MACHINE LEARNING CLASSIFICATION MODELS FOR SEISMIC VULNERABILITY OF BUILDINGS

Athira Anilkumar[1], Shemin T John[1], Pradip Sarkar[1*], Robin Davis[2]

[1]National Institute of Technology Rourkela, India, 769008

[2]National Institute of Technology Calicut, India, 673601

* Corresponding author

doi: https://doi.org/10.21467/proceedings.179.30

ABSTRACT

Predicting structural performance under seismic excitation and damage after the seismic event plays a vital role in life safety and economics. This study examines the vulnerability to earthquakes of a two-story 2D steel moment-resisting frame (SMRF) frame using the OpenSeesPy framework. The study utilizes non-linear static and time history analyses to evaluate the structures' reaction to seismic forces. The pushover analysis offers vital insights into the nonlinear behaviour of structures, whereas time history analysis considers the actual ground motion patterns in real-time. There is a lack of comprehensive assessments of machine learning (ML) progress within different areas of structural engineering. This includes a thorough examination of existing literature that can offer a timely evaluation of methods for assessing risk and resilience in the built environment. This study examines the performance of the selected ML classification techniques, such as Random Forest(RF), Logistic Regression, Decision Tree, K-Nearest Neighbours(KNN), LightGBM, CatBoost, Naïve Bayes, XGBoost, and AdaBoost. Binary classification is done to classify the data set as having low and high inter-storey drift. These predictions are useful to the government and other private companies for preparing an effective methodology to be followed after seismic events, efficient retrofitting and rehabilitation to extend the durability of existing structures, insurance estimates, decision-making in disaster risk reduction, and others.

1. INTRODUCTION

Conducting vulnerability assessments in regions prone to seismic activity is essential for comprehending and reducing the potential consequences of earthquakes. Understanding vulnerability aids in creating and implementing strict building regulations, guaranteeing that buildings are engineered to endure seismic forces. Additionally, it plays a crucial role in preparing for and responding to emergencies, enabling the implementation of focused measures in regions with greater susceptibility and enhancing the efficiency of recovery activities following an earthquake. Integrating OpenSees in seismic vulnerability assessment enhances its capability as it is a powerful tool that facilitates the analysis and understanding of structural responses [1], [2]. Recently, OpenSeesPy, which uses the Python module as its scripting language and has a provision to provide plots using Python commands, was introduced[3]. Machine learning (ML) has gained significant attention and utility in structural engineering. Machine learning methods can improve the precision of predicting and classifying structure responses during seismic occurrences by training models using past earthquake damage data and including various characteristics [4], [5].

The present study targets this deficiency by assessing the effectiveness of machine learning classification methods. The dataset of inter-storey drift obtained from the seismic analysis in Openseespy is subjected to binary classification to categorize it into low and high inter-storey drift groups. The SHAP analysis is done

Proceedings DOI: 10.21467/proceedings.179; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-7-4

to rank the selected random variables. The predictions derived from this research hold significant value for government agencies and private firms in developing efficient post-seismic procedures, retrofitting methods, insurance estimation, catastrophe risk reduction decision-making, and other pertinent domains.

2. RESEARCH SIGNIFICANCE

The seismic vulnerability assessment for buildings using fragility curves is paramount as it enhances public safety, community resilience, and sustainable urban development. Moreover, the findings influence insurance practices, financial preparedness, and international collaboration on standardized approaches to seismic risk reduction. The present study employs a vulnerability assessment framework based on OpenSeesPy. This research integrates advanced computational techniques in seismic engineering. Incorporating machine learning to classify earthquake damage states based on inter-storey drift adds a cutting-edge dimension to traditional seismic vulnerability assessments. It catalyzes policy formulation, regulation updates, and public awareness initiatives, ultimately fostering a more resilient and informed society in the face of seismic challenges.

3. MODELLING AND ANALYSIS

The selected structure is a 2D non-existent two-story SMRF building with plan dimensions of 6 m in the X and a floor height of 3.5 m, as shown in Fig. 1. The cross-section adopted for the steel structure in the W-section is based on nominal dimension. In this structure, the columns are modelled as distributed plasticity elements, and beams are modelled as concentrated plastic elements. The plastic hinges are provided as two nodes at the same coordinates, connected via a rotational spring.

