Geoscience Frontiers

Earthquake risk assessment in NE India using deep learning and geospatial analysis

--Manuscript Draft--

Manuscript Number:	GSF-D-20-00084R1
Article Type:	Research Paper
Keywords:	earthquake; CNN; GIS; hazard; Vulnerability; Risk
Corresponding Author:	Biswajeet Pradhan, PhD University Technology Sydney Sydney, AUSTRALIA
First Author:	Ratiranjan Jena, Msc
Order of Authors:	Ratiranjan Jena, Msc
	Biswajeet Pradhan, PhD
	Sambit Prasanajit Naik, PhD
	Abdullah M. Alamri, PhD
Abstract:	Earthquake prediction is currently the most crucial task required for the probability, hazard, risk mapping, and mitigation purposes. From the last decade, event prediction has attracted increasing research attention from the academia and industries. However, deep learning techniques have been rarely tested for earthquake probability mapping. Therefore, this study developed a convolutional neural network (CNN) model for earthquake probability assessment and then performed vulnerability, hazard, and risk mapping. A prediction task, in which the model predicts magnitudes more than 4 Mw, was first abstracted by considering nine indicators. Prediction results and intensity variation were then used for probability assessment and hazard map production, respectively. Finally, the risk was produced by multiplying hazard, vulnerability, and coping capacity. The vulnerability was prepared by using six vulnerable factors, and the coping capacity was estimated by using the number of hospitals and disaster budget. This study contributes to addressing the problems in the NE region of India, which is becoming a high hazard zone. Prediction of events more than 4 Mw using CNNs is required. The CNN model for a probability distribution is a robust technique that provides good accuracy. In particular, the proposed model was experimentally tested on datasets of NE India and achieved good accuracy. Results show that CNN is superior to the other algorithms, which completed the prediction task with an accuracy of 0.94, precision of 0.98, recall of 0.85, and F1 score of 0.91. These indicators were used for probability mapping, and the total area of hazard, vulnerability, and risk was estimated.

Earthquake risk assessment in NE India using deep learning and geospatial analysis

Ratiranjan Jena [1], Biswajeet Pradhan [1, 2,3]*, Sambit Prasanajit Naik [3], Abdullah M. Alamri[4]

[1]The Center for Advanced Modeling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2007, Australia
[2]Center of Excellence for Climate Change Research, King Abdulaziz University, Jeddah, Saudi Arabia
[3]Department of Energy and Mineral Resources Engineering, Sejong University, Choongmu-gwan, 209, Neungdong-roGwangin-gu, Seoul 05006, Korea
[3]Active Fault and Earthquake Hazard Mitigation Research Institute, Pukyong National University, Busan, 48513, South Korea
[4]Department of Geology & Geophysics, College of Science, King Saud University, Riyadh 11451, Saudi Arabia
*Corresponding author Email. Biswajeet.Pradhan@ues.edu.au

Abstract

Earthquake prediction is currently the most crucial task required for the probability, hazard, risk mapping, and mitigation purposes. Earthquake prediction attracts the researchers' attention from both academia and industries. Traditionally, the risk assessment approaches have used various traditional and machine learning models. However, deep learning techniques have been rarely tested for earthquake probability mapping. Therefore, this study develops a convolutional neural network (CNN) model for earthquake probability assessment in NE India. Then conducts vulnerability using analytical hierarchy process (AHP), Venn's intersection theory for hazard, and integrated model for risk mapping. A prediction of classification task was performed in which the model predicts magnitudes more than 4 Mw that considers nine indicators. Prediction classification results and intensity variation were then used for probability and hazard mapping, respectively. Finally, earthquake risk map was produced by multiplying hazard, vulnerability, and coping capacity. The vulnerability was prepared by using six vulnerable factors, and the coping capacity was estimated by using the number of hospitals and associated variables, including budget available for disaster management. The CNN model for a probability distribution is a robust technique that provides good accuracy. Results show that CNN is superior to the other algorithms, which completed the classification prediction task with an accuracy of 0.94, precision of 0.98, recall of 0.85, and F1 score of 0.91. These indicators were used for probability mapping, and the total area of hazard (21412.94 Km^2), vulnerability (480.98 Km^2), and risk (34586.10 Km^2) was estimated.

Keywords: earthquake; convolutional neural network; geospatial information systems; hazard; vulnerability; risk; north-east India

1. Introduction

Seismic hazard estimation is an important research topic that is concentrated on probabilistic seismic hazard assessment (PSHA) over several countries worldwide. The PSHA estimation frequently accepts new challenges in enhancing the methodology used for its development since its formation (Cornell, 1968; Kebede and Van Eck, 1996; Veneziano et al., 1984; Kijko and Graham, 1999, 2004; Sabetta et al., 2005; Lindholm and Bungum, 2000, 2003). According to King (1986), radon concentration variation could be regarded as evidence of tectonic disturbances in the earth's crust and could be used as precursors for future earthquakes (Kraner et al., 1968; King and Minissale, 1994; Pearson, 1967; Singh et al., 2014; Virk et al., 2012; Walia et al., 2005; Zmazek et al., 2003). The aforementioned parameters, which are used in geophysical processes for seismic hazard assessment, could change the soil characteristics. Several theoretical and empirical algorithms have been used for seismic hazard assessment to determine the effects of these parameters (Zmazek et al., 2003; Ramola et al., 2008; Choubey et al., 2009). Several studies on seismic hazard assessment have been conducted for the Indian sub-continent using numerous algorithms and techniques (Krishnan, 1959; Guha, 1962; Gubin, 1968; Tandon, 1956). These works were emphasized on the concept of intensity-based zoning and micro-zonation (Jaiswal and Sinha, 2007; Bansal et al., 2013; Verma et al., 2013). PSHA is still regarded as a traditional methodology for hazard assessment. Generally, the PSHA model is based on an inappropriately homogenized catalog of events with many associated uncertainties. Numerous studies were conducted by (Bhatia et al., 1999; Desai and Choudhury, 2014a, b, c, 2015; Jaiswal and Sinha, 2007; Mahajan et al., 2010; Nath and Thingbaijam, 2011; Naik and Choudhury, 2015; Parvez et al., 2003; Sharma, 2003; Sharma and Malik, 2006; Shukla and Choudhury, 2012; Sitharam et al., 2006; Anbazhagan et al., 2016; Rout and Das, 2018; Lindholm et al., 2016) to estimate seismic hazards and continuously improve the methodology.

Das et al. (2016) developed the uniform hazard spectra for northeast (NE) India using a probabilistic approach. NE is considered being one of the seismically most active locations worldwide together with the other five largest seismic zones: Turkey, Taiwan, Mexico, California, and Japan. NE India is located at the zone covered by the Burmese arc toward the east and the Himalayan arc in the northern part (Jaishi et al., 2014; Singh et al., 2014). High seismicity has been observed in NE India due to the complicated tectonics that originated from the collision between the Indian and the Eurasian Plates. The subduction zone originated in the eastern part of India along the Indo-Myanmar Range (Dewey and Bird, 1970). In NE India, the main earthquake-generating faults are Disang and Naga fault, which are both thrust in nature (Jaishi et al., 2014). The Bengal Basin seismicity could be generated due to intraplate activities and events observed in Tripura and Mizoram associated with a plate boundary fold belt. Dauki, Sylhet, Hail-Hakula, Tista, Mat, and Tuipui faults are also responsible for the occurrence of several events in E-W, N-E, NE-SW, and NNW. The most prominent fault is Mat fault in Mizoram state (Jaishi et al., 2014). Hence, they studied the radon anomaly monitoring and correlation with the possibility of earthquake occurrences (Jaishi et al., 2014). Numerous authors have predicted earthquakes based on the precursor using primary analysis of soil radon and thoron anomalies. The multiple regression method was used to differentiate the radon anomalies caused only by seismic events rather than meteorological parameters. Several studies on radon anomaly variation were also conducted for monitoring purposes (Jaishi et al., 2014; Singh et al., 2014).

Sitharam et al. (2015) described the surface-level spatial variation of seismic hazard for India covering the latitude and longitude of 6°–38° N and 68°–98° E, respectively. They claimed that the most recent seismic activity knowledge was applied in India for hazard estimation, which is associated with numerous uncertainties along with the seismicity parameters through several modeling techniques. They also presented the surface level hazard by employing many site amplification factors associated with V_{S30} values estimated from the topographic gradient based on slope values. Furthermore, they estimated the peak horizontal acceleration (PHA) using surface-level spatial variation for the return periods of 475 and 2475 years. Lindholm et al. (2016) proposed a novel PSHA approach for the Indian sub-continent. They employed three different recurrence models, namely, a fault model, a seismic zonation model, and a grid model, to perform PSHA. They finally observed that the peak ground acceleration for 10% exceedance in 50 years for Koyna, Kutch, and Gujrat regions are 0.4 and 0.3 g. They also observed higher ground motion amplitudes in Gujarat than those in the Koyna due to high frequency via comparison. Nathe et al. (2014) performed the seismic risk assessment in the city of Kolkata by using vulnerability exposures, such as land use/cover, building typology, population density, and age. They conducted micro-zonation for the city by integrating geological, seismological, and geotechnical thematic layers and vulnerability components following a logic-tree framework. Finally, they estimated the structural and socioeconomic risks. They classified the damage probabilities into five classes.

In recent years, machine learning techniques are being implemented in several applications to solve real-world problems, specifically in earthquake study. Jena et al., (2020a) conducted an earthquake probability assessment for the Indian subcontinent using deep learning. In a separate work, Jena et al., (2020b) implemented the recurrent neural network (RNN) for the earthquake probability estimation in Odisha, India. Alimoradi et al. (2015) analysed ground motion using machine-learning techniques and achieved excellent results. Schaefer and Wenzel (2019) implemented the multi-variate machine learning method for megathrust earthquake hazard assessment. Besides, many machine-learning methods have been used for geotechnical applications such as landslide susceptibility mapping and other environmental applications (Chen and Li, 2020); Chen et al., 2020; Zhao and Chen, 2020a; Zhao and Chen, 2020b (groundwater spring potential mapping); Fanos and Pradhan, 2018, 2019).

Studies on earthquake probability and hazard assessment in NE India are limited, and almost 70% of the assessment is based on traditional techniques. However, researchers have not performed comprehensive investigations on earthquake probability, hazard, vulnerability, and risk assessment in the NE region. Few studies have been conducted using deep learning and geospatial techniques in India however, no comprehensive study in NE India for earthquake risk assessment. However, for the first time, we conducted a study that will help in mitigation planning. Because the NE India is characterized by complicated tectonics, where a large number of events with magnitudes more than five experienced that makes the region a high hazard zone. Therefore, according to the precursor and probabilistic studies, the seismologist and researchers expect the probability of events with magnitude more than 5Mw could hit the NE that could be a disaster. Thus, continuous probability, hazard, vulnerability, and risk, as well as coping capacity mapping, monitoring, and mitigation planning are required for NE India. Hence, the CNN and Analytical Hierarchy Process (AHP) approaches are combined to create an integrated coping capacity risk map followed by probability, hazard, and vulnerability. This study addresses the following questions: 1) Is it possible to achieve good accuracy in probability mapping without considering the earthquake precursors; 2) How the developed model could successfully predict the events and be applied for hazard mapping; 3) How accurate is the developed risk map and how could it be applied for mitigation planning.

2. Data and Methodology

2.1 Study area

NE India is popularly known as a north-eastern region comprising of various states: Arunachal Pradesh, Meghalaya, Assam, Manipur, Nagaland, Sikkim, Mizoram, and Tripura. The total area of NE India is approximately 262,230 km^2. The total population living in this region is approximately 45,772,188, and the density is 170/km^2 (450/sq. mi). NE of India is divided into four seismogenic source zones: Eastern Syntaxis (zone I), Arakan-Yoma Subduction Belt (zone II), Shillong Plateau (zone III), and the two thrusts, namely, Main Central Thrust (MCT) and Main Boundary Thrust (MBT) (zone IV) (Dutta, 1964). These zones are further divided into nine zones based on tectonics, geology, focal mechanism, and event characteristics (Angelier and Baruah, 2009; Das et al., 2016; Jena and Pradhan, 2019). NE India is a mega-earthquake prone zone due to active faults originating from three major plates, namely, Eurasian, Indian, and Burma Plates. Assam (1897) and Assam–Tibet (1950) earthquakes experienced in this region are considered being the two largest earthquakes in the history of (Mw > 8.0) and many more events with (8.0 > Mw > 7.0), respectively. The Asam-Tibet earthquake is still the largest in India. This earthquake received increased attention from scientists for seismic hazard and risk assessments due to its complicated structure and high seismicity. The nine seismic zones classified for NE India are as follows: North-South Indo Burma fold Belt, NE-SW Indo Burma fold Belt, Sagging Fault region, NW-SE trending feature, Tibetan Plateau, Eastern MCT, Shillong Plateau, Sylhet Fault, and NE-SW trending Structure. The lithology of NE India is characterized by sandstone, shale, limestone, quartzite, conglomerates, phyllites, and volcanic rocks. The study area map is presented in Figure 1.

2.2 Data

The basic input data are an appropriate and reliable earthquake catalog for probability assessment. Wason et al. (2012) proposed a magnitude conversion procedure to convert various magnitudes to moment magnitude. Earthquake data were collected from various databases, such as the National Centre for Seismology (NCS), the National Earthquake Information Center (NEIC), the Global Centroid Moment Tensor (GCMT), and the United States Geological Survey (USGS) for NE India, for the historical events from 1897 until 2019 and applied these data for training and validation in the CNN model. In addition, several thematic indicators were obtained in GIS by creating a database. Some events were also collected from seismological bulletins of the Indian Meteorological Department to complete the catalog. Digital elevation model (DEM), administrative boundary, building information, and land use/cover data were acquired from the DIVA-GIS (https://www.diva-gis.org/) and IGIS map site (https://www.igismap.com/). Hazard, probability, vulnerability, and risk maps were generated using the created databases in GIS. Different algorithms, such as Inverse Distance Weighting (IDW), spline, Euclidian distance, kernel density, and buffer, were used to create several layers for risk assessment. Causal factors and importance of vulnerability layers were obtained on the basis of the literature using AHP and experts' opinions. The details of the data sources, raw data, derived data, their importance, and the procedure of layer derivation are presented in Table 1.

2.2.1 Seismic factors

Magnitude density: The likelihood of occurrence of a specific magnitude earthquake can be understood through cluster analysis. Therefore, magnitude density can help in identifying the high probable zone through the probability distribution analysis (Bathrellos et al., 2017).

Epicenter density: Epicenter zone of earthquakes gives a view of the main and several branches of clusters. Epicenter density can also provide the information of high probable zone (Zebardast, 2013). Through this study, large earthquake clusters, rifto-genesis of structures and the propagation of main fractures can be the focus in hazard modeling (Rashed and weeks, 2003).

Distance from epicenter: With the increase in distance from the clustered epicenter zone, the probability of earthquake occurrence decreases. This gives the information that with an increase in distance from the epicenters, the interconnection of fractures and faults decreases (Pourghasemi et al., 2019).

