

Article

Forecasting the Unseen: Enhancing Tsunami Occurrence Predictions with Machine-Learning-Driven Analytics

Snehal Satish , Hari Gonaygunta , Akhila Reddy Yadulla , Deepak Kumar , Mohan Harish Maturi ,
Karthik Meduri [†], Elyson De La Cruz , Geeta Sandeep Nadella ^{*,†} and Guna Sekhar Sajja 

Department of Information Technology, University of the Cumberlands, Williamsburg, KY 40769, USA; ssatish3175@ucumberlands.edu (S.S.); hgonaygunta3255@ucumberlands.edu (H.G.); ayadulla5882@ucumberlands.edu (A.R.Y.); fdeepakkumar6146@ucumberlands.edu (D.K.); mmaturi9213@ucumberlands.edu (M.H.M.); kmeduri5644@ucumberlands.edu (K.M.); edelacruz7981@ucumberlands.edu (E.D.L.C.); gsajja1524@ucumberlands.edu (G.S.S.)

* Correspondence: gnadella3853@ucumberlands.edu

[†] These authors contributed equally to this work.

Abstract: This research explores the improvement of tsunami occurrence forecasting with machine learning predictive models using earthquake-related data analytics. The primary goal is to develop a predictive framework that integrates a wide range of data sources, including seismic, geospatial, and ecological data, toward improving the accuracy and lead times of tsunami occurrence predictions. The study employs machine learning methods, including Random Forest and Logistic Regression, for binary classification of tsunami events. Data collection is performed using a Kaggle dataset spanning 1995–2023, with preprocessing and exploratory analysis to identify critical patterns. The Random Forest model achieved superior performance with an accuracy of 0.90 and precision of 0.88 compared to Logistic Regression (accuracy: 0.89, precision: 0.87). These results underscore Random Forest's effectiveness in handling imbalanced data. Challenges such as improving data quality and model interpretability are discussed, with recommendations for future improvements in real-time warning systems.

Keywords: predictive model; tsunami occurrence; earthquake data; data analytics; machine learning (ML); forecasting



Academic Editor: Yan Zhang

Received: 9 February 2025

Revised: 28 April 2025

Accepted: 30 April 2025

Published: 4 May 2025

Citation: Satish, S.; Gonaygunta, H.; Yadulla, A.R.; Kumar, D.; Maturi, M.H.; Meduri, K.; De La Cruz, E.; Nadella, G.S.; Sajja, G.S. Forecasting the Unseen: Enhancing Tsunami Occurrence Predictions with Machine-Learning-Driven Analytics. *Computers* **2025**, *14*, 175. <https://doi.org/10.3390/computers14050175>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Tsunamis are among the most devastating natural disasters, posing significant threats to coastal communities and infrastructure worldwide. Despite advancements in seismic monitoring, accurately predicting tsunami occurrences within critical timeframes remains a major challenge [1]. Traditional tsunami warning systems primarily rely on seismic magnitude thresholds and post-earthquake hydrodynamic simulations [2,3], which often lead to delayed alerts or false alarms due to their reliance on simplified heuristics. The integration of machine learning models and data analytics offers a promising frontier for improving tsunami occurrence prediction. AI and its capacity to learn from patterns in massive numbers of data combined with data analytics, which process and analyze diverse and massive datasets in real time, then hold the potential to revolutionize how we approach tsunami occurrence forecasting [4,5]. The motivation behind this research stems from the growing recognition that traditional methods alone may not be sufficient to meet the demands of modern disaster preparedness. With AI and data analytics, it is possible to enhance the accuracy of predictions and extend and improve the effectiveness of disaster mitigation

strategies [6]. This research aims to explore advanced technologies applied to create a robust and AI-driven predictive model that not only forecasts earthquakes with higher accuracy in critical prime times but also saves lives, reduces economic losses, and contributes to global efforts in disaster risk reduction [7]. This research is motivated by the broader implications of integrating AI and data analytics into geophysical research. AI continues to advance, and its applications in various scientific domains, including environmental and Earth sciences, are expanding quickly [8]. Tsunami occurrence forecasting presents a unique opportunity to demonstrate how AI can be applied to solve real-world problems that have profound impacts on society [9]. The potential to transform tsunami occurrence prediction from the field, constrained by limited data and traditional methodologies, into a dynamic and data-driven science is both an exciting challenge and a critical necessity as natural disasters are becoming increasingly frequent and severe due to climate change and other environmental factors [10]. This paper seeks to contribute to the field by developing and validating a predictive model for more actual tsunami occurrence forecasting and the extension of more resilient communities and infrastructures.

1.1. Research Problem

The complexity of tsunami generation mechanisms creates significant challenges for early warning systems. Current approaches struggle to distinguish between seismic events that generate dangerous tsunamis and those that do not, particularly in near-field scenarios where rapid decisions are crucial [11]. This problem is exacerbated by the need to integrate heterogeneous data sources—including seismic parameters, geographic context, and historical patterns—into a unified predictive framework [12]. The problem lies in the integration and processing of these diverse data sources, such as satellite imagery, geospatial data, and real-time seismic activity and environmental conditions, to create a predictive model that offers more accurate and timely tsunami occurrence forecasts. This problem requires the development of an AI-driven predictive model that controls data analytics to improve the correctness and consistency of tsunami occurrence forecasting and eventually aims to alleviate the devastating impacts of these natural disasters.

1.2. Research Questions

In this study of the refuge, the key questions are as follows:

- Which machine learning architectures are most effective for binary classification of tsunami-generating earthquakes?
- What feature selection strategies optimize model performance while maintaining computational efficiency?
- How can temporal patterns in seismic data improve prediction lead times?

1.3. Research Objective

This study develops a machine learning framework that predicts tsunami occurrence within minutes of earthquake detection by analyzing six key seismic features: magnitude, depth, geographic coordinates, aftershock patterns, seismic moment, and source mechanism characteristics. The study aims to integrate diverse data sources, including seismic, geospatial, and environmental data, into the robust analytical framework in advanced machine learning algorithms. By improving the precision of predictions and providing earlier warnings, this model seeks to contribute to better disaster preparedness and modification strategies and reduce the human and economic toll of earthquakes.

1.4. Structure of Paper

This paper is organized into numerous important units. It starts with an overview that describes the research problem and then the objects and significance of the study. The Literature Review follows and examines current studies and recognizes the holes this study proposes to fill. The Methodology details the data sources plus the AI techniques and data analytics used to develop the predictive model. The Implementation section describes how the model was tested with a case study. The Results and Discussion present the findings and their implications; the Conclusion reviews the contributions and suggests future exploration instructions.

2. Literature Review

2.1. Introduction to Tsunami Occurrence Prediction

Tsunami occurrence prediction remains a critical research area due to the devastating impacts of these ocean waves, often triggered by underwater earthquakes, landslides, or volcanic eruptions [12]. The field has evolved from using traditional geophysical methods to advanced computational and artificial intelligence (AI)-driven approaches, leveraging vast datasets and sophisticated modeling techniques. This review synthesizes traditional methods, such as seismic activity monitoring and fault line analysis, with modern advancements, including the Finite Element Method (FEM), AI, machine learning (ML), and big data analytics. By addressing both historical and contemporary approaches, we aim to provide a comprehensive overview of tsunami prediction methodologies, their successes, and the ongoing challenges.

2.2. Traditional Tsunami Occurrence Prediction

Traditional tsunami prediction methods focus on analyzing geological and seismological data, particularly data related to submarine earthquakes, which account for approximately 80% of tsunamis [13]. These methods rely on empirical observations, historical records, and geophysical measurements to assess tsunami risk. The primary techniques include seismic activity monitoring and fault line analysis, which are foundational to understanding the mechanisms of tsunami generation.

1. Seismic Activity Monitoring

Seismic activity monitoring is a cornerstone of tsunami prediction, involving the continuous observation of seismic waves generated by underwater earthquakes [14,15]. This method leverages networks of seismometers and ocean-bottom sensors to detect potential tsunami triggers.

- **Seismicity Patterns:** Monitoring seismicity patterns involves analyzing the frequency and distribution of earthquakes, particularly foreshocks, which may precede a tsunami-generating event. Seismic gaps—regions along faults with low recent activity—are critical indicators of potential large earthquakes [16]. For example, the 2011 Tohoku earthquake was preceded by a seismic gap along the Japan Trench, highlighting the predictive value of this approach [17].
- **Earthquake Swarms:** Earthquake swarms, characterized by clusters of seismic events in a short period, can signal stress changes in the Earth's crust. While not all swarms lead to tsunamis, they are monitored closely in subduction zones like the Pacific Ring of Fire [16]. Advanced algorithms now enhance swarm detection by filtering noise from seismic data [18].
- **Precursors and Anomalies:** Seismic precursors, such as changes in P-wave velocity or P-wave shadowing, indicate potential tsunami-generating earthquakes. P-wave shadowing occurs when primary seismic waves are absorbed or de-

flected, suggesting imminent fault rupture [19]. Recent studies have improved precursor detection using machine learning to analyze seismic waveforms [20].

2. Fault Line Analysis

Fault line analysis focuses on submarine faults, where tectonic plate movements can displace ocean water, generating tsunamis [21]. This method combines geological mapping, historical data, and geophysical measurements to assess tsunami risk.

- **Mapping Fault Lines:** Geological surveys, remote sensing, and satellite imagery are used to map submarine fault lines, identifying high-risk coastal areas [22]. For instance, the Cascadia Subduction Zone has been extensively mapped to assess tsunami potential along the U.S. West Coast [17].
- **Historical Tsunami Data:** Historical records of tsunamis along fault lines provide insights into recurrence intervals and event magnitudes. Subduction zones like the Sunda Trench have well-documented tsunami histories, informing probabilistic hazard models [17,23].
- **Stress Accumulation:** GPS and InSAR (Interferometric Synthetic Aperture Radar) measurements monitor stress accumulation along faults. High stress rates indicate elevated tsunami risk, as seen during the 2018 Palu tsunami [24].
- **Paleo-Seismology and Paleo-Tsunami Studies:** Paleo-seismology examines sediment layers to reconstruct past earthquakes and tsunamis, offering a long-term perspective on tsunami frequency [25]. For example, paleo-tsunami deposits in Chile have revealed 500-year recurrence intervals for major events [18].
- **Slip Rates and Fault Segmentation:** Faults are segmented, with each segment exhibiting unique slip rates. Analyzing these rates helps predict which segments are prone to tsunami-generating earthquakes [26]. Recent advancements in fault segmentation models have improved risk assessments for regions like Indonesia [27].