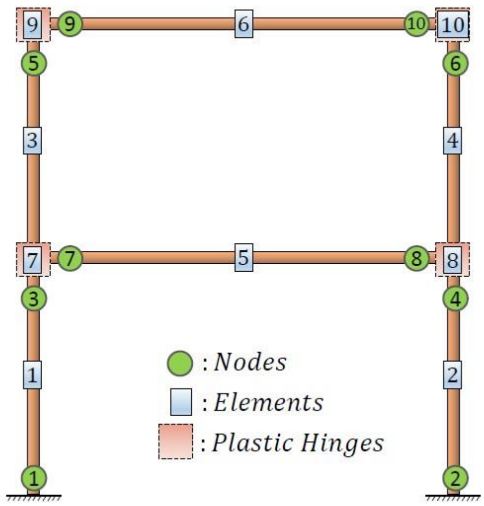


Fig 1: 2D Steel moment resisting frame model

The model is subjected to gravitational forces by utilizing the gravity function, and then modal analysis is conducted. In addition, the time history function performs non-linear time-history analysis (NLTHA) using the provided ground motion. The analysis techniques are performed using OpenSeesPy commands. The code additionally incorporates functions for resetting the analysis environment. In essence, the given code functions as a structure for constructing and evaluating a 2D steel moment-resistant frame construction. This enables users to examine its performance when subjected to different types of loads.

Time history analysis considers the dynamic properties of the structure and the applied loads rather than reducing the seismic forces into a single static load. The NLTHA incorporated the earthquake loading by utilizing a collection of 20 natural ground motion recordings, as listed in Table 1. Each ground motion data consists of one vertical component and two horizontal components.

Table 1: Selected Ground Motions for Time History Analysis

Sl No:	Earthquake	Station	Time Period (sec)
1	Northridge	Beverly Hills	29
2	Duzce Turkey	Bolu	55.9
3	Hector Mine	Hector	45.31
4	Imperial Valley	Delta	100.15
5	Kocaeli Turkey	Duzce	27.185
6	Kocaeli Turkey	Arcelik	30
7	Kobe Japan	Nishi Akashi	40.96
8	Landers	Coolwater	28.002
9	Superstition Hills	El Centro	59.995
10	Superstition Hills	POE road	22.3
11	Landers	Yermo Fire	44
12	Kobe Japan	Shin Osaka	40.96
13	Manjil Iran	Abbar	46
14	Loma Prieta	Capitola	39.005
15	Loma Prieta	Gilroy Array	39.945
16	Cape Mendocino	Rio dell overpass	36
17	Chi Taiwan	CHY-101	90
18	Chi Taiwan	TCU045	90
19	San Fernando	Hollywood LA	79.45
20	Friuli Italy	Tolmezzo	36.345

4. CLASSIFICATION OF ISD USING ML

For the binary classification of the inter-storey drift obtained from the seismic analysis with OpenseesPy, the methods employed were Logistic Regression, Decision Tree (DT), K-Nearest Neighbouring (KNN), Random Forest (RF), XGBoost, LightGBM, CatBoost, Naïve Bayes (NB), Support Vector Machine (SVM) and AdaBoost. These algorithms were selected from recently published studies.[6], [7]. Logistic regression is a statistical method employed for binary classification problems, in which the outcome variable is categorical and consists of two possible classes[8]. Decision Tree recursively partitions the input space into regions based on the feature values, guided by a set of decision rules learned from the training data. The random forest algorithm is a type of ensemble learning approach that builds numerous decision trees during the training process. It then combines the predictions of these trees by voting or averaging in order to enhance accuracy and reduce overfitting.

XGBoost, AdaBoost, CatBoost and LightGBM are boosting algorithms in machine learning, which are ensemble methods that combine weak learners sequentially to create strong learners, with each subsequent model focusing more on instances that previous models misclassified. AdaBoost is susceptible to noisy data and outliers, and its performance is generally poor when the weak learners are complex or lack diversity [9]. CatBoost automatically handles categorical variables by converting them into numerical values using advanced encoding techniques [10]. XGBoost enhances classic gradient-boosting algorithms by integrating regularisation approaches to reduce overfitting and enhance generalization performance[11]. LightGBM uses a tree-based learning algorithm and employs a novel technique called Gradient-based one-sided sampling (GOSS) to select the most informative data samples for training, resulting in faster computation and reduced memory usage [12]. KNN employs the majority class of the k nearest neighbours in the feature space to classify data points. The decision boundary of the KNN algorithm is highly adaptable and can effectively handle nonlinear classification tasks by adjusting to the distribution of the data [13]. Naive Bayes is a probabilistic classifier that applies Bayes' theorem and assumes that the features are mutually independent [14]. The SVM method functions by locating the optimal hyperplane that efficiently separates the data points into separate classes while also maximizing the margin between them [15].

5. DATASET AND METHODOLOGY

A machine learning task was performed utilizing a Python environment with Jupyter Notebook. The ML algorithm was built using Scikit-Learn, a library that offers several tools for model selection, model fitting, data preprocessing, and data evaluation [16]. The steel structure used in this study was previously taken to create the data set. For this, 20 pairs of ground motions were selected (2 horizontal components of each pair), resulting in 40 ground motions. These ground motions were linearly scaled from 0.1 to 1g without altering its natural characteristics[17]. Therefore, in total, there were 400 sets of ground motions. The significant random variables influencing the earthquake responses were identified as input parameters, and their statistical properties are summarized in Table 2. These random variables and ground motions were used as the parameters for the seismic analysis using OpenSeesPy to find the corresponding inter-storey drift (ISD). This data set was randomly divided into training (70%) and testing (30%) data.