PGA density: Ground motion information can be understood from PGA associated with tectonic fractures or faults (Kamranzad et al., 2020). This factor provides the information on ground acceleration linked to the lithology, amplification factor, and source to site distance including the magnitude size.

2.2.2 Geotechnical factors

Slope and elevation: Faults are associated with slopes that give fault slip and seismic information found to be in hilly areas more than the plane lands (Xu et al, 2012; Jena et al., 2020). Similarly, with an increase in elevation complicated structures formed interlinked to slopes and increases the probability of earthquakes.

Fault density: The main source zone of events can be identified through high fault density that indicates the complicated tectonic structure (Jena et al., 2020).

Distance from fault: Generally, spatial probability zones are observed near to active faults and the probability decreases with an increase in distance. In the current study, all active faults were included (Alizadeh et al., 2018).

Lithology and amplification factor: Lithology varies in every seismic prone area. Amplification factor is different for all soil and lithotypes that is associated with grain density, compactness and thickness (Dhar et al., 2017). Hard rock has less amplification factor than loose sedimentary rocks.

2.2.3 Exposure factors

Social and structural characteristics: Buildings, transportation nodes and land use areas will be highly vulnerable if situated near to active faults. Transportation nodes are a key factor in earthquake vulnerability study (Alizadeh et al., 2018). Lowering down the building heights, use of good construction materials, land allocation, equally spaced spatial distribution of buildings and proper development plan can reduce vulnerability. Reinforcement of the old vulnerable structures should be the focus. The exposures are highly vulnerable due to earthquakes in NE India. The weights/priority were calculated and presented in Table 2.

2.3 Methodology

The details of the training process of convolutional neural network (CNN) were described mathematically to explain parameter learning. The description was portrayed using the artificial neural network (ANN) technique. The details can be found in the work by (Mitchell, 1997; Han et al., 2018).

2.3.1 Forward propagation

Convolutional neural network (CNN) comprises fully connected, pulling, and convolutional layers and dropouts (Figure 3). However, CNN is quite different from multilayer perceptron neural network (MLPNN) in terms of architecture. Several convolutional kernels, pooling layer, and dropout were used to compute various feature maps. However, the feature map (*j*th) of convolution kernel (*l*th), $x_j^l x_j^l$, can be calculated as follows:

$$a_j^l = \sum_{i=1}^{N^{l-1}} K_{ij}^l \times X_i^{l-1} + b_j^l, x_j^l = f(a_j^l), \qquad (1)$$

where X_I^{L-1} , (*l* -1)th layer, and *i*th feature map could be observed; therefore, $N^{l-1}$is the total number of feature maps for a particular layer. The convolution kernel K_{ij}^l is analogous to the *i*th map in (*l*-1)th layer and *j*th map in *l*th layer, whereb_j^l is considered to be the bias term of the described kernel, and $f\,(\cdot)$ introduces non-linearity into the multi-layer networks that indicate the element-wise non-linear activation function. Sigmoid, ReLU, and tanh are classic activation functions (Glorot et al., 2011). The pooling layer, which was placed after a convolutional layer, aims to reduce parameters, integrate features, and conduct shift invariance by reducing the resolution of feature maps. The pooling function could be introduced as *downsample* $(\cdot)$, where X_j^l is the feature map, and S_j^l could be presented as follows:

$$S_j^l = downsample(X_j^l). \qquad (2)$$

Two typical pooling operations, such as average and max pooling, are generally applied in CNN (Boureau et al., 2010). The pooling operation works as a k × k matrix and results in a single value, which could be the max or the mean of that region. Several fully connected layers were used to focus on mid-level feature map learning after the convolutional layer, followed by the pooling layers, such as AlexNet, LeNet, and Visual Geometry Group (VGG). However, these layers require a large number of weight parameters for a full connection. The feed-forward process of CNN is similar to that of the ANN model, which is formulated as:

$$a_j^l = \sum_{i=1}^{N^{l-1}} X_i^{l-1} W_{ij}^l + b_j^l, x_j^l = f(a_j^l), \qquad (3)$$

where W_{ij}^l denotes weight vector, and b_{ij}^l indicates bias term for the *l*-th layer and *i*-th filter. In a neural network, Softmax activation is applied to the last dense layer that converts the last dense layer output to a probability distribution. Thus, Softmax is used to predict the class if the target class is two. Let o_i and y_i respectively denote the predicted label and the ground-truth label for input data. The loss function could be calculated by:

$$E = 1/2 \sum_{i=1}^{N^L} \|y_i - o_i\|^2, o_i = X_i^l, \qquad (4)$$

where Lth and output layer N^L are the total number of nodes, and E indicates the classification error of all output nodes. Based on the Euclidian distance, the loss function presented in Eq. (4) is also called Euclidean loss. Several other loss estimation alternatives, such as hinge, contrastive, sigmoid cross entropy, information gain, and Softmax losses, are available. Additional details are provided in the work by (Lowe, 1999).

2.3.2 Backward propagation

The error propagation raised in the output to the input layer could be observed in the backward propagation for the optimized label prediction result. Therefore, bias term and weight vectors could be updated again after other layers to reduce these errors (Han et al., 2018; Hecht-Nielsen, 1992). The update of parameters could be formulated as:

$$W_{ij}^{l} = W_{ij}^{l-1} + \eta \frac{\partial E}{\partial W_{ij}^{l-1}}, b_i^{l} = b_i^{l-1} + \eta \frac{\partial E}{\partial b_i^{l-1}}, \quad (5)$$

where learning rate is ηis, and the partial derivatives of the loss functions are $\frac{\partial E}{\partial W_{ij}^{l}}$ and $\frac{\partial E}{\partial b_i^{l}}$ considering W_{ij}^{l} and b_i^{l}, respectively (Han et al., 2018), which can be presented as:

$$\frac{\partial E}{\partial W_{ij}^{l}} = \frac{\partial E}{\partial a_i^{l}} \frac{\partial a_i^{l}}{\partial W_{ij}^{l}}, \ \frac{\partial E}{\partial b_i^{l}} = \frac{\partial E}{\partial a_i^{l}} \frac{\partial a_i^{l}}{\partial b_i^{l}}. \quad (6)$$

Let δ_i^{l} indicate error term on the *l*-th layer in the first part of the right-hand side of Eq. (6), which combines with the second part result. Eq. (6) could be represented as:

$$\frac{\partial E}{\partial W_{ij}^{l}} = \delta_i^{l+1} f'\left(a_i^{l}\right) x_i^{l}, \ \frac{\partial E}{\partial b_i^{l}} = \delta_i^{l+1} f'\left(a_i^{l}\right). \quad (7)$$

If the output layer is *l* + 1 and the *l*th layer is fully-connected, then the δ_i^{l} as the error term can be computed as follows:

$$\delta_i^{l} = \frac{\partial}{\partial a_i^{l-1}} \frac{1}{2} \sum_{i=1}^{N^{l+1}} \|y_i - o_i\|^2 = -(y_i - X_i^{l}) f'\left(a_i^{l-1}\right), \quad (8)$$

where the derivative of the *l*th layer activation function is $f'\left(x_i^{l}\right)$. If all the convolution layers are presented as *l* and *l* + 1, then the error term δ_i^{l} can be computed by following the chain rule as:

$$\delta_i^{l} = \left(\sum_{j=1}^{N^{l+1}} W_{ji}^{l} \, \delta_j^{l+1}\right) f'\left(a_i^{l-1}\right). \quad (9)$$

If the pooling layer is the *l*-th layer and convolution layer is *l* + 1, then the error δ_i^{l} can be computed as (Goh, 1995):

$$\delta_i^{l} = \left(\sum_{j=1}^{N^{l+1}} K_{ji} \times \delta_j^{l+1}\right) f'\left(x_i^{l}\right), \quad (10)$$

where the pooling function is $f\left(x_i^{l}\right)$ and its derivative is $f'\left(x_i^{l}\right)$; the function is linear. Therefore, the last term of Eq. (10) will disappear if the derivative $f'\left(x_i^{l}\right)$ is 1. If the pooling layer is *l*+1 and the *l*-th layer is a convolutional layer, then the δ_i^{l} can be computed as:

$$\delta_i^{l} = upsample(\delta_i^{l+1}) f'\left(x_j^{l}\right), \quad (11)$$

where the upsampling operation is represented by *upsample*(). If the pooling layer in the CNN model acquires mean pooling, then the error is uniformly distributed among the units through upsampling (Shen et al., 2016). If the pooling layer adopts max pooling, then the max receives all the error. However, input through the particular unit would result in output with small changes. The bias term and weight vector can be updated by following the up-down direction through the previous update.

2.3.3 Performance evaluation

Three-phase procedure for parameter learning involves data point embedding and distant-supervised phase, which is also called a pre-training phase to generate noiseless data and final supervised phase (Jiang et al., 2019). The distant-supervised phase is necessary to improve the accuracy of the output prediction classification or the probability distribution. The pre-training phase for datasets is not mandatory in input embedding and unnecessary if the result obtained by the CNN is acceptable and good. Final supervised training requires numerous epochs while the distant phase needs one epoch to train the model on this dataset. Back-propagation is applied to update the weight vector and bias in distant-supervised and supervised training phases (Han et al., 2018).

The classifier's performance can be presented as *accuracy* (Chicco and Jurman, 2020*)*:

$$\text{Accuracy} = \frac{Number\ of\ correctly\ labelled\ samples}{Number\ of\ all\ testing\ samples}. \quad (12)$$

From the harmonic mean of *precision* and *recall*, F-1 score can be computed as (Chicco and Jurman, 2020*)*:

$$\text{F-1}=\left(\frac{precision^{-1}+recall^{-1}}{2}\right)^{-1}. \quad (13)$$

According to the obtained result, the achieved accuracy was 94%. Therefore, the train and test accuracy and loss values were plotted in figure 4.

3. CNN-AHP model execution

3.1 Probability

A sequential CNN model for earthquake classification prediction and probability distribution was applied in the current research (Figure 5). This model comprises four convolutional layers, and each layer comprised a pooling layer and a dropout (Figure 3). The current model shows that a supervised classifier with 70% (training set) and 30% (testing set) of spectrograms was randomly applied for training, and the performance accuracy was estimated based on two-class classification (Gholamy et al., 2018; Chen et al., 2020). The earthquake data were defined as those without any specific condition and split of a large dataset using a 70/30 ratio while 75/30 and 80/20 ratios provide low accuracy and useful for small data set (Jena et al., 2020; Chen et al., 2020). The CNN model was first developed with convolution kernels, pooling layers, and dropouts in a sequential model to predict earthquake and non-earthquakes as 1 and 0, respectively. Earthquake catalog was collected from different databases and random points generated using GIS to train the CNN classifier. Several thematic layers were used to create a training dataset from DEM, shapefiles, and catalog along with target points. Data splitting was performed by dividing into train and test sets. Different algorithms were then applied for normalization, optimization, variable definition, and compilation. A test dataset was applied to predict the values that can be used for probability assessment. Numerous earthquake events were reported in NE India. However, the events were filtered based on magnitudes more than 4 (Mw) and then used for training because low-magnitude events have less capacity for destruction. Proper inspection and data quality assessment facilitated the database creation of 250 earthquakes for two classifications and probability distribution estimation. Adam optimizer was applied to optimize the output and epochs (10,000); batch size (100), validation split (0.3), and verbose (1) were implemented to avoid overfitting. However, this model learns from the data points of indictors associated with earthquake and non-earthquake events. Digitization could create noise in the multivalued data points derived from thematic layers; thus, the noise could affect the model performance, which can be improved by noise removal and pre-processing. Moreover, the model performs well and provides good accuracy in probability mapping generated from the classification prediction results. Table 3 explains the characteristics of all the trainable parameters.

3.2 Hazard

Hazard is the term associated with the spatial and temporal probability of the events. In this work, the hazard map was prepared based on CNN-based probability and intensity level in the study area (Plaza et al., 2019). The intensity map was created by calculating the intensity values from magnitudes. Then, IDW interpolation technique was implemented to make the intensity variation (Bartier and Keller 1996). Next, the Venn-diagram intersection theory was implemented to find out the very high hazard zones, and the quantile classification technique was implemented to classify the hazard zones. This hazard assessment using a combined approach of artificial intelligence with GIS was conducted for the first time in NE India.

$$Z_p = \frac{\sum_{i=1}^{n}\left(\frac{z_i}{d_i^p}\right)}{\sum_{i=1}^{n}\left(\frac{1}{d_i^p}\right)} \qquad (14)$$

where, Z_p is the estimation value of variable z, z_i is the sample value in point I, d_i^p is the distance between estimated to sample point and n is the coefficient that determines weight. The intersection between two layers A (probability) and B (intensity variation), denoted by $A \cap B$.

$$A \cap B = \{x: x \in A \text{ and } x \in B\} \qquad (15)$$

where x is the element of the intersection and for both layers.

3.3 Vulnerability

Six layers were selected for vulnerability assessment because of data unavailability and consistency issue in the AHP approach (Jena et al., 2020). The layers were described in the data section. The relative importance of the factors used for pair-wise comparison is presented in Table 2. Then, by applying the normalization technique, the weight and rank of all the layers were evaluated.

$$AW = l_{max}W \qquad (16)$$

The matrix of pair-wise comparison is A and W indicates the Eigen-vector. The largest Eigenvector is l_{max} whereas X is the eigenvector of A can be calculated as mathematically presented in Eq. (17). In the next step, the weighted sum tool in the GIS is used to make the vulnerability map.

$$(A - l_{max}W) * X = 0 \qquad (17)$$

The consistency index can be estimated as CI by the expression presented below:

$$CI = \frac{(\lambda_{max} - n)}{n-1} \qquad (18)$$

Where the validation parameter is λ_{max}. The consistency index (CI) was used to estimate the consistency of pairwise comparison. The consistency ratio (CR) that is < 0.1 can be accepted for the priority evaluation and the equation mathematically as follows:

$$CR = CI/RI \qquad (19)$$

Vulnerability map was generated in GIS using the priority values of factors derived from AHP (Table 2).

3.4 Coping capacity

The coping capacity map was developed by using the following two categories of data: the number of hospitals and the disaster budget of NE India. Coping capacity was integrated into the hazard and vulnerability indexes, thereby generating the total risk. Afterward, the integrated coping capacity risk map was created by the categorization of the five classes described in the risk section (Figure 10). Based on the experts' opinion weights disaster budget (50%); Mobile

(20%); district (15%) and sub-divisional hospitals (15%) were implemented in the weighted sum tool to estimate the total capacity in NE India.

