CS 559: MODELING TSUNAMI POTENTIAL FROM GLOBAL SEISMIC PATTERNS IN EARTHQUAKES

Humna Sultan[*] **and Spurthi Setty**[*]
[*]Stevens Institute of Technology
ssetty2@stevens.edu, hsultan@stevens.edu
Fall 2025

ABSTRACT

This study presents a binary classification approach to predict tsunami occurrence based on global earthquake data collected between 2001 and 2022.

1 Introduction

Earthquakes and tsunamis represent two of the most devastating natural hazards facing humanity, with the potential to cause catastrophic loss of life, infrastructure damage, and economic disruption within minutes. The 2004 Indian Ocean tsunami claimed approximately 230,000 lives, while the 2011 Great East Japan Earthquake and tsunami resulted in approximately 22,000 fatalities [12]. These events underscore the critical importance of accurate and timely risk assessment systems.

The fundamental challenge in earthquake and tsunami risk assessment lies in the complex, non-linear relationships between seismic parameters (magnitude, depth, location), geological characteristics, and coastal geometry. Traditional approaches such as Probabilistic Seismic Hazard Analysis (PSHA) rely heavily on physics-based models that require extensive computational resources and struggle to capture intricate variable interactions [15]. Machine learning offers a transformative paradigm by leveraging data-driven approaches to discover complex patterns that traditional methods cannot easily model.

This project develops and evaluates machine learning models for global earthquake and tsunami risk assessment using binary classification to predict tsunami occurrence from seismic events. The analysis will identify the most influential features—such as magnitude, depth, and location—that are most predictive of tsunami generation, contributing to research on hybrid systems that combine physics-based and data-driven techniques for enhanced natural hazard assessment.

2 Background

Traditional earthquake and tsunami risk assessment has relied on Deterministic and Probabilistic Seismic Hazard Analysis (DSHA/PSHA). PSHA, formalized by Cornell [2], integrates multiple earthquake sources with associated probabilities but faces fundamental limitations including unclear physical bases, extensive data requirements, and computational expense [15]. Physics-based tsunami simulations require hours to days for high-resolution models, rendering them unsuitable for real-time warning systems, while conventional earthquake early warning systems suffer from magnitude saturation effects [5, 11].

Machine learning offers promising solutions. Jozinović et al. [7] developed Graph Convolutional Neural Networks achieving strong magnitude prediction performance, while Goda et al. [5] created ML-based tsunami inundation systems reducing computational costs by 99% compared to physics-based models. Tree-based methods like Random Forests and XGBoost naturally capture complex non-linear interactions, with Khorami et al. [8] demonstrating that multi-feature ML models significantly outperform traditional univariate approaches [1].

However, current ML approaches face limitations including limited comprehensive datasets spanning diverse regions, interpretability concerns, and insufficient uncertainty quantification [10, 5]. This project addresses these gaps by systematically evaluating multiple ML algorithms on a global earthquake-tsunami dataset, emphasizing feature importance analysis and performance comparison to provide practical guidance for next-generation risk assessment systems.

3 Data

The dataset selected is titled ”Global Earthquake-Tsunami Risk Assessment” [14] and is a part of the larger ”Seismic Features & Tsunami Classification Dataset for Risk Assessment.” The dataset includes seismic characteristics and tsunami potential indicators for 782 significant earthquakes recorded globally from 2001 to 2022. Each record captures a range of geophysical, intensity-based, and temporal features that collectively inform the likelihood of tsunami generation. The target variable indicated by this dataset is ”tsunami,” represented through binary classification of either ”1” or ”0” to predict tsunami potential. A tsunami value of ”0” indicates that no tsunami occurred following the earthquake, while ”1” indicates that a tsunami was triggered. There is a relatively balanced class distribution of non-tsunami and tsunami occurrences, with approximately 61% (478 records) classifying as non-tsunami events and approximately 39% (304 records) classifying as tsunami-potential events. This relative balance of this dataset allows for suitable binary classification tasks.

There are 13 total features: magnitude, depth, latitude, longitude, CDI (Community Decimal Intensity), MMI (Modified Mercalli Intensity), SIG (Event significance score), NST (Number of seismic monitoring stations), DMIN (Distance to nearest seismic station in degrees), GAP (Azimuthal gap between stations in degrees), year, and month. These features can be placed into categories to create structure for the model. Magnitude, depth, latitude, and longitude provide direct information about seismic strength and location and can be utilized to make predictions regarding whether a tsunami will occur. CDI, MMI, SIG, NST, DMIN, and GAP represent measurement reliability and impact and can be used to predict intensity, significance, and station data, all of which are critical to make predictions and conclusions with concrete evidence. Year and month can be utilized to find temporal patterns, making connections between specific time periods and seasons. The magnitude values in the dataset span from 6.5 to 9.1 on the Richter scale, which quantifies the energy released by an earthquake. The average magnitude across all recorded events is approximately 6.94. There are 28 earthquakes in the dataset with magnitude greater than or equal to 8, including the 2004 (9.1) and 2011 (9.1) mega-earthquakes.