As for the distribution of the obtained ISD values, those were classified as damage state 1 (DS1) and damage state 2 (DS2) drift; that is, the drift value of less than 1% was considered DS1, and a drift value of more than 1% was considered DS2. The proposed framework for the ML classification model is shown in Fig 2. As there are 2 damage states, the classification method adopted is binary classification.

Table 2. Statistical properties of significant random variables

Parameter	Designation	Mean (μ)	COV (%)	Probability Distribution	Source
Yield Strength of Steel (MPa)	f_y	355	7.6	Log-Normal	(Sadowski et al., 2015)
Elastic Modulus of steel (MPa)	E_s	255000	1	Log-Normal	(Anisha et al., 2023.)
Distributed Load (kN/m)	DL	20	5	Normal	Assumed
Concentrated Load (kN)	CL	50	5	Normal	Assumed
Lumped Mass (ton)	m	75	5	Normal	Assumed

The ML models were trained using the training data sets, and the performance of each ML model was assessed using the Confusion Matrix (CM). The confusion matrix is a matrix that visually shows the errors generated by machine learning models, hence indicating their performance. Fig 3 displays the depiction of the confusion matrix. The confusion matrix displays the true and expected classes, with the rows and columns corresponding to these classes. The diagonal elements of the confusion matrix indicate the

Proceedings DOI: 10.21467/proceedings.179
ISBN: 978-81-984081-7-4

Series: AIJR Proceedings
ISSN: 2582-3922

instances that have been accurately classified. The assessment of each machine learning model is quantified using precision, F1 score, recall, and accuracy, which are calculated using the confusion matrix [14].

The accuracy parameter, used to evaluate overall performance, is computed as the ratio of correctly classified examples to the total number of cases. Recall and precision are metrics employed to assess the accuracy of machine learning models in forecasting particular damage states. Precision is the measure of the ML models' ability to reliably diagnose the fraction of damage states.

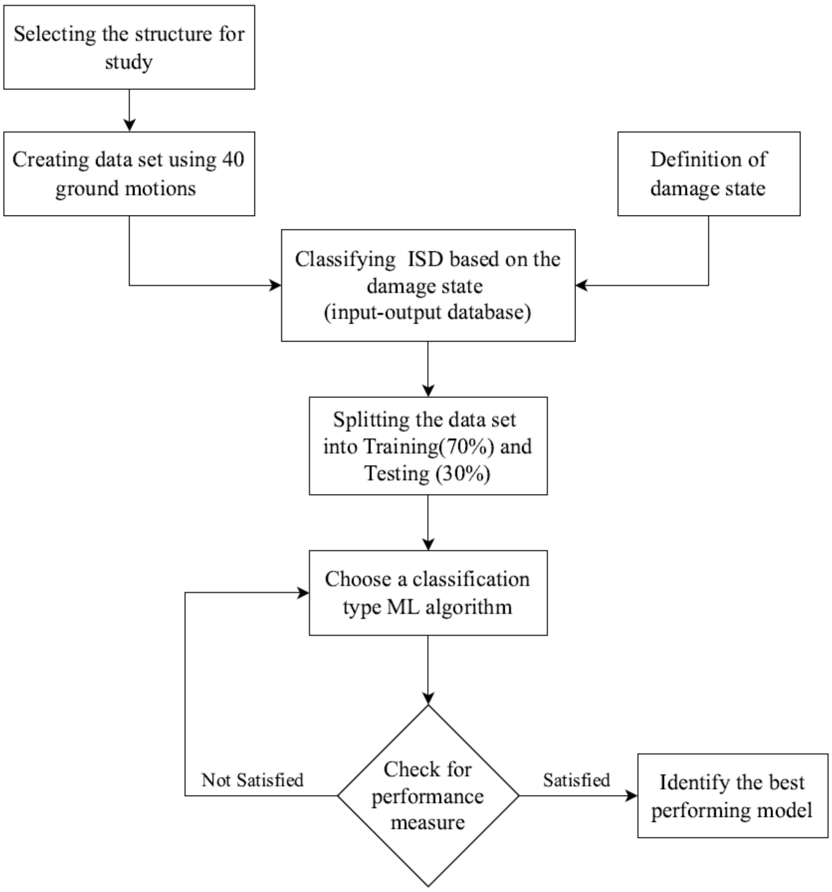
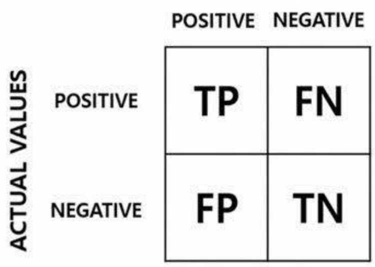


Fig 2. Methodology of ML-based classification

More precisely, accuracy refers to the ratio of accurate positive identifications to the total number of positive identifications. Recall, in contrast, denotes the ratio of correctly identified actual positives. The F1 score is a composite measure that combines recall and precision, offering a well-balanced evaluation of the model's performance. The formulas for all the parameters listed above are provided below.