3.5 Risk

Spatio-temporal probability (hazard) and the specific types of elements at risk were considered to estimate the probability of losses as risk (Jena and Pradhan, 2020). Finally, the risk was estimated by multiplying hazards derived from probability and intensity with vulnerability. The final risk will be the coping capacity-based risk. The detailed process is presented in Figure 4. The expression of risk can be mathematically written as:

$$Risk = \frac{(Hazard * vulnerability)}{coping\ capacity} \quad (20)$$

4. Results

4.1 Probability

The CNN model predicted the probability of occurrence based on two-class classification for future events. The probable areas were estimated and located through GIS, and the percentage of high probable zones is described in Figure 6. Very high to medium probable zones cover the entire NE of India and contribute to the active tectonics of that region. The probability zones were not classified because the probability map indicates that the entire NE India is highly probable for earthquakes presented as 0-1 (low to high). Arunachal Pradesh is the only state that comes under low to high probability. The rest of the states (Assam, Meghalaya, Manipur, Mizoram, Nagaland, and Tripura) fall in high probable zones covering a total population of approximately 45,588,381 living in these zones as per the recent census data. A total of 95% of NE India falls in very high probable zones, while 5% of area covers the low probable zones because of presence of seismically active faults with many earthquake occurrences. The prediction accuracy was 0.94. The model achieved a precision rate of 0.98, recall value of 0.85 and F1 score of 0.91. The published earthquake hazard map of India by the Geological Survey of India (GSI) can be used for validation purpose.

4.2 Hazard

The degree of spatial variation of earthquake hazard in the NE of India was developed. Therefore, an intensity level of more than 5 could be regarded as a hazardous zone. The intensity map is presented in Figure 7. Intensity level is very high in the regions of Bhutan and Myanmar, while the NE and central part of the region is under low to medium category. Next, the hazard map was categorized into five classes based on intensity: very high (>9), high (8–9), moderate (6–8), low (5–6), and very low (<5). Hazard results indicate that approximately 7.6% (21412.94 km^2) of NE India is classified as a very-high hazard zone while 67.37% (189717.97 km^2) is in a high hazard zone (Figure 8) and (Table 4). Conversely, 0.64% (1802.84 km^2) and 7.01% (19745.02 km^2) of the study region were classified as very-low and low hazard zones, accordingly. Most of these areas are located in the south- and north-western parts, while north-eastern parts of NE India are under the very low zone. The entire Manipur state is classified as a very high hazard zone; Mizoram, Assam, Meghalaya, Nagaland are covered by a high hazard zone. However, Arunachal Pradesh and Tripura are covered by moderate to very low hazard zones based on the obtained results.

4.3 Vulnerability

Several criteria were utilized as input data to assess the vulnerability of communities and land use/cover (Figure 2). An earthquake vulnerability map was developed and categorized into five classes based on quantile classification technique (Figure 9). The developed map signifies that approximately 78.57% of the area is under very high to moderate vulnerability, while low and very low areas covered 21.43% of the region. The vulnerability index was obtained from the processing of six criteria. Approximately, 22.57% (6358386.73 km^2) and 0.2% (48097.91 km^2) of the total area are covered by high and very high vulnerable zones, respectively. However, 55.83% (15,720,551.43 km^2), 0.03% (6752.68 km^2), and 21.4% (6,027,317.89 km^2) of the area are respectively covered by moderate, very low, and low vulnerable zones as presented in Table 4.

4.4 Coping capacity

Coping capacity varies state wise in NE India. People in these areas have access to hospitals and are educated in terms of disasters. Some specific states that are under moderate to high coping capacity are Assam, Sikkim, and Arunachal Pradesh; however, other states fall under low to very-low coping capacity. By contrast, Manipur state of the entire NE region of India is characterized by zones of low to very-low coping capacity and falls under a very-high hazard zone. Thus, assimilating the coping capacity is critical to deriving the real risk scenario. By contrast, the areas in Tripura and Manipur with low coping capacity and very-high vulnerability are due to the combined influence of the disaster budget and the total number of hospitals. Figure 10 demonstrates the coping capacity of NE India.

4.5 Risk

The earthquake risk was estimated, mapped, and depicted spatially in Figure 11. The risk map was classified based on quantile classification techniques and presented as very-low, low, moderate, high, and very-high. Risk results indicate that 15.64% (34586.10 km^2) of the area was regarded as a very-high-risk zone while the high-risk zone comprised 26.15% (57856.74 km^2) of the area. The two classes of risk zones are located in the south-eastern and western parts of the study area. High and very-high earthquake risks could be observed in Mizoram, Manipur, Nagaland, Meghalaya, and Sikkim states (Table 4). Medium, very-low, and low risk zones cover approximately 42.40% (9379687.71 km^2), 15.82% (3499487.21 km^2), and 27.29% (6037551.23 km^2) of area, respectively. Assam state could be regarded as moderate risk zones, and some parts of Arunachal Pradesh are under moderate and low-risk zones because it is located in the interior part of the study region. The work of Pandey et al. (2017), which shows the total events and dense clustering in NE India, is adopted for risk map validation (Figure 12). According to their map and the seismic hazard zonation map of India, the risk result is accurate.

5. Discussion

The seismicity rate can be the main indicator to estimate the distribution of earthquake probability. However, the seismicity rate depends on the total number of events in a particular area for a given time. Toda et al. (2008) proposed a method that assumed a time window of 10 years for seismicity rate in an area of 100×100 km^2. The current study used a complete seismicity catalog to train a CNN model to identify the location of earthquake probability based on nine indicators. The earthquake data were defined as those without any specific condition of stress disturbance and split of a large dataset using a 70/30 ratio while 75/30 and 80/20 ratios provide low accuracy and useful for small datasets (Jena et al., 2020; Chen et al., 2020). According to the probability distribution study, 249070 km^2 of the NE region falls under a high probable zone. The reasons could be high epicenter density with several high magnitude events and intensity. Specifically, high fault density, along with folds and active faults and complicated geological structures contribute to the probability of NE region than any other locations in India (Jena and Pradhan, 2020). These areas fall under the eastern part of the Himalayan collision zone, generating several strike-slip faults. Therefore, due to high magnitude events, high intensity events are observed in the central parts and very-high in the SE parts, characterized by sedimentary rocks, ophiolites and populated areas with low elevation, high fault density, and frequent events (Rout and Das, 2018). According to the hazard map, a considerable area in the northern part and coastline of NE India was classified as a very-high hazard zone because of frequent and high-intensity events. The low to very-low classes cover the hilly regions with fewer faults and events in the study site. Therefore, the clear view of some populated locations in the northern part of the region indicates that north-western parts are under a very-high hazard zone. According to the vulnerability results, south and northern parts fall near the active faults and are considered being moderate to highly vulnerable to earthquakes. By contrast, low elevation, high population density, gentle slope, high rail density, and high land use/cover in these areas are responsible for the very high vulnerability in the described zones. The highly vulnerable zones are attributed to a high level of dependent population, high land use density, less distance from land use to the epicenters, unsafe sanitation systems, and railway-dependent population. Areas with low and very low vulnerability comprise good socio-economic conditions. They are not closely exposed to the high magnitude earthquake locations. Nevertheless, coping capacity is a game-changer during earthquake periods (Hoque et al., 2019). An educated society can effectively cope with vulnerability. The coping capacity in the Assam, Sikkim and Arunachal Pradesh is high because of recently established hospitals and mobile hospitals and a good education system after the devastating effects of several earthquakes and active faults. Therefore, the education system in NE is superior to that of the previous condition. This superiority is attributed to the knowledge of the measures that must be taken during and after the events and its application on coping. Without coping capacity, the risk map can still be produced, but the resulting outcome will be different. Furthermore, this outcome cannot be considered as the actual risk. However, low to very-low-risk zones, which have sufficient disaster budgets and hospitals and mitigation measures, could be found in the northern part of the region. Areas close to Myanmar should be the focus of earthquake mitigation planning. Consistency was observed in the spatial distribution of risk assessment results, in which the hazard, vulnerability, and degree of coping capacity were linked. Locations with dense population and land use, low elevation, steep slope, high fault density and epicenter, and magnitude distribution with less coping capacity index fall under high-risk areas in NE India. However, some areas with high risk could be changed to low because of their status mitigation capacity and proper planning (Hoque et al., 2018; Jena et al., 2020). Furthermore, the validation approach and the analysis confirm that the developed model could provide reliable and accurate information on population risk. The coping capacity was integrated with the vulnerability and hazard to produce the total risk.

The advantages and disadvantages of the proposed integrated model deal with the implementation, application type, data quality. This regional earthquake study using a robust technique of CNN model and multi-criteria assessment could provide a detailed and accurate risk result. This model could provide the knowledge to choose the necessary criteria under each component for probability, hazard, vulnerability, and risk assessment through CNN and GIS. The AHP is applied for vulnerability assessment, which is effective for prioritizing criteria based on the multi-criteria decision-making process, to calculate the weights. AHP provides the best solution for priority analysis and the most used multi-criteria decision making (MCDM) in academia and industries. This study gives evidence of comprehensive risk assessment using the integrated geospatial and AHP approaches and efficient for risk assessment at the regional scale to estimate accurate information. However, incorporation of mitigation measures are required for the development of the actual risk map through a proper risk assessment procedure.

A certain number of disadvantages are associated with this model. The CNN model requires large data points for an effective study on earthquake probability distribution. The CNN model is data-dependent and requires a huge number of data points for training and testing purposes. Choosing proper parameters for probability mapping is crucial otherwise may lead to a biased result. The AHP approach is limited to the magic number of 7 (+ or -2) and has consistency issues. Therefore, more than seven criteria cannot be involved in vulnerability assessment.

6. Conclusion

A deep learning-based integrated earthquake risk-mapping model for NE India using a complete earthquake catalog, DEM and shapefile data, and spatial analysis is proposed in this research. The chosen area is NE of India, which is characterized by 262,230 km^2 and falls under the Indian government. This area is selected to test the usefulness and applicability of the proposed approach. The risk mapping approach is validated using the earthquake hazard map created by previous researchers. The risk results indicated that 15.64% (34586.10 km^2) of the area was regarded as a very high-risk zone while the high-risk zone comprised 26.15% (57856.74 km^2) of the area falls under SE and SW parts.

The limitations and challenges of this study associated with acquiring data at a regional scale and processing through deep learning techniques, which is difficult. Therefore, secondary data was used because of the unavailability of the primary data. In the future, high-resolution DEM derived from Light Detection and Ranging (LiDAR) data could be generated for earthquake studies to fulfil the requirement of high-quality data. Curvature is not included in the current research for probability mapping as a “criteria” due to its less accuracy. Similarly, this research is limited to earthquake risk assessment without considering liquefaction factors, soil characteristics, fault characteristics, and precursors due to data unavailability. The aforementioned criteria could be considered for future earthquake prediction and probability assessment. Therefore, future works will be focused on addressing the aforementioned limitations. Despite the drawbacks presented in this study, the proposed method is still considered being effective for earthquake risk assessment and could help in efficient disaster risk reduction measures. This method could also be applied to any other disaster in large-scale data modification. Criteria selection was based on site-specific data types; thus, this model could be tested and validated for any other locations in India. The findings of the current research establish a framework for probability, hazard, vulnerability, coping capacity and risk mapping. Planners, administrators, and decision-makers could use the developed model for prevention and mitigation purposes to minimize expected losses for future risk.

Acknowledgment

This research was fully funded by the Center for Advanced Modeling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney. This research was also supported by Researchers Supporting Project number RSP-2020/14, King Saud University, Riyadh, Saudi Arabia. Authors would like to thank the three anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

References

Alimoradi, A., Beck, J.L., 2015. Machine-learning methods for earthquake ground motion analysis and simulation. *J. Eng. Mech. 141*(4), 04014147.

Alizadeh, M., Ngah, I., Hashim, M., Pradhan, B., Pour, A.B., 2018. A hybrid analytic network process and artificial neural network (ANP-ANN) model for urban earthquake vulnerability assessment. *Remote Sens-Basel 10*(6), 975.

Anbazhagan, P., Sreenivas, M., Ketan, B., Moustafa, S.S.R., Al-Arifi, N.S.N., 2016. Selection of ground motion prediction equations for seismic hazard analysis of peninsular India. *J. Earthq. Eng. 20*, 699-737.

Angelier, J., Baruah, S., 2009. Seismotectonics in Northeast India: a stress analysis of focal mechanism solutions of earthquakes and its kinematic implications. *Geophys. J. Int. 178*, 303-326.

Bansal, B.K., Verma, M. 2013. Science and technology based earthquake risk reduction strategies: The Indian scenario. *Acta Geophys. 61*, 808-821.

Bartier, P.M., Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Comput. Geosci. 22*(7), 795-799.

Bathrellos, G.D., Skilodimou, H.D., Chousianitis, K., Youssef, A.M., Pradhan, B., 2017. Suitability estimation for urban development using multi-hazard assessment map. *Sci. Total Environ. 575*,119-134.

Bhatia, S.C., Kumar, M.R., Gupta, H.K., 1999. A probabilistic seismic hazard map of India and adjoining regions. *Annal di Geofisica.* 42, 1153-1164, http://hdl.handle.net/2122/1382.

Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J., 2010, June. Learning mid-level features for recognition, In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2559-2566.

Chen, W., Li, Y., 2020. GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models. *Catena 195*, 104777.

Chen, W., Zhao, X., Tsangaratos, P., Shahabi, H., Ilia, I., Xue, W., Wang, X., Ahmad, B.B., 2020. Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. *J. Hydrol*. *583*, 124602.

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics 21*(6), 1-13.

Choubey, V.M., Kumar, N., Arora, B.R., 2009. Precursory signatures in the radon and geohydrological borehole data for M4. 9 Kharsali earthquake of Garhwal Himalaya. *Sci. Total Environ. 407*, 5877-5883.

Cornell, C.A., 1968. Engineering seismic risk analysis. *B. Seismol. Soc. Am. 58*, 1583-1606.

Das, R., Sharma, M.L., Wason, H.R., 2016. Probabilistic seismic hazard assessment for northeast India region. *Pure Appl. Geophys*. *173*, 2653-2670.

Desai, S.S., Choudhury, D., 2014a. Spatial variation of probabilistic seismic hazard for Mumbai and surrounding region. *Nat. Hazards 71*, 1873-1898.

Desai, S.S., Choudhury, D., 2014b. Deaggregation of seismic hazard for two ports in Mumbai metropolitan region, In: *Advances in Soil Dynamics and Foundation Engineering, Geo-Shanghai International conference* 2014, Shanghai, China, 62-71.

Desai, S., Choudhury, D., 2014c. Deterministic seismic hazard analysis for greater Mumbai, India, In *Geo-Congress 2014: Geo-characterization and Modeling for Sustainability,* Atlanta, Georgia, United States, 389-398.

Desai, S.S., Choudhury, D., 2015. Site-specific seismic ground response study for nuclear power plants and ports in Mumbai. *Nat. Hazards Rev. 16*, 04015002.

Dewey, J.F., Bird, J.M., 1970. Mountain belts and the new global tectonics. *J. Geophys. Res. 75*, 2625-2647.

Dhar, S., Rai, A.K., Nayak, P., 2017. Estimation of seismic hazard in Odisha by remote sensing and GIS techniques. *Nat. Hazards 86*(2), 695-709.

Dutta, T.K., 1964. Seismicity of Assam-zones of tectonic activity. *Bulletin of Natural Geophysical Research Inst*itute *2*, 152-163.

