Proceedings of the 10th International Conference on Civil Structural and Transportation Engineering (ICCSTE 2025)
July, 2025 | Imperial College London Conference Center, London, United Kingdom
Paper No. 347
DOI: 10.11159/iccste25.347

Integrating GIS and Machine Learning for Seismic Damage Assessment of Liquid Storage Tanks in California

FNU Tabish[1], Iraj H.P. Mamaghani[1], Raja Abubakar Khalid[1], Faisal Ahmed[1]
[1] Dept of Civil Engineering, University of North Dakota, 243 Centennial Drive, Stop 8115, Grand Forks, ND 58202, USA
fnu.tabish@und.edu; iraj.mamghani@und.edu; raja.khalid@und.edu; faisal.ahmed.1@und.edu

Abstract **-** This study explores the integration of Geographic Information Systems (GIS) and Machine Learning (ML) methods to assess seismic-induced damage categories of above-ground liquid storage tanks across California. A real-world damage tank dataset was considered to perform GIS-based spatial analysis, to identify clustering and distribution patterns. The results reveal that Southern California's most significant damage concentrations align with the region's high seismic vulnerability. To predict damage categories, four ML classifiers were evaluated: Decision Tree (DT), Random Forest (RF), XGBoost, and Support Vector Machine (SVM). Initial model performance was limited due to class imbalance in the dataset. The Random Forest model showed relatively better results compared to the others and was further improved using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. The enhanced model significantly improved classification performance, achieving training scores of 0.93 across all evaluation metrics. On the test set, the model attained a maximum precision of 0.75, a recall of 0.73, and an accuracy of 0.73. These findings demonstrate that combining Random Forest with SMOTE can effectively improve predictive accuracy and generalization in imbalanced datasets. Overall, this research highlights the practical application of GIS-based spatial analysis with ML techniques for seismic risk assessment and infrastructure resilience planning.
Keywords**:** Geographic Information Systems, Machine Learning, Storage Tanks, Damage Categories, Confusion Matrix

1. Introduction and Background

Liquid storage tanks are a fundamental element of current infrastructure that are used for a variety of purposes, like providing water supply, firefighting protection, and oil and dangerous matter storage in industrial installations. Because of their large number, many tanks are located in seismic areas [1-2]. Overturning, buckling, damage to the roof, sliding, uplift, and differential settlement are some common failure modes of these structures in several strong earthquakes. This failure can result in catastrophic consequences, such as liquid spilling and hazardous post-earthquake fires. And if not properly maintained or operated, it will cause severe hazards such as economic losses, environmental pollution, and human threats [3]. The timely prediction of earthquake-induced damage in liquid storage tanks is of real importance.

The response modelling and seismic induced damage assessment of liquid storage tanks are particularly challenging since there exists much uncertainty in the structural properties and seismic hazard parameters. High-fidelity numerical simulation tools, e.g., finite element method (FEM), provide a strong foundation to predict the dynamic response of tanks under seismic excitation with confidence [4]. The high computational costs of these models and the fact that risk must be assessed in the presence of several sources of uncertainty make computational risk assessment expensive and very time-consuming. Surrogate models have been developed to mitigate the computational cost in seismic hazard estimation compared to high-fidelity simulations [5]. These models give simple mathematical descriptions of complex physical systems. Various physics-based surrogate models were developed for steel liquid storage tanks. Balakis [6] proposed a performance-based seismic risk assessment approach for fixed-roof steel tanks by employing a surrogate single mass model composed of an elastic and a nonlinear component. The author stresses the importance of fragility analysis and identifies issues such as non-sequential damage propagation, with an emphasis on elephant's foot buckling. A neural-network proxy model was derived by Micheli and Laflamme [7] based on post-earthquake reconnaissance datasets for the estimation of seismic damage in steel storage tanks. The proposed cascade Neural Network (NN) model yielded better prediction performance than conventional models in that the cascade NN compensated for data imbalance and led to higher prediction accuracy at different damage levels. Quinci and Paolacci [8] developed a machine learning–based seismic risk assessment approach for industrial non-structural components using Artificial Neural Network (ANN) surrogate models, calibrated through nonlinear finite element (FEM) simulations. Their methodology integrates seismic hazard and vulnerability analyses

without making prior assumptions about fragility distributions. Tabish et al. [9] proposed a two-step Random Forest (RF) model to predict earthquake-induced damage states in liquid storage tanks. initially, a binary RF classifier predicted damaged and undamaged tanks, attaining training and test accuracies of 0.90 and 0.81, respectively. The second step employed a multi-class RF classifier with the ADASYN technique for data balancing. This enhanced approach yielded a training accuracy of 0.93 across evaluation metrics and a maximum test accuracy of 0.85, outperforming all other predicted classifier models.

Researchers have utilized GIS for spatial analysis and damage assessment in earthquake-affected areas. Hatayama [10] also employed GIS to find a spatial correlation between tsunami inundation height and damage to the oil storage tank during the 2011 Tohoku Earthquake. This geographic analysis allowed us to understand the pattern of damages at the regional scale and to derive impact functions for industrial tanks. Toprak [11] employed the GIS to investigate the spatial damage of the earthquake on lifeline systems, including the underground pipeline system. Their research showed that GIS could be used to map damaged areas and improve post-earthquake infrastructure evaluation. Liu et al [12] employed remote sensing with GIS techniques to accurately extract tsunami-flooded zones and assess damage to buildings following the 2011 Tohoku-Oki earthquake. Their GIS-based spatial analysis enabled detailed mapping of affected areas, facilitating improved post-disaster damage evaluation. This article utilizes GIS and ML techniques to assess the rapid damage risk induced by the earthquake in major cities of the West Coast. California, being a vulnerable earthquake area with a history of earthquake-induced damage to the liquid storage tanks, has been selected to implement the ML learning techniques and spatial analysis using GIS.

2. Database Used and Sources

The analytical database used in this study was compiled from a post-earthquake reconnaissance report [13], which documented observed seismic-induced damage to above-ground liquid storage tanks. The dataset consists of real-world damage records across a wide range of tank structures and seismic scenarios, enabling a robust evaluation of vulnerability patterns and damage mechanisms. The key predictive variables in the dataset include peak ground acceleration (*PGA*), earthquake location, geometric and material properties of tanks (e.g., diameter *D*, height *H*, shell thickness *t*, *H/D* and *D/t* ratios). Additional variables include the anchorage condition (anchored vs. unanchored) and the liquid height at the time of the seismic event. To ensure consistency in evaluating tank damage, each record was classified according to the HAZUS-MH (Hazards U.S. Multi-Hazard) framework into five distinct damage categories (DC0–DC4) based on severity and repair cost, as developed by the American Lifelines Alliance [13]. The damage categories range from DC 0 (no damage) to DC 4 (complete collapse), with intermediate levels capturing common failure modes such as roof deformation, elephant foot buckling, anchorage and piping failures, and significant loss of contents. A summary of these categories is provided in Table 1, while the statistical distribution of tanks across the damage states is illustrated in Figure 1. The dataset comprises a total of 134 tanks, of which approximately 20% (31 tanks) exhibited no damage (DC0). In contrast, complete collapse (DC4) was recorded in 10.2% of cases (16 tanks), representing a diverse and realistic range of damage severities observed in past earthquake events.

3. Methodology

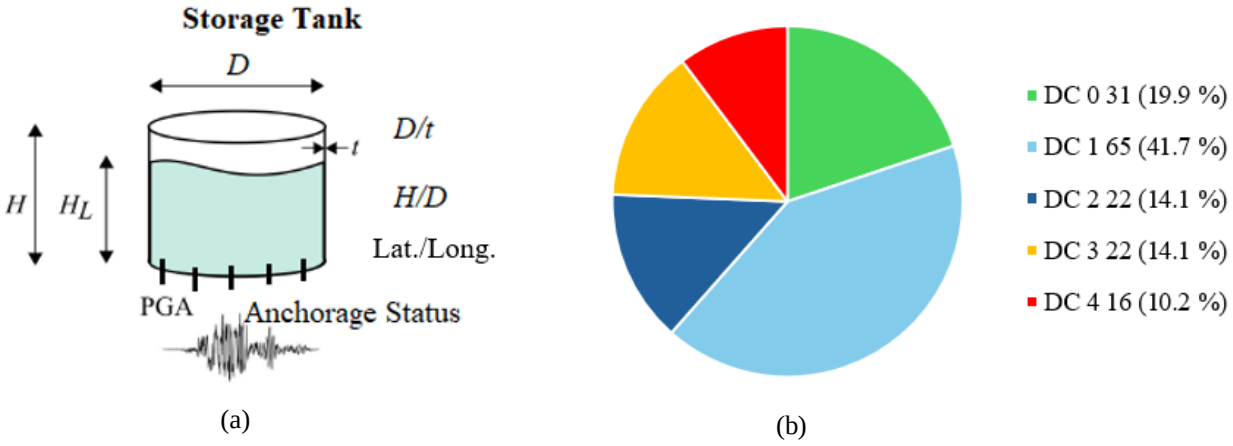
3.1. Geographic Information System (GIS) Analysis

To analyze the spatial clustering and distribution of earthquake-induced damage across California, a GIS-based approach was implemented. The dataset was first processed in ArcGIS Pro, where Excel files were converted into geodatabase tables using the Excel to Table tool. The resulting dataset, named CaliforniaDamage, was then joined to a California counties shapefile using the COUNTY attribute. As county names were unique within the dataset, the join produced an accurate spatial representation of damage across the state. To visualize the results, proportional pie symbols were used, with each pie slice representing the proportion of damage categories (DC 0 to DC 4) within each county. The overall size of each symbol corresponded to the total number of reported damage incidents, enabling a clear visual comparison of damage severity across regions. Symbol scaling and opacity were adjusted to improve boundary visibility

and enhance interpretability. As shown in Figure 2, the resulting spatial distribution map clearly illustrates that the most significant damage clusters are located in Southern California. Several counties show high proportions of severe damage (DC 3 and DC 4), indicated by the dominance of orange and red segments in the pie charts. In contrast, northern and central regions display fewer or less severe incidents, often dominated by green and blue segments (DC 0 and DC 1). This pattern aligns with known seismically active zones and population densities.

Table 1: Damage Category Description based on HAZUS.

Damage Category	Description
DC 0	No damage
DC 1	Damage to the roof, minor loss of content, minor shell damage, damage to attached pipes, no elephant foot failure
DC 2	Elephant foot buckling with no leak or minor loss of contents
DC 3	Elephant foot buckling with major loss of content, severe damage
DC 4	Complete collapse



(a) (b)

Fig. 1: (a) Storage tank input parameters; (b) Distribution of tank data among the mapped damage categories.

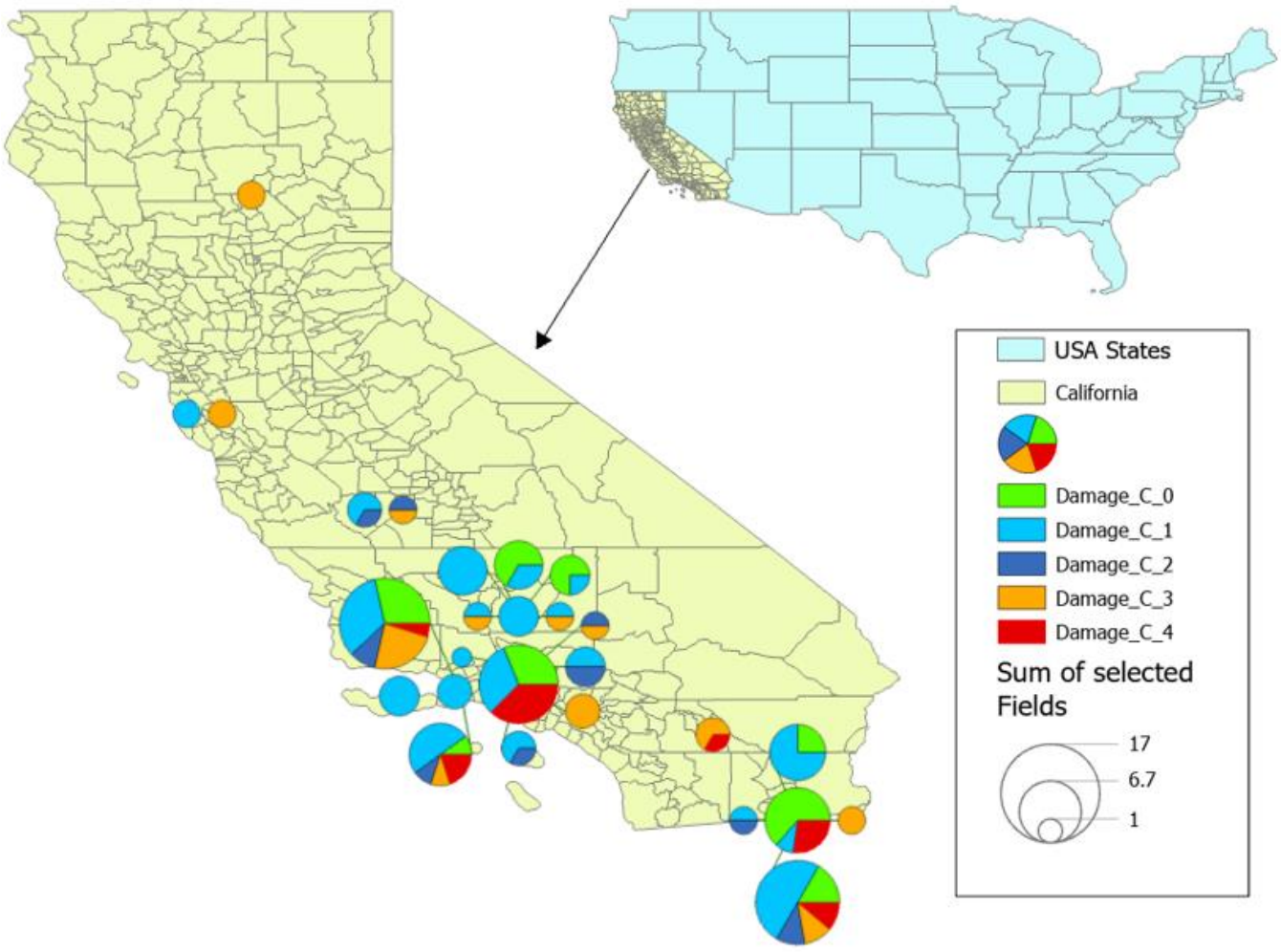


Fig. 2: Spatial clustering or distribution of damage concerning individual earthquakes.

3.2. ML Classifier Models and Balancing Technique

Four ML models: DT, RF, XGBoost, and SVC, were used to classify the five tank failure modes (DC 0-DC 4). The full dataset was subjected to hyperparameter tuning to identify the best settings for each classifier model. Based on this tuning, a maximum depth of 5 was selected for the DT, RF, and XGBoost models. Moreover, 100 estimators were chosen in both RF and XGBoost to ensure a trade-off between performance and generalization. In case of the SVC model, the model architecture consisted of a Support Vector Classifier using a radial basis function (RBF) kernel, with a regularization parameter C=1.0 and automatic kernel coefficient gamma='scale'.

Based on the results discussed in the next section, the Random Forest classifier model was identified as the best-performing model in terms of balancing accuracy and generalizability, despite its lower test scores compared to some alternatives. To further enhance its performance, particularly in addressing class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE works by generating synthetic examples for the minority class, thereby producing a more balanced training dataset.

3.3. Performance Evaluation

The four aforementioned classifier models were applied to predict damage categories from the dataset, using a 70-30 split for training and testing. Model performance and generalization to unseen data were evaluated using confusion matrices, which summarize the correct and incorrect predictions made by a classification model. The diagonal elements represent correct predictions (true positives and true negatives), while the off-diagonal elements indicate misclassifications (false positives and false negatives).

In this study, model performance was assessed using precision, recall, and accuracy. Accuracy reflects the overall effectiveness of the model, whereas precision and recall provide insight into the model's performance for each specific failure mode. High values across these metrics indicate a robust and reliable classification model. Representative confusion matrices for this multi-class classification task are presented in Figure 3.

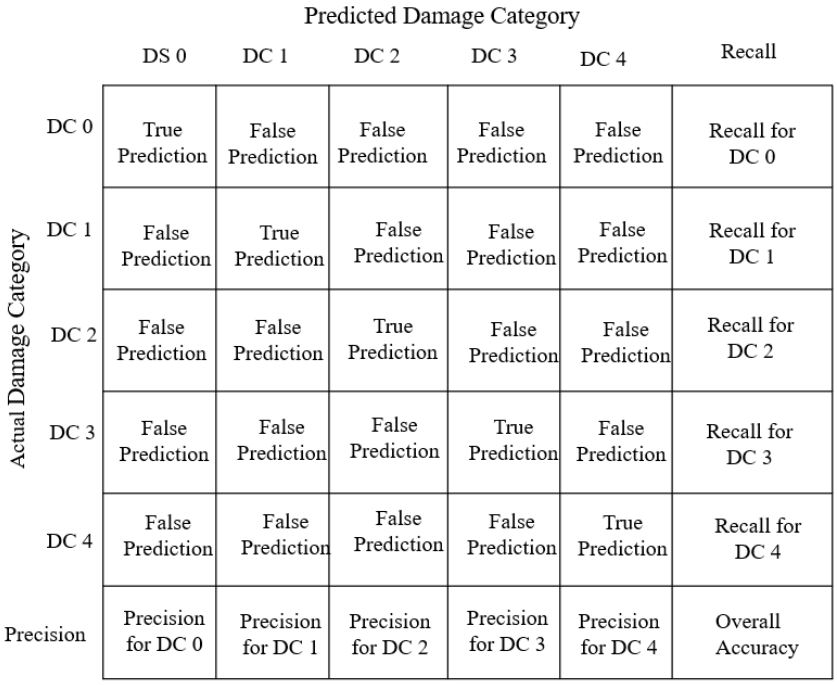


Fig. 3: Typical confusion matrix for multi-classification.

4. Analysis and Results

The model evaluation results presented in Table 3 show that the XGBoost model achieved the highest performance in the training set (Precision = 1.0, Recall = 1.0, and Accuracy = 1.0), indicating it could memorize the training data well. However, the test set results (Precision = 0.58, Recall = 0.59, and Accuracy = 0.59) demonstrate its ability to memorize the training data effectively. However, its performance dropped significantly on the test set (Precision = 0.58, Recall = 0.59, Accuracy = 0.59), clearly indicating overfitting. This suggests that although the model fit the training data exceptionally well, it failed to generalize to unseen data, likely due to class imbalance. Similarly, the Decision Tree (DT) model exhibited poor performance, which may be attributed to insufficient feature learning or its sensitivity to the imbalanced dataset, as listed in Table 3. Random Forest model showed moderate generalization ability with an accuracy score of 0.94 (train) and 0.55 (test), as listed in Table 3. Compared to XGBoost, it demonstrated less prone to overfitting than XGBoost. The underperformance of all models is likely due to the imbalanced class distribution, as illustrated in Figure 1(b).

To address this issue, the SMOTE balancing technique was applied in conjunction with the Random Forest classifier. This approach led to improved performance, with training scores reaching 0.93 across all evaluation metrics and test set metrics

improving to a maximum of Precision = 0.75, Recall = 0.73, and Accuracy = 0.73. These findings suggest that integrating SMOTE with the Random Forest classifier enhances model generalization and effectively mitigates class imbalance. The final confusion matrices for both the training and test datasets, shown in Figure 4, provide a detailed breakdown of each model's classification performance across different damage states.

Table 3: Results comparison of different ML classifier models.

Model	Training Set			Test set		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
DT Classifier	0.83	0.81	0.81	0.43	0.48	0.48
Support Vector Classifier	0.62	0.66	0.66	0.38	0.45	0.45
XGBoost Classifier	1.0	1.0	1.0	0.58	0.59	0.59
RF Classifier	0.95	0.94	0.94	0.54	0.54	0.55
SMOTE-RF Classifier	0.93	0.93	0.93	0.75	0.73	0.73

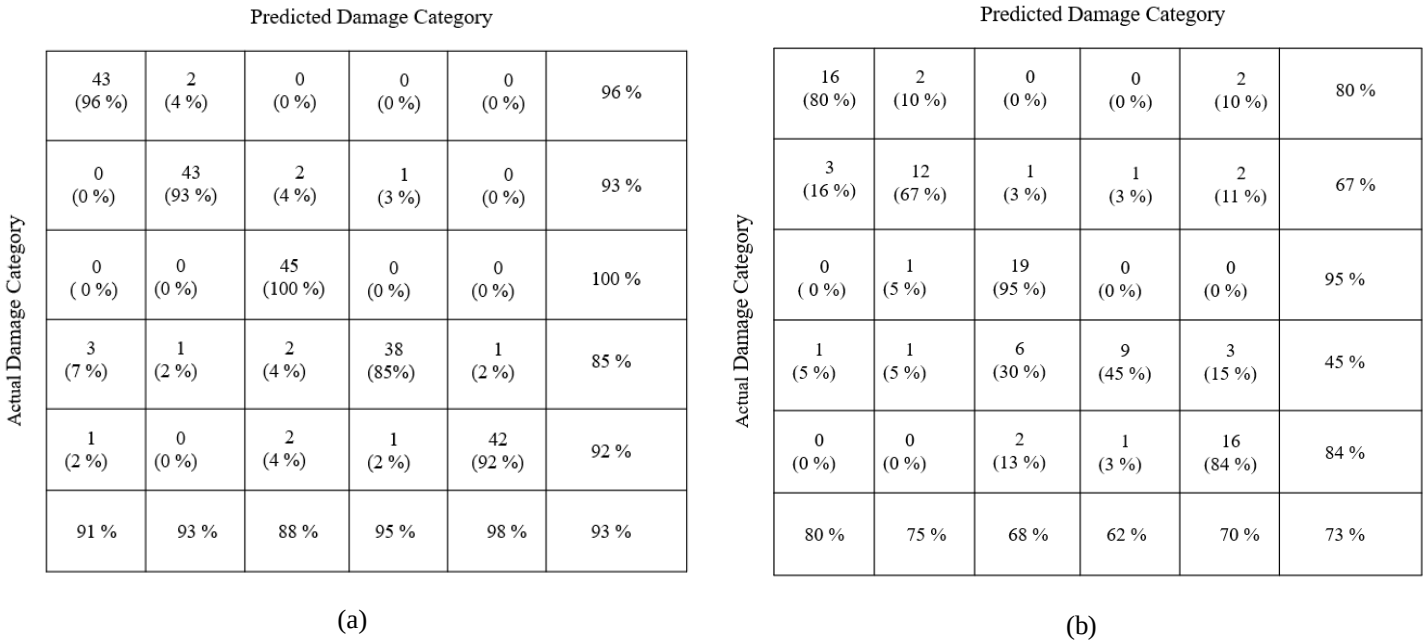


Fig. 4: SMOTE-RF confusion matrix for (a) Training set (b) Test set.

5. Conclusion

This study demonstrates the effectiveness of integrating Geographic Information Systems (GIS) and Machine Learning (ML) techniques to assess seismic-induced damage in above-ground liquid storage tanks. A real-world dataset was employed, comprising ground motion characteristics, tank dimensions, and actual damage classifications documented in post-earthquake field reports. The damage states of the selected tanks were standardized using the HAZUS classification system for storage tank damage. A GIS-based spatial analysis was conducted to identify clustering and distribution patterns of earthquake-induced damage across California. The resulting spatial distribution map indicates that the most significant damage clusters are concentrated in Southern California, likely due to the region's high seismic vulnerability. Four ML classifiers, Decision Tree (DT), Random Forest (RF), XGBoost, and Support Vector Machine (SVM), were implemented to predict damage categories. Among these ML models, the Random Forest classifier, when paired with the Synthetic Minority Over-sampling Technique (SMOTE), showed the most promise in handling imbalance dataset and improving prediction accuracy. This combination yielding training scores of 0.93 across all evaluation metrics and test set metrics of Precision = 0.75, Recall = 0.73, and Accuracy = 0.73. These results indicate that integrating SMOTE with the Random Forest classifier effectively enhances model generalization and mitigates the adverse effects of class imbalance. Overall, this research underscores the value of applying advanced analytical techniques for infrastructure safety assessment and contributes to ongoing efforts in disaster risk reduction. Future research may investigate the integration of additional spatial and structural data, temporal modeling of seismic activity, and validation across diverse infrastructure systems and geographic areas.

References

[1] R. J. Merino, E. Brunesi, and R. Nascimbene, “Probabilistic evaluation of earthquake-induced sloshing wave height in above-ground liquid storage tanks,” Eng. Struct., vol. 202, p. 109870, 2020.

[2] F. Tabish, I.H.P. Mamaghani, S. Ullah, M.H. Karim, and R. Godasu, “Application of machine learning to classify seismic induced damage of storage tanks,” CSCE-SCGC, Structures Specialty Conference, ST-584, pp.1-7, 2025.

[3] M. Farajian, M. I. Khodakarami, and D.-P. N. Kontoni, “Evaluation of soil-structure interaction on the seismic response of liquid storage tanks under earthquake ground motions,” Computation, vol. 5, no. 1, p. 17, 2017.

[4] S. Lee, B. Kim, and Y.-J. Lee, “Seismic fragility analysis of steel liquid storage tanks using earthquake ground motions recorded in Korea,” Math. Probl. Eng., vol. 2019, no. 1, p. 6190159, 2019.

[5] l. Micheli, H. Nguyen, L.c. Chang, and M. Faytarouni, “Machine learning for seismic-induced damage estimation of steel tanks,” Proc. Int. Struct. Eng. Constr., vol. 8, p. 1, 2021.

[6] K. Bakalis, D. Vamvatsikos, and M. Fragiadakis, “Seismic risk assessment of liquid storage tanks via a nonlinear surrogate model,” Earthq. Eng. Struct. Dyn., vol. 46, no. 15, pp. 2851–2868, 2017.

[7] L. Micheli, L. Cao, and S. Laflamme, “Surrogate-based performance evaluation strategy for high performance control systems under uncertainties,” in Active and Passive Smart Structures and Integrated Systems XIV, SPIE, 2020, pp. 398–411.

[8] G. Quinci and F. Paolacci, “A novel machine learning based framework for the seismic risk assessment of industrial plant,” Pressure Vessels and Piping Conference, American Society of Mechanical Engineers, 2023, p. V007T08A010.

[9] F. Tabish, I.H.P. Mamaghani, M.H. Karim, S. Ullah, and R. Godasu, “Seismic damage prediction of liquid storage tanks using machine learning and balancing techniques,” Journal of Structural Design and Construction Practice, DOI: 10.1061/JSDCCC/SCENG-1894. American Society of Civil Engineers (ASCE), 2025.

[10] K. Hatayama, “Damage to oil storage tanks from the 2011 Mw 9.0 Tohoku-Oki tsunami,” Earthq. Spectra, vol. 31, no. 2, pp. 1103–1124, 2015.

[11] S. Toprak and I. Tutuncu, “GIS characterization of spatially distributed lifeline damage,” in Technical Council on Lifeline Earthquake Engineering Monograph, 1999, pp. 110–119.

[12] W. Liu, F. Yamazaki, H. Gokon, and S. Koshimura, “Extraction of tsunami-flooded areas and damaged buildings in the 2011 Tohoku-oki earthquake from TerraSAR-X intensity images,” Earthq. Spectra, vol. 29, no. 1_suppl, pp. 183–200, 2013.

[13] A.L. Alliance, A. L, “Seismic fragility formulations for water systems, Part 2-Appendices”, Washington, DC., FEMA, American Society of Civil Engineers, USA, 2001.