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

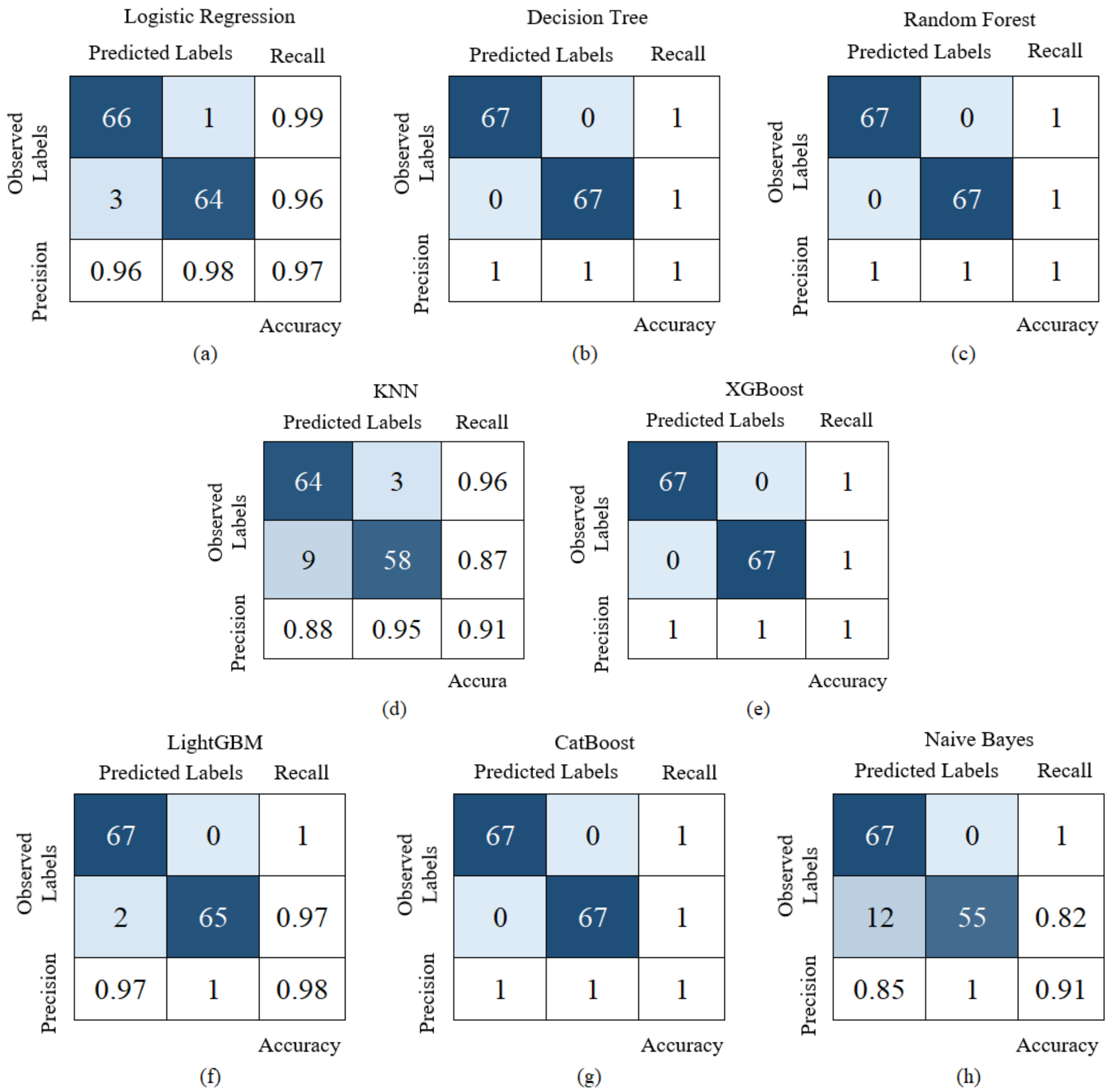
$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Fig 3: Concept of the confusion matrix

Proceedings DOI: 10.21467/proceedings.179
ISBN: 978-81-984081-7-4

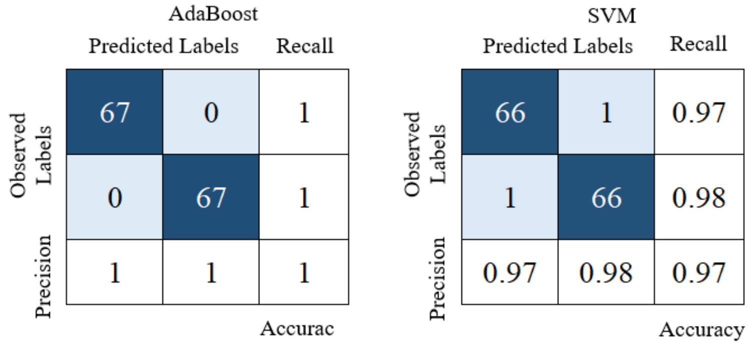
Series: AIJR Proceedings
ISSN: 2582-3922

Fig 4 illustrates the performance of the machine learning models on the testing data. RF and DT models demonstrated a remarkable accuracy rate of 100%. The logistic regression and support vector machine (SVM) models achieved considerable accuracy, accurately predicting 97% of the outcomes. It was observed that all other boosting models, including XGBoost, CatBoost, and AdaBoost, achieved a higher accuracy rate of 100%, except for LightGBM, which had an accuracy rate of 98%. The KNN and Naive Bayes models had the lowest accuracy, achieving a rate of 91% when compared to the other 10 machine learning models. The performance metrics for each machine-learning model are presented in Table 3. Tree-based machine learning models exhibited superior accuracy compared to the other models.



Proceedings DOI: 10.21467/proceedings.179
ISBN: 978-81-984081-7-4

Series: AIJR Proceedings
ISSN: 2582-3922



(i) (j)

Fig 4. Performance of ML models – (a) Logistic Curve, (b) DT, (c) RF, (d) k-NN, (e) XGB, (f) LGBM, (g) CB, (h) NB, (i) AdaBoost (j) SVM

Table 3: Summary of the performance of ML models based on the testing dataset.

Classifier	Accuracy	Precision	Recall	F1-score	ROC AUC
Logistic Regression	0.97	0.98	0.95	0.96	0.99
Decision Trees	1	1	1	1	1
k-NN	0.91	0.91	0.91	0.90	0.95
Random Forest	1	1	1	1	1
XGBoost	1	1	1	1	1
LightGBM	0.98	1	0.97	0.98	0.99
CatBoost	1	1	1	1	1
Naive Bayes	0.91	1	0.82	0.90	0.99
AdaBoost	1	1	1	1	1
SVM	0.97	0.98	0.98	0.97	0.97

The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are commonly used in binary classification models to evaluate the model's performance and to visualize the trade-off between sensitivity and specificity [18]. These curves were generated by plotting the true positive rate (TPR) against the false positive rate (FPR). The performance improves as the curve approaches the abscissa at $x = 0$ and the ordinate at $y = 1$. The AUC is a numerical measure that indicates how close the curve is to the two axes. As demonstrated, the models that utilize tree-based algorithms and boosting techniques exhibit superior performance, achieving the highest AUC scores. ROC curve obtained for each ML models are shown in Fig 5.

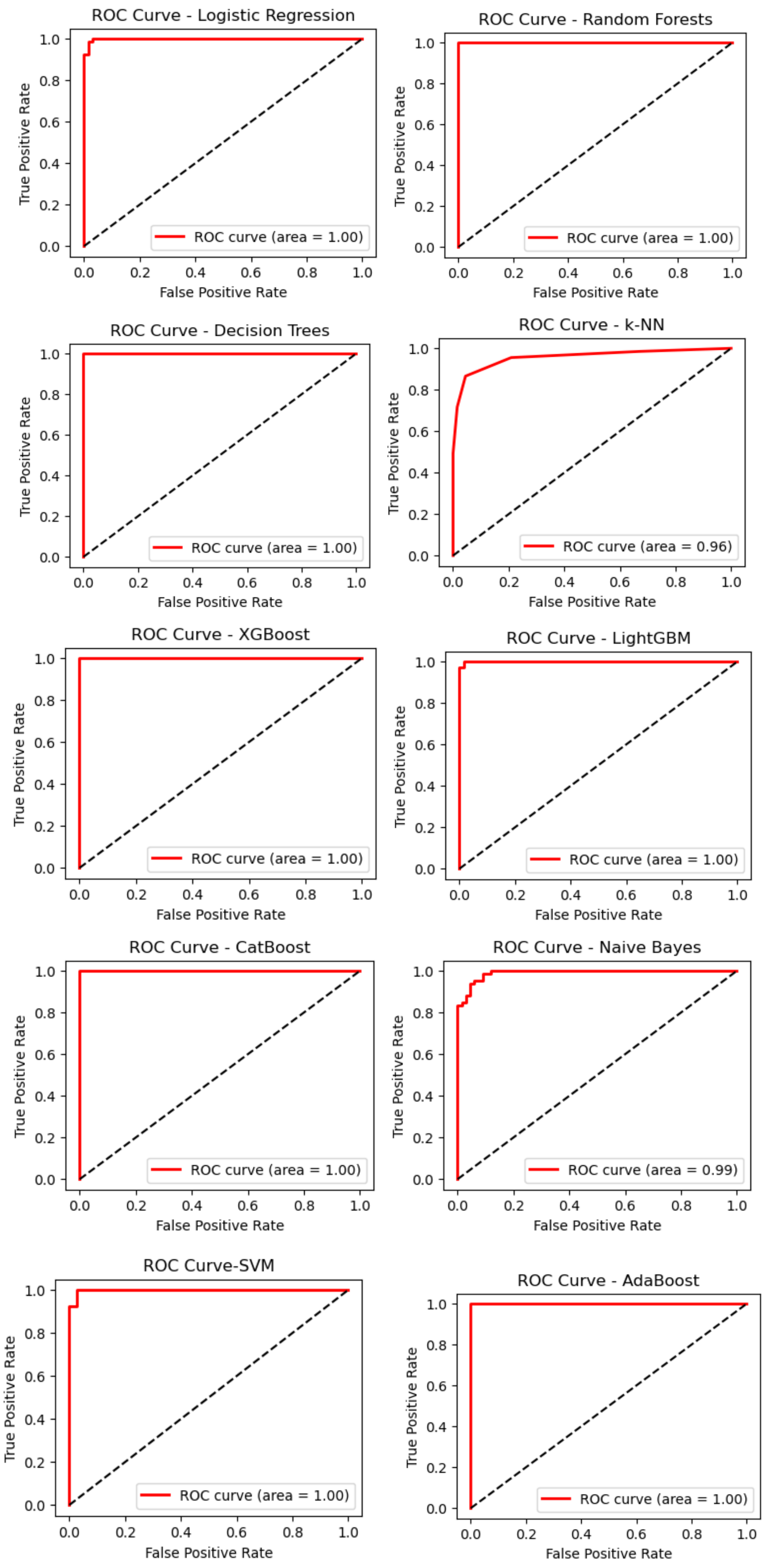


Fig 5. ROC Curve of each ML model.

Proceedings DOI: 10.21467/proceedings.179
ISBN: 978-81-984081-7-4

Series: AIJR Proceedings
ISSN: 2582-3922

6. SHAP ANALYSIS

The SHapley Additive exPlanations (SHAP) method, presented by Lundberg and Lee, is a highly effective and rigorous approach to interpreting machine learning models. The SHAP technique utilizes conditional expectation and game theory to reveal the impact of different input features on each seismic damage grade [19]. This method helps comprehend the ML model's inclination toward determining the damage states. The boosting ensemble utilizes the SHAP approach to analyze its operational pattern and the correlation between input features and seismic damage further.

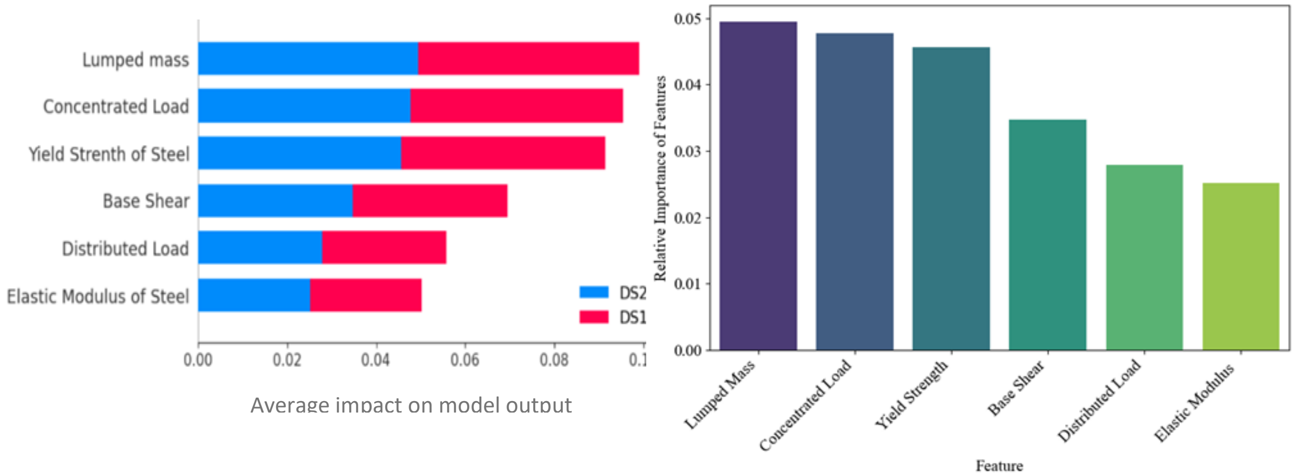
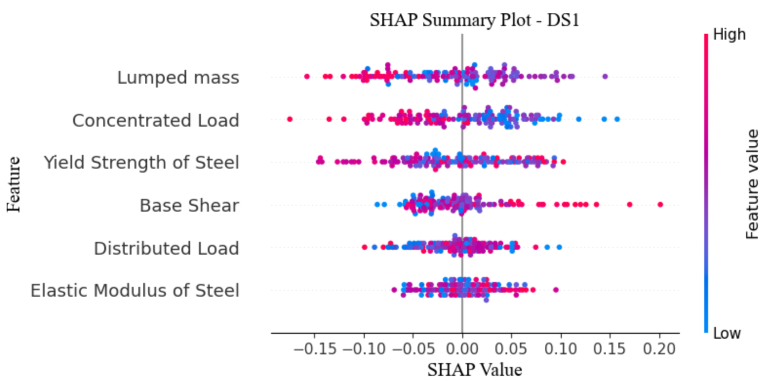


