SEISMIC VULNERABILITY MODELING USING MACHINE LEARNING AND GIS

SHALU[a], TWINKLE ACHARYA[a], DHWANILNATH GHAREKHAN[a,*], DIPAK SAMAL[b]

[a]*Faculty of Technology, CEPT University, Kasturbhai Lalbhai Campus, University Road, Navrangpura, Ahmedabad - 380009, Gujarat, India*
[b]*Tata Institute of Social Science, VN Purav Marg, Deonar, Mumbai, 400088, Maharashtra, India*
**Corresponding author: dhwanilnath@gmail.com*

(Received February 05, 2024; revised version accepted October 10, 2024)

ABSTRACT

Seismic vulnerability modeling is critical to seismic risk assessment, enabling decision-makers to identify and prioritize areas and structures most susceptible to earthquake damage. The use of machine learning (ML) algorithms and Geographic Information Systems (GIS) has surfaced as an encouraging approach for seismic vulnerability modeling due to their ability to integrate and analyze large volumes of data. In this abstract, we present a novel approach to seismic vulnerability modeling that leverages the power of ML and GIS. Using Artificial Neural Networks and Random Forest algorithms, the damage intensity values for an earthquake event with the help of various factors like the location, depth, land cover, distance from major roads, rivers, soil type, population density, and distance from fault lines were predicted. The resulting damage intensity values were classified, keeping the Modified Mercalli Intensity Scale as a reference. The ANN and Random Forest algorithms performed very well in this study, and both the models' accuracy was above 95% for training and testing data. Utilizing the damage intensity values map, the global seismic hazard map, and other socio-physiological parameters were utilized to generate an exposure grid zonation map. Applying this approach to a case study in the Satara district of Maharashtra highlights the model's effectiveness in identifying vulnerable buildings and improving seismic risk assessment. This approach provides a valuable tool for disaster management and urban planning decision-makers to develop effective mitigation strategies, prioritize resources, and improve overall disaster resilience.

KEY WORDS: Machine Learning, Earthquake, Artificial Neural Network, Random Forest, Seismic Vulnerability

0963-0651/23/$5.00

INTRODUCTION

An earthquake is defined as any abrupt shaking of the earth's surface induced by the energy released due to the passage of seismic waves. Earthquakes are believed to be one of the mostcatastrophic natural disasters. The impact of earthquakes can lead to extensive and uncontrollable devastation to the environment and society globally, resulting in substantial physical and economic harm. The repercussions of earthquakes include loss of human life and property, remodeling of the river course and mud fountains, for example, the 1934 Bihar earthquake when the agricultural fields were engrossed with mud, and fire risks near gas pipelines or electric infrastructure (Manish). 9

As the plate tectonics theory states, the earth is divided into slabs of solid rock masses referred to as "plates" or tectonic plates which are always in motion. These tectonic plates may be continental or oceanic and are in slow continuous motion, and their movement forms three different types of tectonic boundaries. When two plates come together, it is called a convergent boundary, but when they move apart, they are divergent. And when the plates move side by side, they form a transform boundary. The financial damage caused by earthquakes is approximately $787 billion. Disaster management before earthquakes happen is a vital strategy to reduce earthquake-induced damage. The earthquakes' exact time, magnitude, and place of occurrence are still unforeseeable (Lee, Saro, et al.). 19

Over the past few years, scientists have investigated a specific region's susceptibility from various perspectives, such as geotechnical, structural, and socioeconomic factors. Researchers have employed a range of multi-criteria decision-making (MCDM) techniques to assess seismic vulnerability, such as the analytic hierarchy process (AHP) and fuzzy logic. Developing decision-making methods that can quickly fulfill demands requires expert opinions, which can lead to bias and error. To address this issue, artificial intelligence algorithms, including evolutionary algorithms and adaptive neuro-fuzzy inference systems (ANFIS), have been implemented in geological research, specifically for evaluating seismic vulnerability. (Peyman Yariyan). 28

Effective disaster risk reduction and management (DRRM) requires a comprehensive understanding of risk, hazards, vulnerability, and interconnectedness (M.J.D. De Los Santos). Geographic Information System (GIS) is a powerful technology that can visualize, map, and analyze the interrelationships among these elements in DRRM. However, the success of DRRM-related mapping projects depends on adequate and dependable information. Remote sensing has become a valuable operational tool in DRRM as it can provide a substantial amount of data. Recent studies on scenarios have proven useful in promoting awareness and formulating policies (Ravi Sinha). 37

Disaster scenarios can sensitize stakeholders, identify vulnerable areas and population groups, and evaluate the effectiveness of various disaster management interventions. Urban areas are particularly susceptible to earthquakes as they typically have a high population density and contain significant infrastructure and resources. Seismic hazard assessment involves evaluating the expected damage and losses resulting from an earthquake in a specific region for a particular hazard event, such as an earthquake of a certain magnitude at a specific location. Risk assessment is a methodology used to estimate the consequences of scenario earthquakes. Furthermore, this evaluation estimates the number of injuries, casualties, and possible economic damage. That is why disaster management before the event is necessary. Factors like building information, altitude, lithology, land use, elevation, distance from streams, roads, and population density are considered for assessing the ability of a place or a building to withstand seismic waves. To predict seismic vulnerabilities, various machine learning algorithms such as Support Vector Machine, K-Nearest Neighbor, Bagging, Radial Basis Function, Logistic Regression, Artificial Neural Networks (ANN), and Random Forest were employed. 8

However, it was observed that they had been conducted on a region-specific scale. No geospatial study for seismic vulnerability has been done for India and is focused on predicting seismic vulnerability. Still, no study has considered an earthquake's potential damage intensity. Across the globe, damage or seismic intensity has conventionally been utilized to gauge the shaking pattern and the scale of the destruction caused by earthquakes. (David J. Wald). Thus, with this study, the damage intensity values for any earthquake event dependent on its location, magnitude, and other socio-physical characteristics have been predicted, and utilizing that information, a risk assessment for the Satara district of Maharashtra state has been conducted using the Artificial Neural Network and Random Forest Algorithms. 18 This article highlights the significance of utilizing GIS technology to conduct disaster scenario studies in promoting awareness, informing policy decisions, and formulating effective disaster management plans.

STUDY AREA

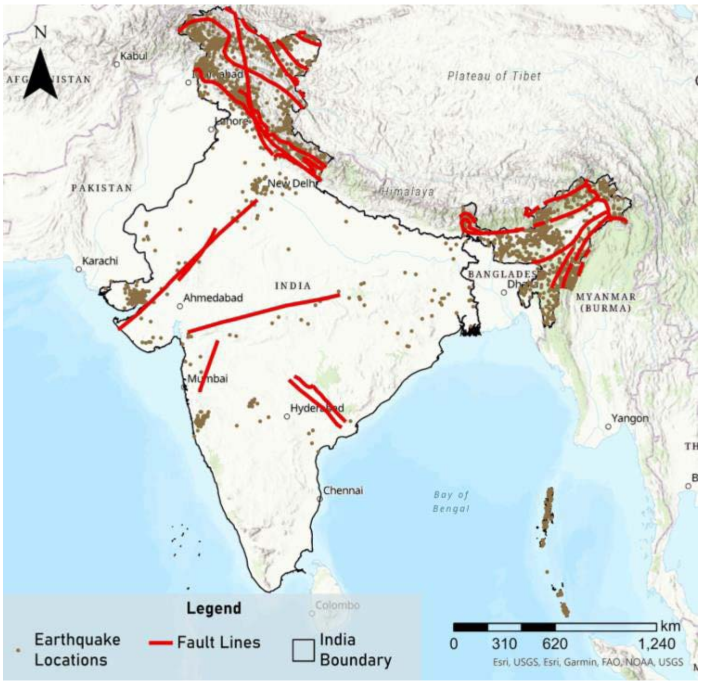


Figure 1. Study area map of India with earthquake event locations and fault line 4