Fanos, A.M., Pradhan, B., 2019. A Novel Hybrid Machine Learning-Based Model for Rockfall

Source Identification in Presence of Other Landslide Types Using LiDAR and GIS. *Earth Syst. Environ. 3*(3), 491-506.

Fanos, A.M., Pradhan, B., 2018. Laser scanning systems and techniques in rockfall source identification and risk assessment: a critical review. *Earth Syst. Environ. 2*(2), 163-182.

Gholamy, A., Kreinovich, V., Kosheleva, O., 2018. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (Cs)*, University of Texas at El Paso, pp.6.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315-323.

Gubin, I. E., 1968. Seismic zoning of Indian Peninsula, *Bulletin of International Institute of Seismological Earthquake Engineering* 5, 109-139.

Guha, S. K., 1962. Seismic regionalization of India, In: Proceedings of 2nd Symposium, *Earthquake Engineering*, Roorkee, 191-207.

Goh, A.T., 1995. Back-propagation neural networks for modeling complex systems. *Artif. Intell. Eng*. *9*(3), 143-151.

Han, D., Liu, Q., Fan, W., 2018. A new image classification method using CNN transfer learning and web data augmentation. *Expert Syst. Appl. 95*, 43-56.

Hecht-Nielsen, R., 1992. Theory of the backpropagation neural network, In: *Neural networks for perception*, Academic Press, 65-93.

Hoque, M.A.A., Tasfia, S., Ahmed, N., Pradhan, B., 2019. Assessing spatial flood vulnerability at Kalapara Upazila in Bangladesh using an analytic hierarchy process. *Sensors-Basel 19*(6), 1302.

Jaishi, H.P., Singh, S., Tiwari, R.P., Tiwari, R.C., 2014. Correlation of radon anomalies with seismic events along Mat fault in Serchhip District, Mizoram, India. *Appl. Radiat. Isotopes 86*, 79-84.

Jaiswal, K., Sinha, R., 2007. Probabilistic seismic-hazard estimation for peninsular India. *B. Seismol. Soc Am. 97*(1B), 318-330.

Jena, R., Pradhan, B., 2019. A Model for Visual Assessment of Fault Plane Solutions and Active Tectonics Analysis Using the Global Centroid Moment Tensor Catalog. *Earth Syst. Environ*. 4,197–211.

Jena, R., Pradhan, B., 2020. Integrated ANN-cross-validation and AHP-TOPSIS model to improve earthquake risk assessment. *Int. J. Disast. Risk Re.* 101723.

Jena, R., Pradhan, B., Al-Amri, A., Lee, CW., Park H-J., 2020. Earthquake Probability Assessment for the Indian Subcontinent Using Deep Learning. *Sensors-Basel* 20(16), 4369.

Jena, R., Pradhan, B., Alamri, AM., 2020. Susceptibility to Seismic Amplification and Earthquake Probability Estimation Using Recurrent Neural Network (RNN) Model in Odisha, India. *Appl. Sci-Basel* 10(15), 5355.

Jena, R., Pradhan, B., Beydoun, G., Sofyan, H., Affan, M., 2020. Integrated model for earthquake risk assessment using neural network and analytic hierarchy process: Aceh province, Indonesia. *Geosci. Front. 11*(2), 613-634.

Jiang, X., Wang, Y., Liu, W., Li, S., Liu, J., 2019. Capsnet, cnn, fcn: Comparative performance evaluation for image classification. *Int. J. Mach. Learn. Com. 9*(6), 840-848.

Kamranzad, F., Memarian, H., Zare, M., 2020. Earthquake Risk Assessment for Tehran, Iran. *ISPRS Int. J. Geoinf. 9*(7), 430.

Kebede, F., Van Eck, T., 1997. Probabilistic seismic hazard assessment for the Horn of Africa based on seismotectonic regionalisation. *Tectonophysics 270*, 221-237.

Kijko, A., Graham, G., 1999. "Parametric-historic" Procedure for Probabilistic Seismic Hazard Analysis Part II: Assessment of Seismic Hazard at Specified Site. *Pure Appl. Geophys. 154*, 1-22.

King, C.Y., 1986. Gas geochemistry applied to earthquake prediction: An overview. *J. Geophys. Res-Sol Ea. 91*, 12269-12281.

King, C.Y., Minissale, A., 1994. Seasonal variability of soil-gas radon concentration in central California. *Radiat. Meas. 23*, 683-692.

Kraner, H.W., Schroeder, G.L., Evans, R.D., 1968. Measurements of the effects of atmospheric variables on $ sup 222$ rn flux and solid-gas concentrations, In: *Aip Conf Proc*., Massachusetts Inst. of Tech., Cambridge, 191-215

Krishna, J., 1959. February. Seismic zoning of India, In: *Earthquake Engineering Seminar, Roorkee University, India*, 32-38.

Lindholm, C.D., Bungum, H., 2000. Probabilistic seismic hazard: a review of the seismological frame of reference with examples from Norway. *Soil Dyn. Earthq. Eng. 20*, 27-38.

Lindholm, C., Parvez, I.A., Kühn, D., 2016. Probabilistic earthquake hazard assessment for Peninsular India. *J. Seismol. 20*, 629-653.

Lowe, D.G.,1999. September. Object recognition from local scale-invariant features, In: *Proceedings of the International Conference on Computer Vision*, Kerkyra, Corfu, Greece, 1150-1157.

Mahajan, A.K., Thakur, V.C., Sharma, M.L., Chauhan, M., 2010. Probabilistic seismic hazard map of NW Himalaya and its adjoining area, India. *Nat. Hazards 53*, 443-457.

Mitchell, T.M., 1997. Artificial neural networks. *Mach. Learn. 45*, 81-127.

Nath, S.K., Adhikari, M.D., Maiti, S.K., Devaraj, N., Srivastava, N., Mohapatra, L.D., 2014. Earthquake scenario in West Bengal with emphasis on seismic hazard microzonation of the city of Kolkata, India. *Nat. Hazard Earth Sys. 14*, 2549.

Nath, S.K., Thingbaijam, K.K.S., 2011. Peak ground motion predictions in India: an appraisal for rock sites. *J. Seismol. 15*, 295-315.

Naik, N., Choudhury, D., 2015. Deterministic seismic hazard analysis considering different seismicity levels for the state of Goa, India. *Nat. Hazards 75*, 557-580.

Pandey, A.K., Chingtham, P., Roy, P.N.S., 2017. Homogeneous earthquake catalogue for Northeast region of India using robust statistical approaches. *Geomat. Nat. Haz. Risk 8*, 1477-1491.

Parvez, I.A., Vaccari, F., Panza, G.F., 2003. A deterministic seismic hazard map of India and adjacent areas. *Geophys J Int. 155*, 489-508.

Pearson, J.E., 1967. Natural environmental radioactivity from radon 222. *US National Center for Radiological Health, USA, 1-31.*

Plaza, F., Salas, R., Nicolis, O., 2019. Assessing Seismic Hazard in Chile Using Deep Neural Networks, *In: Natural Hazards-Risk, Exposure, Response, and Resilience*, IntechOpen. DOI: 10.5772/intechopen.83403.

Pourghasemi, H.R., Gayen, A., Panahi, M., Rezaie, F., Blaschke, T., 2019. Multi-hazard probability assessment and mapping in Iran. *Sci. Total Environ. 692*, 556-571.

Pradhan, B., Abokharima, M.H., Jebur, M.N., Tehrany, M.S., 2014. Land subsidence susceptibility mapping at Kinta Valley (Malaysia) using the evidential belief function model in GIS. *Nat. Hazards* 73(2), 1019-1042.

Ramola, R.C., Prasad, Y., Prasad, G., Kumar, S., Choubey, V.M., 2008. Soil-gas radon as seismotectonic indicator in Garhwal Himalaya. *Appl. Radiat Isotopes 66*, 1523-1530.

Rashed, T., Weeks, J., 2003. Assessing vulnerability to earthquake hazards through spatial multicriteria analysis of urban areas. *Int. J. Geogr. Inf. Sci. 17*(6), 547-576.

Rout, M.M., Das, J., 2018. Probabilistic seismic hazard for Himalayan region using kernel estimation method (zone-free method). *Nat. Hazards 93*, 967-985.

Sabetta, F., Lucantoni, A., Bungum, H., Bommer, J.J., 2005. Sensitivity of PSHA results to ground motion prediction relations and logic-tree weights. *Soil Dyn. Earthq. Eng. 25*, 317-329.

Schaefer, A.M., Wenzel, F., 2019. Global megathrust earthquake hazard-maximum magnitude assessment using multi-variate machine learning. *Front. Earth Sci. 7*, p.136.

Sharma, M.L., 2003. Seismic hazard in the northern India region. *Seismol. Res. Lett. 74*, 141-147.

Sharma, M.L., Malik, S., 2006. Probabilistic seismic hazard analysis and estimation of spectral strong ground motion on bed rock in northeast India, In: *4th international conference on earthquake engineering, Taipei, Taiwan*, 12-13.

Shukla, J., Choudhury, D., 2012. Estimation of seismic ground motions using deterministic approach for major cities of Gujarat. *Nat. Hazard Earth Sys. 12*, 2019.

Shen, L., Lin, Z., Huang, Q., 2016. Relay backpropagation for effective learning of deep convolutional neural networks, In: *European Conference on Computer Vision*, 467-482.

Singh, S., Prasad J.H., Tiwari, R., Chandra T, R., 2014. Variations of soil radon concentrations along Chite Fault in Aizawl district, Mizoram, India. *Radiat Prot. Dosim. 162*, 73-77.

Sitharam, T.G., Anbazhagan, P., Raj, K.G., 2006. Use of remote sensing and seismotectonic parameters for seismic hazard analysis of Bangalore. *Nat. Hazard Earth Sys*. 6, 927-939.

Sitharam, T.G., Kolathayar, S., James, N., 2015. Probabilistic assessment of surface level seismic hazard in India using topographic gradient as a proxy for site condition. *Geosci. Front. 6*, 847-859.

Toda, S., Lin, J., Meghraoui, M., Stein, R.S., 2008. 12 May 2008 M= 7.9 Wenchuan, China, earthquake calculated to increase failure stress and seismicity rate on three major fault systems. *Geophys. Res. Lett. 35*, L17305, doi:10.1029/2008GL034903.

Veneziano, D., Cornell, C.A., O'hara, T., 1984. Historic method for seismic hazard analysis. *Electric Power Research Institute, Report, Palo Alto, CA, NP-3438*.

Verma, M., Bansal, B.K., 2013. Seismic hazard assessment and mitigation in India: an overview. *Int. J. Earth Sci. 102*, 1203-1218.

Virk, H.S., Walia, V., Sharma, A.K., Kumar, N., Kumar, R., 2012. Correlation of radon anomalies with microseismic events in Kangra and Chamba valleys of NW Himalaya. *Geofis. Int. 39*, 221-227.

Walia, V., Virk, H.S., Yang, T.F., Mahajan, S., Walia, M., Bajwa, B.S., 2005. Earthquake prediction studies using radon as a precursor in NW Himalayas, India: a case study. *Terr. Atmos. Ocean Sci. 16*, 775.

Wason, H.R., Das, R., Sharma, M.L., 2012. Magnitude conversion problem using general orthogonal regression. *Geophys. J. Int. 190*, 1091-1096.

Xu, C., Dai, F., Xu, X., Lee, Y.H., 2012. GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. *Geomorphology 145*, 70-80.

Zebardast, E., 2013. Constructing a social vulnerability index to earthquake hazards using a hybrid factor analysis and analytic network process (F'ANP) model. *Nat. Hazards*. *65*(3), 1331-1359.

Zhao, X., Chen, W., 2020a. GIS-based evaluation of landslide susceptibility models using certainty factors and functional trees-based ensemble techniques. *Appl. Sci-Basel*. *10*(1), 16.

Zhao, X., Chen, W., 2020b. Optimization of Computational Intelligence Models for Landslide Susceptibility Evaluation. *Remote Sens-Basel*. *12*(14), 2180.

Zmazek, B., Todorovski, L., Džeroski, S., Vaupotič, J., Kobal, I., 2003. Application of decision trees to the analysis of soil radon data for earthquake prediction. *Appl. Radiat. Isotopes. 58*, 697-706.

Figure captions

Figure 1. a) Indian subcontinent b) NE India with districts, c) Location of the NE India showing the tectonics and detailed geology (JTr: Triassic and Jurassic rocks, Jms: Jurassic metamorphic and sedimentary rocks, Jks: Jurassic and Cretaceous sedimentary rocks, Ks: Cretaceous sedimentary rocks, MzPz: Paleozoic and Mesozoic metamorphic rocks, Mzi: Mesozoic intrusive rocks, N: Neogene sedimentary rocks, Osm: Ordovician metamorphic and sedimentary rocks, Pg: Paleogene sedimentary rocks, Pr: Permian rocks, Pz: undifferentiated

Paleozoic rocks, Pzi: Paleozoic igneous rocks, Pzl: Lower Paleozoic rocks, Pzu: Upper Paleozoic metamorphic rocks, PzPc: Paleozoic undivided Precambrian rocks, Q: Quaternary sediments, Qs: Quaternary sand, S: Silurian rocks, TKim: Cretaceous and Tertiary igneous and metamorphic rocks, TKs: Cretaceous and Tertiary sedimentary rocks, TKv: Cretaceous and Tertiary volcanic rocks, Ti: Tertiary igneous rocks, TrCs: Upper Caboniferious–Lower Triassic sedimentary rocks, Trms: Triassic igneous and sedimentary rocks, Ts: Tertiary sedimentary rocks, and Pc: Precambrian rocks). *(Data source: USGS).*

Figure 2. Criteria used for earthquake vulnerability map.

Figure 3. Architecture of the proposed CNN model.

Figure 4. Accuracy and loss for training and testing.

Figure 5. Methodological flowchart of the proposed method for earthquake risk assessment.

Figure 6. Earthquake probability map.

Figure 7. Intensity variation in NE India.

Figure 8. Earthquake hazard map of NE India.

Figure 9. Earthquake vulnerability map.

Figure 10. Coping capacity of NE India.

Figure 11. Earthquake risk map.

Figure 12. Correlation between a) earthquake risk in NE India and b) earthquake cluster zones with large events (Adopted from Pandey et al., 2017).

Table captions

Table 1. Description of parameters and data source.

Table 2. Priority and rank estimation for all the parameters of vulnerability.

Table 3. Parameters used for CNN method and accuracy in probability mapping.

Table 4. Hazard, vulnerability and risk areas in NE India.