Fig 6: Important score and Relative significance of Input parameters in RF

Fig 6 presents a comprehensive overview of the significance of the score and the relative significance of each input characteristic in predicting outcomes using the RF model, as measured by the mean absolute SHAP value. The lumped mass is shown to be the most influential factor in determining the damage condition.

Figure 7 shows the SHAP summary plot for each tag, illustrating the significance and pattern of each input variable in predicting each damage state. The horizontal axis reflects the probability of prediction outcomes. The vertical axis represents the ordering of significant input variables, arranged in descending order from top to bottom. Positive values indicate a greater probability of prediction results, whereas negative values indicate a lower probability of prediction results. Based on the SHAP analysis, it is evident that the lumped mass is the most influential parameter, with the concentrated load at the nodes of the steel frame being the next significant factor. The importance score of the Elastic modulus of Steel was found to be the lowest.



Proceedings DOI: 10.21467/proceedings.179
ISBN: 978-81-984081-7-4

Series: AIJR Proceedings
ISSN: 2582-3922

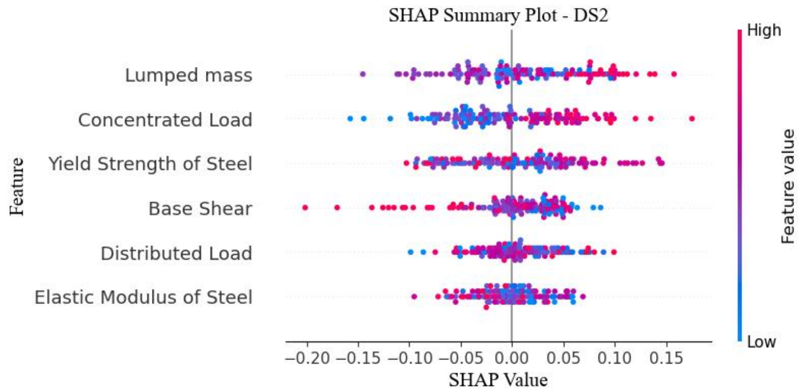


Fig 7: SHAP Summary plot for DS1 and DS2

7. CONCLUSION

This study employs ten machine learning algorithms to predict and classify structures into various damage states (DS1 and DS2). Among these algorithms, tree-based models such as Random Forest (RF), Decision Trees (DT), XGBoost, CatBoost, and AdaBoost demonstrated superior accuracy compared to other models. In contrast, K-Nearest Neighbors (KNN) and Naive Bayes models exhibited the lowest accuracy. A significant part of the study involved SHAP analysis of the RF model. This analysis revealed that the lumped mass is the primary parameter significantly impacting inter-storey drift, which is a critical factor in assessing structural damage. The SHAP summary image visually represents the precise impact of each variable on estimating the damage status, providing a clear and interpretable insight into the model's decision-making process. Overall, the study highlights the effectiveness of tree-based machine learning models in accurately predicting structural damage states and underscores the importance of lumped mass in influencing inter-storey drift. The use of SHAP analysis further enhances the interpretability of the model, making it easier to understand the contribution of different variables to the prediction outcomes.

Conflict of interest

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

REFERENCES

[1] A. Bhosale, N. P. Zade, R. Davis, and P. Sarkar, "Experimental Investigation of Autoclaved Aerated Concrete Masonry," *Journal of Materials in Civil Engineering*, vol. 31, no. 7, Jul. 2019, doi: 10.1061/(asce)mt.1943-5533.0002762.

[2] P. K. Dhir, N. P. Zade, A. Basu, R. Davis, and P. Sarkar, "Implications of Importance Factor on Seismic Design from 2000 SAC-FEMA Perspective," *ASCE ASME J Risk Uncertain Eng Syst A Civ Eng*, vol. 6, no. 2, Jun. 2020, doi: 10.1061/ajrua6.0001048.

