SEISMIC DAMAGE ASSESSMENT OF BUILDINGS USING RANDOM FOREST

Md. Ismail Monsury[1], Sohel Rana*[2], A. H. M Shahidul Islam[3], Shafayat Bin Ali[4] and M. Abdur Rahman Bhuiyan[5]

[1] *Former Research Assistant, Institute of Earthquake Engineering Research, Chittagong University of Engineering & Technology, Bangladesh, e-mail: ismailcuet21@gmail.com*
[2] *Assistant Professor, Institute of Earthquake Engineering Research, Chittagong University of Engineering & Technology, Bangladesh, e-mail: ranasohel@cuet.ac.bd*
[3] *Research Assistant, Institute of Earthquake Engineering Research, Chittagong University of Engineering & Technology, Bangladesh, e-mail: engr.ahmshahidul@gmail.com*
[4] *Associate Professor, Department of Disaster Engineering and Management, Chittagong University of Engineering & Technology, Bangladesh, e-mail: shafayat@cuet.ac.bd*
[5] *Professor, Department of Civil Engineering, Chittagong University of Engineering & Technology, Bangladesh, e-mail: arbhuiyance@cuet.ac.bd*

***Corresponding Author**

ABSTRACT

Seismic damage assessment of buildings is an essential task to predict the physical damage scenario of buildings to reduce socio-economic loss in a future earthquake. In this regard, machine learning (ML) is an emerging new horizon with the possibility to assess the damage state of a building due to an earthquake, depending on the parameters used to train ML models. Incorporating regional geology, site characteristics, various seismic vulnerability parameters, and earthquake characteristics will result in better prediction of seismic damage to buildings. This study incorporates these characteristics to train an ML model employing the Random Forest algorithm and attempts to predict the damage state of the buildings. The building data and information, including the damage states, were collected from the 1999 Duzce earthquake damage database. The trained model showed an accuracy rate of 55%. This study illustrates the application of ML techniques in seismic damage assessment, which will help to reduce future seismic damage and associated losses due to the damage of buildings.

Keywords: *Seismic damage, Random Forest, Duzce earthquake database*

1. INTRODUCTION

Seismic damage assessment of buildings is a pivotal task to enhance seismic resilience and pre-earthquake preparedness. In recent decades, urbanization, especially in seismically active regions, has significantly increased the risk of extensive structural damage and associated socioeconomic losses, due to unplanned urbanization, lack of quality control in construction, poor materials, in developing countries. To reduce seismic damage of buildings and social and economic losses due to the damage, it is necessary to assess the extent of structural damage due to any future earthquake for emergency response, recovery planning, and long-term seismic resilience. Traditional post-earthquake damage assessments rely heavily on manual inspections and resources to accumulate data. This approach is fruitful, but inherently time-consuming and labour-intensive. This method becomes a mammoth task when densely populated urban areas with large building portfolios need to be assessed. Over the past decade, advancements in artificial intelligence, particularly machine learning (ML), have opened a new avenue for automating and enhancing seismic damage assessment. ML models have shown promising capabilities in predicting structural performance based on building attributes, seismic inputs, and environmental conditions.

Application of ML methods to predict seismic damage is widely explored in the scientific community. Morfidis & Kostinakis, (2017) demonstrated that an artificial neural network (ANN) can be used for the predictive damage assessment of reinforced concrete (RC) buildings, both in pre- and post-earthquake scenarios. The study considered 11-14 seismic parameters and found that only 5 input parameters can provide a satisfactory result. A combination of fuzzy logic and ANN was employed in a study (Sánchez-Silva & García, 2001) for seismic damage assessment considering earthquake severity, soil conditions, and structural properties. Hwang et al. (2021) proposed a two-step ML method using simulation data, where at first a regression model was used to predict engineering demand parameter (e.g., story drift), followed by a classification model to identify collapse potential. Their case study on a 8-story RC building found XGBoost as a good ML model to predict damage. A rapid post-earthquake seismic damage evaluation procedure based on a convolutional neural network (CNN) was proposed (Lu et al., 2021). They utilized data-driven time-frequency distributions (TFDs) of ground motions and a building inventory for corresponding damage levels. Kim & Song, (2022) proposed deep neural network based on a variational autoencoder (VAE) for unsupervised, near real-time seismic damage identification. The model was trained on flexibility matrices derived from operational modal analysis and demonstrated satisfactory performance in recognizing both single and multiple seismic damage cases.

An ensemble learning method like random forest (RF) combines multiple trees that split data into random subsets and trains each tree on a random subset, leading to higher accuracy in prediction. It utilizes bagging and feature randomness to create a robust module (Breiman, 2001; Wang et al., 2022). The effectiveness of RF in the seismic damage classification problem is widely studied. Roeslin et al. (2020) used the post-earthquake survey data from the 2017 Puebla-Morelos earthquake to construct an RF model that achieved good accuracy in seismic damage prediction. Mangalathu et al. (2020) utilized seismic damage data from the South Napa earthquake, and their model had moderate accuracy using Random Forest.

From the review of previous studies, it has been found that most of the studies ignored important parameters, e.g., earthquake and site characteristics, to assess seismic damage. It is necessary to include earthquake characteristics and site soil condition along with important building parameters to predict seismic damage of buildings. Given the infinite possibilities of ML as an emerging technology for intelligent seismic damage assessment, this study aims to train an RF model using a proper earthquake damage inspection dataset. The case study considered building characteristics, vulnerability parameters, regional geology and soil characteristics to assess effectiveness and accuracy of RF to classify seismic damage state of the buildings. The outcome of the current study will be helpful for the earthquake engineers to utilize ML models to predict seismic damage of buildings.

2. METHODOLOGY

Earthquake damage data for the study were collected from the 1999 Düzce earthquake damage database, Structural Engineering Research Unit (SERU), Middle East Technical University, Ankara, Turkey (SERU). It contains data on 484 buildings, including their location, building characteristics and earthquake response, for the 1999 Düzce earthquake. The datasets were classified into 5 different classes according to damage state, namely collapsed/removed (C/R), none (N), light (L), moderate (M) and severe (S).

2.1 Data Pre-Processing

Twelve (12) parameters were taken as input features, namely: i) Building Height, ii) Building Age, iii) Torsional Irregularity, iv) Projection in Plan, v) Minimum Normalized Lateral Stiffness Index, vi) Minimum Normalized Lateral Strength Index, vii) Overhanging Ratio, viii) Normalized Redundancy Score, ix) Priority Index, x) Peak Ground Acceleration, xi) Peak Ground Velocity/ Peak Ground Acceleration, xii) Effective Spectral acceleration -Sa(T1). Any non-standard values for any of the parameters under consideration were replaced with “NaN " (Not a Number) so that the pandas data frame could detect them as missing data. After that, any missing data for the parameters “Building age” and “Overhanging ratio” were replaced with the mean of the column of the respective parameters using the SimpleImputer class from scikit-learn imputation tools. For a class column with a NaN value, the entire row was removed from the parameters under consideration, as a row with a missing class label cannot be used for supervised learning. Consequently, the dataset reduces to 483 nos of data.

2.1.1 Outlier Identification and Removal

Outliers are data points that vary significantly from the majority of the dataset. While training ML models, outliers may lead to inaccurate pattern finding, resulting in poor predictions (Kashyap, 2024). The Z-score method was used to detect outliers for the normally distributed feature, the minimum normalized lateral stiffness index (Figure 1), and the interquartile method was used to detect outliers for the skewed feature, the priority index (Figure 2). In the Z-score method, the three-sigma rule was employed to detect outliers. Any data point that was more than 3 standard deviations away from the mean was considered an outlier (GeeksforGeeks, 2025). Table 1 shows the parameters and output of the three sigma rules. The interquartile range was used to detect outliers in the priority index, where the interquartile range (IQR) was defined as the difference between the third quartile (Q3) and the first quartile (Q1). Then, the lower limit was the difference between Q1 and 1.5IQR, and the upper limit was the sum of Q3 and 1.5IQR (Khan Academy, n.d.). The values that exceeded these limits were considered outliers. Table 2 shows the parameters and limits of the interquartile range method.

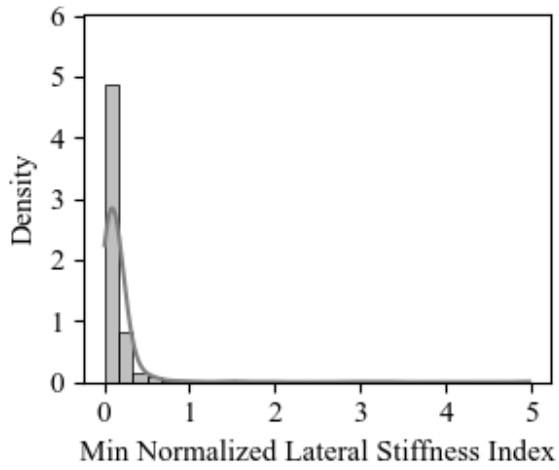

Figure 1: Distribution of minimum normalized lateral stiffness index

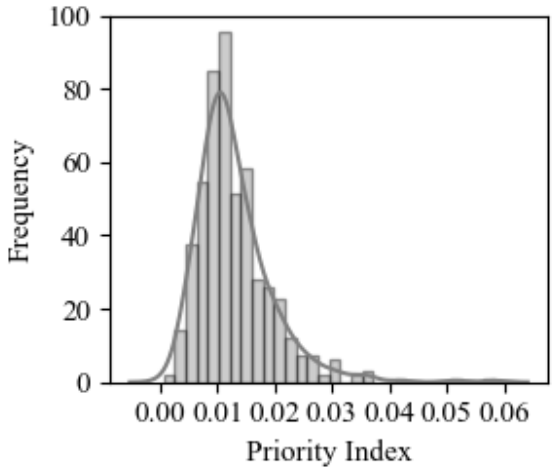

Figure 2: Distribution of priority index

Table 1: Parameters and output of the three-sigma rule

Mean, μ	Standard Deviation, σ	Upper limit, μ+3σ	Lower limit, μ-3σ
0.1546	0.3345	1.1581	-0.8489

Table 2: Parameters and limits of the interquartile range method

Third quartile (Q3)	First quartile (Q1)	Interquartile range (IQR)	Upper limit (Q3+1.5IQR)	Lower limit (Q1-1.5IQR)
0.0155	0.0087	0.0068	0.0257	-0.0015

After finding the upper and lower limits for the corresponding input features, any values that exceeded these threshold values were replaced with the upper or lower limit. This was done both for the minimum normalized lateral stiffness index and the priority index.

2.2 Random Forest Model Preparation

Two variables, x and y, were defined, where the 12 input parameters were stored in the x variable, and the class categories were stored in y via integer location-based indexing (iloc). The train variable and the test variable were set such that 80% of the data were used for training the ML model and 20% of the data were used for testing the ML model. This split was set to be random to get a representative sample of the whole dataset, and the random seed was fixed to 56 to get the same random shuffle every single run of the model. This also ensures the reproducibility of the ML model using the same dataset and indicates any changes in performance are due to the model or data updates, not because of randomness (Kaggle, n.d.; scikit-learn, 2025). The damage class definition (y) is a multiclass target, which is a categorical value that needs to be converted into an equivalent numerical value. This was done by the LabelEncoder class from sklearn. preprocessing Python library (scikit-learn, 2025).

The 12 input features used in this study were not on the same scale. Scaling of the features is required so that ML algorithms do not heavily favour features with larger values. Also, the scaling of the input features was done using the StandardScaler tool from sklearn.preprocessing library after the splitting of the dataset to prevent data leakage. These scales the input features such that each feature has a mean of zero and a standard deviation of one (scikit-learn, 2025).