Figure 1a

Click here to access/download;Figure;Figure 1(a).jpg

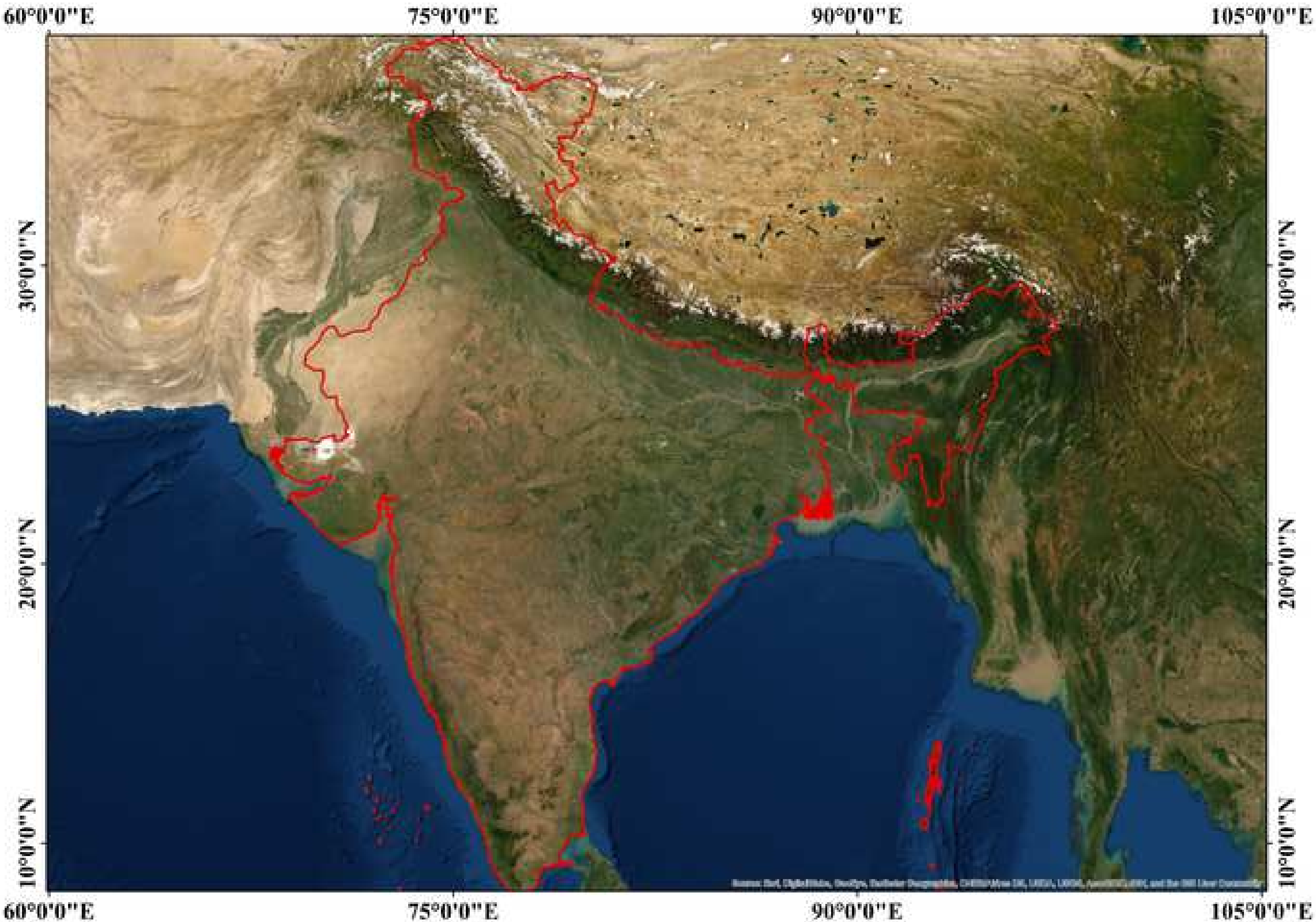


Figure 1b

Click here to access/download;Figure;Figure 1(b).jpg

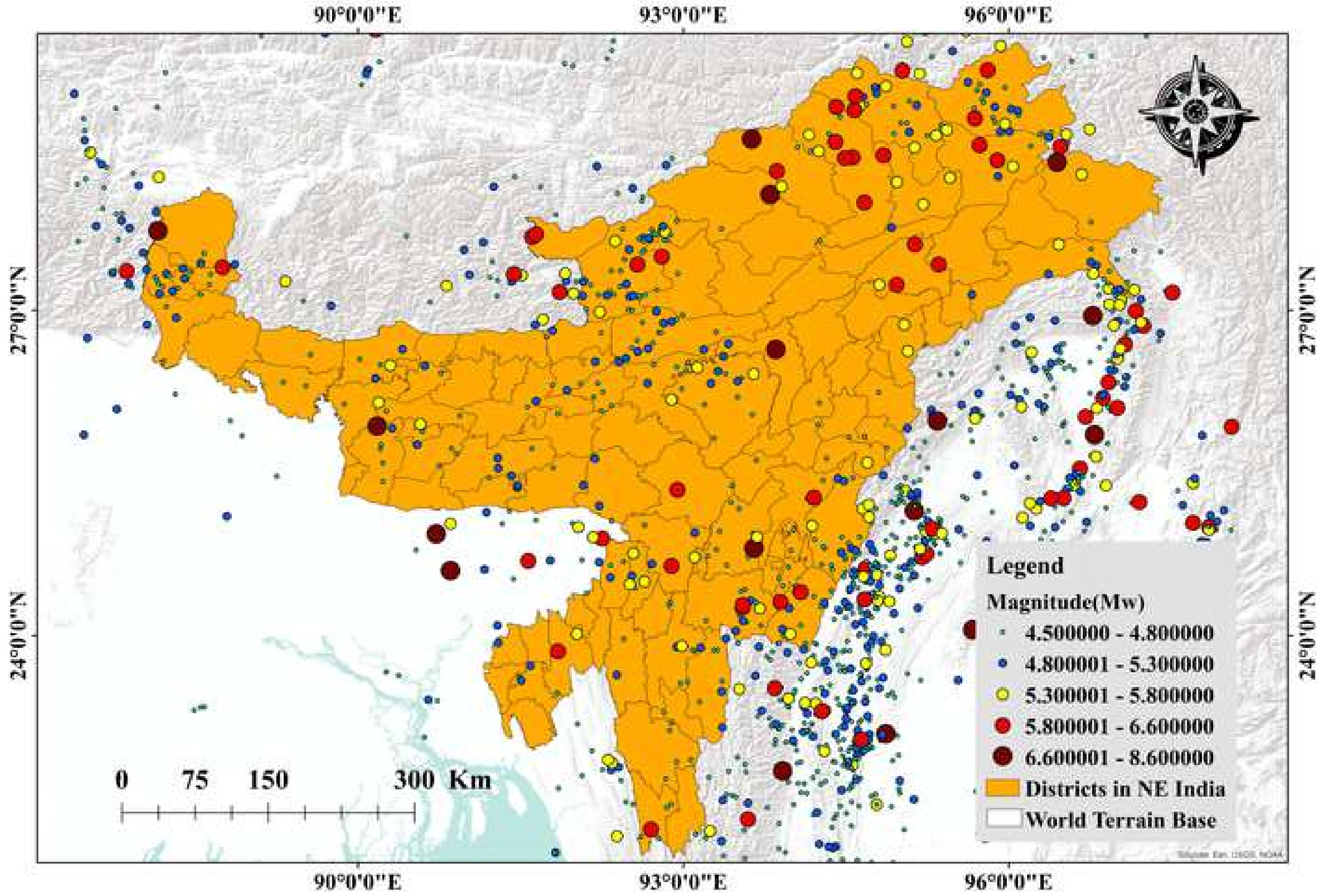


Figure 1c

Click here to access/download;Figure;Figure 1(c).jpg

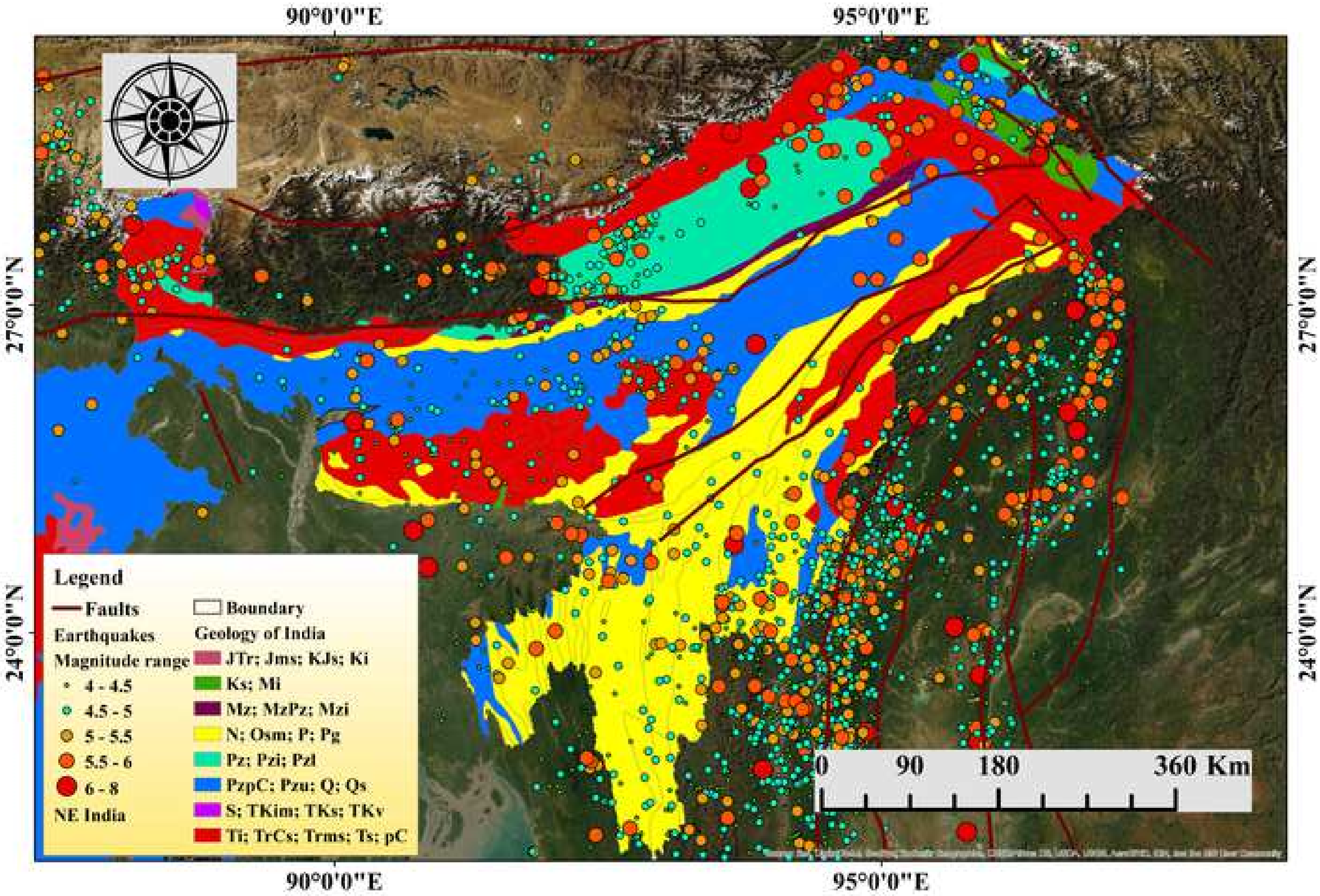


Figure 2

Click here to access/download;Figure;Figure 2.jpg

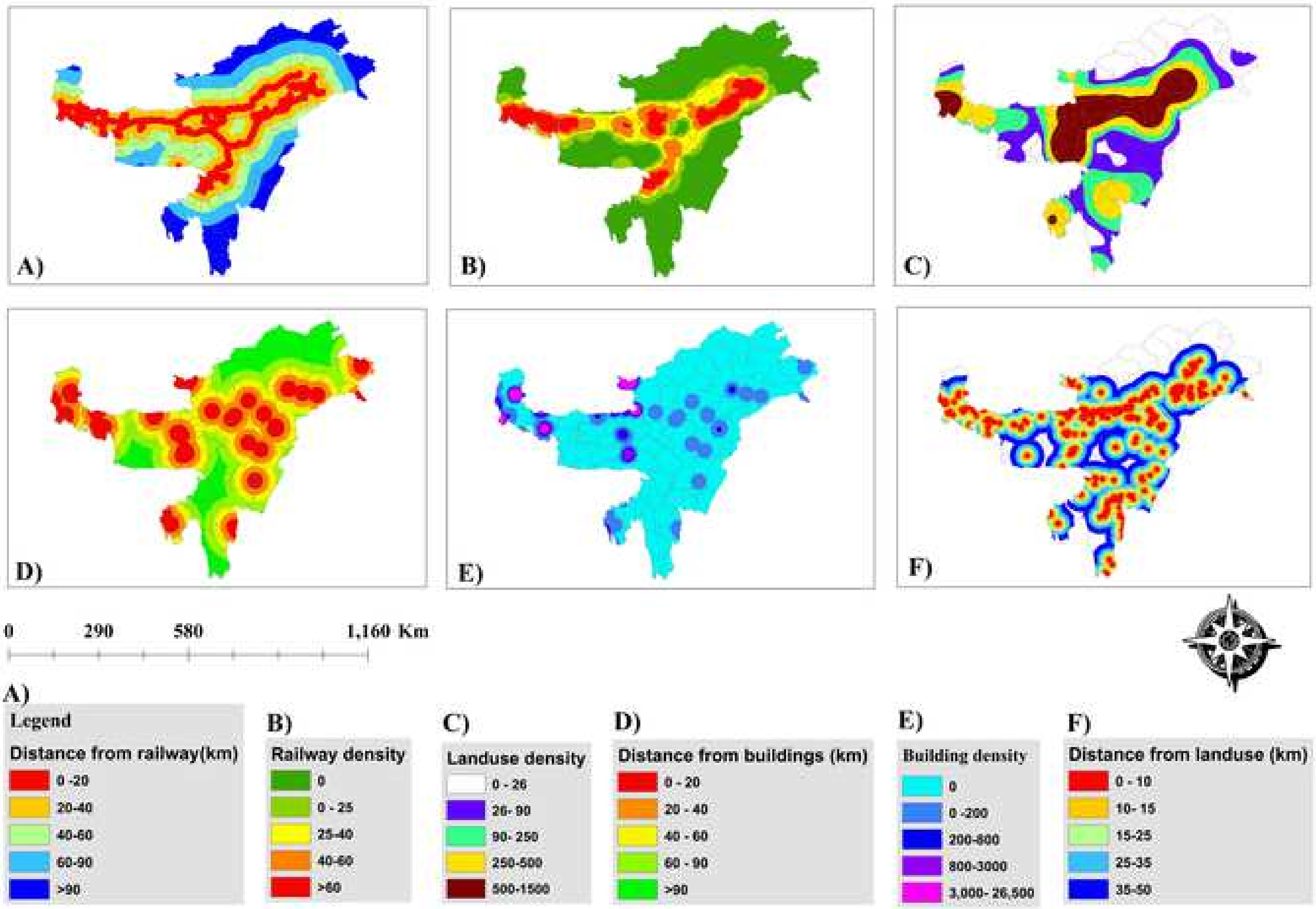