[3] F. McKenna, "OpenSees: A Framework for Earthquake Engineering Simulation," *Comput Sci Eng*, vol. 13, no. 4, pp. 58–66, Jul. 2011, doi: 10.1109/MCSE.2011.66.

[4] P. Kourehpaz and C. Molina Hutt, "Machine Learning for Enhanced Regional Seismic Risk Assessments," *Journal of Structural Engineering*, vol. 148, no. 9, Sep. 2022, doi: 10.1061/(ASCE)ST.1943-541X.0003421.

[5] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," *Water (Switzerland)*, vol. 10, no. 11, pp. 1–40, 2018, doi: 10.3390/w10111536.

[6] J. Kiani, C. Camp, and S. Pezeshk, "On the application of machine learning techniques to derive seismic fragility curves," *Comput Struct*, vol. 218, pp. 108–122, Jul. 2019, doi: 10.1016/j.compstruc.2019.03.004.

[7] S. Tesfamariam and Z. Liu, "Earthquake induced damage classification for reinforced concrete buildings," *Structural Safety*, vol. 32, no. 2, pp. 154–164, Mar. 2010, doi: 10.1016/j.strusafe.2009.10.002.

[8] Y. Xie, M. Ebad Sichani, J. E. Padgett, and R. DesRoches, "The promise of implementing machine learning in earthquake engineering: A state-of-the-art review," *Earthquake Spectra*, vol. 36, no. 4, pp. 1769–1801, Nov. 2020, doi: 10.1177/8755293020919419.

[9] S. Cui, Y. Yin, D. Wang, Z. Li, and Y. Wang, "A stacking-based ensemble learning method for earthquake casualty prediction," *Appl Soft Comput*, vol. 101, Mar. 2021, doi: 10.1016/j.asoc.2020.107038.

[10] Y. Zhao and D. Gorse, "Earthquake prediction from seismic indicators using tree-based ensemble learning," *Natural Hazards*, vol. 120, no. 3, pp. 2283–2309, Feb. 2024, doi: 10.1007/s11069-023-06221-5.

[11] S. N. Somala, S. Chanda, M. AlHamaydeh, and S. Mangalathu, "Explainable XGBoost–SHAP Machine-Learning Model for Prediction of Ground Motion Duration in New Zealand," *Nat Hazards Rev*, vol. 25, no. 2, May 2024, doi: 10.1061/nhrefo.nheng-1837.

[12] C. Yang, K. Zhang, G. Chen, Y. Pan, L. Zhang, and L. Qu, "Application of Machine Learning to Determine Earthquake Hypocenter Location in Earthquake Early Warning," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2023.3348107.

[13] P. Hajibabaee, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "Dimensionality reduction techniques in structural and earthquake engineering," *Eng Struct*, vol. 278, Mar. 2023, doi: 10.1016/j.engstruct.2022.115485.

[14] H. D. Nguyen, J. M. LaFave, Y. J. Lee, and M. Shin, "Rapid seismic damage-state assessment of steel moment frames using machine learning," *Eng Struct*, vol. 252, Feb. 2022, doi: 10.1016/j.engstruct.2021.113737.

[15] L. MacEda, J. Llovido, and A. Satuito, "Categorization of earthquake-related tweets using machine learning approaches," in *Proceedings - 2018 International Symposium on Computer, Consumer and Control, IS3C 2018*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 229–232. doi: 10.1109/IS3C.2018.00065.

[16] S. Bhatta and J. Dang, "Seismic damage prediction of RC buildings using machine learning," *Earthq Eng Struct Dyn*, vol. 52, no. 11, pp. 3504–3527, Sep. 2023, doi: 10.1002/eqe.3907.

[17] N. P. Zade, B. Das, P. Sarkar, and R. Davis, "Seismic Performance of a New Capacity Design Scheme for RC Framed Building," 2022, *Taylor and Francis Ltd.* doi: 10.1080/13632469.2020.1838968.

[18] M. S. Abdalzaher, S. S. R. Moustafa, M. Abd-Elnaby, and M. Elwekeil, "Comparative Performance Assessments of Machine-Learning Methods for Artificial Seismic Sources Discrimination," *IEEE Access*, vol. 9, pp. 65524–65535, 2021, doi: 10.1109/ACCESS.2021.3076119.

[19] S. Mangalathu, S. H. Hwang, and J. S. Jeon, "Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach," *Eng Struct*, vol. 219, Sep. 2020, doi: 10.1016/j.engstruct.2020.110927.

Proceedings DOI: 10.21467/proceedings.179
ISBN: 978-81-984081-7-4

Series: AIJR Proceedings
ISSN: 2582-3922