4 Methodologies

To perform calculations with the features provided, data preprocessing is essential. The normalization of continuous features such as magnitude, depth, and significance scores is important to ensure consistent scaling across models. This prevents any single feature from disproportionately influencing the learning process and is also necessary for distance-based algorithms and gradient-based optimization methods; in these methods, feature magnitude directly affects model behavior, making scaled and normalized values a necessity for appropriate outputs. Temporal features, particularly year and month can be encoded to preserve cyclical patterns. Techniques such as sine-cosine transformation or one-hot encoding can help capture these periodic relationships without introducing artificial linearity. Categorical features in the dataset can be transformed as needed depending on the structure of the model. Combining features to make predictions is critical in this scenario; feature selection to evaluate the individual and combined impact of geophysical, intensity-based, and temporal features is important to understand model performance under different variables and considerations. In addition to standard normalization and encoding, effective preprocessing involves addressing outliers, assessing multicollinearity among features, and maintaining class balance across the target variable. Outlier detection helps prevent extreme values from skewing model behavior, while multicollinearity checks ensure that highly correlated predictors do not distort feature importance or inflate variance in model coefficients. While manipulating the dataset, it is imperative to maintain the balanced distribution between tsunami and non-tsunami events to prevent bias and ensure fair evaluation for accuracy metrics such as precision and recall.

Several machine learning models can be deployed to make predictions for this dataset; we will focus on deploying and evaluating each of these methods and conclude with selecting the most accurate and robust model for this scenario. Logistic regression can be utilized to offer insight into the linear relationships between features such as magnitude and tsunami risk. Additionally, the random forest methodology can be utilized to capture nonlinear interactions and provide feature importance rankings.

5 Evaluation Metrics + Comparison Methods

To evaluate the performance of the classification models, a comprehensive set of metrics will be used to ensure both overall accuracy and class-specific reliability. Accuracy will provide a general measure of how often the model correctly predicts tsunami and non-tsunami events. Precision, recall, and F1-score will also be prioritized and calculated in order to provide additional information about the positive class, when tsunami = 1. Precision will measure the proportion of predicted tsunami events that are actually correct, while recall will assess the model's ability to detect all true tsunami events. The F1-score, which balances precision and recall, will offer a more nuanced view of the model's effectiveness in handling both false positives and false negatives. Confusion matrices will also be analyzed to visualize prediction outcomes and identify patterns in misclassification. Understanding these metrics is critical for model selection and refinement; evaluating each will help provide a comprehensive understanding of performance of each model, allowing for us to choose the most effective solution.

Multiple ML algorithms—including Logistic Regression, Random Forest, XGBoost, and potentially neural networks—will be trained and compared using the metrics described above. Model performance will be assessed across all evaluation metrics, with particular emphasis on recall to minimize false negatives (missed tsunami events), as these have the most severe consequences for disaster preparedness. Additionally, we will analyze computational efficiency by comparing training time and prediction latency across models. Feature importance rankings from tree-based models will be compared to identify consistent predictors of tsunami occurrence. The final model selection will balance predictive performance, interpretability, and computational efficiency to ensure practical applicability for risk assessment systems.

References

[1] Aránguiz, R., et al. Beyond tsunami fragility functions: experimental assessment for building damage estimation. *Scientific Reports*, 13:13975, 2023.

[2] Cornell, C. A. Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58(5):1583–1606, 1968.

[3] Cornell, C. A. Probabilistic analysis of damage to structures under seismic loads. In *Dynamic Waves in Civil Engineering*, pages 473–488, 1971.

[4] Ellingwood, B. R. and Kinali, K. Earthquake risk assessment of building structures. *Reliability Engineering & System Safety*, 74(3):251–262, 2001.

[5] Goda, K., et al. Machine learning-based tsunami inundation prediction derived from offshore observations. *Nature Communications*, 13:5489, 2022.

[6] Jena, R., Pradhan, B., and Beydoun, G. Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia. *Science of the Total Environment*, 749:141582, 2020.

[7] Jozinović, D., et al. An early warning system for earthquake prediction from seismic data using batch normalized graph convolutional neural network with attention mechanism (BNGCNNATT). *Sensors*, 22(17):6482, 2022.

[8] Khorami, M., et al. Interpretable machine learning based tsunami bridge fragility assessment. *International Journal of Disaster Risk Reduction*, 115:104864, 2025.

[9] Lee, K., et al. Static analysis-based rapid fire-following earthquake risk assessment method using simple building and GIS information. *Scientific Reports*, 14:21468, 2024.

[10] McGuire, J. J., et al. Earthquake and tsunami prediction enhanced by deep-learning model. Technical report, Los Alamos National Laboratory, 2022.

[11] McInnes, H. L. and Shirzaei, M. Earthquake and tsunami prediction enhanced by deep-learning model. *Engineering & Technology Magazine*, 2022.

[12] Mori, N., Satake, K., Cox, D., et al. Giant tsunami monitoring, early warning and hazard assessment. *Nature Reviews Earth & Environment*, 3(9):557–572, 2022.

[13] Mulia, I. E., Gusman, A. R., and Satake, K. Applying a deep learning algorithm to tsunami inundation database of megathrust earthquakes. *Journal of Geophysical Research: Solid Earth*, 125(9):e2020JB019690, 2020.

[14] Uzaki, A. Global Earthquake-Tsunami Risk Assessment Dataset. Kaggle, 2024. Available at: `https://www.kaggle.com/datasets/ahmeduzaki/global-earthquake-tsunami-risk-assessment-dataset`.

[15] Wang, Z., Ormsbee, L., and Govindaraju, R. S. Seismic hazard assessment: Issues and alternatives. *Journal of Earthquake Engineering*, 15(4):633–651, 2011.