Figure 3

Click here to access/download;Figure;Figure 3.jpg

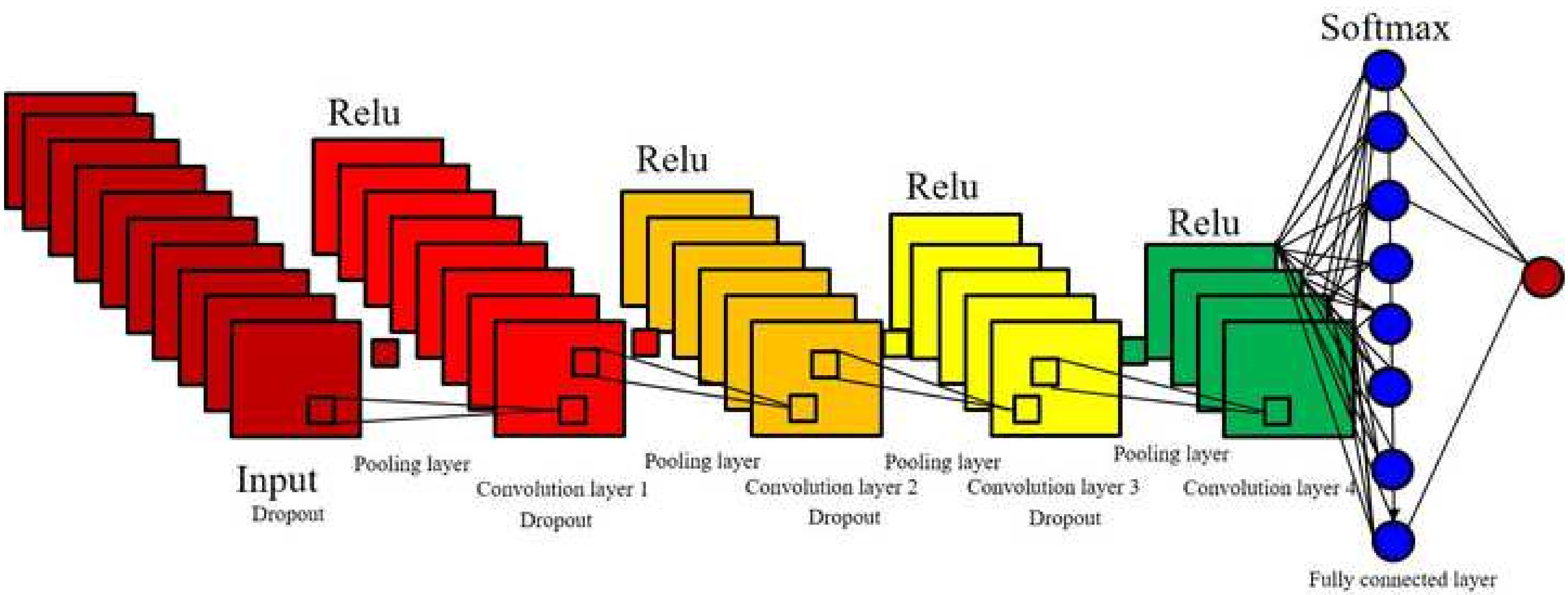


Figure 4a

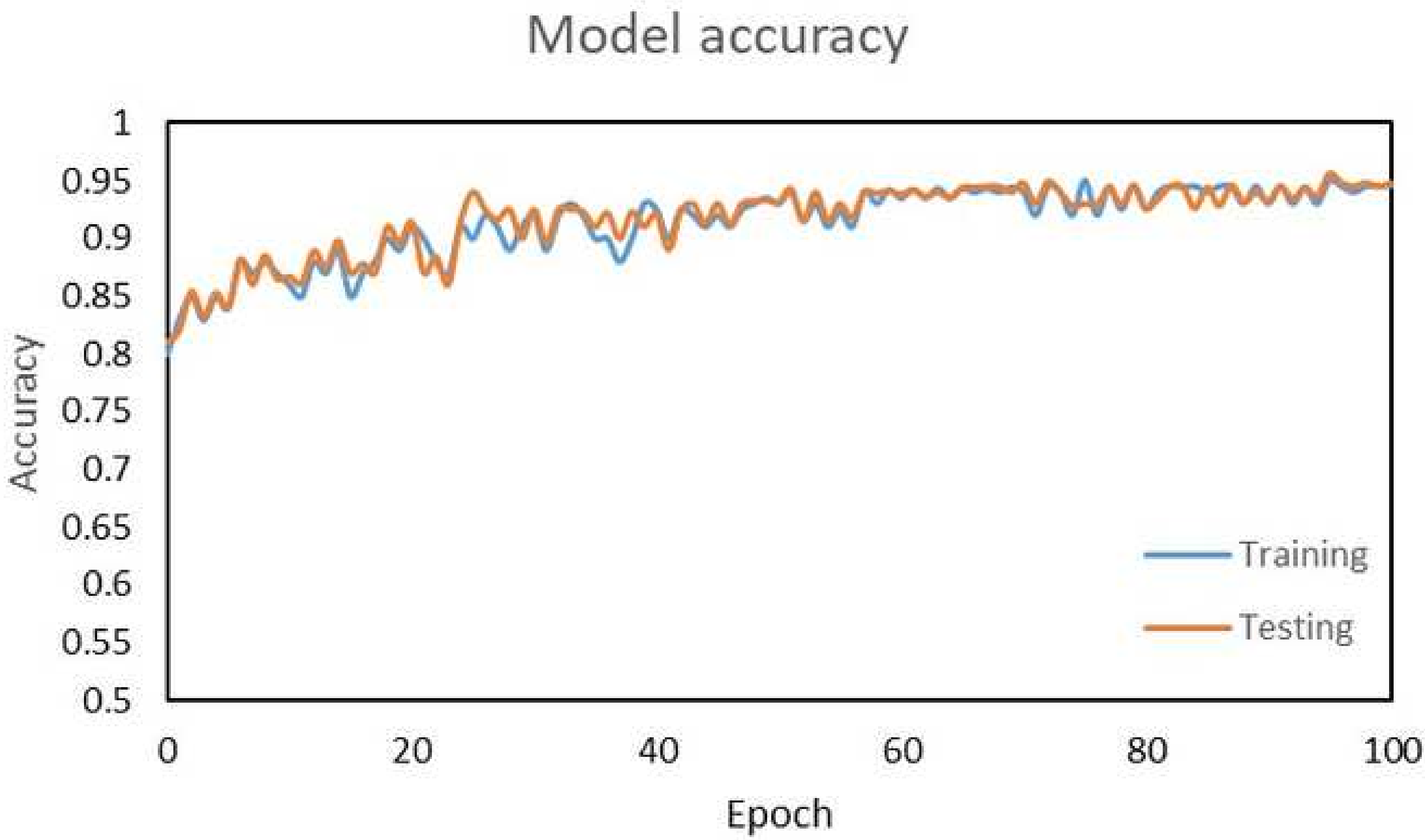
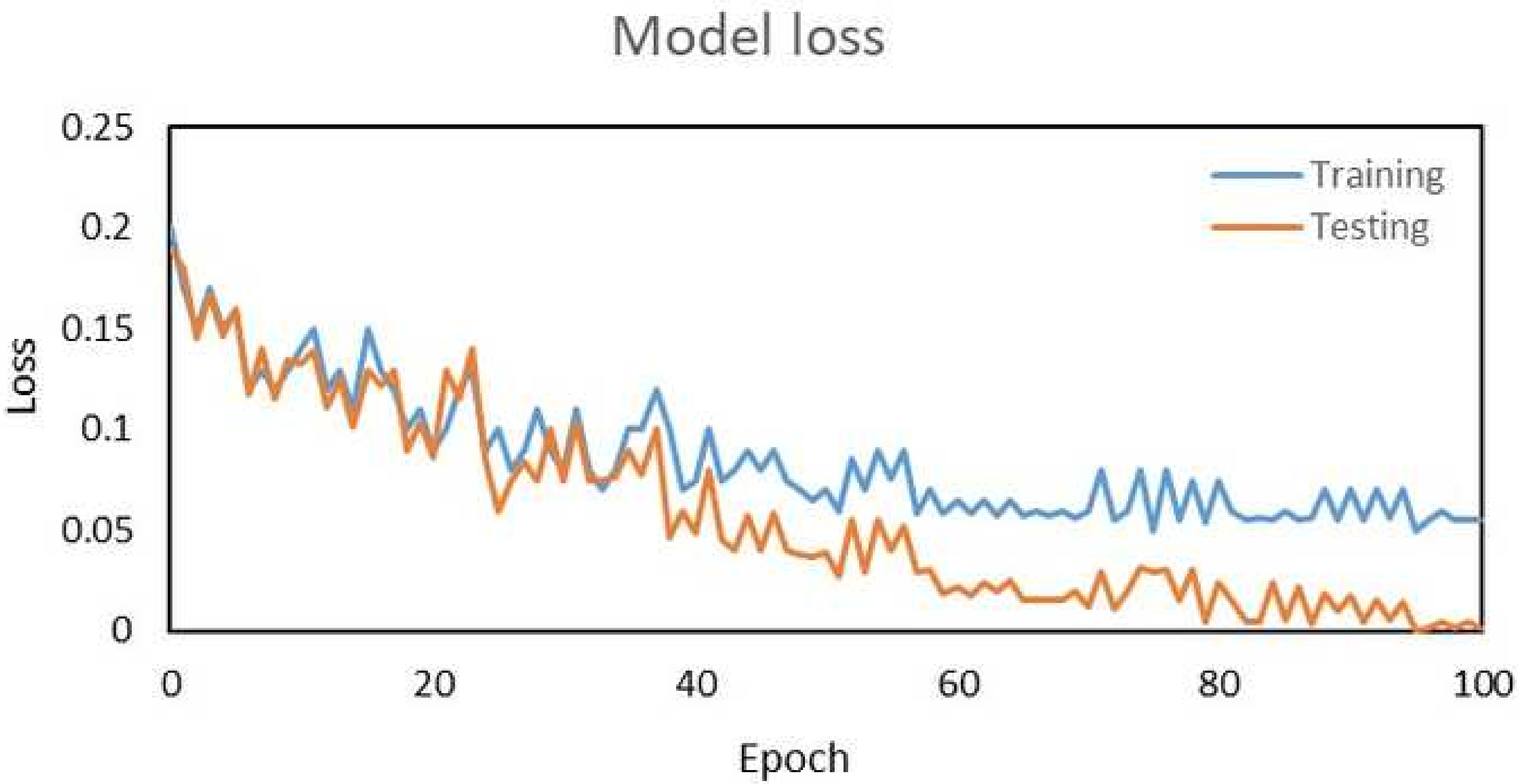

Model accuracy
Accuracy
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
Epoch
Training
Testing

Figure 4b



Model loss
0.25
0.2
0.15
0.1
0.05
0
Loss
0
20
40
60
80
100
Epoch
Training
Testing