India is located above the equator in the northern hemisphere between the latitudes 8°4’ and 37°6’ and longitude of 68° 7’and 97°25’. India’s total geographic area is around 3.28 million square kilometers, which makes it the 7th largest country with 29 states and eight union territories. India is one of the most earthquake-prone countries in the world, with a long history of devastating earthquakes. This is primarily due to the subduction of the Indian plate beneath the Asian plate, which creates much tectonic activity. The Indian plate moves at a rate of 33 (+-6) mm per year, making it one of the fastest-moving plates in the world. As a result, India has different levels of seismicity, with the southern part of the country experiencing strong earthquakes and the northern part experiencing large, tremendous, and mega earthquakes. To help mitigate the effects of earthquakes, the Bureau of Indian Standards created a seismic zoning map that divides the country into four seismic zones based on how likely they are to experience earthquakes. Zone V is the most active, with the highest likelihood of earthquakes, while Zone II is the least active. This map is based on historical seismic activities and ground motion. In the last 100 years, the number and strength of earthquakes in India has increased significantly. Some experts believe this is due to the changing climate, while others suggest it may be due to increased urbanization and population growth. India has over 66 active faults, which are fractures or zones of fractures between two blocks of rock. The movement of these blocks of rock releases energy, which travels in the form of waves and causes earthquakes. The Himalayan belt is one of the most active areas in seismic activity, divided by 15 major active faults. The Northern part of India has 16 tectonically active faults, while Southern India has about 30 neotectonic faults. 25

The Andaman and Nicobar Islands are at an exceptionally high risk of earthquakes, falling under the very high hazard zone of the seismic activity map. In addition, many hidden faults throughout India contribute to the country’s seismicity.

DATA USED

USGS Earthquake Hazards Program monitors, reports, and researches earthquakes and hazards. The USGS Earthquake Hazards Program of the U.S. Geological Survey (USGS) is part of the National Earthquake Hazards Reduction Program (NEHRP) led by the National Institute of Standards and Technology (NIST). Under this program, a database with the earthquake events has been curated that contains the location information, depth value, and magnitude value of each earthquake event that has occurred globally. The dataset has been utilized to extract natural earthquake events, not those caused by nuclear activities. The other variable values for the event points were extracted per the sources in the table below.

Satellite-Based and Other Products

DATA	SOURCES	UNITS	RATIONALE
Latitude & Longitude	USGS Earthquake Data	Degree Decimal	Location of past activities to understand the trend
Depth	USGS Historical Data	Kilometers (km)	The lower the depth, the more destructive power.
Magnitude	USGS Historical Data	Moment Magnitude Scale	The higher the magnitude, the more the area of damage
Elevation	SRTM	Meters (m)	The landslide post-seismic activity increases with an increase in elevation.
Population Density	CENSUS	Persons per km^2	The greater the population density, the more the chances of casualties.
Land Cover	Sentinel-2	--	Different categories of land have different susceptibility.
Lithology	ESRI Data Catalog	--	The more complex the geological formation of minerals, the lower the earthquake wave and the weaker the destructive power.
Distance from Stream	Open Street Map	Meters (m)	The area closer to streams is more likely to be damaged.
Distance from Faults	ESRI Data Catalog	Meters (m)	Areas nearer to faults have experienced more earthquakes in the past year.
Distance from Roads	Open Street Map	Meters (m)	To ensure proper evacuation, the closer the road, the easier it is.

Table 1. Data inputs

The training and testing dataset had records from the year 1900-01-01 to 2020-12-31.

In-Situ Calculations

The study used the magnitude values of the past earthquakes between 1900-2022, which were collected from the USGS Earthquake Data Catalog. It associated those different magnitude values (regional, moment, body, etc.) to surface magnitude value using the following formulae:

$M_S = m_b\text{-}2.74/0.46$

$M_S = M_W\text{-}2.07/0.67$

Where M_S is the surface magnitude, M_W is Moment Magnitude, and m_b is the body magnitude. A magnitude based on the amplitude of Rayleigh surface waves measured at a period near 20 sec. Ms is primarily valuable for large (>6) shallow events, providing secondary confirmation on their size. After gathering the surface magnitude value, the peak ground acceleration value for each earthquake incident was calculated using Donovan’s

Formula:

$$PGA = 1080\ e^{0.5\ Ms}\ (R+25)^{-1.32}$$

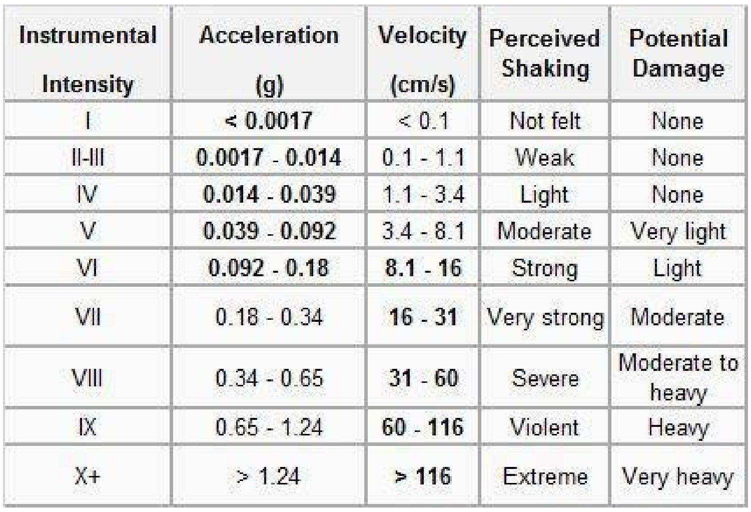
Where Ms is the surface magnitude, R = Distance from the hypocentre to the event's site (in Kilometers), and peak ground acceleration (PGA) equals the maximum ground acceleration during earthquake shaking at a location. PGA is a measure of how much ground shakes at a particular location during an earthquake event. It is calculated by looking at the highest acceleration record on an accelerogram device. To understand the severity of an earthquake event, damage intensity values are considered, which helps in correlating the damages caused by an event with the magnitude.

$$I_{mm} = 2.20 \log (PGA) + 1.00 \text{ [for } M_S \text{ values 3.5 to 5]}$$

$$I_{mm} = 3.66 \log (PGA) - 1.66 \text{ [for } M_S \text{ values 5+]}$$

Each damage intensity value can be correlated to the modified Mercalli intensity values. Earthquakes cause different effects on the earth's surface, known as the earthquake's intensity. A scale has been developed to measure this intensity by considering the different observations of people who have experienced that event. This scale is called the Modified Mercalli Intensity scale and helps everyone understand the potential damage caused by the event.

All these data points were segregated into a 70:30 ratio for training (19040 records) & testing data (8160 records) and another validation dataset (another 786 records) with a temporal gap of 6 months before the building of models. 19



Instrumental Intensity	Acceleration (g)	Velocity (cm/s)	Perceived Shaking	Potential Damage
I	< 0.0017	< 0.1	Not felt	None
II-III	0.0017 - 0.014	0.1 - 1.1	Weak	None
IV	0.014 - 0.039	1.1 - 3.4	Light	None
V	0.039 - 0.092	3.4 - 8.1	Moderate	Very light
VI	0.092 - 0.18	8.1 - 16	Strong	Light
VII	0.18 - 0.34	16 - 31	Very strong	Moderate
VIII	0.34 - 0.65	31 - 60	Severe	Moderate to heavy
IX	0.65 - 1.24	60 - 116	Violent	Heavy
X+	> 1.24	> 116	Extreme	Very heavy

Table 2. Modified Mercalli Intensity (MMI) scale. 24

Exploratory Data Analysis (EDA)

EDA is the first step in any modeling study. With the help of EDA, the relationship between various factors is established, and data patterns are also analyzed. It gives us a basic understanding of how and which factor affects the predictor variable the most. We are exploring the influence of various parameters before the architecture leads to a better understanding. While directly influencing the desired target parameter, the inputs do not account for the interaction between them.

In this study, the damage intensity values for each of the earthquake events above magnitude 3.5 were calculated between the year 1900-2022; using the USGS earthquake explorer data, the location (latitude, longitude), magnitude, and depth (distance from hypocentre, in km) was extracted. Other datasets like land cover, population density, elevation, distance from roads, significant rivers, distance from fault lines, and lithology type information were compiled from the abovementioned sources. The earthquake incidents were categorized years, and their trend was studied, also the trend of earthquake incidents above magnitude 7. 18

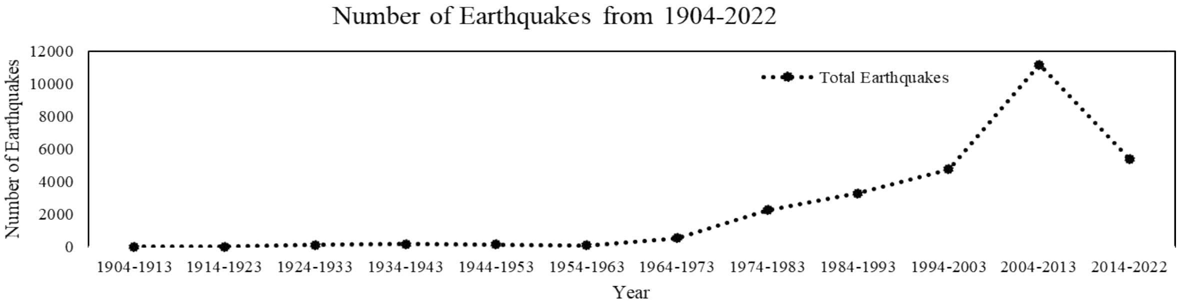


Figure 2. Number of earthquakes from 1904-2022

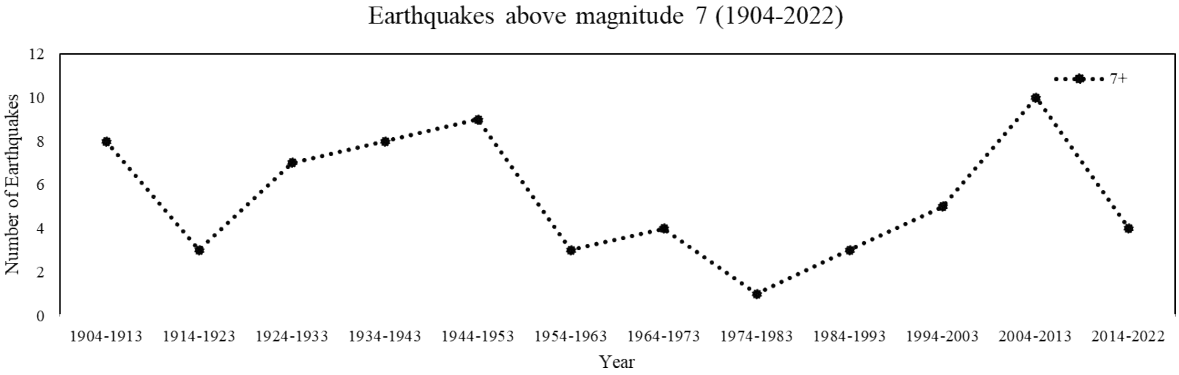


Figure 3. Graph for earthquake above magnitude 7(1904-2022)

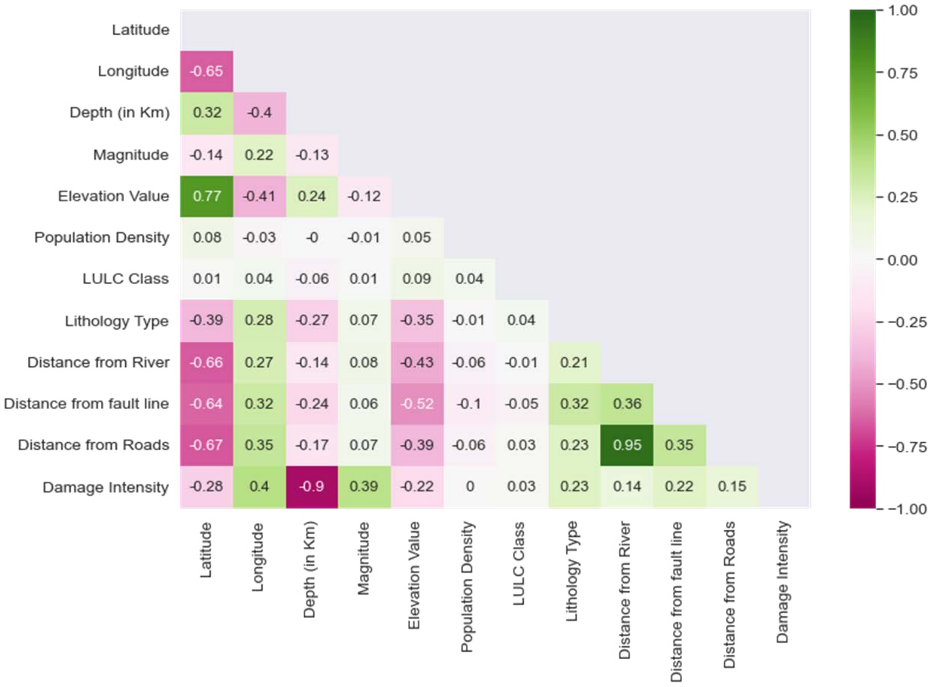


Figure 4. Correlation Matrix

A correlation matrix was generated to understand the influence of the variables on each other by displaying the correlation coefficients for different variables. The correlation matrix depicted in Fig. 4 describes the potential level of influence between different parameters on a normalized scale. The image depicts the degree of influence between all the possible pairs of values used in this study.

The above correlation matrix shows that "distance from the fault line" negatively influences Z. This understanding is associated with higher elevations, like the Himalayan range, which tends to have fewer faults. Similarly, depth is inversely correlated to damage intensity at an extreme level. It can be confirmed that the more profound the epicenter, the shockwaves become weaker before it touches the surface.

MODEL DEVELOPMENT

Artificial Neural Network (ANN)

"Artificial neural networks (ANNs) are biologically inspired computational networks (Y.-S. Park)." Artificial neural networks are computer programs that simulate how the human brain and nervous system process information. These networks comprise individual processing units called neurons connected through weighted connections known as synaptic weights. The neurons process information received from other neurons to generate an output signal, which is achieved using an activation function. Neural networks come in two main types: feed-forward and feed-back.

They must be trained using an algorithm to make neural networks effective in their respective tasks. One popular algorithm is the Rectified Linear Unit (ReLu) activation function. ReLu is a piecewise linear function that outputs the input directly if it is positive and zero if it is negative. This activation function is beneficial when dealing with nonlinear functions and is easily trained with multilayer Perceptron and convolutional neural networks. 16

The gradient descent algorithm is another important aspect of neural network training. It is an optimization algorithm used to solve machine-learning problems. This algorithm approaches the optimal solution of the objective function by obtaining the minimum loss function and related parameters. There are two types of gradient descent algorithms: batch gradient descent and stochastic gradient descent. Batch gradient descent calculates gradients for the whole dataset, which can be time-consuming for large datasets. On the other hand, stochastic gradient descent performs one update at a time, which makes it much faster. However, it has a higher variance that causes the objective function to fluctuate heavily. 25

The present study applies a Stochastic Gradient Descent transfer function with ReLu activation for the estimation of the seismic vulnerability of India.

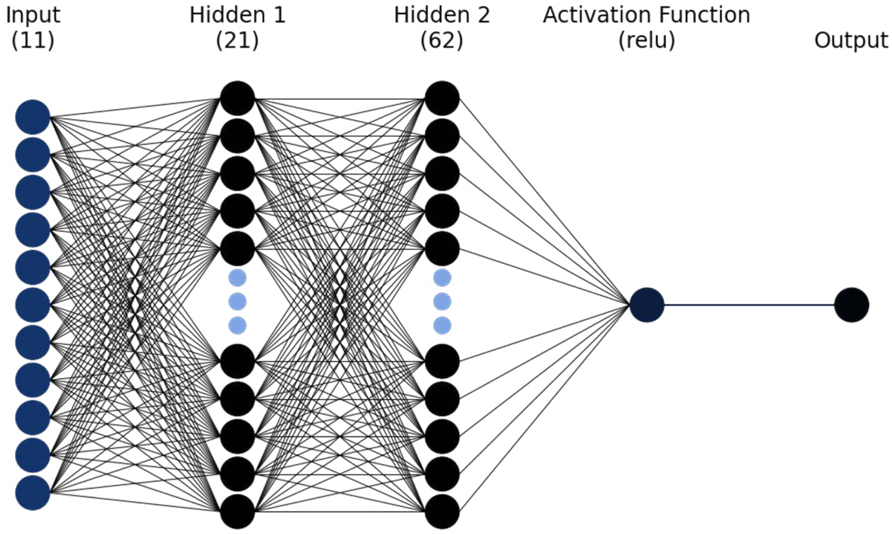


Figure 5. ANN Architecture

Eleven inputs (mentioned in the table –) have been used in the model. Over multiple iterations, 16 combinations of input hyperparameters were fine-tuned to provide the optimized parameterized model. The number of hidden layers was kept as two as it performed better than the multi-layer perceptron model. The study adapts the standard conditions by Heaton, which follows the 2n-1 rule for the first hidden layer when selecting the number of nodes within each hidden layer, where n is the number of inputs and the 3n-1 rule for the second layer, making the first layer have 21 nodes and second layer have 62 nodes. 11

A two-layered artificial neural network was utilized for this study. The created datasets were segregated into a 70:30 ratio for training (19040 records) & testing data (8160 records). The model was trained and fine-tuned to optimize the results. 15

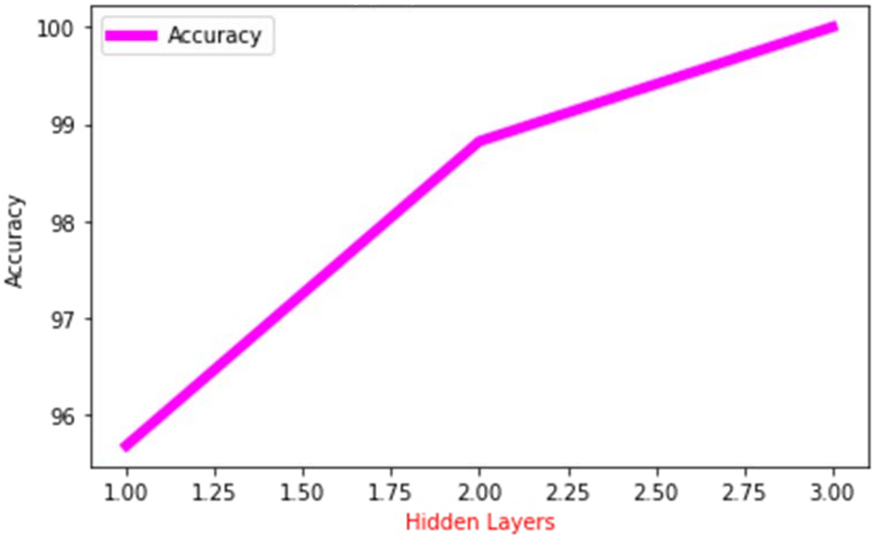


Figure 6. Hyperparameter curve for Hidden layers 2

The number of hidden layers in a neural network is decided, keeping the complexity of datasets in mind. For this study, it was noticed that increasing the hidden layers beyond 2 saturates the model's accuracy, thus resulting in overfitting of the model.

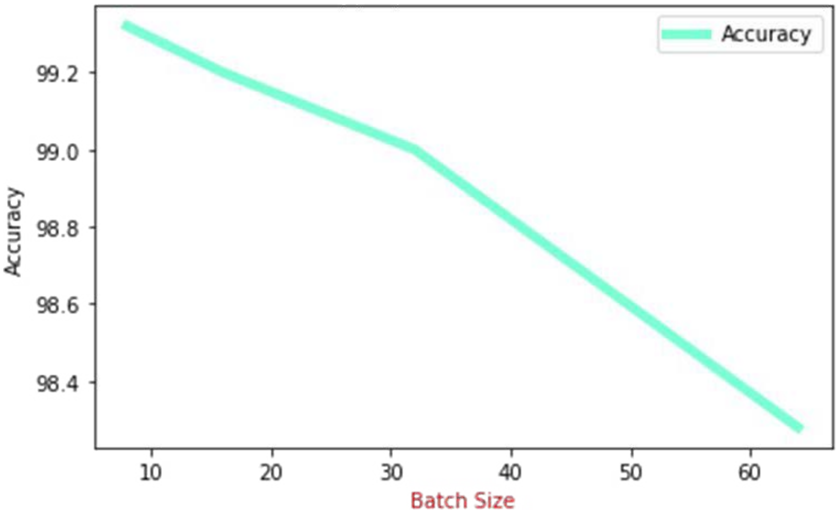


Figure 7. Hyperparameter curve for Batch size 10

The model training started with taking a batch size of 64 and checking the accuracy against the same. Simultaneously, the batch size was decreased to 16; and it was observed that the accuracy achieved was the best in this case. Keeping a batch size of 16 meant that the entire training data would pass through the model in batches of 16 observations at a point while training the model.

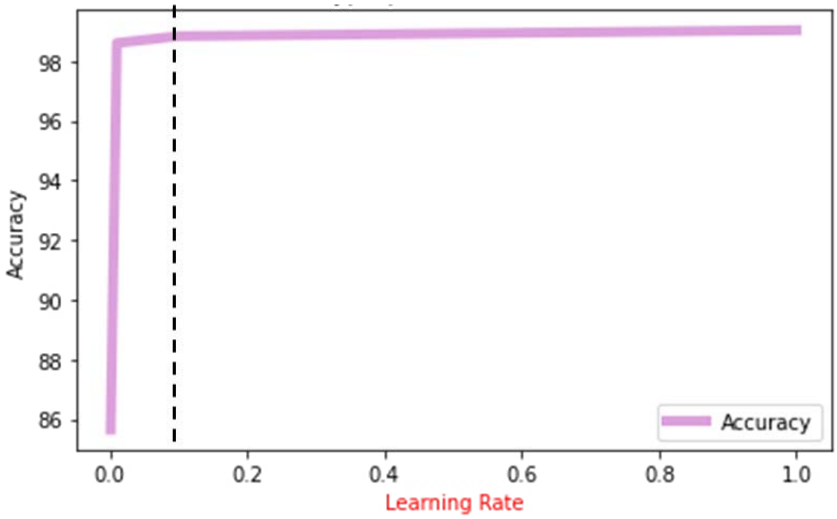


Figure 8. Hyperparameter curve for learning rate 3

The hyperparameter controls the rate of learning or speed at which the model learns. It regulates the number of allocated errors with which the model's weights are updated. The learning rate value is in the range of 0.0 - 1.0. This study's accuracy value started saturating upon increasing the learning rate value beyond 0.1.

Random Forest (RF)

A random forest is a type of classifier that comprises a set of tree-structured classifiers {h(x, Θk), k=1, ...}. Each tree in the random forest casts a unit vote for the most commonly occurring class for the given input x. Moreover, the {Θk} represents independently and identically distributed random vectors (Breiman, Leo). Random Forest is a computer program that helps classify or make predictions based on data. It is commonly used in many applications such as predicting whether a customer will buy a product or identifying whether an email is spam or not. It works by using many small decision trees together to make a final decision, rather than relying on just one tree. The algorithm creates different training subsets from the sample training data with replacement, meaning that it can use the same data points more than once, making it more accurate. These subsets are selected randomly from the dataset and are called bootstrap samples. Therefore, each decision tree or model is produced using samples from the original data, with replacement, in a process called Bootstrapping. Each model is trained independently, generating individual results. The outcome is then formed by calculating the average output of all the decision trees. This step is called Aggregating.

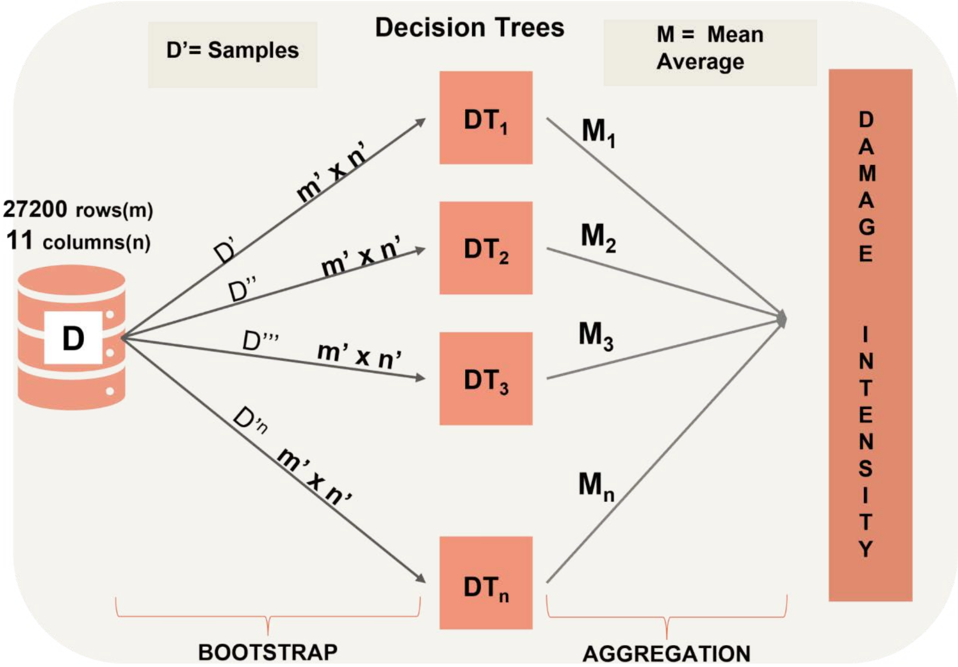


Figure 9. RF Architecture

In this study, specific model parameters called hyperparameters are fine-tuned to improve the model's performance. Random forest functions on the combination of multiple trees of varied degrees and levels. Each one is responsible for driving the model in an optimized form. The parameter *"n_ estimators"* defines the number of trees within the algorithm based on the model tuning. Typically, increasing the number of trees in the random forest model leads to more generalized results. However, this also increases the time complexity of the model. The model's performance increases with the number of trees but levels off after a certain point. The *"max_depth"* parameter is crucial, as it determines the longest path between the root and leaf nodes in each decision tree. Setting a large *max_depth* may result in overfitting. The *"max_features"* parameter determines the maximum number of input variables provided to each tree. The default value, which is the square root of the number of features in the dataset, is usually a good choice to consider.

A random forest was used to build several iterations of an RF model. Like ANN, the datasets created were randomly segregated into a 70:30 training and testing data ratio The process of optimizing the model involves adjusting certain parameters to improve its accuracy. These parameters include *"n_estimators,"* which refers to the number of decision trees generated, *"max_depth,"* which is the longest distance between the root node and the leaf node, and *"max_features,"* which is the number of variables randomly selected as candidates at each node. The parameters were adjusted over several iterations of the model to assess how overall accuracy was affected by each.

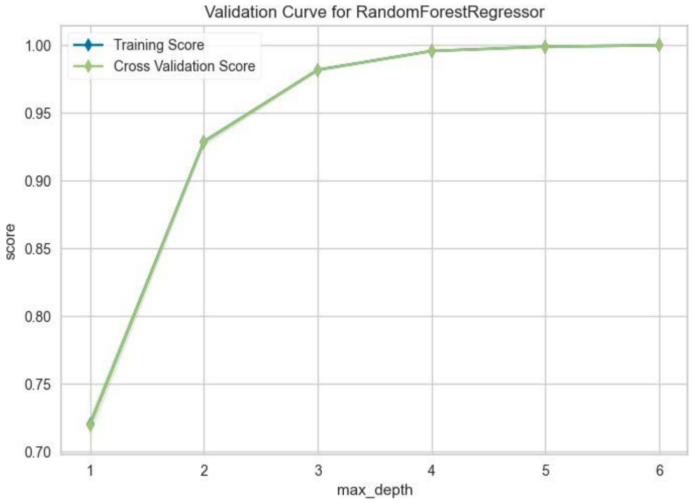


Figure 10. Validation curve for max_depth 6

It is essential to remember that *max_depth* is not the same thing as the depth of a decision tree. *max_depth* is a way to pre-prune a decision tree. This study observed that the model is achieving a threshold after max_depth 4. 10

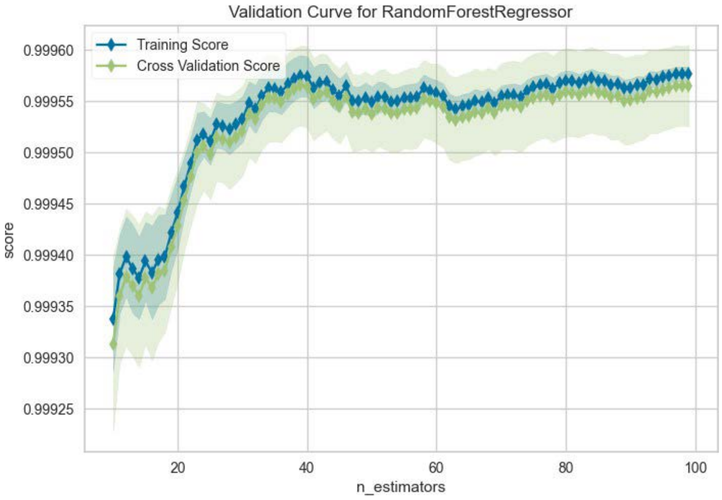


Figure 11. Validation curve for n_estimators 13

A higher number of decision trees gives better results but slows the processing. The model attained maximum accuracy at 40 *n_estimators* (number of decision trees).

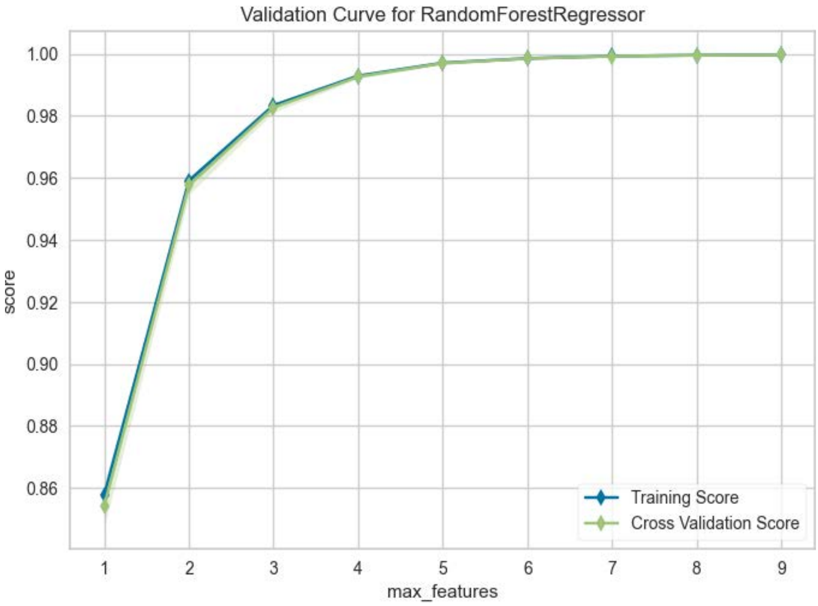


Figure 12. Validation curve for max_feature 11

The *max_features,* if not specified, considers all the parameters. There are 11 input parameters, and the model attains maximum accuracy by using any random 5 of them. After this, it is sustained. Like ANN, the input and temporal periods are the same for the input dataset. This provides a better comparative capacity between the models. Within RF, the following fine- tuned parameterized values were identified for the final model. The following fine-tuned parameters are described in Table 3 –.

Hyperparameter	Value
n_estimators	40
max_depth	4
max_features	5

Table 3 . Hyperparameters for RF

SENSITIVITY ANALYSIS

Sensitivity analysis is the label used for a collection of methods for evaluating how sensitive model output is to changes in parameter values (Franceschini S). Sensitivity analysis identifies which input variables are essential in contributing to the prediction of the output variable. It quantifies how the changes in the input parameters' values alter the outcome variable's value (Dowlatabadi) (Muriel Gevreya).

The one-dimensional sensitivity focuses on varying one parameter while keeping the remaining constant. This can provide an understanding of the level of influence between the parameter and the output scale. For both the models, the mean values of input parameters were taken; for instance, in the case of depth, the mean value is 54.8 km, and the deviation of this value with 50% on either side of the mean value, i.e., plus and minus 25 km's and the result shows that it shifted damage intensity from -1.42 to 1.48. The parameters like streams, roads & population showed minimal sensitivity toward the damage intensity. This is known as one-way sensitivity analysis since only one parameter is changed simultaneously. The analysis was repeated on different parameters at different times, and the values have been plotted in the graph below.

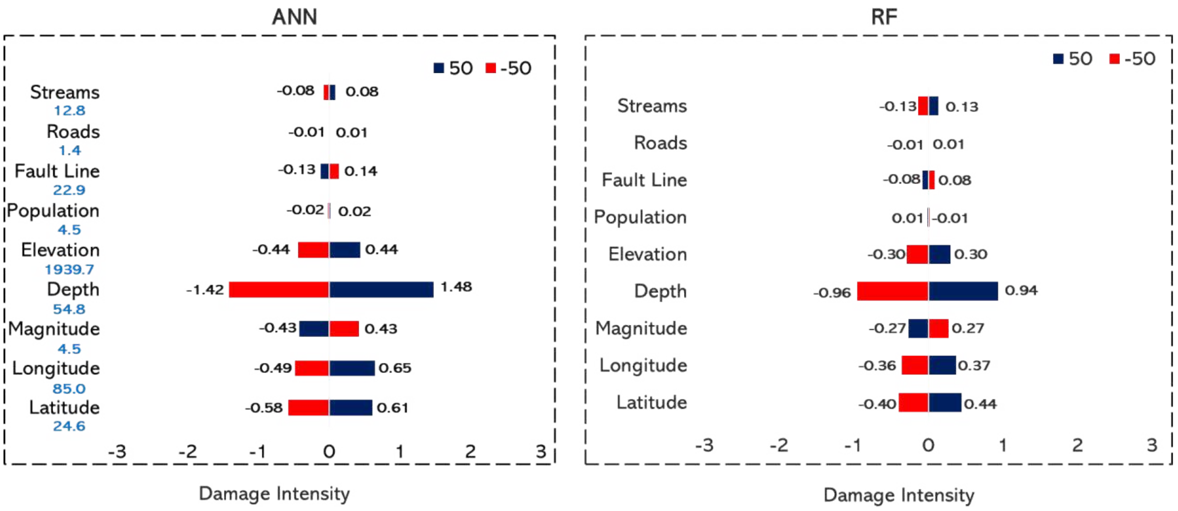
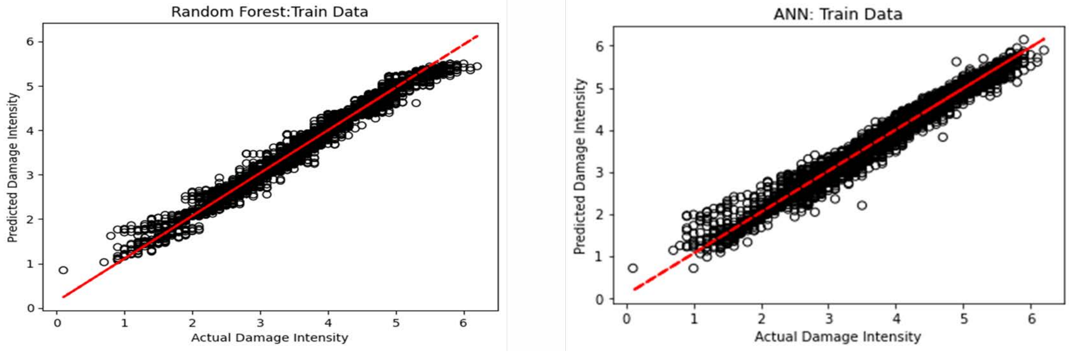


Figure 13. Sensitivity analysis for (a) ANN, (b) RF 20

RESULT

The training stage performance of both the models has been depicted in Fig —.



(a)

(b)

Figure 14. Training data scatter plot for (a) RF, (b) ANN 4

Data Points: 19040, (70% of the dataset)	RF	ANN
R^2 (Coefficient of Correlation)	0.98	0.98
MAE (Mean Absolute Error)	0.08	0.07
PRMSE (Percentage Root Mean Square Error)	6.4%	6.08%

Table 4. Model performance (training dataset) 8

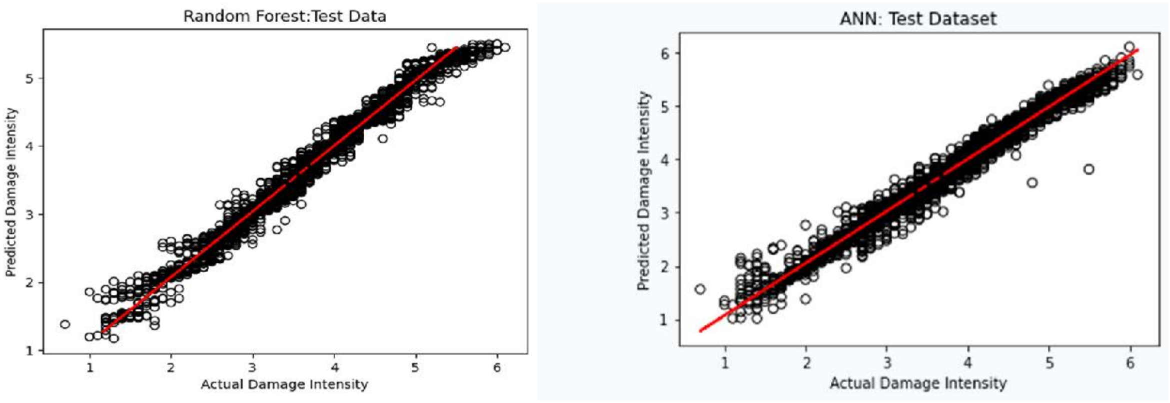


Figure 15. Test data scatter plot for (a) RF, (b) ANN 3

Data Points: 8160, (30% of the dataset)	RF	ANN
R (Coefficient of Determination)	0.99	0.99
MAE (Mean Absolute Error)	0.08	0.08
PRMSE (Percentage Root Mean Square Error)	4.02%	7.49%

Table 5. Model performance (testing dataset) 8

From the scatter plots, the values obtained from the model are close to the actual values, along with some outlier values. Similar results were seen upon testing the model with the remaining 8160 points of the datasets. Other metrics like the coefficient of determination, mean absolute error, and percentage root mean square error was utilized to understand the model's accuracy. From the table, it can be seen that the model is performing exceptionally well for both the training and testing dataset.

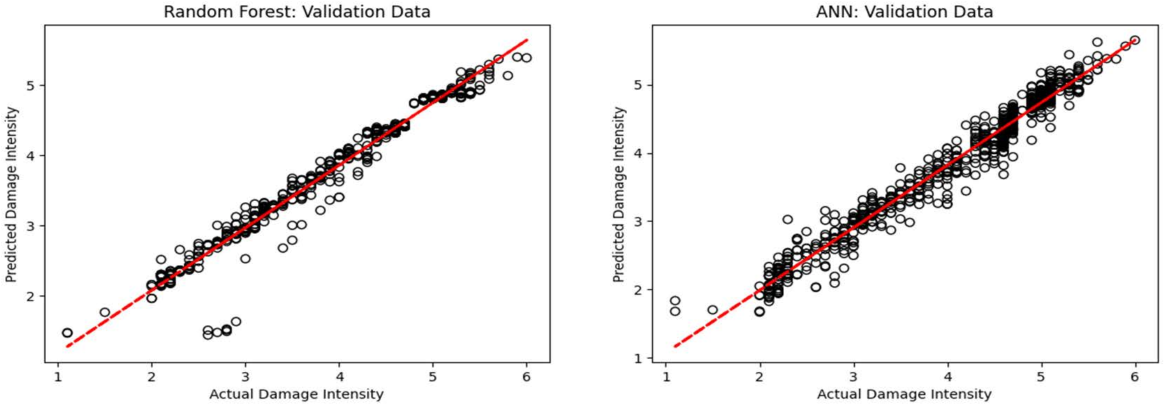


Figure 16. Validation data scatter plot for (a) RF, (b) ANN 18

Validation Dataset: June 2021 – September 2022 Data Points: 789	**RF**	**ANN**
R (Coefficient of Determination)	0.99	0.96
MAE (Mean Absolute Error)	0.08	0.22
PRMSE (Percentage Root Mean Square Error)	4.02%	7.51%

Table 6. Model performance (validation dataset) 4

Lastly, to rule out any temporal dependencies of the model on the dataset, a new dataset with six months of the temporal gap was run through the model. The training and testing datasets had events until December 2020, and the new dataset (Validation Dataset) with 789 data points consisted of the events post June 2021 until September 2022. For the validation dataset as well, it can be observed from the scatter plot and the metrics results that the model has performed exceptionally well.

Temporal Trend Analysis

The models performed fairly close; although the ANN model captured the peaks and dips of the damage intensity values, Random Forest performance was better in accuracy as it provided the average of the damage intensity values generated through each decision tree. Furthermore, a temporal analysis of the test dataset was plotted in Fig –to depict the trend comparison between the actual and model values.

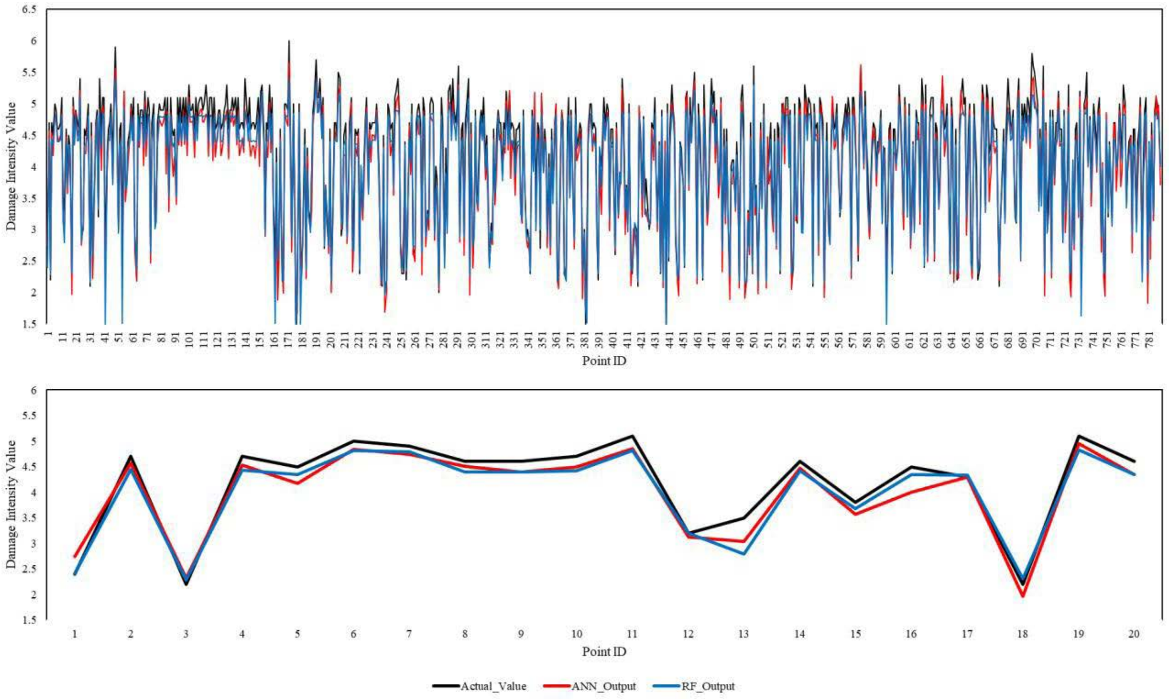


Figure 17. Validation Curves

Upscaling to Spatial Scale

The resulting models were used to interpolate and upscale to a spatial scale of a 2 km grid. This approach will provide a good understanding of the spatial pattern and distribution over India. The classes were defined using the Modified Mercalli Intensity Scale as the reference to quantify the vulnerability. 8

Vulnerability Level	MMI Scale	Description
Very Low	II	Minor- rarely felt
Low	III	Minor- noticed by a few people
Moderate	IV	Light- felt by many people, minor damage
High	V	Significant damage
Very High	VI-VII	Damage variables depend upon building construction and substrate

Table 7. MMI Scale levels

The following maps show the different vulnerability levels in India. It can be observed that even though the events of earthquakes are more frequent along the Northern belt, the vulnerability level is higher along the Southern coast. This means that if there is an earthquake of around the same magnitude along the northern belt and the southern coast, it would be more devastating along the southern coast.

In the figure below, both the model outputs can be seen on the spatial front. It can be observed that most regions depicted as high and very high vulnerability areas are the regions where the frequency of earthquakes is lesser compared to the areas classified as shallow and low-risk areas. This means that an earthquake of magnitude will be more devastating for the regions along southern and central India, mainly because there is a lot of population and infrastructure concentrated in that area, thus resulting in more damage. 11

The following maps show the different vulnerability levels in India. It can be observed that even. 13

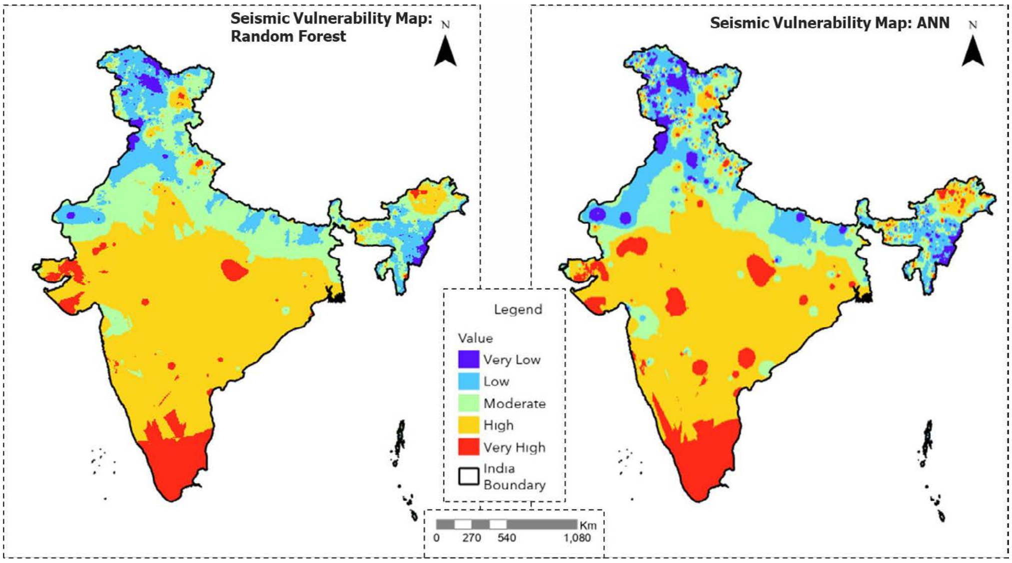
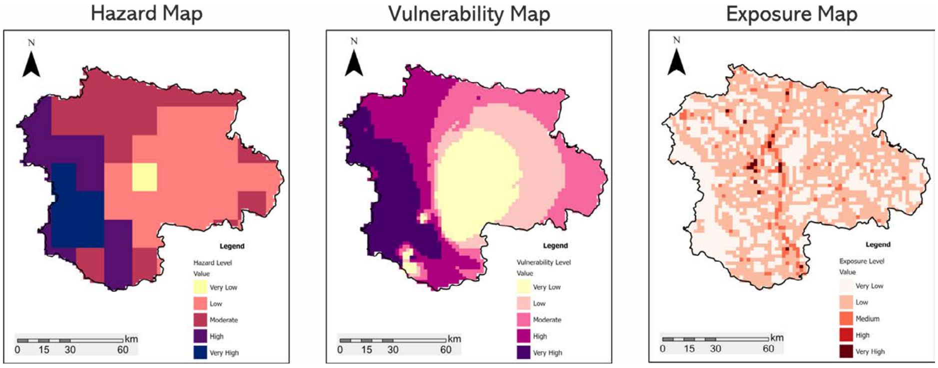


Figure 18. Seismic Vulnerability map for India (Model outputs)

CASE STUDY EXAMPLE

A risk map for the entire India can be generated using the Seismic vulnerability map. For example, a risk map for Satara district, Maharashtra, was generated as the application area of this study. The Satara district of Maharashtra experienced an earthquake of 5.0 magnitude on 16