The model was created with RandomForestClassifier with 75% of the training data for training each tree in the forest and a random seed of 42. Then, predictions were made on the test dataset, and the accuracy score was determined from the similarity of the test class and the predicted class.
The 4 classes defined in the dataset did not have an equal number of data points for each class. The imbalanced dataset was cross-validated using stratified k-fold cross-validation, ensuring the same percentage of representative data of each class in each fold. The data was split into 10 folds, such that each fold has the same proportion of class labels. The model was trained and tested 10 times, using a different fold as the test set each time.

Distinguished parameters that play a vital role in training outcomes of ML models are known as hyperparameters. Tuning of these parameters can highly boost the accuracy of the ML models (Hossain et al., 2021). A param_grid dictionary with 4 estimators, 3 maximum features, 3 maximum depth, and 3 maximum samples was defined, resulting in 108 combinations. Then, GridSearchCV, a hyperparameter optimization technique, was used to tune the ML model. Each combination was evaluated using 10-fold cross-validation, totalling 1080 fits. After that, the model's performance for each class was evaluated.

3. RESULTS AND DISCUSSIONS

The accuracy for the classification of damage classes using the RF model was 0.43. Table 3 shows the classification report for the untuned RF model. It was observed that the random forest model failed to recall the severe class accurately. Also, the recall score for collapsed/removed (C/R), light (L), and none (N) was less than 50 percent. Despite an accuracy rate of 0.43, the failure of this model to recall the severe class completely and the low rate of recall of other classes depict the poor performance of this ML model. Additionally, all the F1-scores were lower than 50%, resulting in a greater imbalance between precision and recall. This is primarily due to the class imbalance dataset used in this study. The confusion matrix is shown in Figure 3.

Table 3: Random Forest classification report before tuning

	Class	Encoded Label	Precision	Recall	F1-score	Support
	C/R	0	0.42	0.42	0.42	12
	L	1	0.47	0.44	0.45	32
	M	2	0.43	0.56	0.49	34
	N	3	0.44	0.40	0.42	10
	S	4	0.00	0.00	0.00	9
Accuracy					0.43	97
Macro Average			0.35	0.36	0.36	97
Weighted Average			0.4	0.43	0.41	97

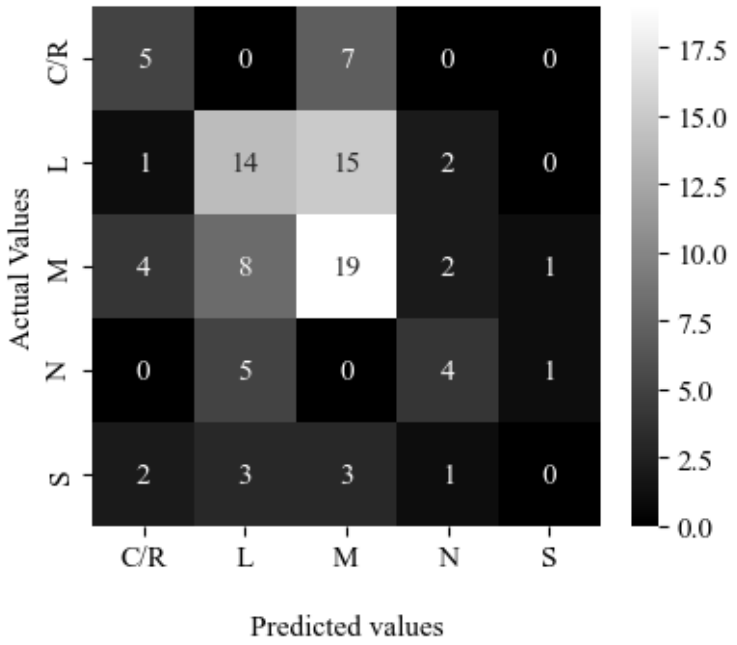

Figure 3: Confusion matrix before tuning

Cross-validation hyperparameter tuning with GridSearchCV modified this accuracy to 0.55. This score was obtained for parameters: maximum depth = none, maximum feature = 0.2, maximum sample = 1, and number of estimators = 100. The classification report is shown in Table 4.

Table 4: Random Forest classification report after tuning

	Class	Encoded Label	Precision	Recall	F1-score	Support
	C/R	0	0.50	0.50	0.50	12
	L	1	0.58	0.59	0.58	32
	M	2	0.52	0.65	0.58	34
	N	3	0.67	0.40	0.50	10
	S	4	0.50	0.22	0.31	9
Accuracy					0.55	97
Macro Average			0.55	0.47	0.49	97
Weighted Average			0.55	0.55	0.54	97

The overall precision, as well as the precision for the individual classes, increased significantly, all above 50%. Also, tuning the model improved the recall rate and was successful able to recalling each class. The F1-score also improved for all the classes. The confusion matrix is shown in Figure 4.

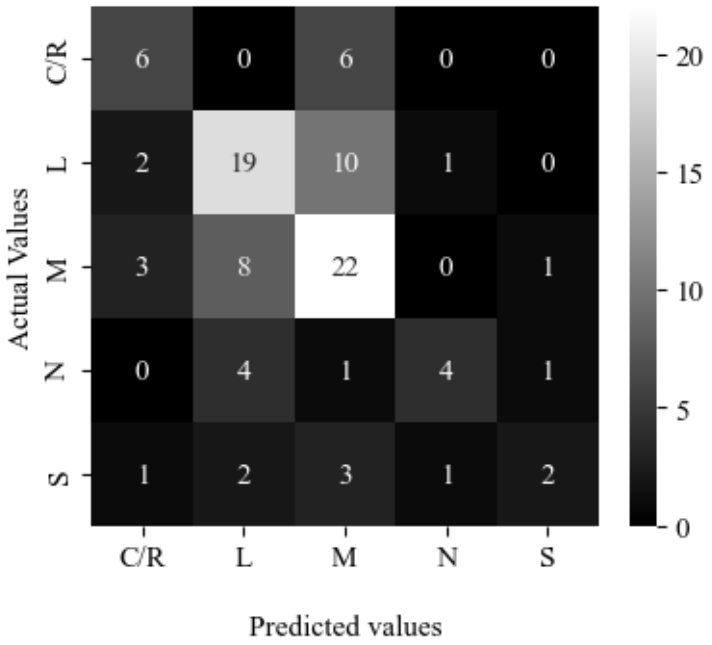

Figure 4: Confusion Matrix After Tuning

4. CONCLUSIONS

This study aimed to train a random forest model to predict seismic damage using the Düzce database. Selected seismic performance or vulnerability parameters as assessed after the Duzce earthquake and seismic intensity data for the earthquake were used to utilize the data in seismic damage prediction. Initially, the untuned model showed an accuracy of 0.433 with very poor recall for several damage classes considered. However, hyperparameter tuning boosted this accuracy of 0.55 with a good recall rate and precision greater than 50% for all the damage classes. In the study small dataset of real earthquake damage of building was used, use of larger dataset with adequate no. of data of each damage class may provide better result. This can be done mergering building damage from other post-earthquake inpection, or numerialaly investigated building data.

This highlights the suitability of machine learning techniques like random forest in predicting the seismic damage of buildings using the proper input parameters. This eventually opens a greater possibility for rapid pre-earthquake evaluation and swift action from the authorities, minimizing the future loss of life and resources. However, improvement of the results must be made, if possible, to get reliable outcome, which may provide more realistic damage performance as depicted in physical, analytical or numerical investigation. In addition, use of other ML techniques can also be utilize with proper consider of parameters and associated data in seismic damage assessment of buildings. Application of machine learning kind of technique is really useful not only considering their rapid assessment, but also considering the overcrowded cities like Dhaka, where on-site investigation is a very difficult task. However, reliable justification must be made with analytical or numerical assessment of buildings for such cases.

ACKNOWLEDGEMENTS

The current study is a part of a study that was funded by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET). The authors are thankful to the funding authority to provide grants to conduct to the study.