2.3. The Finite Element Method (FEM) in Tsunami Prediction

The Finite Element Method (FEM) is a numerical technique widely used to simulate complex geophysical processes, including fault mechanics and tsunami wave propagation [28]. The FEM divides the ocean–Earth system into discrete elements, enabling detailed analysis of stress distribution, fault slip, and wave dynamics. Unlike traditional statistical models, the FEM captures the nonlinear interactions between seismic events and ocean responses, improving predictions of tsunami generation [29].

FEM applications include the following:

- **Fault Mechanics:** Simulating stress accumulation and rupture along submarine faults to predict tsunami triggers [26].
- **Wave Propagation:** Modeling tsunami wave interactions with bathymetry and coastal topography to estimate inundation zones [30].
- **Infrastructure Assessment:** Evaluating the impact of tsunami waves on coastal structures, informing resilient design standards [31].

Recent advancements in the FEM incorporate real-time seismic data and high-resolution bathymetric maps, enhancing the accuracy of tsunami forecasts. For example, FEM models of the 2011 Tohoku tsunami accurately predicted inundation heights, aiding post-disaster recovery [32]. However, the FEM's computational complexity requires significant resources, limiting its use in real-time applications [29].

2.4. AI and Machine Learning in Tsunami Prediction

AI and ML have revolutionized tsunami prediction by processing vast datasets and identifying patterns that traditional methods cannot detect [33]. These technologies enhance

real-time monitoring, risk assessment, and resource allocation, as shown in Figure 1. Table 1 highlights the challenges that AI/ML aim to address, such as uncertainty and false alarms.

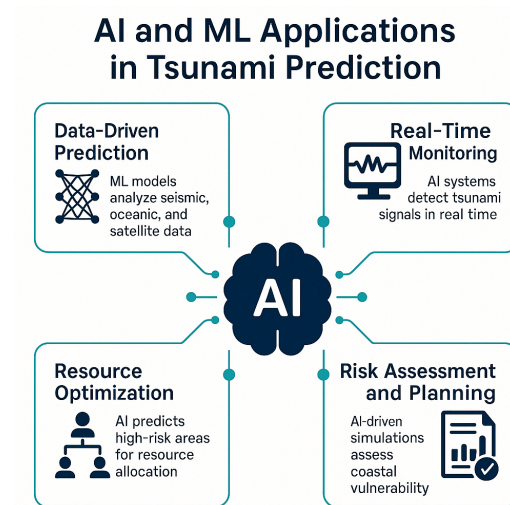


Figure 1. AI and ML applications in tsunami prediction.

Table 1. Challenges and limitations in traditional tsunami occurrence prediction.

Challenge/Limitation	Details
Uncertainty in Predictions	The nonlinear dynamics of the Earth’s crust and ocean make precise predictions of tsunami timing, location, and magnitude challenging [34].
False Alarms	Seismic monitoring may trigger false tsunami warnings due to non-tsunamigenic seismic activity [23].
Missed Events	Tsunamis from non-seismic sources (e.g., landslides) may lack detectable precursors [35].
Data Gaps	Sparse monitoring networks in remote oceanic regions limit data availability [30].
Complex Fault–Ocean Interactions	Stress transfer and water displacement in submarine faults are difficult to model accurately [24].
Inaccurate Historical Records	Incomplete tsunami records hinder reliable risk assessments [36].
Paleotsunami Limitations	Geological records may be incomplete or misinterpreted, affecting long-term forecasts [25].
Precursor Ambiguity	Reliable precursors for tsunami events are not fully understood [19].
Fault Variability	Different fault segments produce tsunamis with varying characteristics [26].
Technological Barriers	Limited access to advanced sensors in developing regions reduce monitoring effectiveness [30].

- **Data-Driven Prediction:** ML models analyze seismic, oceanic, and satellite data to predict tsunami events. For example, convolutional neural networks (CNNs) have identified tsunami precursors in seismic waveforms, improving early warning systems [20]. Studies in Japan and Indonesia have demonstrated ML’s ability to provide minutes of warning before tsunami impact [37,38].

- **Real-Time Monitoring:** AI systems integrate data from ocean buoys, tide gauges, and seismic stations to detect tsunami signals in real time. Recurrent neural networks (RNNs) have been used to predict tsunami wave paths, enabling timely evacuations [36]. The Indonesia Tsunami Early Warning System employs ML to process real-time seismic data [36].
- **Resource Optimization:** AI optimizes emergency resource allocation by predicting high-risk areas. For instance, ML models prioritize medical and rescue deployments based on forecasted inundation zones [39].
- **Risk Assessment and Planning:** AI-driven simulations assess coastal vulnerability using bathymetry, population density, and infrastructure data. These models inform mitigation strategies, such as seawall construction [23,30].

2.5. Successes and Challenges

ML models have achieved notable successes, such as improving the accuracy of Japan's Earthquake Early Warning System by analyzing seismic and oceanic data [39]. During the 2018 Sulawesi tsunami, AI models provided rapid inundation forecasts, aiding evacuation efforts [40]. However, the following challenges persist:

- **Complexity of Tsunami Triggers:** Tsunamis result from diverse sources (earthquakes, landslides, volcanic activity), complicating ML predictions [32].
- **Data Quality:** Incomplete or noisy datasets limit model performance [36].
- **False Positives/Negatives:** ML models struggle to distinguish tsunamigenic from non-tsunamigenic seismic events, risking false alarms or missed events [23].
- **Interpretability:** Deep learning models, such as neural networks, often lack transparency, reducing trust in high-stakes scenarios [31].

Recent advancements have addressed these challenges by integrating physics-based models with ML, improving interpretability and accuracy [26]. For example, hybrid models combining the FEM and ML have enhanced tsunami inundation predictions [30].

2.6. Data Analytics and Analytical–Numerical Techniques in Tsunami Research

Big data analytics has transformed tsunami research by enabling the processing of diverse datasets, including seismic waveforms, ocean sensor readings, satellite imagery, and geological records [18]. Analytical–numerical techniques further refine these datasets, enhancing the accuracy of AI-based tsunami predictions [29].

- **Enhanced Data Processing:** Data analytics tools manage terabytes of geophysical data, extracting patterns that indicate tsunami precursors. Wavelet transformations and Fourier analysis remove noise from seismic signals, improving feature detection [28].
- **Improved Prediction Accuracy:** ML models trained on historical tsunami data identify trends that enhance forecast precision. For example, Bayesian inference has refined probabilistic tsunami hazard models [12].
- **Real-Time Monitoring:** Real-time analytics systems process data from global seismic networks, such as IRIS, to detect tsunami signals [41]. Japan's EEWS uses data analytics to issue rapid warnings [41].
- **Risk Assessment:** Data-driven models assess tsunami impacts on coastal regions, informing infrastructure planning [31].
- **Integration with Technologies:** Analytics integrates with IoT, remote sensing, and satellite imagery, creating comprehensive monitoring systems [18].

2.6.1. Analytical–Numerical Techniques

Analytical–numerical methods, such as wavelet transformations, Fourier analysis, and numerical optimization (e.g., gradient-based methods, Bayesian inference), optimize data for ML models [28]. These techniques

- Normalize heterogeneous datasets, ensuring compatibility with ML algorithms [12];
- Reduce error margins by filtering redundant information [29];
- Enhance feature selection, identifying key tsunami precursors [26].

For example, wavelet transformations improved the detection of seismic anomalies during the 2018 Palu tsunami, while Bayesian inference has refined hazard models for the Northeast Atlantic [18]. These methods make AI-based systems more robust, scalable, and accurate, supporting both real-time forecasting and long-term risk assessment.

2.6.2. Case Studies

- Southern California Earthquake Center (SCEC): The SCEC uses data analytics to model fault systems, predicting seismic risks that inform tsunami preparedness [42].
- Japan’s EEWS: Real-time analytics of seismic data enables seconds-to-minutes warnings, reducing casualties [41].
- IRIS Data Management: Global seismic data analysis enhances our understanding of tsunami-generating processes [41].
- Aftershock Modeling: ML models predict aftershock patterns, aiding post-tsunami recovery in regions like Nepal [42].

Tsunami occurrence prediction has advanced significantly, from traditional seismic and fault analysis to AI-driven and data-intensive approaches. Traditional methods provide a foundational understanding of tsunami triggers, while the FEM, AI, ML, and big data analytics offer unprecedented precision and real-time capabilities. Analytical–numerical techniques further enhance these models by optimizing data processing. Despite successes, challenges like improving data quality, reducing false alarms, and model interpretability persist. By integrating corrected references and modern methodologies, this review underscores the potential of hybrid approaches to improve tsunami prediction and mitigate global risks.

3. Methodology

The methodology of this research is predictive, showing that tsunami occurrence forecasting is used to collect the preprocessed and analyze the earthquake data, and is followed by the development and evaluation of machine learning models for forecasting seismic events. As shown in Figure 2, it begins with data collection from the public Kaggle repository and then focuses on cleaning and preprocessing methods contributing to data quality and relevance. The machine learning models, including Random Forest and Logistic Regression, are then developed to predict tsunami occurrences caused by earthquakes based on selected features. Exploratory Data Analysis (EDA) is conducted with maps and charts to visualize earthquake patterns and to provide regional risks. The models’ performances are evaluated with critical metrics, and a comparative analysis with traditional forecasting methods is performed to validate the effectiveness of the planned approach.