September 2008. It has an area of 10,480 km^2 and a population of 3,003,741, of which 14.17% were urban (as of 2011). Although the magnitude wasn't much, the damage was alarming. Nearly 606 buildings were severely damaged in 110 villages, and another 573 buildings in 92 villages in Patan taluka also inflicted minor damage within 100 kilometers of the epicenter. This engenders a need to be prepared if such an event happens again. For this purpose, a hazard map was generated from the Global Seismic Hazard map. The exposure map was created after overlaying maps of different areas like schools, banks, and other public spaces. Finally, a risk map was created. For Satara district to understand the potential losses in terms of lives, health, economy, and livelihood. The risk map was categorized into five risk levels, and the at-risk population was estimated. The risk map depicts that around 40 % of the district falls under moderate to high-risk levels, with 44 % of the population at risk. 17



(a) (b) (c)

Figure 19. (a) A hazard map,(b) a Vulnerability map, and (c) an Exposure map for the Satara district of Maharashtra

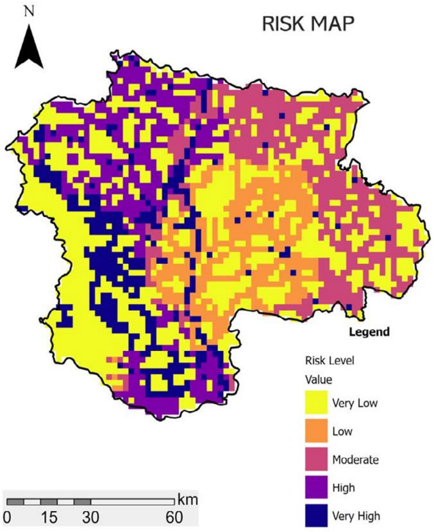


Figure 20. Satara district Risk level map

DISCUSSION

Both the models, Artificial Neural Network (ANN) and Random Forest (RF), worked quite well for this study. The performance of the models was similar in terms of accuracy; however, RF performed slightly better than the ANN model. As seen from the graphs and maps, the ANN model could capture the peaks and dips of the data values closely, giving a better geospatial representation.

However, in the case of the Random Forest model, the output values generated were closer to the actual values. Still, this model could not capture the high and low values as a random forest is a bagging ensemble model that takes the output values from all the decision trees and provides an average of those values as a final output.

As this study has been done on a coarser resolution of 2 km X 2 km, moving forward, finer resolution datasets can be taken to understand the seismic vulnerability of an area. More parameters for the models can be added, like building materials, age-wise population, etc., can be taken to make the study more holistic. Lastly, other models like LSTM (Long Short Term Memory), TLBO (Teaching Learning Based Optimization), etc., can also be used for similar studies.

CONCLUSION

Earthquakes pose significant risks to human lives, infrastructure, and economies, making them one of the most dangerous and unpredictable natural hazards. Recent studies have focused on assessing the potential impact of a future earthquake in the Himalayan region, specifically in terms of magnitude and the resulting consequences. This study utilized a Machine Learning approach to model India’s earthquake vulnerability, considering various geological, physical, and social factors that contribute to seismic vulnerability. To estimate the potential impact, the study considered earthquakes in the past and calculated the damage intensity associated with each event. Machine Learning algorithms, specifically Artificial Neural Networks and Random

Forests, are employed to map the distribution of seismically vulnerable areas into five categories: Very High, High, Moderate, Low, and Very Low. Both the Artificial Neural Networks and Random Forest models demonstrated similar accuracy in predicting seismic vulnerability. The models used the collected data to generate a zonation map, categorizing regions based on their vulnerability to earthquakes. The map provides a visual representation of the exposure grid, indicating areas with varying degrees of vulnerability and potential damage intensity.

This analysis and vulnerability index map can be valuable tools for prioritizing regions that require immediate risk reduction interventions. Additionally, the vulnerability index map facilitates a comprehensive understanding of risk metrics associated with different areas, allowing for more targeted and effective risk reduction strategies. The findings can guide decision-makers in allocating resources and implementing measures to minimize the potential loss of human lives and financial damage caused by future earthquakes. 22

ACKNOWLEDGMENT

We want to thank the Faculty of Technology, CEPT University, and CEPT University management for providing the infrastructure and support to call out this study.

AUTHORS STATEMENT

Shalu did the data acquisition and data processing, developed the ANN model, conducted validation and analysis, and drafted the manuscript. Twinkle Acharya developed the Random forest model and edited the manuscript. Dhwanilnath Gharekhan guided and supervised the analysis and model development and drafted and edited the manuscript. Dipak R Samal guided and supervised the analysis and model development.

REFERENCES

Bureau of Indian Standards. *Bureau of Indian Standards*. 2002. Document. 7 10 2021.

Choudhary, Srishti. *mint*. 25 September 2019. Article. 7 10 2021.

David J. Wald, V. Q. (1999). Relationships between Peak Ground Acceleration, Peak

Ground Velocity, and Modified Mercalli Intensity in California. *SAGE*, 557-564. doi:10.1193/1.1586058

Dowlatabadi, S. M. Blower and H. "Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: An HIV Model, as an Example." *International Statistical Review (ISR)* (1994): 229-243. Article.

Ehsan Harirchian, V. K. (2021). A Synthesized Study Based on Machine Learning

Approaches for Rapid Classifying Earthquake Damage Grades to RC Buildings. *Applied Sciences, MDPI*, 1-33. doi:10.3390/app11167540

Franceschini S, Tancioni L, Lorenzoni M,Mattei F, Scardi M. "An ecologically constrained procedure for sensitivity analysis of Artificial Neural Networks and other empirical models." *PLOS One* (2019). Research Article.

Ju Han, A. S.-Y. (2021). Improvement of Earthquake Risk Awareness and Seismic. *Remote Sensing, MDPI*. doi:10.3390/rs13071365

K.S. Valdiya, Jaishri Sanwal. "The Dynamic Indian Crust." *Elsevier* (2017): 1-14. Document.

M. J. D. De Los Santos, J. A. (2021). GIS-BASED RAPID EARTHQUAKE EXPOSURE AND VULNERABILITY MAPPING. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVI-4*. doi:10.5194/isprs- archives-XLVI-4-W6-2021-125-2021 Manish. *STUDYIQ*. 13 February 2023. Article. 07 March 2023. <https://www.studyiq.com/articles/earthquakes-in-india/>.

Mithila Verma, Brijesh K. Bansal. "Active fault research in India: achievements and future perspective." *Geomatics, Natural Hazard and Risk* VII (2013):

65-84. Document.

Muriel Gevreya, Ioannis Dimopoulosb, Sovan Lek. “Two-way interaction of input variables in the sensitivity analysis of neural network models.” *Elsevier* (2006): 43-50. Journal Article.

Murthy, C. (2005). Earthquake Tips: Learning Earthquake Design and Construction. In C. Murthy, *IITK-bmTPC Earthquake Tips: Learning Earthquake Design and Construction* (pp. 27-28). Kanpur: National Programme on Earthquake Engineering Education.

Peyman Yariyan, M. A. (2020). Earthquake Vulnerability Mapping Using Different Hybrid Models. *Symmetry, MDPI*. doi:10.3390/sym12030405

Ravi Sinha, K.S.P. Aditya and Achin Gupta. “GIS-BASED URBAN SEISMIC RISK ASSESSMENT USING RISK.IITB.” *ISET Journal of Earthquake Technology* (2008): 3-4. Document.

Ruder, Sebastian. “An overview of gradient descent optimization.” (2017). Article.

Sazli, Murat H. “A brief review of feed-forward neural networks.” (2006): 11-17. Article

Saro Lee, M. P. (2019). SEVUCAS: A Novel GIS-Based Machine Learning Software for

Seismic Vulnerability Assessment. *Applied Sciences, MDPI*. doi:10.3390/app9173495

Tati Zera, A. R. (2021). Mapping of Peak Ground Acceleration Values using the Donovan Model for Sumatran. 9th Asian Physics Symposium (pp. 1-6). Journal of Physics. doi:10.1088/1742-6596/2243/1/012031

Y.-S. Park, S. Lek. “Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling.” Developments in Environmental Modelling (2016): 123-140. Chapter.