in Hindukush region using machine learning techniques,” *Natural Hazards*, vol. 85, no. 1, pp. 471–486, Jan. 2017, doi: 10.1007/s11069-016-2579-3.

[15] B. Gutenberg and C. F. Richter, “FREQUENCY OF EARTHQUAKES IN CALIFORNIA*.”

[16] X. Wang *et al.*, “Small Earthquakes Can Help Predict Large Earthquakes: A Machine Learning Perspective,” *Applied Sciences (Switzerland)*, vol. 13, no. 11, Jun. 2023, doi: 10.3390/app13116424.

[17] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, “Machine Learning Predicts Laboratory Earthquakes,” *Geophys Res Lett*, vol. 44, no. 18, pp. 9276–9282, Sep. 2017, doi: 10.1002/2017GL074677.

[18] Z. Li, M. A. Meier, E. Hauksson, Z. Zhan, and J. Andrews, “Machine Learning Seismic Wave Discrimination: Application to Earthquake Early Warning,” *Geophys Res Lett*, vol. 45, no. 10, pp. 4773–4779, May 2018, doi: 10.1029/2018GL077870.

[19] L. Breiman, “Random Forests,” 2001.

[20] E. Scornet, G. Biau, and J. P. Vert, “Consistency of random forests,” *Ann Stat*, vol. 43, no. 4, pp. 1716–1741, Aug. 2015, doi: 10.1214/15-AOS1321.

[21] A. Mignan and M. Broccardo, “Neural network applications in earthquake prediction (1994-2019): Meta-analytic and statistical insights on their limitations,” *Seismological Research Letters*, vol. 91, no. 4, pp. 2330–2342, Jul. 2020, doi: 10.1785/0220200021.

[22] S. Hochreiter and J. ¨ Urgen Schmidhuber, “Long Short-Term Memory.”