REFERENCES

Alcantara, E. A. M. ;, Saito, T., Barrera, E., Azuara, G., Alcantara, A. M., & Saito, T. (2023). Machine Learning-Based Rapid Post-Earthquake Damage Detection of RC Resisting-Moment Frame Buildings. *Sensors, 23*(10), 4694. https://doi.org/10.3390/S23104694

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Demir, A., Sahin, E. K., & Demir, S. (2024). Advanced tree-based machine learning methods for predicting the seismic response of regular and irregular RC frames. *Structures*, *64*, 106524. https://doi.org/10.1016/j.istruc.2024.106524

GeeksforGeeks. (2025). *Z score for Outlier Detection*. Available at https://www.geeksforgeeks.org/machine-learning/z-score-for-outlier-detection-python/, Last Assessed on October, 2025.

Hossain, Md. R., Timmer, D., & Moya, H. (2021). Machine learning model optimization with hyper-parameter tuning approach. *International Conference on Advanced Engineering, Technology and Applications (ICAETA)*, August 2021, Istanbul, Turkey.

Hwang, S. H., Mangalathu, S., Shin, J., & Jeon, J. S. (2021). Machine learning-based approaches for seismic demand and collapse of ductile reinforced concrete building frames. *Journal of Building Engineering*, *34*, 101905. https://doi.org/10.1016/j.jobe.2020.101905

Kaggle. (n.d.). *What actually random_state does?* Retrieved on September 22, 2025, from https://www.kaggle.com/discussions/questions-and-answers/465033

Kashyap, P. (2024). *Outlier Detection and Removal in Machine Learning*. Last Assessed on October, 2025 at https://medium.com/@piyushkashyap045/outlier-detection-and-removal-in-machine-learning-a-beginners-guide-c3e0c43af6cf

Khan Academy. (n.d.). *Identifying outliers with the 1.5xIQR rule*. Retrieved on September 21, 2025, from https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule

Kim, M., & Song, J. (2022). Near-Real-Time Identification of Seismic Damage Using Unsupervised Deep Neural Network. *Journal of Engineering Mechanics*, *148*(3), 04022006. https://doi.org/10.1061/(ASCE)EM.1943-7889.0002066Lu, X., Xu, Y., Tian, Y., Cetiner, B., & Taciroglu, E. (2021). A deep learning approach to rapid regional post-event seismic damage assessment using time-frequency distributions of ground motions. *Earthquake Engineering & Structural Dynamics*, *50*(6), 1612–1627. https://doi.org/10.1002/eqe.3415

Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. (2020). Classifying earthquake damage to buildings using machine learning. *Earthquake Spectra*, *36*(1), 183–208. https://doi.org/10.1177/8755293019878137

Morfidis, K., & Kostinakis, K. (2017). Seismic parameters' combinations for the optimum prediction of the damage state of R/C buildings using neural networks. *Advances in Engineering Software*, *106*, 1–16. https://doi.org/10.1016/j.advengsoft.2017.01.001

Roeslin, S., Ma, Q., Juárez-Garcia, H., Gómez-Bernal, A., Wicker, J., & Wotherspoon, L. (2020). A machine learning damage prediction model for the 2017 Puebla-Morelos, Mexico, earthquake. *Earthquake Spectra*, *36*(2_suppl), 314–339. https://doi.org/10.1177/8755293020936714

Sánchez-Silva, M., & García, L. (2001). Earthquake damage assessment based on fuzzy loǵic and neural networks. *Earthquake Spectra*, *17*(1), 89–112. https://doi.org/10.1193/1.1586168

scikit-learn. (2025). *Machine learning in Python — scikit-learn 1.7.2 documentation*. https://scikit-learn.org/stable/. Last acessed on September, 2025.

SERU. *METU Civil Engineering Department Structural Engineering Research Unit*. Retrieved on December, 2023, from https://seru.metu.edu.tr/archives.html

Wang, X., Mazumder, R. K., Salarieh, B., Salman, A. M., Shafieezadeh, A., & Li, Y. (2022). Machine Learning for Risk and Resilience Assessment in Structural Engineering: Progress and Future Trends. *Journal of Structural Engineering*, *148*(8). https://doi.org/10.1061/(ASCE)ST.1943-541X.0003392