Figure 5

Click here to access/download;Figure;Figure 5.jpg

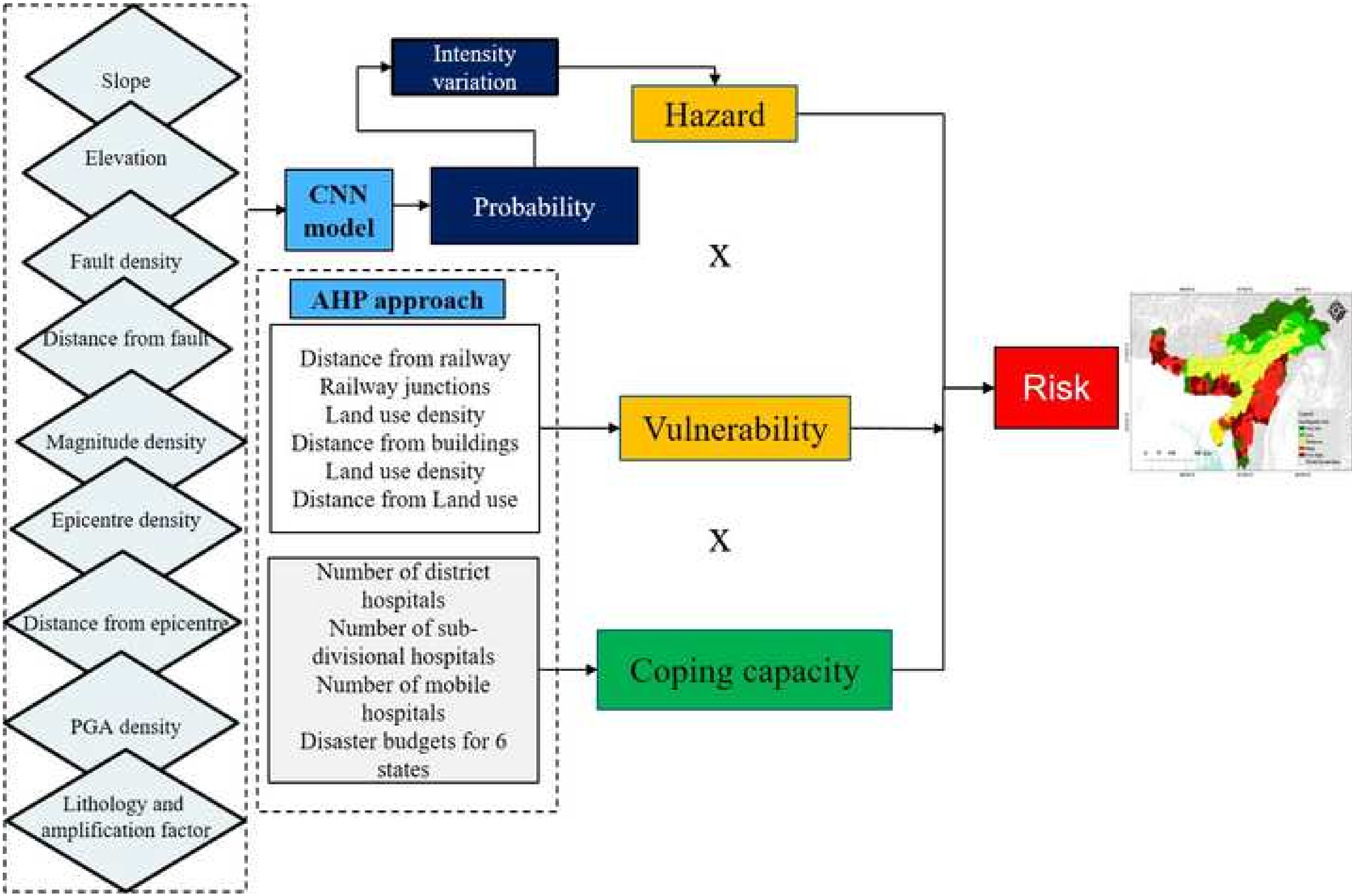


Figure 6 Click here to access/download;Figure;Figure 6.jpg

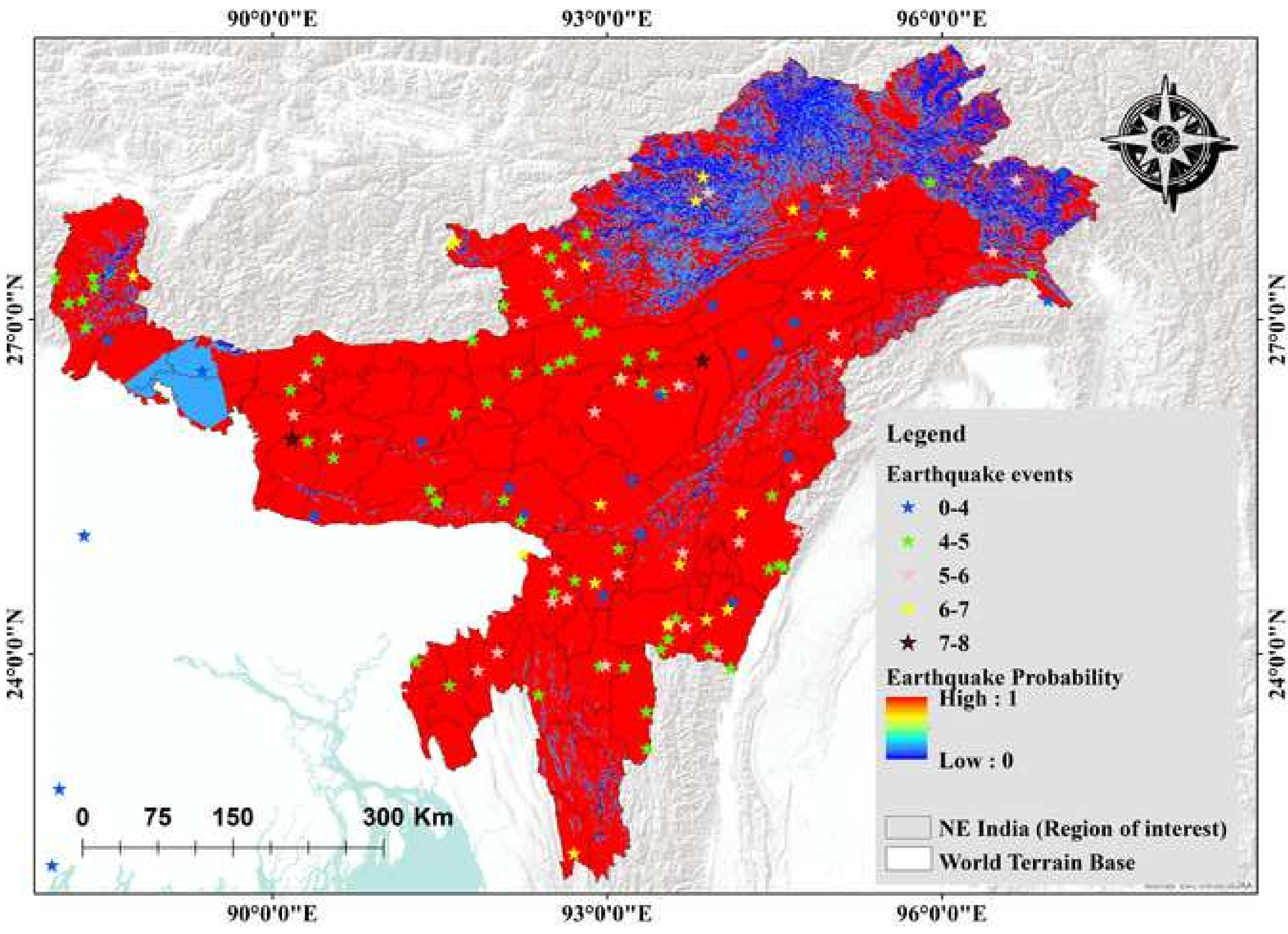


Figure 7

Click here to access/download;Figure;Figure 7.jpg

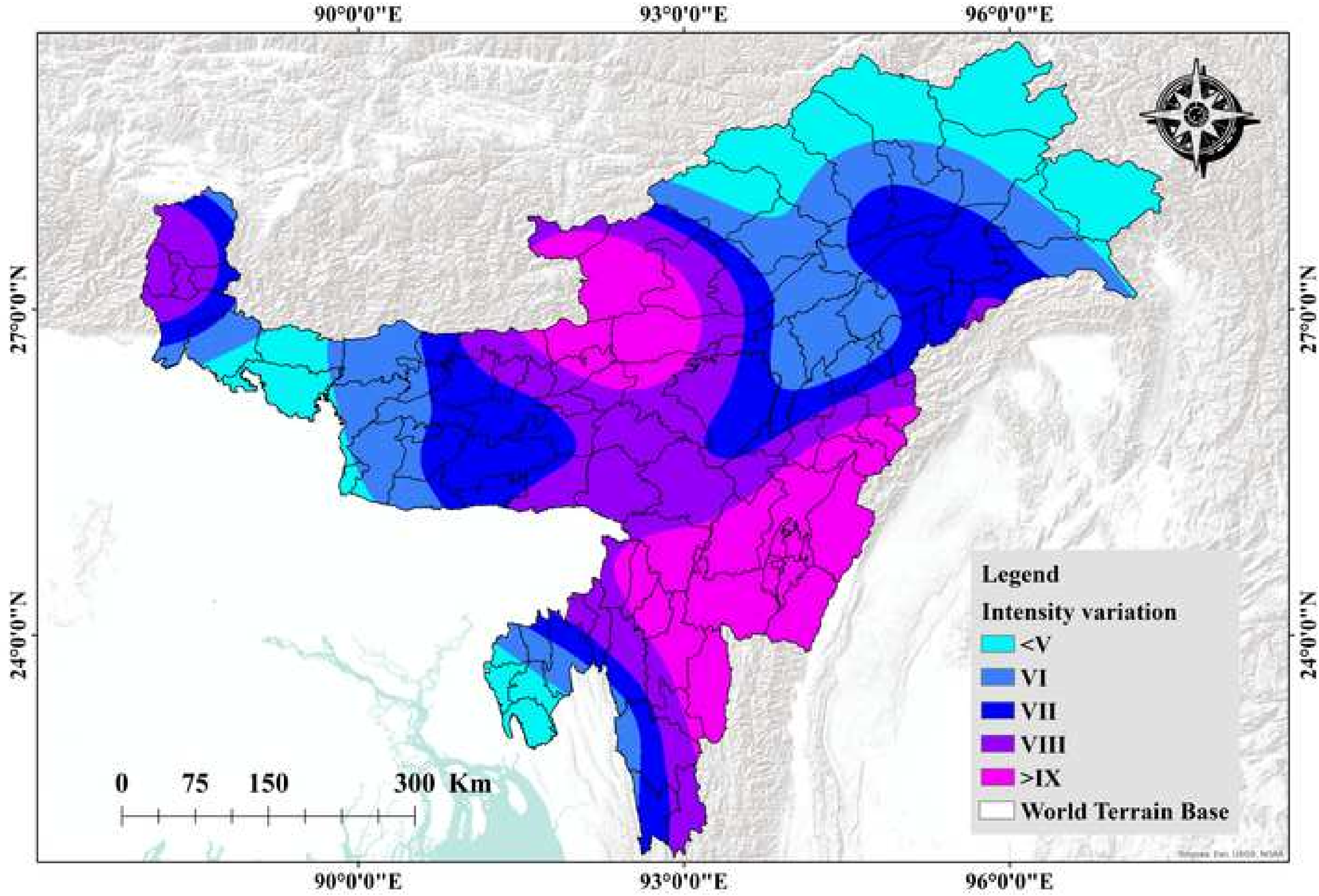


Figure 8

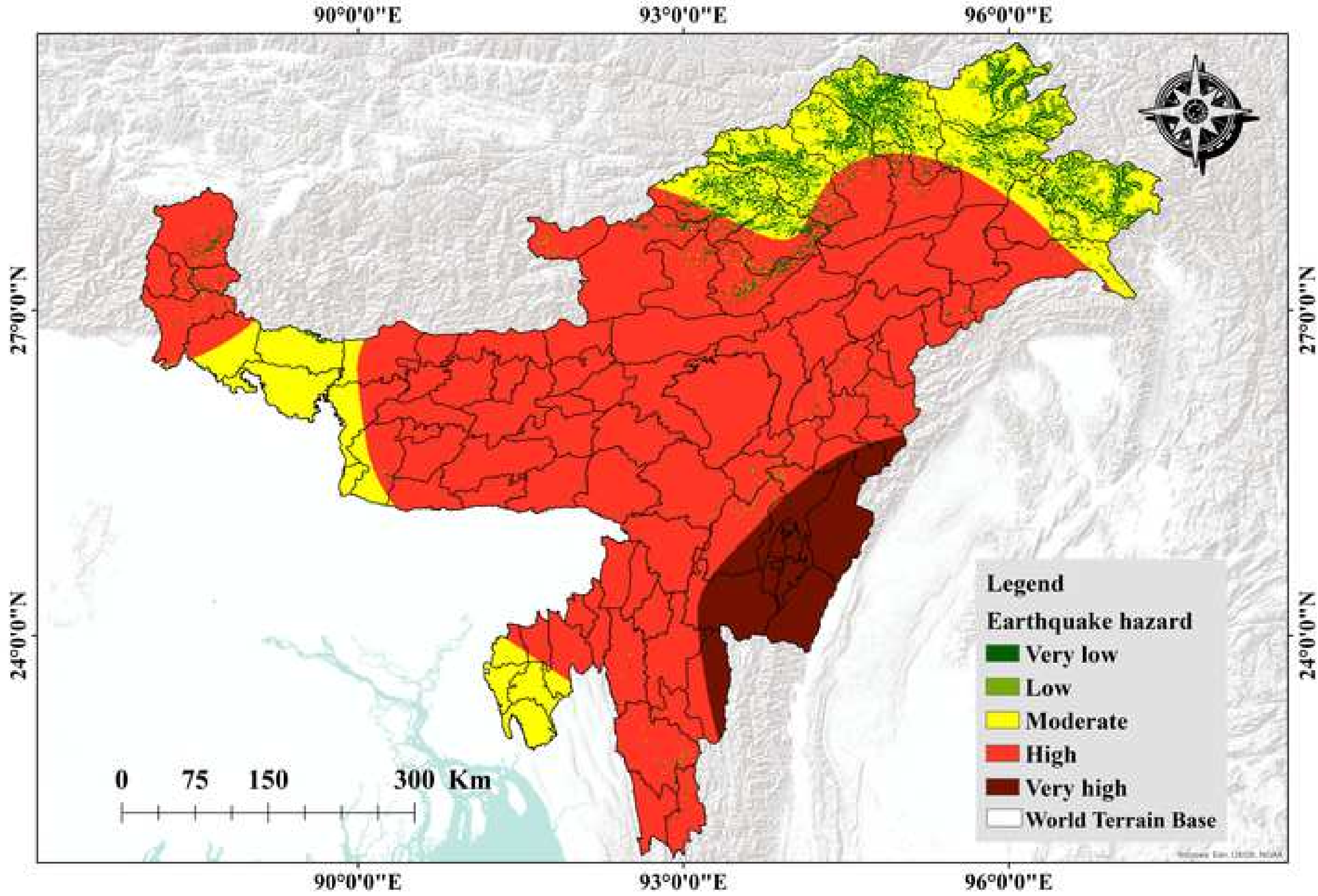

90°0'0"E
93°0'0"E
96°0'0"E
27°0'0"N
24°0'0"N
Legend
Earthquake hazard
Very low
Low
Moderate
High
Very high
World Terrain Base
0
75
150
300 Km