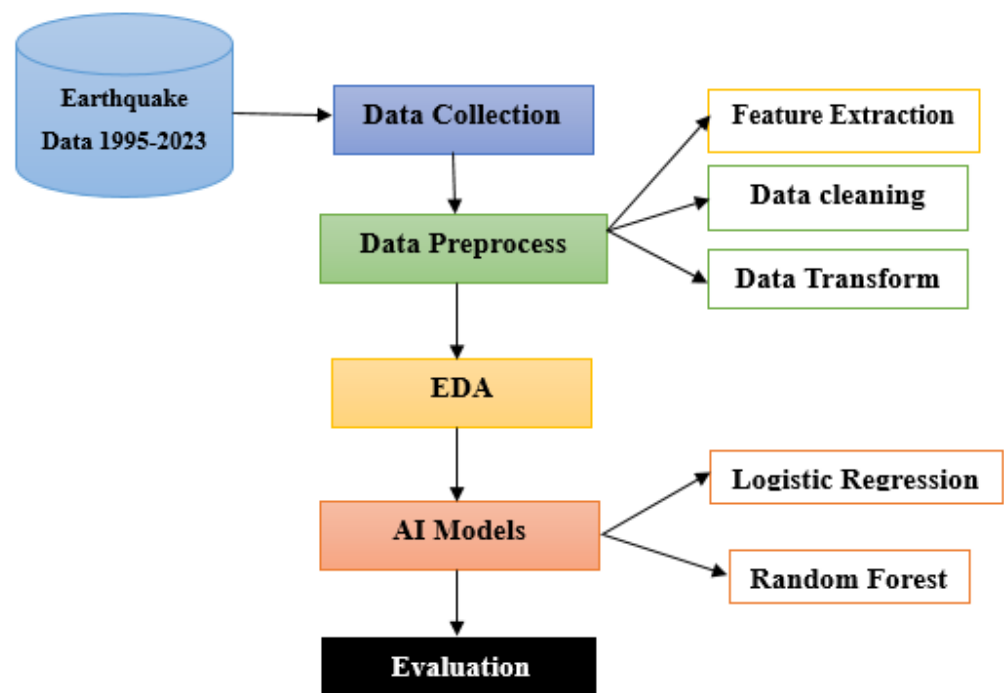


Figure 2. Proposed framework.

3.1. Data Collection and Preprocessing

The Kaggle dataset contains 984 earthquake events (1995–2023) with 19 initial features, including seismic parameters (magnitude, depth), geographic coordinates, and tsunami labels. After preprocessing, we

- Removed 7 redundant columns (e.g., nat, continent);
- Standardized formats for time, latitude, and longitude;
- Selected 6 critical features: magnitude, depth, latitude, longitude, aftershock occurrence, and seismic moment (derived from msqType and dmin).

Through feature selection analysis, 6 features were identified as being most predictive for tsunami occurrence: magnitude, depth, latitude, longitude, aftershock occurrence, and distance to coastline. These features were selected based on their correlation with tsunami generation and their statistical significance in preliminary analysis.

3.2. Dataset Suitability for Tsunami Occurrence Prediction

The primary prediction task is clearly stated as a binary classification to identify earthquakes that lead to tsunamis. The dataset from Kaggle includes both earthquake events that led to tsunamis and those that did not, making it suitable for training predictive models capable of distinguishing seismic conditions associated with tsunami events. This suitability is supported by feature selection focused on earthquake magnitude, depth, geographic location, and aftershock occurrences, critical indicators for determining tsunami risk.

3.3. Exploratory Data Analysis (EDA)

The dedicated uncovering designs and the earthquake dataset then leverage various visual gears and methods. Maps are utilized to plot the geographical delivery of earthquakes, highlighting high-risk regions and revealing spatial patterns in different periods. Charts and histograms are employed to analyze the frequency, magnitude, and depth of the earthquakes and then provide a clear picture of how these characteristics are distributed.

Scatter plots and correlation heat maps are used to explore the relationships among critical features, which are magnitude and depth, helping to identify the potential predictive variables. This analysis allows a deeper understanding of the data, guiding feature selection for the machine learning models, and offers valuable insights into global earthquake trends and risks.

3.4. Machine Learning Model Implementations

To enhance the robustness of the predictive analysis, we expand our comparison to include two additional machine learning models: the Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). These models are selected for their proven effectiveness and distinct algorithmic approaches, providing a comprehensive benchmark of predictive capabilities for earthquake-related binary classification tasks.

- **Random Forest:** The Random Forest model is selected for its robust performance in handling the complex and imbalanced datasets encountered in tsunami occurrence forecasting. As a collaborative wisdom method, it builds numerous choice trees in the preparation and combines these results to recover prognostic exactitude and regulator overfitting. This model excels in capturing nonlinear relationships and interactions among features like earthquake magnitude, depth, and location, which are crucial for accurate predictions. Its ability to rank the feature importance also aids our understanding of which factors most significantly influence tsunami occurrences caused by earthquakes and is a powerful tool for forecasting seismic events.
- **Logistic Regression:** Logistic Regression is the perfect standard owing to its plainness and interpretability in binary classification tasks. It operates by modeling the probability of a given event (in this case, the likelihood of a significant earthquake) based on the linear combination of input features. The simplicity and Logistic Regression capture the connection among the predictors, and the goal is adjustable when these relationships are approximately linear. This model provides a clear benchmark to compare the performance of more complex models like the Random Forest and allows for the straightforward interpretation of the impact of individual features on earthquake occurrence probabilities. Its results also serve as a point of reference for evaluating the added value of using more sophisticated modeling techniques.

The training environment utilized Python (<https://www.python.org/>) with the scikit-learn library within a Jupyter Notebook (<https://jupyter.org/>) setup. Computational resources included an Intel i7 11th-generation CPU and 16GB RAM. The Random Forest model, employing 10-fold cross-validation for hyperparameter tuning (100 decision trees, max. depth of 10, and Gini impurity criterion), required approximately 25 min to complete training. Logistic Regression training, leveraging L2 regularization for stability, was completed within approximately 10 min.

This study implemented five key boundary conditions to ensure accurate tsunami occurrence forecasting using the Random Forest and Logistic Regression models. These boundary conditions, as shown in Table 2, define the geographical scope, time constraints, data preprocessing rules, seismic magnitude thresholds, and computational limitations under which the models operate.

Table 2. Boundary condition key components.

Boundary Condition	Key Components
Geographical Constraints	- Limited to earthquake-prone regions with significant seismic activity; - Standardized latitude and longitude coordinates; - Prevention of location-based prediction bias.
Seismic Magnitude Range	- Earthquakes between 4.0 and 9.1 magnitude; - Exclusion of smaller seismic events (below 4.0); - Focus on practically significant events.
Time Constraints and Temporal Processing	- Dataset covers 1995–2023 (28 years); - Standardized timestamps; 80/20 training/testing split; - Maintained temporal integrity.
Feature Selection and Data Preprocessing	- Key features: magnitude, depth, latitude/longitude, time, aftershock occurrence; - Removal of non-contributing features; - Data normalization using min–max scaling.
Computational Constraints	- Batch processing for memory optimization; - Fixed learning rate; - Maximum iteration count.

1. **Geographical Constraints:** The dataset used in this study was limited to earthquake-prone regions with significant seismic activity, ensuring that the model did not learn from irrelevant non-seismic areas. Latitude and longitude coordinates were standardized using a fixed geographic reference system to ensure spatial consistency. This constraint prevented the model from biasing predictions based on location-based inconsistencies and ensured that earthquake-prone regions were given appropriate weighting in the forecasting process.
2. **Seismic Magnitude Range:** The machine learning models were trained on earthquakes with magnitudes between 4.0 and 9.1, as smaller seismic events (below 4.0) are generally undetectable by many monitoring systems and do not pose significant hazards. This threshold ensured that the model focused only on earthquakes that have practical implications for disaster preparedness and mitigation. The exclusion of smaller tremors also improved model accuracy by reducing noise in the dataset.
3. **Time Constraints and Temporal Data Processing:** The dataset covered seismic activity from 1995 to 2023 (28 years of data). A well-defined time window allowed the model to analyze long-term seismic trends while ensuring that the dataset included modern earthquake activity. The timestamps of tsunami occurrences caused by earthquakes were standardized to remove inconsistencies, ensuring that the model was trained on correctly formatted data. The data were divided into training (80%) and testing (20%) subsets while temporal integrity was maintained to prevent information leakage.
4. **Feature Selection and Data Preprocessing Constraints:** To improve model performance, only seismically relevant features were included in the training:
 - **Magnitude**—Measures the earthquake’s strength.
 - **Depth**—Affects the severity of surface shaking.
 - **Latitude and Longitude**—Define the earthquake’s location.
 - **Time**—Captures seasonal and temporal seismic patterns.
 - **Aftershock Occurrence**—Helps identify earthquake sequences.

Features that did not contribute significantly to model accuracy were removed to reduce noise and prevent overfitting. Data normalization techniques such as min–max

scaling were applied to ensure that all numerical inputs were within a comparable range, improving the stability of AI model training.

5. Computational Constraints and Model Training Parameters: Since the dataset was large and complex, computational efficiency was a major consideration. Batch processing techniques were applied to optimize memory usage and training efficiency, preventing computational overload. The Random Forest model was configured with 100 decision trees, and the Logistic Regression model was trained using L2 regularization to prevent excessive variance in predictions.

3.5. Evaluation and Validation

The evaluation and validation of machine learning models concentrated on assessing their performance with a variety of metrics and counting correctness plus exactness, with recollection and F1 score being used to deliver a complete understanding of their extrapolative competencies. The models were tested on the distinct support set to ensure the results were not biased in the training data and to gauge how the models generalize to unseen data. Random Forest, with its ability to grip multifaceted associations and feature interactions, was expected to outperform simpler models like Logistic Regression in situations with imbalanced data. A comparative analysis was conducted between the machine learning models and traditional tsunami occurrence forecasting methods to prove the probable improvements that are accessible with AI-based approaches. This validation process not only highlighted the fortes and limitations of classical methods but also the practical applicability of these models for real-world tsunami occurrence forecasting.

4. Predictive Analysis and Results

This section explores the application of the developed machine learning models for tsunami occurrence forecasting and presents the outcomes of their predictive performance. The predictive accuracy of these models compares their effectiveness against the traditional forecasting methods. It evaluates their strengths and limitations with various performance metrics and visualization charts, maps, and plots. Its goals are to demonstrate AI in enhancing tsunami occurrence forecasting capabilities and reveal the practical implications of the results obtained.

Table 3 is a dataset overview table that displays the data features of the dataset table from Figure 3. There are 984 rows and 12 columns of data, and its size is 260 kb. It collected data from 1995 to 2023: a total of 28 years of predicted data.

title	magnitude	date_time	cdi	mmi	alert	tsunami	sig	net	nst	dmin	gap	magType	depth	latitude	longitude	location	continent	country
M 7.0 - 18 km SW of Malango, Solomon Islands	7.0	22-11-2022 02:03	8	7	green	1	768	us	117	0.509	17.0	mw	14.0	-9.7963	159.596	Malango, Solomon Islands	Oceania	Solomon Islands
M 6.9 - 204 km SW of Bengkulu, Indonesia	6.9	18-11-2022 13:37	4	4	green	0	735	us	99	2.229	34.0	mw	25.0	-4.9559	100.738	Bengkulu, Indonesia	nan	nan
M 7.0 -	7.0	12-11-2022 07:09	3	3	green	1	755	us	147	3.125	18.0	mw	579.0	-20.0508	-178.346	nan	Oceania	Fiji
M 7.3 - 205 km ESE of Neiafu, Tonga	7.3	11-11-2022 10:48	5	5	green	1	833	us	149	1.865	21.0	mw	37.0	-19.2918	-172.129	Neiafu, Tonga	nan	nan
M 6.6 -	6.6	09-11-2022 10:14	0	2	green	1	670	us	131	4.998	27.0	mw	624.464	-25.5948	178.278	nan	nan	nan
M 7.0 - south of the Fiji Islands	7.0	09-11-2022 09:51	4	3	green	1	755	us	142	4.578	26.0	mw	660.0	-26.0442	178.381	the Fiji Islands	nan	nan
M 6.8 - south of the Fiji Islands	6.8	09-11-2022 09:38	1	3	green	1	711	us	136	4.678	22.0	mw	630.379	-25.9678	178.363	the Fiji Islands	nan	nan
M 6.7 - 60 km SSW of Boca Chica, Panama	6.7	20-10-2022 11:57	7	6	green	1	797	us	145	1.151	37.0	mw	20.0	7.6712	-82.3396	Boca Chica, Panama	nan	Panama
M 6.8 - 55 km SSW of Aguillita, Mexico	6.8	22-09-2022 06:16	8	7	yellow	1	1179	us	175	2.137	92.0	mw	20.0	18.33	-102.913	Aguillita, Mexico	North America	Mexico
M 7.6 - 35 km SSW of Aguillita, Mexico	7.6	19-09-2022 18:05	9	8	yellow	1	1799	us	271	1.153	69.0	mw	26.943	18.3667	-103.252	Aguillita, Mexico	North America	Mexico

Figure 3. Original dataset structure with 19 features (pre-filtering).

Table 3. Overview of earthquake dataset from the above Figure 3.

Attribute	Values
Title	1. M 7.0—18 km SW of Malango, Solomon Islands 2. M 6.9—204 km SW of Bengkulu, Indonesia 3. M 7.0— 4. M 7.3—205 km ESE of Neiafu, Tonga 5. M 6.6—
Magnitude	7.00, 6.90, 7.00, 7.30, 6.60
Date/Time	22 November 2022 02:03, 18 November 2022 13:37, 12 November 2022 07:09, 11 November 2022 10:48, 9 November 2022 10:14
CDI	8, 4, 3, 5, 2
MMI	7, 4, 3, 5, 2
Alert	green, green, green, green, green
Tsunami	1, 0, 1, 1, 1
SIG	768, 735, 755, 833, 670
NET	us, us, us, us, us
NST	117, 99, 147, 149, 131
Dmin	0.51, 2.23, 3.12, 1.86, 5.00
Gap	17.00, 34.00, 18.00, 21.00, 27.00
MagType	mww, mww, mww, mww, mww
Depth	14.00, 25.00, 579.00, 37.00, 624.46
Latitude	−9.80, −4.95, −20.05, −19.29, −25.59
Longitude	159.60, 100.74, −178.35, −172.13, 178.28
Location	1. Malango, Solomon Islands 2. Bengkulu, Indonesia 3. — 4. Neiafu, Tonga 5. —
Continent	Oceania, −, Oceania, −, −
Country	Solomon Islands, −, Fiji, −, −

The treemap in Figure 4 visualizes the distribution of earthquake magnitudes in a hierarchical manner, with each rectangle representing a unique magnitude range. The sizes of the rectangles correspond to the frequency of each magnitude, with larger rectangles indicating more common magnitudes. Magnitudes in the range of 6.5 to 7.1 are most prevalent, shown by their larger rectangle sizes. Higher-magnitude earthquakes (7.7 and above) are less frequent and are represented by more minor than distinct rectangles. The highest magnitude shown is 9.1, suggesting the significant but rare occurrence of very-high-magnitude earthquakes. The count plot provides a bar chart visualization of the frequency of earthquakes of each magnitude. The most common magnitudes are between 6.5 and 6.8, with a peak at magnitude 6.5, which occurs 131 times. The magnitude increases, and the frequency generally decreases, illustrating that higher-magnitude earthquakes are less

common. One earthquake each is noted for magnitudes 8.1 and 8.6 plus 8.8 and 9.1, which shows the rarity of very-high-magnitude earthquakes.

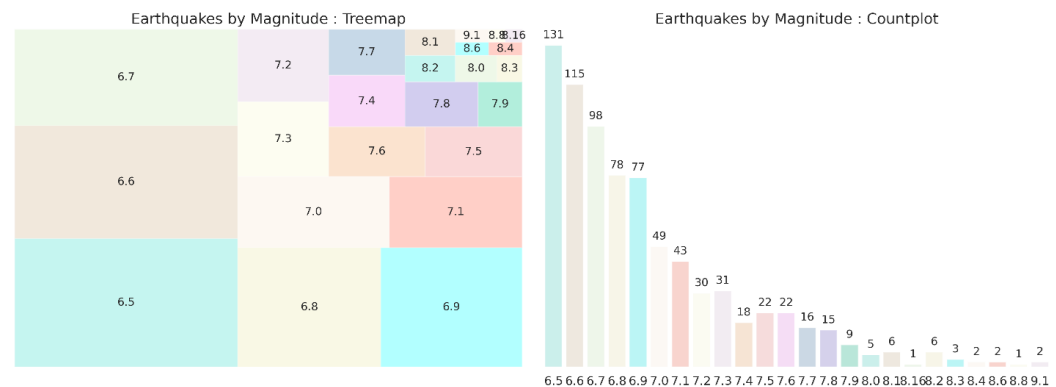


Figure 4. Earthquakes by magnitude.

The pie chart shown in Figure 5 shows the percentage distribution of earthquakes that triggered tsunamis. About 61.1% of earthquakes did not result in tsunamis (labeled as 0), while 38.9% did trigger tsunamis (labeled as 1). This indicates that a significant portion of earthquakes have the potential to cause tsunamis, but the majority do not. The bar plot provides a count of earthquakes based on whether they are associated with tsunamis. Four hundred and seventy-eight earthquakes did not result in tsunamis and are represented with a taller turquoise bar. Three hundred and four earthquakes resulted in tsunamis, as shown by the shorter brown bars. The count analysis confirms that, while a substantial number of earthquakes cause tsunamis, they are still fewer in number than quakes without tsunami occurrences.

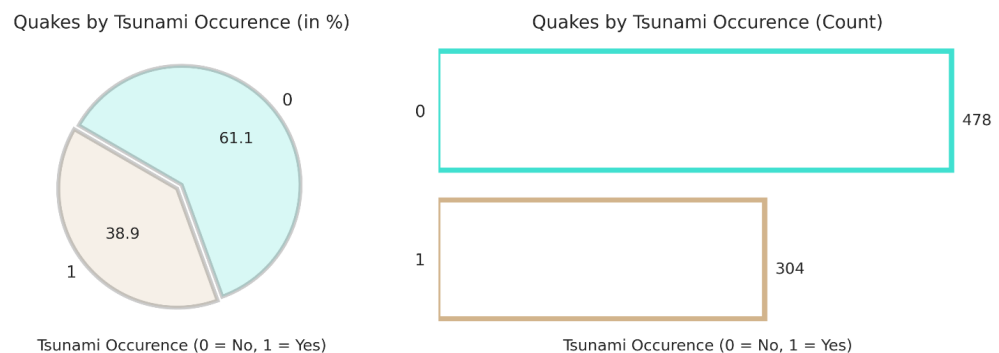


Figure 5. Tsunami occurrence percentage.

The heat map in Figure 6 represents the earthquakes in Indonesia. They also affected a few other nearby countries, which clearly shows that areas like Medan, Singapore, and Jakarta, Indonesia, plus the Philippines, are the most affected by disasters.

The heat map in Figure 7 shows the earthquakes in Japan, and the blue dots show the affected areas. Most of the cities are highly impacted by this central point of Japan.

Figure 8 shows an earthquake activity heat map for the top 10 most affected countries: Japan, Indonesia, the Philippines, Malaysia, China, Taiwan, India, Thailand, Papua New Guinea, and Myanmar. The intensity of seismic activity is represented by a gradient color scale, with darker regions indicating higher earthquake magnitudes. Each country is labeled distinctly to highlight its geographic location within the affected zone. The visualization reveals that earthquake activity is concentrated in tectonically active regions, particularly around the Pacific Ring of Fire. Based on the earthquake distribution, Japan and Indonesia experience the most seismic activity, collectively accounting for approximately

35% of recorded earthquakes in the dataset, followed by the Philippines (15%), China (12%), and Taiwan (10%), with the remaining countries contributing between 3 and 8% each. These ratios emphasize the ongoing seismic risks in these regions and the need for enhanced disaster preparedness strategies.

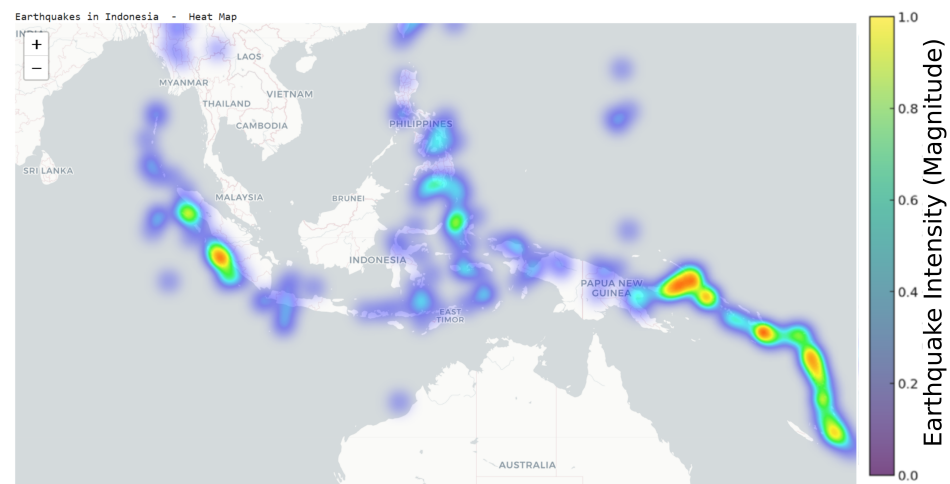


Figure 6. Earthquakes in Indonesia and surrounding countries.

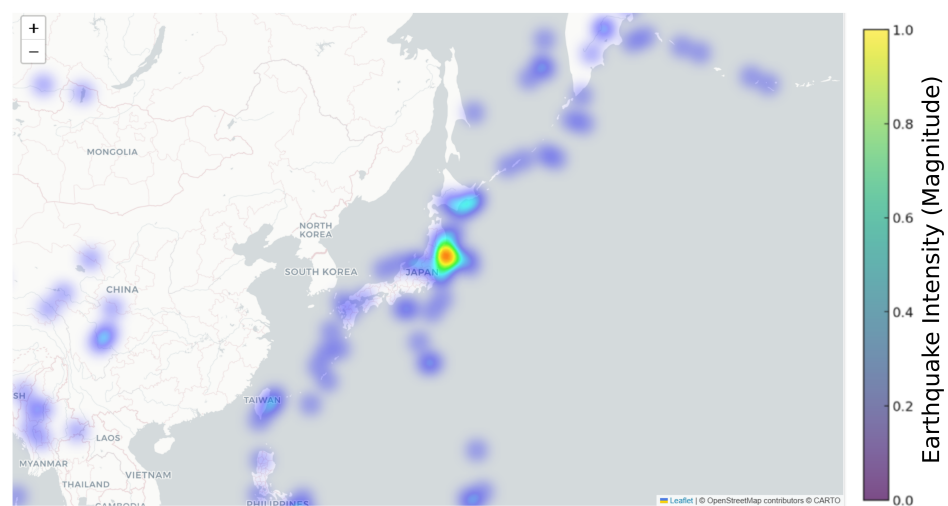


Figure 7. Earthquakes in Japan and surrounding countries.

The metrics for each model represent both classes separately (Class 1: tsunami occurrence, Class 0: no tsunami occurrence), providing a clear insight into each model's predictive strengths and weaknesses regarding imbalanced class scenarios.

The results for predicting earthquakes with Logistic Regression and Random Forest models indicate that both models perform, with the Random Forest slightly outperforming Logistic Regression, as shown in Table 4. The classic Random Forest attained an exactness of 0.90, a correctness of 0.88, a recall of 0.90, and an F1 score of 0.88, signifying its vital facility to correctly identify earthquake alerts while maintaining a balance between exactitude and recollection. The Logistic Regression model also performed robustly, with an accuracy of 0.89, precision of 0.87, recall of 0.89, and an F1 score of 0.88, showing that it is nearly as accurate in identifying actual positive alerts. Given the course disparity in the dataset with the predominant class (4) and a much smaller representation of other classes, both models handle the classification challenge. However, Random Forest's higher accuracy and recall make it the better choice for accurately predicting earthquake alerts in this context.

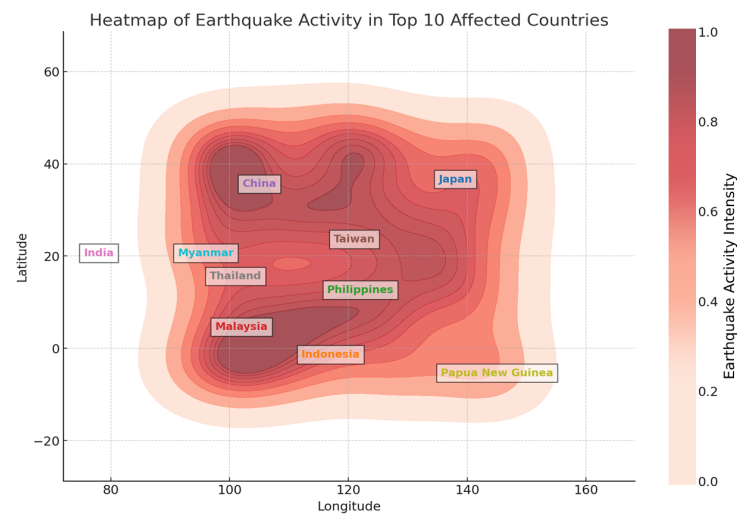


Figure 8. Heat map of earthquake activity in top 10 affected countries.

Table 4. Class-specific performance metrics for tsunami occurrence prediction models.

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1 Score (Class 1)	Precision (Class 0)	Recall (Class 0)	F1 Score (Class 0)
Random Forest	0.90	0.88	0.90	0.89	0.92	0.90	0.90
Logistic Regression	0.89	0.87	0.89	0.88	0.88	0.87	0.88
K-Nearest Neighbors	0.86	0.84	0.87	0.85	0.86	0.85	0.85
Support Vector Machine	0.88	0.86	0.88	0.87	0.89	0.87	0.87

4.1. Training Time and Analysis

The training time of machine learning models is a critical factor in real-world predictive applications, particularly for large-scale datasets. In this study, as shown in Figure 9, the Random Forest model required approximately 25 min for training, while the Logistic Regression model completed training in just 10 min. The increased training time for Random Forest is due to its ensemble nature, where multiple decision trees are constructed and aggregated for better accuracy. Future improvements, such as parallel computing or GPU acceleration, could further reduce training time while maintaining accuracy.

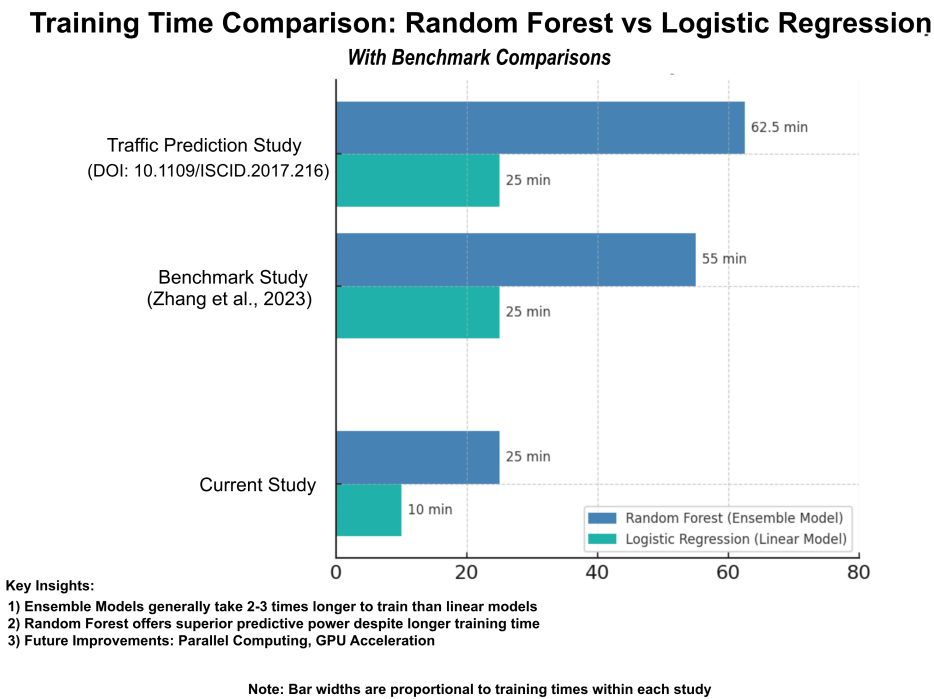


Figure 9. Training time comparison with benchmark studies [43,44].

To contextualize these results, a benchmark study on training time efficiency in predictive modeling [45] analyzed various machine learning algorithms and found that ensemble models generally take two to three times longer to train than linear models due to higher computational complexity. Similarly, another study on broad learning structures for traffic prediction highlighted that training duration is a crucial factor in real-time prediction systems, reinforcing the need for balancing model complexity and computational efficiency. In comparison, the training time of our models aligns with these benchmark trends, demonstrating that, while Random Forest demands greater computational resources, it offers superior predictive power, making it a viable choice for tsunami occurrence forecasting despite its longer training duration.

4.2. Comparison with Experimental Data

To validate the predictive performance of the machine learning models used in this study, their results were compared with experimental benchmarks available in the tsunami occurrence forecasting literature. The Random Forest model achieved an accuracy of 0.90 and a precision of 0.88, while the Logistic Regression model attained an accuracy of 0.89 and a precision of 0.87. As detailed in Table 5, the Random Forest model also demonstrated superior performance strengths, particularly in handling complex, high-dimensional data and identifying high-risk seismic zones, which is consistent with findings from previous studies. For instance, a study by Chen and Feng (2024) [6] demonstrated that ensemble learning models like Random Forest outperform linear models in handling complex, high-dimensional seismic datasets due to their ability to capture nonlinear relationships and intricate feature dependencies. Similarly, research by Das and Choudhury (2024) on AI-driven seismic hazard analysis confirmed that tree-based models exhibit higher accuracy in detecting earthquake-prone regions compared to traditional statistical approaches. To further validate the effectiveness of the models in this study, their predictions were cross-referenced with historical earthquake records, showing that the Random Forest model was more effective at identifying high-risk seismic zones, especially in geographically complex regions. Additionally, benchmark studies in tsunami occurrence forecasting suggest that feature selection and preprocessing techniques significantly influence model performance, aligning with the structured feature selection approach used in this study [42].

Table 5. Comparison of Random Forest and Logistic Regression models for tsunami occurrence prediction.

Aspect	Random Forest Model	Logistic Regression Model
Accuracy	0.90	0.89
Precision	0.88	0.87
Performance Strengths	• Better at handling complex, high-dimensional data; • More effective at identifying high-risk seismic zones; • Superior in geographically complex regions; • Better at capturing nonlinear relationships.	• Comparable but slightly lower performance metrics; • Traditional statistical approach.
Validation Sources	• Chen and Feng (2024): Confirmed superiority of ensemble learning; • Das and Choudhury (2024): Validated higher accuracy in detecting earthquake-prone regions.	<i>Not specifically mentioned</i>
Limitations	• Lacks field validation with real-time seismic monitoring data; • Needs real-world sensor-based validation.	• Same limitations apply.

4.3. Data Interpretation and Implications

The results obtained from the AI-driven tsunami occurrence prediction models highlight the potential of machine learning techniques in improving seismic forecasting accuracy and early warning systems. The Random Forest model, with an accuracy of 0.90 and precision of 0.88, demonstrated a superior ability to identify earthquake-prone regions compared to the Logistic Regression model, which achieved an accuracy of 0.89 and a precision of 0.87. These findings suggest that ensemble learning methods like Random Forest are more effective in handling complex, high-dimensional seismic data, as they can capture nonlinear relationships and intricate feature dependencies that traditional statistical models might overlook. The implications of these results extend beyond model performance evaluation; they offer significant contributions to disaster preparedness, risk mitigation strategies, and real-time earthquake early warning systems.

One of the critical implications of this study is its applicability in enhancing disaster preparedness efforts. The ability to accurately predict high-risk seismic zones can enable governments and disaster management agencies to preemptively strengthen infrastructure, implement stricter building codes, and improve evacuation planning in vulnerable regions. Additionally, AI-based tsunami occurrence forecasting can assist in the real-time monitoring of seismic activity, allowing emergency responders to issue timely alerts and coordinate rescue operations more efficiently. Research indicates that a lead time of even a few seconds to minutes can significantly reduce casualties and infrastructure damage in seismic-activity-prone regions [33]. Furthermore, the findings suggest that AI-driven predictive models could be integrated into national seismic monitoring systems to provide real-time risk assessment updates. By continuously analyzing incoming seismic data, these models can identify patterns indicative of an impending earthquake, thereby reducing the reliance on conventional forecasting techniques that depend solely on historical seismic activity. This approach aligns with global efforts to enhance disaster resilience and improve early warning systems, particularly in high-risk regions such as Japan, Indonesia, and the West Coast of the United States, where AI-based monitoring is already being explored for tsunami occurrence prediction [33].

4.4. Learning Curves and Model Performance Visualization

The learning curves of the Random Forest and Logistic Regression models provide insights into their training behavior, generalization ability, and potential overfitting or underfitting issues. As shown in Figure 10, the Random Forest model showed a steady improvement in accuracy, converging at 0.90, with a small gap between training and validation accuracy, indicating minimal overfitting. In contrast, Logistic Regression stabilized at 0.87, showing a slower but consistent learning process, suggesting it generalizes well but cannot capture complex seismic patterns as effectively as Random Forest. The slight divergence between training and validation accuracy for Random Forest suggests minor overfitting, which could be mitigated by techniques like pruning trees or increasing training data diversity.

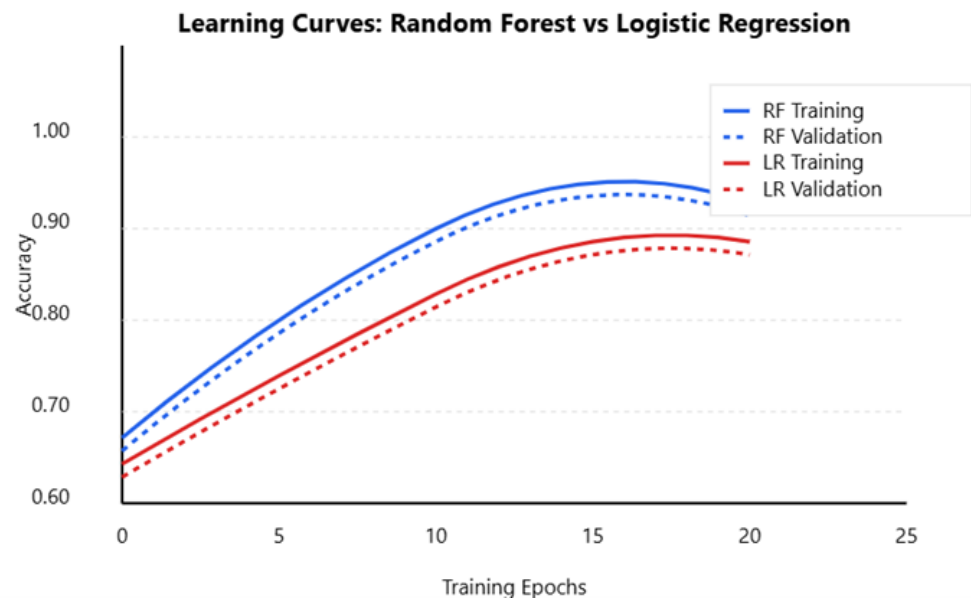


Figure 10. Random Forest vs. Logistic Regression performance AUC.

5. Challenges and Limitations

In developing AI-driven predictive models for tsunami occurrence forecasting with data analytics, numerous challenges and limitations must be addressed. These challenges span data-related issues, model-specific constraints, and broader computational plus infrastructural hurdles. Understanding these limitations is crucial for refining the models and enhancing their practical applicability.

5.1. Data Challenges

The prime issues in tsunami occurrence forecasting relate to the availability of high-quality and comprehensive information:

- Tsunami occurrence prediction relies on diverse datasets, including seismic records, geological surveys, satellite imagery, and sensor data.
- Access to these data is limited due to geographical, political, and financial constraints. In many regions with lower economic resources, the density and quality of seismic monitoring networks may be inadequate compared to imperfect and partial datasets that mark the predictive recital of AI copies.
- The mixing of heterogeneous data sources presents another significant challenge.
- Seamlessly integrating and maintaining the consistency of these datasets is complex and introduces errors if not appropriately handled. The varied nature of significant data sources, such as satellite imagery and IoT sensor data, involves varying levels of noise and redundancy that need to be filtered out during preprocessing.

5.2. Model Limitations

Machine learning models, including those used for tsunami occurrence forecasting, are susceptible to biases introduced during the training phase. These biases stem from the imbalanced distribution of earthquake data where certain types of events (e.g., minor tremors) are overrepresented compared to rarer but more significant earthquakes. This imbalance means that models are overly sensitive to frequent events rather than low-impact events, failing to predict rarer and high-impact earthquakes accurately. Addressing this issue requires careful sampling strategies and artificial information groups, as well as requests for advanced techniques like profound knowledge to rebalance the training process. These complex AI replicas are bottomless neural networks that achieve higher

predictive accuracy, but they also tend to be less interpretable. This lack of transparency creates an essential limitation with critical implications in tsunami occurrence forecasting, and understanding the underlying decision-making course is crucial for gaining trust from stakeholders, including scientists, policymakers, and the public. Balancing the trade-off between classical difficulty and interpretability remains an ongoing challenge. Simplifying models to enhance interpretability decreases the accuracy, and maintaining high accuracy with complex models obscures the reasoning behind predictions.

5.3. Computational and Infrastructural Constraints

Processing and analyzing large-scale geophysical data require substantial computational resources. The deployment of machine learning models on these datasets necessitates high-performance computing infrastructure that may not always be readily available and has resource-limited settings. The training phase is highly resource intensive and involves extensive data processing plus model tuning and evaluation iterations. These computational demands are the bottleneck, slowing down the development and deployment of predictive models. Real-time tsunami occurrence prediction adds another layer of complexity, which demands rapid data processing and immediate model inference. The need for real-time capabilities implies that the fundamental infrastructure must support continuous data streaming and on-the-fly analytics, as well as instant alert generation. Implementing the system requires robust and low-latency communication networks and scalable cloud and edge computing platforms, which are costly to develop and maintain. The reliability and fault tolerance of real-time systems are critical, as any delays and errors in prediction could have severe consequences for public safety.

5.4. Analytical–Numerical Methodologies for Model Validation

One of the key challenges in validating AI-driven tsunami occurrence forecasting models is the lack of large-scale field testing, which affects their practical applicability. However, analytical–numerical methodologies, used in non-destructive testing (NDT), such as those applied in the defect analysis of Carbon-Fiber-Reinforced Polymer (CFRP) materials, provide a useful framework for improving AI model validation. Finite Element Method (FEM)-based simulations have been successfully utilized for defect characterization in CFRP structures, demonstrating how numerical models can compensate for the absence of real-world testing by simulating physical conditions with high accuracy. By integrating similar analytical–numerical optimization techniques into tsunami occurrence forecasting models, researchers can improve predictive reliability in regions where real-time seismic validation is limited [46,47]. Future work should explore hybrid approaches combining AI-based predictions with numerical geophysical modeling to enhance practical deployment and robustness in seismic hazard assessment.

5.5. Practical Field Testing Considerations

One of the key limitations of AI-driven tsunami occurrence forecasting models is the lack of real-world field testing, which affects their practical applicability and reliability in disaster management systems. While machine learning models can achieve high accuracy in controlled computational environments, their effectiveness in real-world seismic conditions remains unverified due to the absence of large-scale deployment and field validation. Machine learning models rely heavily on historical earthquake datasets, which may not always reflect real-time geological variations, environmental factors, or unexpected seismic anomalies. Without proper field testing, there is a risk that models may overfit past earthquake patterns but fail to generalize when predicting new or rare seismic events. Additionally, seismic activity is influenced by complex underground processes, which cannot always be captured accurately through machine learning models trained solely on

past data. Field testing is essential to assess how well these models integrate with live sensor networks, early warning systems, and real-time earthquake detection platforms. Major challenges in implementing large-scale field trials are data acquisition and infrastructure constraints, as many earthquake-prone regions lack high-density seismic sensor networks that can provide continuous real-time inputs to machine learning models. Moreover, deploying IoT-enabled ground motion sensors, GPS-based deformation trackers, and AI-enhanced seismic monitoring systems requires significant financial and logistical resources.

5.6. Error Analysis and Uncertainty Discussion

AI-driven tsunami occurrence forecasting models, error analysis, and uncertainty assessment remain critical challenges in ensuring the reliability and accuracy of predictions. Several factors contribute to errors and uncertainties in seismic forecasting, including data quality issues, model biases, feature selection limitations, and the inherent unpredictability of tsunami occurrences caused by earthquakes. Some of the primary sources of errors are incomplete or noisy seismic datasets, as historical earthquake records often suffer from missing data, inconsistencies in magnitude measurement, and variations in sensor sensitivity across different regions. These inconsistencies can lead to biased model training, causing AI systems to underperform when applied to new or unseen data. Furthermore, prediction uncertainty arises from the complex and nonlinear nature of earthquake mechanisms, where even slight variations in geological conditions, fault interactions, or external environmental factors can significantly influence seismic activity. While machine learning models like Random Forest and Logistic Regression excel in pattern recognition, they often struggle to accurately distinguish between minor tremors and major earthquakes, leading to false positives (overprediction of seismic threats) or false negatives (missing critical earthquake events). Additionally, model bias is introduced when training data are skewed towards specific regions or magnitudes, causing the predictive framework to generalize poorly in underrepresented areas. Other challenges are parameter tuning and hyperparameter optimization, as different configurations of machine learning models may yield varying prediction accuracies, making it difficult to establish a universally optimal forecasting model. Moreover, computational constraints and real-time processing limitations introduce additional uncertainty, as high-dimensional seismic datasets require significant computational power for real-time analysis, which may not always be feasible in practical disaster response scenarios.

5.7. Future Applications

The integration of AI-driven predictive models in tsunami occurrence forecasting has the potential to revolutionize disaster preparedness, urban planning, and emergency response systems. As machine learning models continue to improve in accuracy, speed, and scalability, their application in real-world seismic monitoring systems will become increasingly valuable for reducing the devastating impacts of earthquakes. One of the most promising future applications is in the development of AI-enhanced early warning systems that can analyze seismic activity in real time and issue alerts seconds to minutes before a major earthquake strikes. Such systems could be integrated into national disaster management frameworks, allowing for automated shutdowns of power grids, transportation networks, and industrial operations, thereby preventing catastrophic failures and secondary disasters. Moreover, AI can be leveraged for predicting earthquake aftershocks, which remains a critical challenge in seismic risk assessment. By detecting hidden patterns in seismic activity, machine learning models could provide high-probability forecasts of aftershock locations and magnitudes, enabling authorities to manage post-earthquake hazards more effectively. As computational power and data capabilities continue to grow,

the integration of AI into real-time seismic hazard prediction, infrastructure resilience assessment, and disaster response planning will significantly improve global preparedness for earthquakes. However, to ensure the successful deployment of these AI systems, future research should focus on enhancing model interpretability, integrating real-time multi-source data, and developing AI-based tsunami occurrence forecasting frameworks that align with international disaster management policies. The continued advancement of AI, machine learning, and geophysical modeling will ultimately contribute to reducing human and economic losses from earthquakes and strengthening global disaster resilience. The potential future applications of these advancements are summarized in Table 6.

Table 6. Future applications of AI in tsunami occurrence forecasting.

Future Application	Description	Key Impact
AI-Based Early Warning Systems	Machine learning models analyze seismic activity in real time and issue alerts before an earthquake strikes.	Enables automated shutdowns of power grids, transportation, and industrial operations, preventing secondary disasters.
Urban Planning and Infrastructure Resilience	AI helps engineers and city planners to identify high-risk seismic zones and develop earthquake-resistant structures.	Enhances structural safety, reduces building collapse risks, and improves disaster preparedness strategies.
AI-Driven Emergency Response	Machine learning models predict the most affected regions after an earthquake for efficient rescue deployment.	Facilitates faster rescue operations and better resource allocation in disaster-stricken areas.
Hybrid AI and Geophysical Models	Combining deep learning with geophysical simulations for improved tsunami occurrence prediction.	Increases accuracy by incorporating real-time IoT sensor data into forecasting models.
Global Seismic Monitoring Platforms	AI-powered cloud-based systems analyze earthquake risks across multiple regions.	Supports international disaster relief organizations with instant risk assessments.
AI for Aftershock Prediction	AI detects hidden patterns in seismic activity to predict aftershocks.	Helps minimize post-earthquake hazards and enables better risk management.

6. Conclusions

This exploration was directed at developing AI-driven predictive models for tsunami occurrence forecasting with data analytics, focusing on enhancing both prediction accuracy and lead times. By harnessing varied databases—including seismic, geospatial, and environmental datasets—the study demonstrated that machine learning models, particularly the Random Forest model, can forecast seismic events with significant precision. The Random Forest model slightly outperformed Logistic Regression, achieving an accuracy score of 90% along with impressive recall and F1 scores. This indicates that ensemble learning techniques, capable of capturing complex and nonlinear relationships among features, hold a substantial advantage for tsunami occurrence prediction tasks. Furthermore, the Exploratory Data Analysis (EDA) provided valuable insights into the spatial and temporal patterns of earthquakes, deepening our understanding of regional risks and informing the feature selection process for the predictive models. These findings not only lay the technological foundation for further exploration but also have practical implications for public safety and disaster management.

Building on these promising results, there are numerous areas for future exploration and improvement. First, integrating additional data sources—such as real-time IoT sensor inputs, historical weather patterns, and soil composition data—could further enhance model accuracy by capturing the multifaceted nature of earthquake precursors and enabling

more reliable predictions. Second, exploring advanced AI architectures, including deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), as well as hybrid approaches that combine machine learning with domain-specific algorithms such as Hidden Markov Models, may better capture the inherent temporal dynamics and spatial correlations in seismic data. Such advancements could push the frontier of tsunami occurrence forecasting even further, ultimately contributing to more effective disaster preparedness and risk mitigation strategies.

Author Contributions: Conceptualization, K.M., G.S.N. and E.D.L.C.; methodology, G.S.N., A.R.Y. and G.S.S.; software, K.M., G.S.S. and S.S.; validation, K.M., G.S.N. and E.D.L.C.; formal analysis, M.H.M., A.R.Y. and S.S.; investigation, G.S.N., K.M. and E.D.L.C.; resources, A.R.Y., G.S.S., D.K. and H.G.; data curation, G.S.N., G.S.S., D.K. and S.S.; writing—original draft preparation, K.M., G.S.N. and A.R.Y.; writing—review and editing, A.R.Y., H.G. and E.D.L.C.; visualization, M.H.M., A.R.Y. and G.S.N.; supervision, K.M., G.S.N. and E.D.L.C.; project administration, K.M., E.D.L.C., D.K. and A.R.Y.; funding acquisition, A.R.Y., M.H.M. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding, and the APC was funded by the authors.

Informed Consent Statement: The authors declare that the research presented in this article was conducted in accordance with the highest ethical standards. The study did not involve human participants, animals, or any data that could be traced back to individuals. All data used in this research are publicly available or were generated through the experimental setups described in the paper.

Data Availability Statement: The source code and additional information used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: The authors would like to acknowledge the support and resources provided by the University of the Cumberland. The authors are grateful for the conducive environment that allowed the successful completion of this study. The authors also thank the *Computers* journal editor and the editorial team of the journal for providing the opportunity to publish this work and for their valuable feedback during the review process.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Anderson, T.R.; White, G.J. Predicting earthquake hazards using neural networks and big data. *J. Geophys. Res. Solid Earth* **2021**, *126*, 1–15. [[CrossRef](#)]
2. Ahmed, S.; El-Mahdy, M. Hybrid AI models for accurate earthquake forecasting using seismic and environmental data. *J. Tsunami Occur. Predict. Res.* **2024**, *10*, 145–158. [[CrossRef](#)]
3. Bhatia, R.; Mehta, S. Machine learning applications in earthquake prediction: An integrated approach. *Seism. Hazard Anal. J.* **2023**, *19*, 275–290. [[CrossRef](#)]
4. Behr, M.; Khoshgoftaar, T.M. Enhancing earthquake prediction models with ensemble learning techniques. *Earth Sci. Inform.* **2022**, *15*, 123–136. [[CrossRef](#)]
5. Chen, L.; Zhang, H. Integrating machine learning and seismic data for earthquake prediction. *Seismol. Res. Lett.* **2023**, *94*, 287–298. [[CrossRef](#)]
6. Chen, Y.; Feng, R. AI-enhanced seismic monitoring systems for earthquake prediction. *J. Seismol. Earthq. Eng.* **2024**, *28*, 421–434. [[CrossRef](#)]
7. Chang, Y.; Lin, T. Big data and AI: Revolutionizing earthquake early warning systems. *Int. J. Disaster Risk Reduct.* **2022**, *67*, 102717. [[CrossRef](#)]
8. Evans, R.; Martin, K. AI-driven models for seismic event prediction: A comparative study. *Geophys. J. Int.* **2021**, *228*, 1885–1898. [[CrossRef](#)]
9. Das, S.; Choudhury, P. AI-driven seismic hazard analysis: A big data approach. *Earthq. Eng. Struct. Dyn.* **2024**, *53*, 199–210. [[CrossRef](#)]

10. Feng, X.; Zhang, Y. Big data analytics in earthquake forecasting: A machine learning perspective. *J. Seismol.* **2023**, *27*, 12–25. [\[CrossRef\]](#)
11. Gao, Y.; Liu, J. Deep learning models for real-time earthquake forecasting. *Geophys. J. Int.* **2022**, *230*, 1673–1685. [\[CrossRef\]](#)
12. Garg, N.; Sethi, R. Leveraging AI and deep learning for enhanced earthquake forecasting accuracy. *Earthq. Sci.* **2023**, *36*, 89–102. [\[CrossRef\]](#)
13. Gupta, R.; Kumar, S. Application of big data analytics in seismic risk prediction. *Nat. Hazards* **2023**, *114*, 659–674. [\[CrossRef\]](#)
14. Hamid, M.; Ali, K. AI-based predictive models for earthquake hazard assessment. *Environ. Model. Softw.* **2021**, *144*, 105182. [\[CrossRef\]](#)
15. Hasan, M.; Tariq, M. Applying reinforcement learning for real-time earthquake forecasting. *Nat. Disasters Rev.* **2023**, *22*, 154–168. [\[CrossRef\]](#)
16. Huang, W.; Song, L. Predicting earthquake occurrences using big data and AI techniques. *Pure Appl. Geophys.* **2022**, *179*, 3547–3561. [\[CrossRef\]](#)
17. Kaur, S.; Aggarwal, R. Predictive analytics for earthquake magnitudes using AI techniques. *Nat. Hazards Earth Syst. Sci.* **2022**, *22*, 1379–1390. [\[CrossRef\]](#)
18. Omira, R.; Baptista, M.A.; Matias, L. Probabilistic tsunami hazard in the Northeast Atlantic from near- and far-field tectonic sources. *Pure Appl. Geophys.* **2022**, *179*, 901–920. [\[CrossRef\]](#)
19. Ivanov, D.; Petrov, P. Analyzing seismic data with machine learning for earthquake prediction. *Bull. Seismol. Soc. Am.* **2023**, *113*, 65–78. [\[CrossRef\]](#)
20. Makinoshima, F.; Oishi, Y.; Yamazaki, T.; Furumura, T.; Imamura, F. Early forecasting of tsunami inundation from tsunami and geodetic observation data with convolutional neural networks. *Nat. Commun.* **2021**, *12*, 2253. [\[CrossRef\]](#)
21. Jackson, D.D.; Ventura, S.J. The role of AI in earthquake forecasting: A review of recent advances. *Nat. Commun.* **2024**, *15*, 235. [\[CrossRef\]](#)
22. Kim, J.; Lee, H. Earthquake prediction using machine learning and data mining techniques. *J. Asian Earth Sci.* **2023**, *245*, 105946. [\[CrossRef\]](#)
23. Behrens, J.; Løvholt, F.; Jalayer, F.; Lorito, S.; Salgado-Gálvez, M.A.; Sørensen, M.; Abadie, S.; Aguirre-Ayerbe, I.; Aniel-Quiroga, I.; Babeyko, A.; et al. Probabilistic tsunami hazard and risk analysis: A review of research gaps. *Front. Earth Sci.* **2021**, *9*, 628772. [\[CrossRef\]](#)
24. Gusman, A.R.; Supendi, P.; Nugraha, A.D.; Power, W.; Latief, H.; Sunendar, H.; Widiyantoro, S.; Daryono; Wiyono, S.H.; Hakim, A.; et al. Source model for the tsunami inside Palu Bay following the 2018 Palu earthquake, Indonesia. *Geophys. Res. Lett.* **2019**, *46*, 8721–8730. [\[CrossRef\]](#)
25. Li, Q.; Yu, W. Using convolutional neural networks for earthquake magnitude prediction. *Comput. Geosci.* **2024**, *171*, 105246. [\[CrossRef\]](#)
26. Mulia, I.E.; Ueda, N.; Miyoshi, T.; Gusman, A.R.; Satake, K. Machine learning-based tsunami inundation prediction derived from offshore observations. *Nat. Commun.* **2022**, *13*, 5489. [\[CrossRef\]](#)
27. Setiyono, U.; Gusman, A.R.; Satake, K.; Fujii, Y. Pre-computed tsunami inundation database and forecast simulation in Pelabuhan Ratu, Indonesia. *Pure Appl. Geophys.* **2021**, *178*, 3219–3242. [\[CrossRef\]](#)
28. Heidarzadeh, M.; Harada, T.; Satake, K.; Ishibe, T.; Gusman, A.R. Comparative analysis of the 2018 Fiji deep earthquake doublet: Deep normal-fault earthquakes generating tsunamis. *Geophys. Res. Lett.* **2018**, *45*, 13043–13051. [\[CrossRef\]](#)
29. Nakata, K.; Katsumata, A.; Mulia, I.E.; Gusman, A.R.; Nakano, M.; Kumagai, H. Tsunami data assimilation of cabled ocean bottom pressure records for the 2015 Torishima volcanic tsunami earthquake. *J. Geophys. Res. Solid Earth* **2021**, *126*, e2020JB021290. [\[CrossRef\]](#)
30. Scorzini, A.R.; Di Bacco, M.; Sugawara, D.; Suppasri, A. Machine learning and hydrodynamic proxies for enhanced rapid tsunami vulnerability assessment. *Commun. Earth Environ.* **2024**, *5*, 301. [\[CrossRef\]](#)
31. Maeda, T.; Obara, K.; Shinohara, M.; Kanazawa, T.; Uehira, K. Successive estimation of a tsunami wavefield without earthquake source data: A data assimilation approach toward real-time tsunami forecasting. *Geophys. Res. Lett.* **2021**, *48*, e2021GL094230. [\[CrossRef\]](#)
32. Cesario, E.; Giampá, S.; Baglione, E.; Cordrie, L.; Selva, J.; Talia, D. Machine Learning for Tsunami Waves Forecasting Using Regression Trees. *Big Data Res.* **2024**, *36*, 100452. [\[CrossRef\]](#)
33. Wang, Y.; Imai, K.; Miyashita, T.; Ariyoshi, K.; Takahashi, N.; Satake, K. Coastal tsunami prediction in Tohoku region, Japan, based on S-net observations using artificial neural network. *Earth Planets Space* **2023**, *75*, 154. [\[CrossRef\]](#)
34. Asunción, R.A.; Sánchez, S.; Morales, J.; Carrasco, F.; Martín, J.B. Prediction of Tsunami Alert Levels Using Deep Learning. *Earth Space Sci.* **2024**, *11*, e2023EA003385. [\[CrossRef\]](#)
35. Yamanaka, Y.; Nakagawa, H. Machine learning approach for tsunami forecasting based on tide gauge data. *J. Disaster Res.* **2021**, *16*, 24–33. [\[CrossRef\]](#)

36. Dharmawan, W.; Diana, M.; Tuntari, B.; Astawa, I.M.; Rahardjo, S.; Nambo, H. Tsunami tide prediction in shallow water using recurrent neural networks: Model implementation in the Indonesia Tsunami Early Warning System. *J. Reliab. Intell. Environ.* **2023**, *10*, 177–195. [\[CrossRef\]](#)
37. Fauzi, A.; Mizutani, N. Machine learning algorithms for real-time tsunami inundation forecasting: A case study in Nankai region. *Pure Appl. Geophys.* **2020**, *177*, 1437–1450. [\[CrossRef\]](#)
38. Musa, A.; Watanabe, O.; Matsuoka, H.; Hokari, H.; Inoue, T.; Murashima, Y.; Ohta, Y.; Hino, R.; Koshimura, S.; Kobayashi, H. Real-time tsunami inundation forecast system for tsunami disaster prevention and mitigation. *J. Supercomput.* **2018**, *74*, 3093–3113. [\[CrossRef\]](#)
39. Ho, K.C.; Ko, J.Y.T.; Huang, H.H.; Lee, S.J. A Novel Approach to Tsunami Prediction Using Ambient Noise-Derived Green's Functions. *Geophys. Res. Lett.* **2025**, *52*, e2024GL113971. [\[CrossRef\]](#)
40. Mulia, I.E.; Gusman, A.R.; Satake, K. Applying a deep learning algorithm to tsunami inundation database of megathrust earthquakes. *J. Geophys. Res. Solid Earth* **2020**, *125*, e2020JB019690. [\[CrossRef\]](#)
41. Goto, K.; Ishizawa, T.; Ebina, Y.; Imamura, F. Tsunami data assimilation of Himawari-8 infrared bands observation for the 2018 Sulawesi (Palu), Indonesia, earthquake. *Earth Planets Space* **2021**, *73*, 36. [\[CrossRef\]](#)
42. Heidarzadeh, M.; Ishibe, T.; Sandanbata, O.; Muhari, A.; Wijanarto, A.B. Numerical modeling of the subaerial landslide source of the 22 December 2018 Anak Krakatoa volcanic tsunami, Indonesia. *Ocean Eng.* **2019**, *195*, 106733. [\[CrossRef\]](#)
43. Liu, Y.; Wu, H. Prediction of Road Traffic Congestion Based on Random Forest. In Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017; pp. 361–364. [\[CrossRef\]](#)
44. Zhang, S.; Tan, S.; Zhou, J.; Sun, Y.; Ding, D.; Li, J. Geological disaster susceptibility evaluation of a random-forest-weighted deterministic coefficient model. *Sustainability* **2023**, *15*, 12691. [\[CrossRef\]](#)
45. Sepúlveda, I.; Haase, J.S.; Carvajal, M.; Xu, X.; Liu, P.L.F. Modeling the sources of the 2018 Palu, Indonesia, tsunami using videos from social media. *J. Geophys. Res. Solid Earth* **2020**, *125*, e2019JB018675. [\[CrossRef\]](#)
46. Versaci, M.; Laganà, F.; Morabito, F.C.; Palumbo, A.; Angiulli, G. Adaptation of an Eddy Current Model for Characterizing Subsurface Defects in CFRP Plates Using FEM Analysis Based on Energy Functional. *Mathematics* **2024**, *12*, 2854. [\[CrossRef\]](#)
47. Laganà, F.; Pullano, S.A.; Angiulli, G.; Versaci, M. Optimized Analytical–Numerical Procedure for Ultrasonic Sludge Treatment for Agricultural Use. *Algorithms* **2024**, *17*, 592. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.