Figure 9

Click here to access/download;Figure;Figure 9.jpg

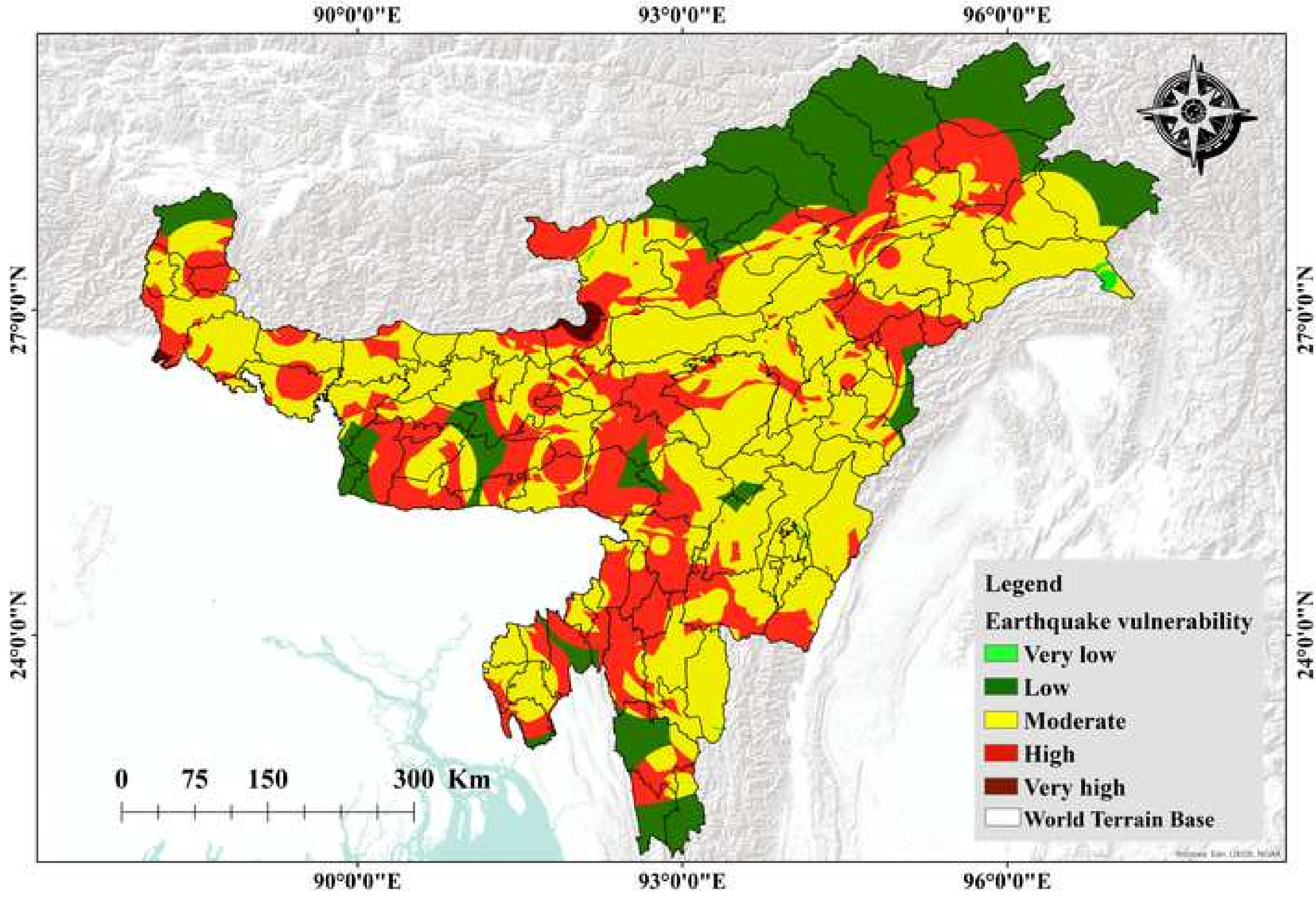


Figure 10 Click here to access/download;Figure;Figure 10.jpg

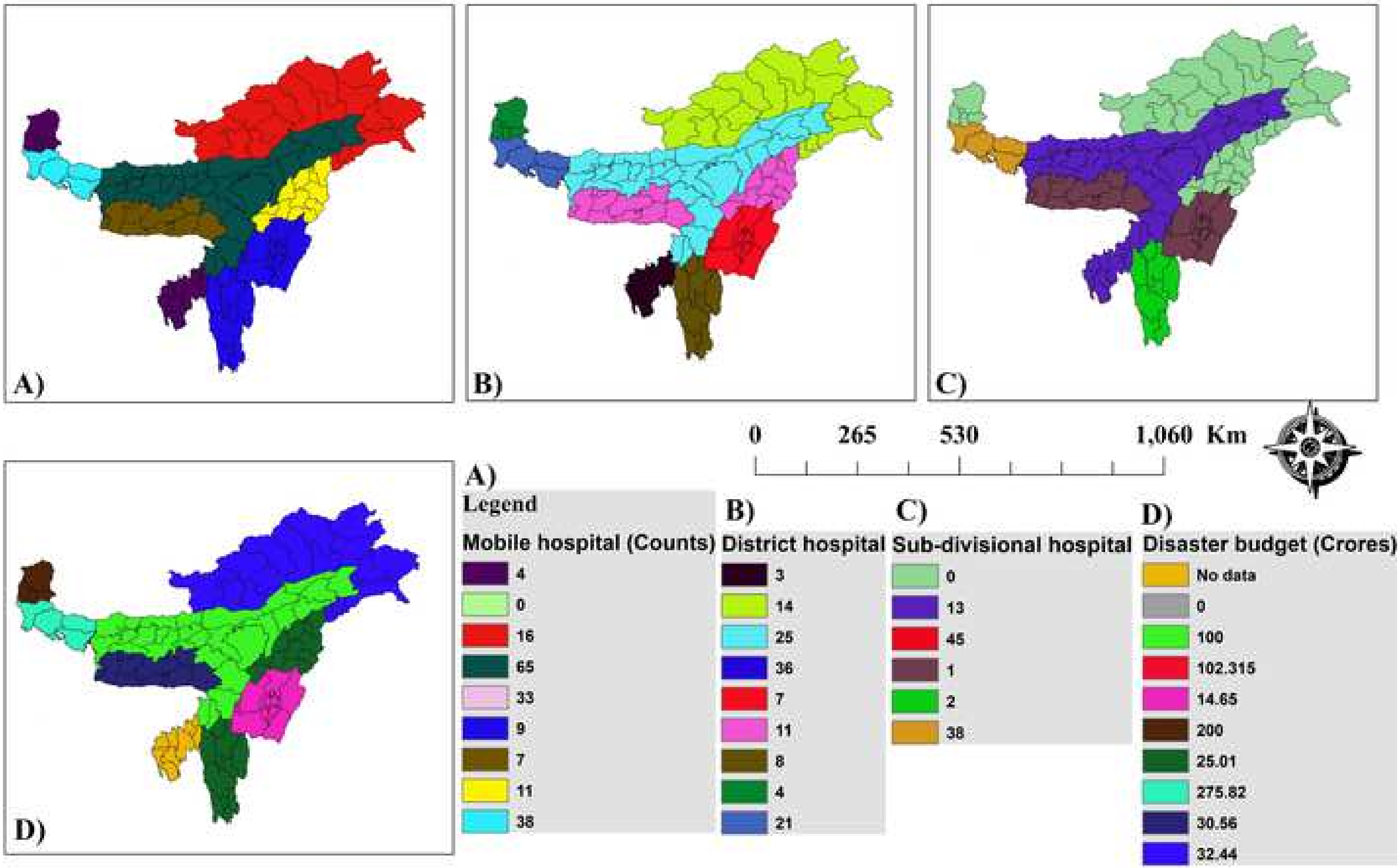


Figure 11

Click here to access/download;Figure;Figure 11.jpg

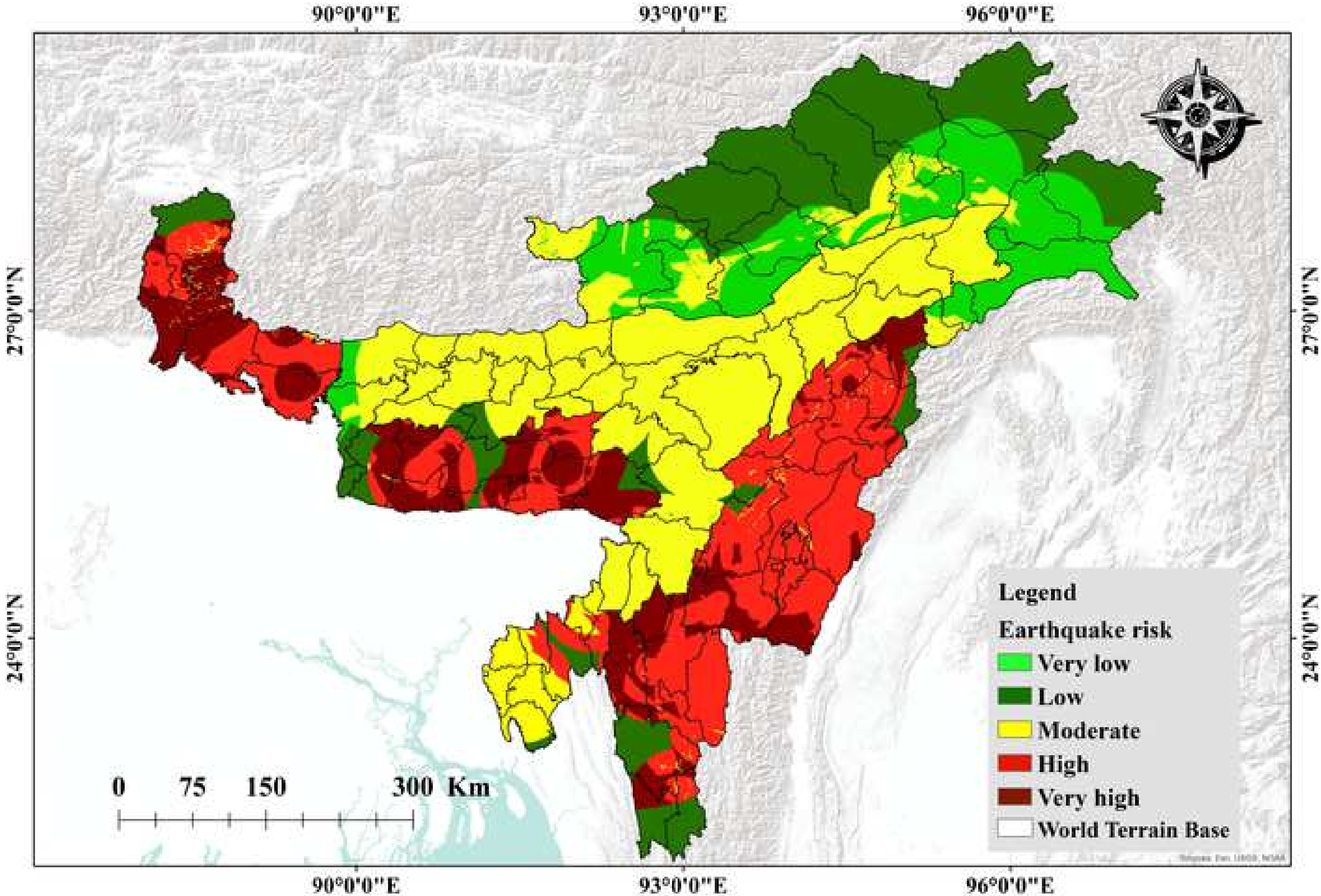


Figure 12

Click here to access/download;Figure;Figure 12.jpg

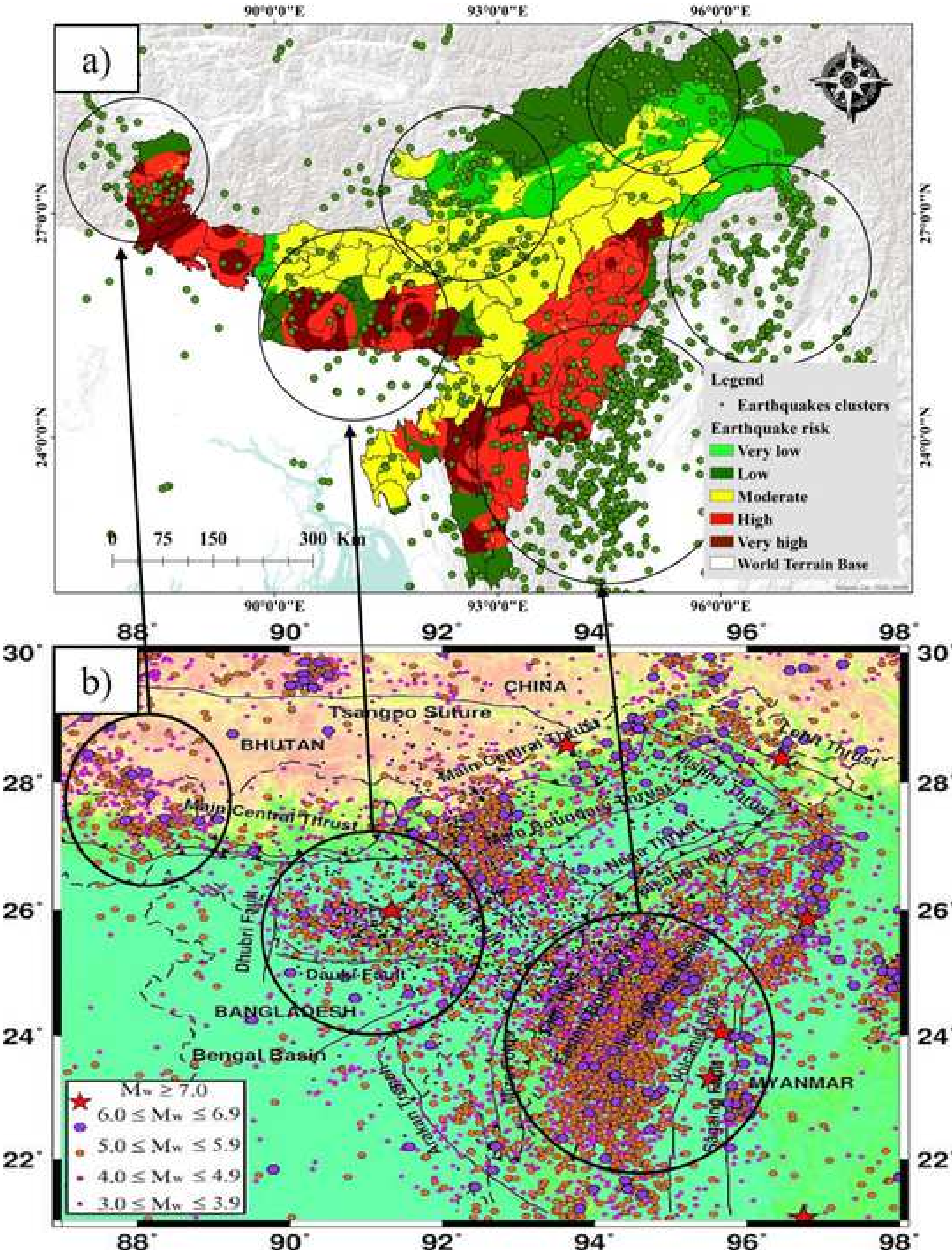


Table 1

Parameters	Data source	Scale and resolution
Earthquake catalogue	Collected from USGS and magnitude conversion conducted based on Wason et al. (2012)	1:3,000,000 and (30m)
Slope Elevation	Derived from SRTM (USGS) https://earthexplorer.usgs.gov/	
Fault density Distance from fault	Using digitisation obtained from Geological map of India, GSI	
Magnitude density Epicentre density Distance from epicentre	Joyner & Boore-1981 and Campbell-1981 attenuation equations were implemented on collected USGS earthquake catalogue	
PGA density	Derived from the catalogue using the equation: MMI=1/0.3×(LOG 10(PGA×980)-0.014)	
Lithology and amplification factor Distance from buildings Land use density Distance from land use Distance from railway Railway density	Geological map of India, GSI Derived from raster data of DIVA GIS and administrative data from shape files. Euclidean distance and kernel density were applied to estimate several parameters.	

Table 2

	1	2	3	4	5	6
1	1	3	2	3	5	4
2	0.33	1	1	1	4	3
3	0.5	1	1	2	4	3
4	0.33	1	0.5	1	3	3
5	0.2	0.25	0.25	0.33	1	2
6	0.25	0.33	0.33	0.33	0.5	1

Category		Priority	Rank	(+)	(-)
1	Building density	36.30%	1	9.60%	9.60%
2	Distance from buildings	17.10%	3	4.40%	4.40%
3	Land use density	20.30%	2	4.70%	4.70%
4	Distance from land use	14.40%	4	3.20%	3.20%
5	Distance from railway	6.30%	5	2.50%	2.50%
6	Railway density	5.70%	6	1.90%	1.90%

Number of comparisons = 15
Consistency Ratio CR = 3.4%
Principal Eigen value = 6.213
Eigenvector solution: 4 iterations, delta = 7.5E-8

Table 3

Layer	Kernel size	Number of kernels	Biases	Total	Activation
Conv1	3×3	200	200	2000	Relu
Conv2	14×14	200	1000	40200	Relu
Conv3	14×14	200	1000	40200	Relu
Conv4	14 ×14	200	1000	40200	Relu
FCL	14 ×14	2	10	402	Softmax
Kernel_regularizer=12(0.0001)					
Accuracy of 0.94, Precision (0.98), Recall (0.85) F1 score is (0.91)					
Total				**123,002**	

Table 4

Hazard Classes	Class no	Area(Km^2)	Area (Hectare)	Area (%)
Very low	1	19745.02	1974502.28	7.02
Low	2	1802.85	180284.26	0.65
Moderate	3	48932.23	4893222.05	17.38
High	4	189717.97	18971797.23	67.37
Very high	5	21412.94	2141293.84	7.61
Total			**28161099.65**	**100**

Vulnerability Classes	Class no	Area(Km^2)	Area (Hectare)	Area (%)
Very high	1	480.98	48097.92	0.17
High	2	63583.87	6358386.73	22.58
Moderate	3	157205.52	15720551.43	55.82
Low	4	60273.11	6027310.89	21.4
Very low	5	67.53	6752.69	0.02
Total			**28161099.65**	**100**

Risk Classes	Class no	Area(Km^2)	Area (Hectare)	Area (%)
Very low	1	34994.88	3499487.21	15.82
Low	2	60375.51	6037551.23	27.29
Medium	3	93796.88	9379687.71	42.4
High	4	57856.74	5785674.28	26.15
Very high	5	34586.1	3458699.22	15.64
Total			**28161099.65**	**100**

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: