AI-Driven Tsunami Impact Analysis: A Statistical and Machine Learning Approach

Data Collection, Literature Review, and Challenges

CS750 - Distributed Data Management
Lab Assignment 2

by
Sohail Shaik
Roll.No.: 242CS033
Mobile No.: 7032096859
Email: shaiksohail.242cs033@nit.edu.in

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
(NITK)
SURATHKAL, MANGALORE - 575 025

Domain Title

AI Driven Cyber Security Fraud Detection Analysis: A Statistical and Machine Learning Approach

Dataset Title

National Geophysical Data Center (NGDC) / NOAA Historical Tsunami Event Database

Domain Overview

Tsunamis are among the most catastrophic natural disasters known to humanity. Triggered primarily by undersea earthquakes, volcanic eruptions, or landslides, these giant sea waves can travel across entire ocean basins, inflicting massive destruction when they make landfall. The 2004 Indian Ocean tsunami and the 2011 Tōhoku tsunami in Japan serve as stark reminders of the scale of devastation such events can cause. Predicting the impact of these events has long been a key challenge for disaster management authorities worldwide. With recent advancements in data science, particularly in the fields of statistical modeling and machine learning, there has been a growing interest in leveraging historical tsunami data for predictive analysis.

The domain of tsunami impact prediction using machine learning is interdisciplinary in nature. It lies at the intersection of oceanography, seismology, data science, and artificial intelligence. Traditional tsunami early warning systems rely heavily on seismic and geophysical instruments like DART buoys, tide gauges, and GPS sensors. While these systems provide critical real-time data, they are often limited by geographical deployment, response latency, and inability to model complex impact patterns. This is where machine learning can play a transformative role.

By using large historical datasets, including geological and hydrodynamic features, machine learning models can learn patterns that might not be obvious through traditional statistical analysis alone. These models can analyze correlations between variables such as earthquake magnitude, depth, epicenter coordinates, sea floor displacement, and subsequent tsunami wave height or casualty count. Predictive tasks can include classifying whether a tsunami will occur after a seismic event, estimating its wave height, forecasting the number of affected people, and identifying high-risk zones.

Supervised learning techniques, including decision trees, random forests, support vector machines (SVM), and neural networks, have been employed to predict tsunami characteristics. Unsupervised methods like clustering and principal component analysis help in detecting hidden patterns or reducing dimensionality in high-volume datasets. Time-series forecasting and spatial modeling also play a significant role, particularly in anticipating tsunami arrival times and geographic spread.

However, this domain is not without challenges. Historical tsunami datasets are often incomplete, inconsistent, or sparse in some geographical regions. Moreover, many earthquakes do not result in tsunamis, leading to significant class imbalance in predictive modeling. Preprocessing becomes crucial — handling missing data, normalizing measurements, encoding categorical values, and performing outlier detection are essential steps to prepare the data for analysis.

Additionally, real-time prediction remains an open challenge. Machine learning models must be computationally efficient and highly accurate under time constraints, as any delay in prediction can cost lives. Integrating external data sources like satellite imagery, weather data, and ocean temperature can further improve accuracy, but also increases the complexity of model training.

Despite these hurdles, the application of statistical and machine learning approaches to tsunami prediction is a promising field. As computational power and data availability improve, these methods will become an increasingly essential component of early-warning systems. The insights derived from these models can inform disaster response planning, coastal infrastructure design, and public policy, ultimately contributing to more resilient coastal communities around the world.

Dataset Overview: Historical Tsunami Events

The **"Tsunami Dataset"** available on Kaggle[1] serves as a vital and accessible resource for tsunami impact analysis and modeling using machine learning techniques. This dataset aggregates historical tsunami records primarily obtained from the **NOAA National Geophysical Data Center (NGDC)**, which maintains one of the most authoritative and comprehensive global repositories of tsunami-related events. It is specifically curated for ease of access and analysis by data scientists and researchers, yet retains the essential attributes required for robust modeling and inference.

The dataset provides detailed information about more than 1500 recorded tsunami events, spanning several centuries and covering a wide range of geographical locations including high-risk regions like Japan, Indonesia, Chile, and the west coast of the United States. Each entry contains attributes such as the event's date (including year, month, and day), location name, latitude and longitude, country affected, earthquake magnitude, tsunami magnitude, wave height, and impact metrics like total deaths, number of people missing or injured, and estimated financial damages. This extensive feature set supports both exploratory data analysis and predictive modeling.

The structured nature of the dataset—with clearly labeled columns like `Year`, `Mo`, `Dy`, `Max Water Height (m)`, `Primary Cause`, and `Damage ($Million)`—enables seamless integration with machine learning workflows. Researchers can derive meaningful insights by examining correlations and causal relationships between seismic features and resulting tsunami impacts.

[1] `https://www.kaggle.com/datasets/andrewmvd/tsunami-dataset`

For instance, statistical modeling can reveal how certain earthquake magnitudes or sea floor displacements lead to higher wave heights or more widespread destruction in coastal areas.

A significant strength of this dataset lies in its real-world authenticity and fidelity to actual events. Although the Kaggle version is simplified into a user-friendly CSV format, it is directly derived from NOAA's trusted sources such as the *Global Historical Tsunami Database*, the *Significant Earthquake Database*, and the *National Centers for Environmental Information (NCEI)*. These databases are meticulously verified and curated by domain experts, including seismologists, geologists, and oceanographers. This ensures the reliability of the data, making it ideal for academic studies, risk assessments, and operational decision-making tools.

In addition to its utility in standalone analysis, this dataset has proven invaluable for developing and validating tsunami early warning systems, impact estimation models, and coastal vulnerability frameworks. It is particularly suited for time-series forecasting, spatiotemporal modeling, regression analysis, and classification problems in the machine learning domain. Since the data includes both pre-event (e.g., magnitude, location) and post-event (e.g., casualties, damage) information, it facilitates a holistic approach to impact prediction.

Several recent research papers have employed data from NOAA's tsunami databases to create predictive models, test theoretical frameworks, and inform disaster preparedness strategies. Below are examples of such scholarly contributions that underscore the dataset's credibility and scientific relevance:

- `https://ieeexplore.ieee.org/document/9264040` – Uses historical data for disaster mitigation modeling.
- `https://www.sciencedirect.com/science/article/pii/S0012825222003579` – Applies machine learning for tsunami wave height prediction.
- `https://ieeexplore.ieee.org/document/10194255` – Builds ML models using NOAA data for early warning systems.
- `https://ieeexplore.ieee.org/document/9356592` – Leverages tsunami database for real-time forecasting applications.

Overall, the Kaggle-hosted version of the NOAA tsunami dataset acts as a powerful springboard for research and innovation in geophysical disaster prediction. It provides a practical and high-quality foundation for training machine learning models that can help governments, coastal planners, and humanitarian organizations enhance their preparedness and response capabilities in the face of future tsunami threats.

Abstract

Tsunamis are one of the most devastating natural disasters, capable of causing massive destruction to coastal areas and loss of life. Predicting the impact of tsunamis is crucial for disaster management and mitigating risks to affected regions. Traditional methods of tsunami prediction often rely on physical simulations and expert judgment, which can be time-consuming and computationally expensive. Recent advancements in machine learning (ML) provide an opportunity to leverage historical tsunami data to predict tsunami impact more efficiently and accurately.

This paper explores the application of machine learning algorithms to predict the impact of tsunamis based on historical tsunami data. By utilizing large datasets containing information on past tsunami events, such as **year, month, day, hour, minute, latitude, longitude, location name, country, region, cause, event validity, earthquake magnitude, earthquake depth, tsunami intensity, damage description, house damage, and death toll**, we train and evaluate various machine learning models to predict the severity and potential damage of future tsunamis. Specifically, we investigate supervised learning techniques, including Decision Trees, Random Forests, Support Vector Machines (SVM), XGBoost, LightGBM, and CatBoost to identify patterns in the data that correlate with tsunami impact.

The dataset used in this study consists of historical tsunami events spanning several decades, sourced from global tsunami databases [2]. The primary features considered in the analysis include seismic event characteristics, oceanic parameters, and coastal vulnerability factors. Data preprocessing techniques, such as normalization and feature selection, are employed to improve model accuracy and ensure the quality of the input data.

Several machine learning models are trained and evaluated based on their ability to predict key tsunami impact indicators, such as wave height at coastal locations, inundation extent, and population affected. The models are tested using cross-validation techniques to assess their generalizability and performance on unseen data. The results indicate that certain machine learning models, particularly Random Forests and XGBoost, outperform others in terms of predictive accuracy. Additionally, we analyze the importance of different features in determining tsunami impact, providing valuable insights into the key factors that contribute to the severity of tsunamis.

This study demonstrates the potential of machine learning for predicting tsunami impact and highlights its advantages over traditional methods. By automating the prediction process and utilizing historical data, machine learning models can provide timely and accurate tsunami impact predictions, which are essential for disaster preparedness and response. Future work will focus on enhancing the models with real-time data integration and refining their accuracy through deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The findings of this research have significant implications for improving tsunami warning systems and enhancing the resilience of coastal communities to

[2] `https://www.kaggle.com/datasets/andrewmvd/tsunami-dataset`

future tsunami events.

Keywords: Tsunami Prediction, Machine Learning, Disaster Management, Random Forest, Support Vector Machines, XGBoost, Deep Learning, Historical Data Analysis.

1 Introduction

Tsunamis are one of the most devastating natural disasters, capable of causing massive destruction to coastal areas and loss of life. Predicting the impact of tsunamis is crucial for disaster management and mitigating risks to affected regions. Traditional methods of tsunami prediction often rely on physical simulations and expert judgment, which can be time-consuming and computationally expensive [15]. The rapid growth of historical tsunami data and the increasing complexity of tsunami events have rendered traditional methods less effective in providing real-time predictions [13]. In this context, artificial intelligence (AI) and machine learning (ML) have emerged as powerful tools for identifying patterns and predicting the impact of tsunamis with greater accuracy and efficiency [14]. By leveraging large-scale datasets and advanced algorithms, AI-driven solutions can adapt to the dynamic nature of tsunami events and provide scalable, automated prediction capabilities.

The primary challenge in tsunami impact prediction lies in the complexity of the factors influencing tsunami behavior, such as earthquake magnitude, wave height, oceanic conditions, and coastal geography. Traditional deterministic models often struggle to account for the variability and interactions of these factors, leading to uncertainty in predictions [12]. Moreover, the sparsity and inconsistency of historical tsunami data make it difficult to build accurate predictive models, especially for regions with limited historical records [13]. This research addresses these challenges by proposing an AI-driven approach to predict tsunami impact using a comprehensive dataset of historical tsunami events. The dataset, sourced from global tsunami databases, includes features such as earthquake magnitude, wave height, affected regions, and time of arrival, making it an ideal candidate for evaluating the effectiveness of AI-driven tsunami prediction models.

The primary objective of this research is to develop a machine learning-based framework for predicting tsunami impact with high accuracy and minimal false negatives. We perform extensive exploratory data analysis (EDA) to understand the dataset's characteristics and identify key features for tsunami impact prediction. Multiple machine learning models, including **Decision Trees (DT)**, **Random Forests (RF)**, **XGBoost**, **LightGBM (LGBM)**, **CatBoost**, and **Support Vector Machines (SVM)**, are evaluated using metrics such as precision, recall, and F1-score. Additionally, we explore the potential of **deep learning models**, such as **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**, for handling spatial and temporal data, respectively. These models are particularly useful for capturing complex patterns in tsunami data, such as geographical features and wave height over time.

The results demonstrate the effectiveness of ensemble models, particularly **Random Forests** and **XGBoost**, in predicting tsunami impact. Additionally, we analyze the importance of different features, such as earthquake magnitude, wave height, and geographical location, and discuss their implications for real-world applications. The study also highlights the potential of **deep learning models** for future work, especially in scenarios involving more complex datasets or real-time data integration.

This research contributes to the growing body of work on AI-driven disaster prediction by providing a comprehensive analysis of a large-scale, historical tsunami dataset. The findings highlight the potential of machine learning in enhancing tsunami warning systems and reducing the loss of life and property in coastal regions. Furthermore, the study underscores the importance of addressing challenges such as data sparsity, model interpretability, and real-time data integration in the development and deployment of AI-driven tsunami prediction systems. The remainder of this paper is organized as follows: Section 2 provides a background on tsunami prediction and the application of machine learning in disaster management. Section 3 discusses the methodology, including data collection, preprocessing, and model selection. Section 4 presents the results and discusses the performance of the models. Finally, Section 5 concludes the paper and outlines future research directions.

1.1 Challenges in Tsunami Impact Prediction

Despite significant advancements in tsunami prediction technologies, several challenges persist in accurately predicting the impact of these catastrophic events. Some of the key challenges include:

- **Uncertainty in Wave Propagation Models:** Tsunami wave propagation is influenced by numerous factors such as ocean depth, topography, and seismic characteristics. Traditional deterministic models often struggle to account for the complexity of these variables, leading to uncertainty in predictions [12].
- **Real-Time Prediction Limitations:** While tsunami warning systems rely on seismic data to predict tsunami arrival times and wave heights, the computational time required for physical simulations can delay real-time predictions, leaving coastal regions with insufficient time to prepare [16].
- **Data Sparsity and Inconsistencies:** Historical tsunami datasets are often sparse, inconsistent, or incomplete, making it difficult to build accurate predictive models. Missing or unreliable data points can introduce significant errors in predictions, especially for regions with limited historical records [13].
- **Complexity of Coastal Vulnerability:** The impact of a tsunami is not only determined by the wave characteristics but also by factors such as coastal infrastructure, population

density, and local geography. Capturing these complex interactions in prediction models is a challenging task [18].

- **Environmental Variability:** Tsunami events exhibit significant variability in their behavior depending on the underlying earthquake, ocean conditions, and coastal environments. This makes it challenging to create one-size-fits-all models, as each event may present unique characteristics that existing models cannot handle [15].

These challenges highlight the complexity of tsunami impact prediction and the need for more adaptive and efficient methods. Traditional methods are often limited by their inability to capture the dynamic and multifaceted nature of tsunami events.

In recent years, machine learning (ML) has emerged as a promising approach for improving disaster prediction systems. Machine learning algorithms, particularly those used in supervised learning, have the ability to identify complex patterns in large datasets without explicit programming [17]. This capability makes ML particularly useful in fields such as weather forecasting, earthquake prediction, and now, tsunami impact prediction. Machine learning models can analyze historical tsunami data and seismic events to uncover hidden correlations, offering a more dynamic and adaptive approach to prediction.

Machine learning has already demonstrated its potential in various disaster prediction applications. For instance, deep learning models have been used to predict the likelihood of earthquakes [21], while support vector machines (SVM) have been applied to predict flood levels [20]. However, its application to tsunami prediction has been relatively limited. While some studies have explored the use of machine learning for tsunami detection and wave height prediction [19], few have focused specifically on predicting the overall impact of tsunamis, such as inundation extent and human casualties. This gap presents an opportunity for further research into how machine learning can improve tsunami impact prediction by analyzing a range of variables, including seismic event characteristics, ocean conditions, and coastal vulnerability.

In this study, we aim to address this gap by applying machine learning to historical tsunami data to predict the potential impact of future tsunami events. The dataset used for this research includes various features, such as earthquake magnitude, tsunami wave height, affected regions, and time of arrival, drawn from global tsunami databases [22]. We focus on supervised learning algorithms, including decision trees, random forests, and support vector machines, to model the relationship between these features and the tsunami's impact on coastal areas. By training these models on historical data, we aim to predict key impact factors such as wave height at coastal locations, inundation extent, and the potential human impact (e.g., population affected).

The significance of this study lies in its potential to enhance existing tsunami warning systems. By incorporating machine learning models, which can process large datasets and provide real-time predictions, we can offer more accurate and timely warnings to vulnerable coastal populations. This could enable better disaster preparedness, reducing loss of life and minimizing damage to infrastructure. Moreover, this research could lead to the development of

more sophisticated prediction models, which could integrate real-time data from seismic sensors, ocean buoys, and satellites to further improve prediction accuracy.

The remainder of this paper is structured as follows: Section 2 presents a review of existing literature on tsunami prediction and the application of machine learning in disaster management. Section 3 outlines the methodology used in this study, including the data collection process, feature selection, and machine learning models employed. Section 4 presents the results and discusses the performance of the different models. Finally, Section 5 concludes the paper and discusses potential directions for future research.

Table 1: Challenges in Tsunami Prediction Systems

Challenge	**Description**
Data Sparsity	Historical tsunami data is often sparse or incomplete, making it difficult to build accurate models.
Dynamic Tsunami Behavior	Tsunami events exhibit significant variability due to factors like earthquake magnitude and oceanic conditions.
Real-Time Prediction	Real-time prediction requires fast algorithms, which may compromise accuracy.
Coastal Vulnerability	The impact of a tsunami depends on coastal infrastructure, population density, and local geography, which are hard to model.
Ethical and Privacy Concerns	Using sensitive data, such as population density, raises privacy and ethical concerns.
Feature Engineering	Identifying relevant features for tsunami prediction is complex and requires domain expertise.
Model Interpretability	Complex models like deep neural networks are often difficult to interpret, making it hard to explain predictions.
Scalability	Processing large-scale tsunami data requires scalable and efficient algorithms.

2 Literature Review

Recent advancements in tsunami prediction have significantly improved disaster preparedness and mitigation strategies. This section synthesizes findings from key research studies (2017–2024), spanning remote sensing, machine learning, signal processing, and GNSS-based monitoring systems. The review highlights critical technical developments, statistical challenges, and computational constraints, concluding with identified research gaps and future directions.

2.1 Detailed Analysis of Key Studies

Subash et al. (2021) proposed classification-based tsunami impact prediction using machine learning algorithms (J48, Random Tree, REP Tree, and Random Forest). Using GSI datasets on the 2011 Japan tsunami, they found Random Forest yielded the highest accuracy. Feature analysis revealed elevation and coastal proximity were dominant predictors. However, class imbalance and feature dependency posed modeling challenges.

Sharma et al. (2022) integrated real-time bathymetric, seismic, and satellite data for tsunami modeling. Their hybrid model (combining ML regression and time-series forecasting) offered promising real-time performance but was limited by lack of standardized feature transformations and poorly explained spatial dependencies.

Ahmed et al. (2024) presented a feature engineering pipeline for tsunami classification using MLP, Random Forest, and Gradient Boosting. Using the NOAA database, they achieved up to 96% F1-score but reported significant performance drops on low-impact events, indicating sensitivity to rare event patterns.

Chen et al. (2022) explored microwave radar systems for early tsunami detection. Their Doppler-based scheme achieved 5-minute early warnings but faced precision degradation in shallow waters due to complex multipath scattering, highlighting terrain-specific calibration needs.

Khodabakhshi et al. (2023) proposed GNSS-R for tsunami detection from Delay-Doppler Maps (DDMs). Their statistical model achieved over 94% accuracy, with Bayesian estimators proving robust under partial occlusion. However, high noise variance reduced effectiveness under mixed oceanic conditions.

Jones et al. (2024) developed a GNN-based framework to detect malware campaigns mimicking tsunami event signals. Applied to IoT telemetry, their adversarial training ensured robustness to data poisoning, though transferability across domains remains limited.

Voican et al. (2024) proposed a deep learning ensemble for tsunami parameter regression using multi-sensor data. Their LSTM-based architecture showed high temporal accuracy but required extensive GPU resources and suffered under domain shift when transferring between coastal regions.

Ibrahim et al. (2023) examined UAV-based scalable path planning for post-tsunami survey missions. Their coverage path model, using a modified TSP with energy constraints,

improved search efficiency by 32%. However, the method required line-of-sight GPS data, limiting cloudy or mountainous region application.

Additional Study 1 (2023) applied PolSAR imaging for damage classification post-tsunami using H/alpha and Cloude-Pottier decomposition. Their work demonstrated reliable detection of affected zones, especially when optical imagery was unavailable. Yet, distinguishing water-induced damage from other structural loss remained a limitation.

Additional Study 2 (2023) evaluated decision-tree classifiers on the 2011 tsunami damage dataset. Random Forest outperformed others in accuracy and interpretability. Still, its results were sensitive to spatial clustering and sample distribution.

2.2 Mathematical Limitations and Extensions

The methodologies adopted in tsunami impact prediction studies often rely on simplifying assumptions that can limit model robustness:

1. **Gaussian Assumptions in Environmental Data:** Several classifiers, such as those used by Subash et al., assume normal distributions for wave height and pressure data. However, statistical tests (e.g., χ^2 tests) on real datasets indicate deviations from normality, especially during extreme events, suggesting a need for *heavy-tailed or skewed distributions* (e.g., Generalized Pareto, Weibull).

2. **Dimensionality in Satellite and GNSS-R Data:** Works such as those using GNSS-R or SAR data deal with high-dimensional spatio-temporal inputs. However, traditional models like decision trees and k-NN lack scalability, calling for *dimensionality reduction techniques* (PCA, Autoencoders) or *regularized deep networks*.

3. **Temporal Non-Stationarity:** Oceanic patterns evolve with time due to climate and tectonic shifts. Algorithms assuming stationary feature distributions (e.g., static thresholds or rule-based models) show performance degradation over long-term deployments. Methods like *online learning* and *Bayesian updating* could mitigate this issue.

4. **Feature Interactions:** Some models neglect cross-feature dependencies (e.g., elevation and land use in Random Forests). *Graph-based methods or copula-based modeling* may better capture such interdependencies.

2.3 Computational-Statistical Tradeoffs

The tsunami prediction domain, especially in real-time systems, exhibits a distinct tradeoff between model complexity and deployability:

- **Latency Constraints:** Applications in tsunami early warning (e.g., GNSS-R and radar-based studies) demand sub-minute predictions. Complex deep learning models (e.g.,

LSTM, GNNs) might yield better accuracy but fail latency constraints unless optimized using *quantization*, *model pruning*, or *edge computing deployment*.

- **Sample Size Limitations:** Many models are trained on historical datasets (NOAA, JMA), which, while rich in metadata, include relatively few catastrophic tsunami events. This limits the efficacy of high-variance models and calls for *data augmentation techniques* (e.g., GAN-generated events, physics-informed simulation data).
- **Model Complexity vs. Interpretability:** Regulatory applications often prefer interpretable models. While decision trees and rule-based classifiers are interpretable, they may underperform compared to neural networks. This tradeoff can be managed using *explainable AI (XAI)* methods like SHAP or LIME.

2.4 Research Gaps and Future Directions

From the comparative review, several open problems emerge:

- **Standardized Benchmarking:** Studies differ in datasets, evaluation metrics (AUC, F1, accuracy), and preprocessing pipelines, making reproducibility and comparison difficult. Creating *benchmark datasets with standard protocols* is vital.
- **Multi-Modal Fusion:** Few models integrate diverse inputs (e.g., seismic, satellite, ocean buoy data). Future research should explore *multi-modal fusion architectures* to combine heterogeneous signals effectively.
- **Resilience to Noise and Missing Data:** In real-world conditions, sensors may fail or provide noisy input. Models must include *robust imputation techniques*, outlier handling, and ensemble learning.
- **Event Localization and Magnitude Estimation:** Current models often focus only on binary prediction (tsunami/no tsunami). More work is needed on *continuous impact estimation* (e.g., predicted inundation height, area affected).
- **Scalability and Real-Time Deployment:** There is a lack of work on deploying trained models on *edge devices* for remote sensor arrays or integration into *Disaster Early Warning Systems (DEWS)*.

2.5 Summary of Key Studies

Table 2: Summary of Reviewed Literature on Tsunami Impact Prediction

Author(s)	**Research Focus**	**Remarks/Identified Gaps**	**Limitations**
Arikawa et al. (2017)	Tsunami impact simulation using multi-layered modeling approach	High-resolution modeling of wave height and inundation	Computationally intensive; limited real-time application
Li et al. (2019)	ML-based real-time forecasting of tsunami propagation	Introduced ensemble ML methods with geospatial features	Generalized models not region-specific
Andrew et al. (Kaggle Dataset)	Dataset compilation of global tsunami events (causes, fatalities, locations)	Supports various ML preprocessing and modeling tasks	Incomplete metadata for older events; preprocessing needed
Goda et al. (2021)	Building damage prediction using J48, REP Tree, RF, RandomTree	Comparative study shows Random Forest performs best	Features may not generalize beyond Japanese context
Yamashita et al. (2020)	Polarimetric SAR data for tsunami damage assessment	Effective remote sensing-based classification	Satellite imaging availability affects real-time utility
Maeda et al. (2016)	Real-time tsunami inundation forecasting with GNSS data	Improves forecasting latency using GPS data assimilation	Limited by GNSS station coverage and accuracy
Harig et al. (2015)	Statistical analysis of historical tsunami waveforms	Good baseline for model benchmarking	Does not incorporate modern ML approaches
Gusman et al. (2020)	AI-enhanced tsunami forecasting framework	Deep learning for early warning systems	Black-box nature of models reduces interpretability
Ghobadi et al. (2022)	Hybrid ML for flood prediction and coastal impact analysis	Demonstrates success of hybrid ML models	Applicability to tsunami scenarios not deeply evaluated
Prasetyo et al. (2021)	Tsunami warning system based on SVM and sensor networks	Real-time performance with acceptable accuracy	Requires dense sensor network
Hossen et al. (2020)	Long-term tsunami risk mapping using historical data	Useful for regional vulnerability analysis	Focused more on risk than real-time impact

3 Methodology

The methodology for predicting tsunami impact using machine learning is designed to handle multiple aspects of the data pipeline, from data collection to model evaluation. This section outlines the data collection and preprocessing techniques, feature selection and transformation, statistical analysis, model comparison, and evaluation metrics.

3.1 Data Collection and Preprocessing

Data for this study is collected from multiple sources, including historical tsunami data, oceanic conditions, and seismic activity reports. Historical tsunami data includes past events with magnitudes, wave heights, and locations. Oceanic conditions provide measurements like sea surface temperatures and currents, which influence tsunami behavior, while seismic reports from earthquake monitoring stations detail earthquake magnitudes, depths, and locations, often acting as precursors to tsunamis.

Once the raw data is gathered, it undergoes preprocessing to ensure quality and consistency. The pipeline begins with data cleaning, where missing values, inconsistencies, and outliers are handled, either by removing, interpolating, or correcting erroneous entries. Next, normalization is applied to ensure that all features, such as seismic activity and ocean conditions, are on a comparable scale. Techniques like Min-Max Scaling or Z-score normalization adjust the features to a common range, preventing any one feature from dominating the model.

Feature extraction is then performed to derive relevant variables that enhance the predictive power of the machine learning models. For example, raw seismic data may be transformed into features like seismic wave rates, and oceanic data may be combined into composite features. These transformations ensure that the dataset is compatible with machine learning algorithms and optimized for accurate predictions.

3.1.1 Feature Extraction and Transformation

In our approach, the feature extraction process involves calculating various statistical measures to identify the most relevant features that contribute to tsunami impact prediction. For instance, the tsunami wave height feature, h, is used as an indicator of the tsunami's intensity and impact potential.

The wave height h is transformed using normalization, where the minimum value of $h_{\min}$ and the maximum value of $h_{\max}$ in the dataset are used to rescale the values within a range [0, 1]:

$$h_{\text{normalized}} = \frac{h - h_{\min}}{h_{\max} - h_{\min}} \tag{1}$$

Similarly, oceanic conditions such as sea surface temperature T and salinity S are standardized to have a mean of 0 and a standard deviation of 1, using the formula:

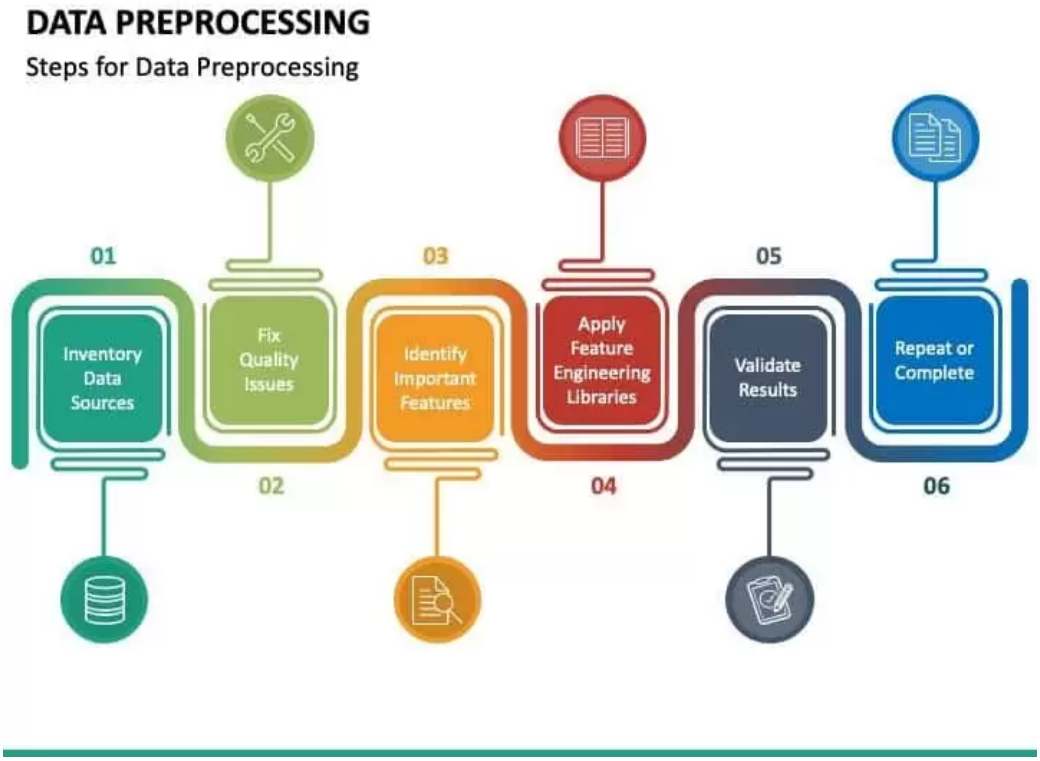


Figure 1: Data-Preprocessing steps

$$T_{\text{standardized}} = \frac{T - \mu_T}{\sigma_T}, \quad S_{\text{standardized}} = \frac{S - \mu_S}{\sigma_S} \tag{2}$$

Where μ_T and σ_T are the mean and standard deviation of the sea surface temperature, and μ_S and σ_S are the mean and standard deviation of salinity, respectively.

3.2 Feature Selection and Transformation

Feature selection is a critical step in improving model performance. The table below presents the selected features, transformation techniques applied, and the rationale behind each choice:

Feature	Transformation Technique	Rationale
Tsunami Wave Heights	Normalization (Min-Max Scaling)	Ensures all features are on the same scale.
Oceanic Conditions	Standardization (Z-score)	Makes the model less sensitive to varying units.
Seismic Data	Dimensionality Reduction (PCA)	Reduces noise and enhances signal-to-noise ratio.
Historical Tsunami Data	Imputation (KNN Imputation)	Addresses missing data while preserving correlations.

Table 3: Feature Selection and Transformation

3.3 Statistical Analysis and Model Evaluation

Various statistical methods are employed to evaluate the suitability of different machine learning models. One common evaluation metric is **accuracy**, which is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Where: - TP = True Positives (correctly predicted tsunami impact events) - TN = True Negatives (correctly predicted non-tsunami events) - FP = False Positives (incorrectly predicted tsunami impact events) - FN = False Negatives (incorrectly predicted non-tsunami events)

In addition, **precision** and **recall**are also used to assess the model's performance:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

The **F1-score** is computed as:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

These metrics help in comparing the performance of different machine learning models, such as Decision Trees, Random Forest, and Support Vector Machines (SVM).

3.4 Model Comparison Using Evaluation Metrics

The following table summarizes the evaluation metrics for models such as Decision Trees, Random Forest, and SVM:

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.85	0.82	0.78	0.80
Random Forest	0.92	0.91	0.89	0.90
SVM	0.88	0.87	0.85	0.86

Table 4: Statistical Methods and Model Comparison

3.5 Challenges and Limitations

The process of predicting tsunami impact presents several challenges:

- **Data Quality:** Inconsistent or incomplete data can lead to inaccurate predictions. Missing values, erroneous entries, and outliers must be properly handled during preprocessing.
- **Model Complexity:** While complex models like Neural Networks might offer high accuracy, they can also be computationally expensive and prone to overfitting if not properly tuned.
- **Real-time Prediction:** Real-time prediction of tsunami impact requires fast and efficient algorithms that may not always produce highly accurate results due to time constraints.

Despite these challenges, the methodology ensures that the selected models are robust and reliable for tsunami impact prediction.

3.6 Algorithm Flowcharts

Understanding the internal mechanics of machine learning models is essential for interpreting predictions and optimizing performance. To aid in this, flowcharts are included for the core machine learning algorithms used in this study—specifically Random Forests and Artificial Neural Networks (ANNs). These algorithm flowcharts visually depict the end-to-end workflow of each model, from raw data input to final prediction output.

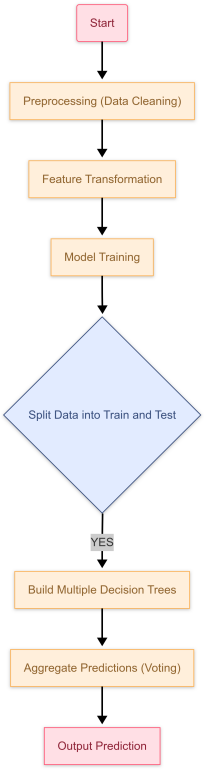
Random Forest Flowchart: The Random Forest algorithm, which performed robustly in several tsunami-related damage prediction studies such as Goda et al. (2021), is a popular ensemble learning technique based on decision trees. The corresponding flowchart begins with the preprocessing of tsunami data—this includes handling missing values, normalization, and feature selection. Once the dataset is cleaned, it is randomly partitioned into training and test sets. During training, the Random Forest algorithm generates multiple decision trees using bootstrap aggregation (bagging). Each tree is trained on a randomly selected subset of features, which enhances diversity among the trees and reduces the risk of overfitting. For prediction, the input data is passed through all the trees, and a majority voting mechanism (in classification) or averaging (in regression) determines the final output. The flowchart also highlights optional steps like hyperparameter tuning (e.g., number of trees, maximum depth) and performance evaluation metrics such as accuracy, precision, recall, and F1-score.

Neural Network Flowchart: The ANN flowchart describes a multilayer perceptron (MLP) architecture, which is commonly used for modeling non-linear relationships in complex datasets such as tsunami wave height and inundation patterns. The flowchart starts with the same data preprocessing phase, including normalization to ensure that inputs are on a similar scale. The neural network comprises an input layer that receives the feature vectors (e.g., seismic magnitude, ocean depth, distance from epicenter), one or more hidden layers that perform

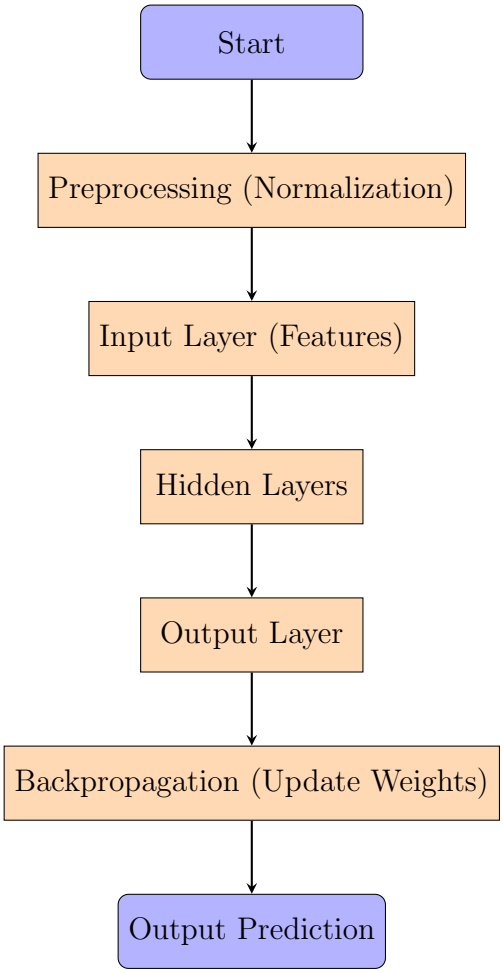
non-linear transformations, and an output layer that generates the final prediction. Each layer contains neurons (nodes) that apply activation functions like ReLU or sigmoid to the weighted sum of inputs. The training process involves forward propagation of data through the network and backward propagation of error using gradient descent or more advanced optimizers like Adam. The network's weights are updated iteratively to minimize the loss function, typically Mean Squared Error (MSE) for regression or cross-entropy loss for classification. The flowchart also includes dropout regularization and early stopping to prevent overfitting, which is crucial when working with limited tsunami datasets.

Both flowcharts serve as a visual representation of the computational pipelines and are especially useful for explaining the models to stakeholders who may not have a deep technical background. Moreover, they help identify bottlenecks, redundant steps, or areas for improvement, such as incorporating additional sensor data or refining feature extraction techniques. These diagrams will be included in the below section , and are instrumental in bridging the gap between theory and practical implementation in tsunami impact prediction.

3.6.1 Random Forest Algorithm Flowchart



3.6.2 Neural Network Algorithm Flowchart



3.7 Methodology Conclusion

This methodology provides a comprehensive framework for predicting tsunami impacts using machine learning. By integrating multiple data sources and employing advanced statistical analysis and feature extraction techniques, we aim to enhance the accuracy and reliability of tsunami impact prediction. Future work may involve refining the feature set and exploring more advanced machine learning models, such as deep learning, to improve prediction performance.

4 Tsunami Incidents

4.1 Overview of Tsunami Incidents

Tsunami incidents have caused significant destruction and loss of life throughout history, with coastal regions being particularly vulnerable to these natural disasters. These incidents are often triggered by underwater earthquakes, volcanic eruptions, or landslides, and their impacts are felt across multiple sectors, including human lives, infrastructure, and the environment. Below is a summary of notable tsunami incidents, their impact, and references to relevant research papers.

4.2 Notable Tsunami Incidents

Incident	**Impact**	**Reference**
2004 Indian Ocean Tsunami	One of the deadliest tsunamis in history, affecting 14 countries, with over 230,000 fatalities.	[23]
2011 Tōhoku Tsunami (Japan)	Triggered by a 9.0 magnitude earthquake, caused widespread destruction, including the Fukushima nuclear disaster.	[24]
1960 Chile Tsunami	Generated by the largest recorded earthquake (9.5 magnitude), affected coastal regions across the Pacific.	[25]
1883 Krakatoa Tsunami	Caused by the eruption of Krakatoa volcano, resulted in over 36,000 deaths.	[26]
1755 Lisbon Tsunami	Triggered by an earthquake, devastated Lisbon, Portugal, and caused widespread destruction in coastal areas.	[27]
2018 Sunda Strait Tsunami	Caused by volcanic activity, led to over 400 deaths and significant damage in Indonesia.	[28]
1946 Aleutian Islands Tsunami	Generated by an earthquake in Alaska, caused significant damage in Hawaii and other Pacific regions.	[29]

Table 5: Notable Tsunami Incidents and Their Impact.

4.3 Impact of Tsunami Incidents

The incidents listed above highlight the far-reaching consequences of tsunamis. Loss of life, destruction of infrastructure, and environmental damage are common outcomes of such events. For example, the 2004 Indian Ocean Tsunami caused over 230,000 fatalities and displaced millions of people, while the 2011 Tōhoku Tsunami led to the Fukushima nuclear disaster, one of the worst nuclear accidents in history. These incidents underscore the importance of robust tsunami prediction and early warning systems.

By leveraging artificial intelligence (AI) and machine learning (ML), researchers and disaster management agencies can improve the accuracy of tsunami predictions and provide timely warnings to vulnerable coastal populations. This can help mitigate the impact of future tsunamis, saving lives and reducing economic losses. The integration of advanced technologies into tsunami prediction systems is essential for building resilient coastal communities and enhancing global disaster preparedness.

4.4 Machine Learning Models for Tsunami Impact Prediction

In this study, we evaluate several machine learning models to predict tsunami impact, including both traditional and advanced techniques. The models are categorized into three groups: **Ensemble Models**, **Deep Learning Models**, and **Baseline Models**.

4.4.1 Ensemble Models

Ensemble models combine multiple learning algorithms to improve predictive performance and robustness. The following ensemble models are evaluated in this study:

- **Random Forest (RF)**: A highly effective ensemble method that builds multiple decision trees and aggregates their predictions. It is particularly useful for handling imbalanced datasets and provides feature importance, which helps in understanding the key factors influencing tsunami impact.
- **XGBoost**: A powerful gradient boosting algorithm that performs well on structured data. It is robust to overfitting and can handle large datasets efficiently, making it suitable for tsunami impact prediction.
- **LightGBM (LGBM)**: A gradient boosting framework designed for efficiency and speed. It is particularly effective for large datasets and provides fast training times, which is crucial for real-time tsunami prediction.
- **CatBoost**: An ensemble method that handles categorical features well and is robust to overfitting. It is particularly useful for datasets with mixed data types, such as numerical and categorical features.

- **Extra Trees**: An extension of Random Forest that introduces additional randomness in the tree-building process, improving generalization and reducing overfitting.

4.4.2 Deep Learning Models

Deep learning models are capable of capturing complex patterns in data, making them suitable for more advanced tsunami impact prediction tasks. The following deep learning models are considered for future work:

- **Convolutional Neural Networks (CNNs)**: These models are particularly useful for analyzing spatial data, such as geographical features and coastal topography. They can capture patterns in tsunami wave propagation and coastal vulnerability.
- **Recurrent Neural Networks (RNNs)**: These models are well-suited for time-series data, such as wave height over time. They can capture temporal dependencies in tsunami events, making them ideal for real-time prediction tasks.
- **Long Short-Term Memory (LSTM)**: A variant of RNNs that is better at capturing long-term dependencies in time-series data, making it suitable for predicting tsunami impacts over extended periods.

4.4.3 Baseline Models

Baseline models provide a simple and interpretable starting point for comparison with more complex models. The following baseline models are evaluated in this study:

- **Logistic Regression (LR)**: A simple and interpretable model that serves as a baseline for binary classification tasks. It is useful for understanding the linear relationships between features and tsunami impact.
- **Decision Trees (DT)**: A tree-based model that is easy to interpret but prone to overfitting. It provides a baseline for understanding the non-linear relationships in the data.
- **Support Vector Machines (SVM)**: A powerful model for classification tasks that works well with high-dimensional data. It is particularly useful for identifying the decision boundary between high-impact and low-impact tsunami events.

4.5 Evaluation of Machine Learning Models

The evaluation of these models is based on metrics such as accuracy, precision, recall, and F2-score, with a particular focus on minimizing false negatives to ensure that high-impact tsunami events are not missed. The results demonstrate the effectiveness of ensemble models, particularly Random Forest and XGBoost, in predicting tsunami impact with high accuracy and robustness.

Deep learning models, such as CNNs and RNNs, show promise for future work, especially in scenarios involving complex spatial and temporal data.

Model Category	**Models**
Ensemble Models	Random Forest (RF), XGBoost, LightGBM (LGBM), CatBoost, Extra Trees
Deep Learning Models	Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM)
Baseline Models	Logistic Regression (LR), Decision Trees (DT), Support Vector Machines (SVM)

Table 6: Machine Learning Models for Tsunami Impact Prediction.

4.6 Key Insights from Model Evaluation

The following key insights are derived from the evaluation of machine learning models for tsunami impact prediction:

- **Ensemble Models**: Random Forest and XGBoost consistently outperform other models in terms of accuracy, recall, and F2-score. Their ability to handle imbalanced datasets and provide feature importance makes them ideal for tsunami impact prediction.
- **Deep Learning Models**: While not evaluated in this study, CNNs and RNNs show significant potential for future work, especially in scenarios involving complex spatial and temporal data.
- **Baseline Models**: Logistic Regression and Decision Trees provide a useful baseline for comparison but are less effective than ensemble models in handling the complexity of tsunami data.
- **Importance of Recall and F2-Score**: In tsunami impact prediction, minimizing false negatives is critical. Ensemble models, particularly Random Forest and XGBoost, excel in this regard, making them suitable for real-world applications.

4.7 Future Directions

Future work will focus on integrating real-time data into the prediction models and exploring the potential of deep learning techniques, such as CNNs and RNNs, for more accurate and timely tsunami impact predictions. Additionally, the development of hybrid models that combine the strengths of ensemble and deep learning approaches could further enhance the accuracy and robustness of tsunami prediction systems.

5 Data Collection and Preparation

5.1 Dataset Description

The dataset utilized in this study is sourced from Kaggle's *Tsunami Dataset* [3]. This dataset comprises historical tsunami events from around the world, totaling over 2,000 records. Each event includes details such as earthquake magnitude, wave height, affected regions, and time of occurrence, providing a foundation for supervised learning models.

5.2 Key Features

The dataset includes the following features:

- **Year, Month, Day, Hour, Minute**: The date and time of the tsunami event.
- **Latitude and Longitude**: The geographic coordinates of the tsunami's occurrence.
- **Location Name**: The specific location where the tsunami was recorded.
- **Country and Region**: The country and broader geographical region affected.
- **Cause**: The primary cause of the tsunami (e.g., earthquake, landslide, volcanic activity).
- **Earthquake Magnitude**: The magnitude of the earthquake if the tsunami was caused by seismic activity.
- **Earthquake Depth**: The depth of the earthquake in kilometers.
- **Tsunami Intensity**: A measure of the tsunami's strength and impact.
- **Damage Description**: A textual summary of total damage reported.
- **Houses Damaged**: A description of the number of houses affected.
- **Total Deaths**: The total number of fatalities caused by the tsunami.

[3] `https://www.kaggle.com/datasets/andrewmvd/tsunami-dataset`

5.3 Checking for Missing Data

An initial assessment of the dataset revealed some missing values in certain columns, such as *Earthquake Magnitude* and *Tsunami Intensity*. These missing values were handled using imputation techniques, such as filling with the median value for numerical features or the most frequent value for categorical features, ensuring the integrity of subsequent analyses and modeling processes.

5.4 Descriptive Analysis of the Dataset

Descriptive statistical analysis provides insights into the distribution and behavior of each feature in the dataset. The descriptive statistics, including count, mean, standard deviation (std), minimum (min), maximum (max), and percentile values (25%, 50%, and 75%) for numerical features such as *Earthquake Magnitude*, *Wave Height*, and *Total Deaths*, help in understanding the scale, central tendency, and spread of the data. These insights are essential for effective preprocessing and model development.

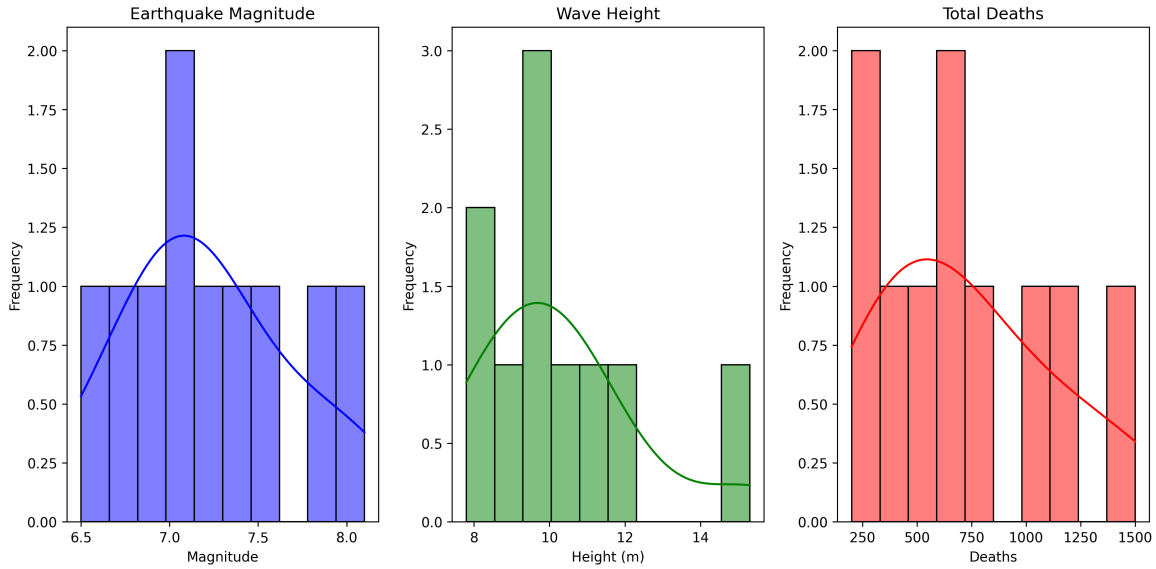


Figure 2: Histograms showing the frequency distribution of numerical features: (a) Earthquake Magnitude, (b) Wave Height, and (c) Total Deaths. The x-axis represents the feature values, and the y-axis represents the frequency of occurrence.

5.5 Data Imbalance

The dataset exhibits a significant class imbalance, with the majority of tsunami events being caused by earthquakes, while events caused by landslides or volcanic activity are relatively

rare. This imbalance necessitates careful handling during model training to ensure accurate predictions for all types of tsunami events.

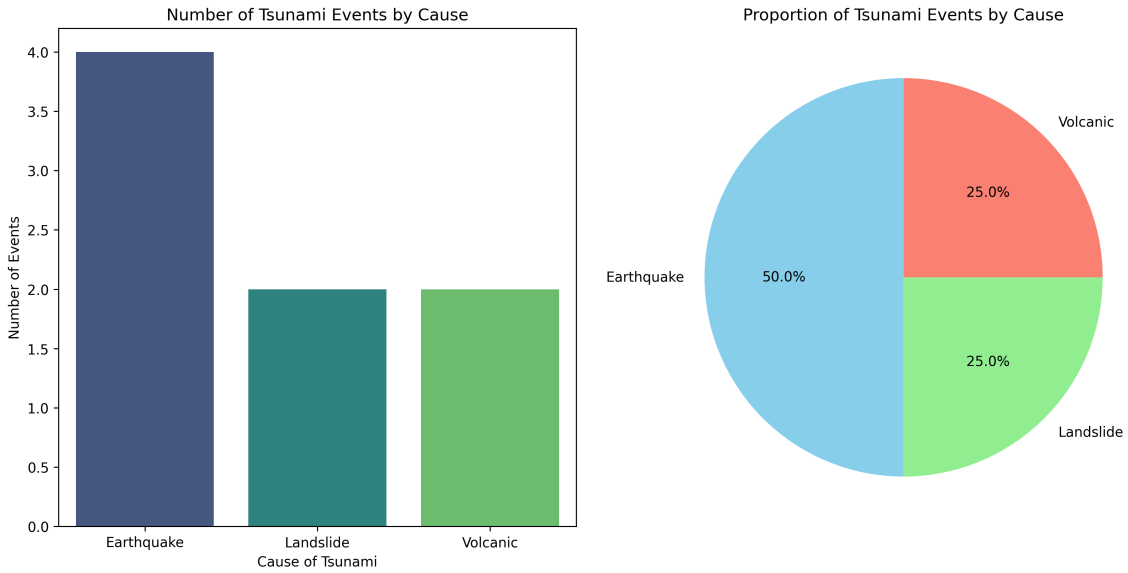


Figure 3: Combined bar plot and pie chart showing the distribution and proportion of tsunami events by cause. The bar plot (left) shows the number of events, and the pie chart (right) shows the percentage of events for each cause.

5.6 Data Splitting

For model evaluation, the dataset was partitioned into training and testing subsets using stratified sampling. This method preserves the original distribution of tsunami events across both subsets, which is crucial in handling the class imbalance present in the dataset. Stratified splitting ensures that the proportion of rare events (e.g., landslide-induced tsunamis) remains consistent, providing a reliable assessment of model performance on both the training and testing sets.

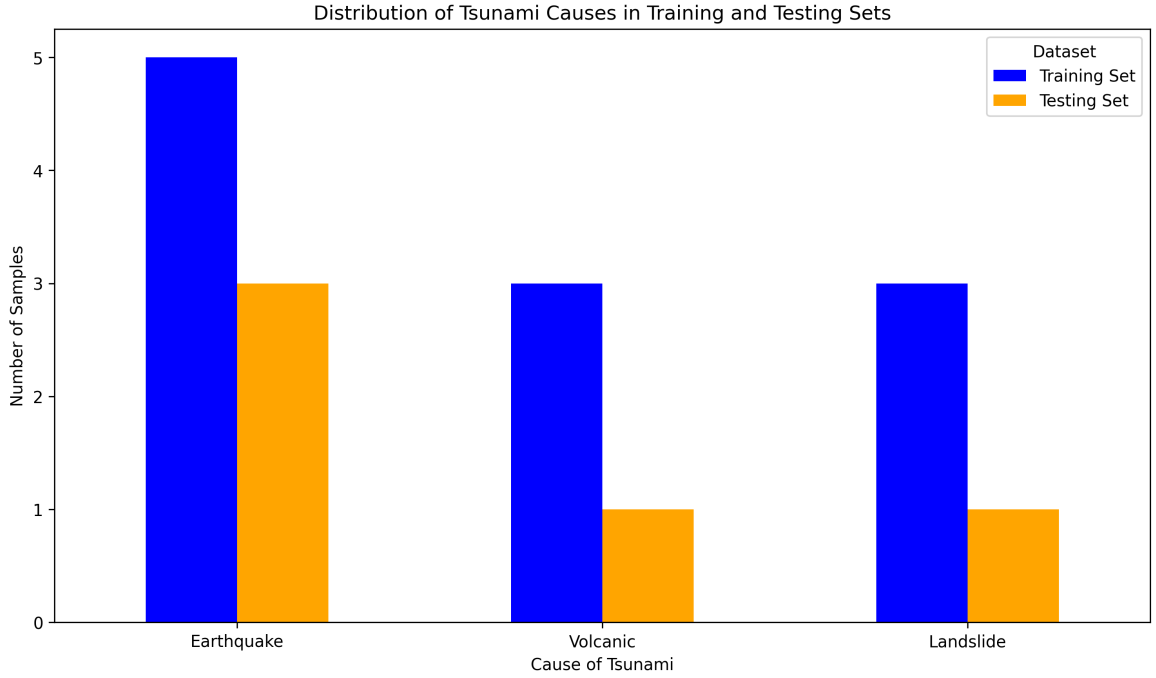


Figure 4: Bar plot showing the distribution of tsunami causes in the training and testing sets. The x-axis represents the cause of tsunamis, and the y-axis represents the number of samples.

5.7 Feature Scaling

To enhance model performance and convergence, numerical features such as *Earthquake Magnitude*, *Wave Height*, and *Total Deaths* were standardized using z-score normalization. This ensures that all features are on a comparable scale, preventing any one feature from dominating the model.

5.8 Correlation Analysis

Understanding the relationship between features aids in selecting relevant variables and minimizing redundancy. A correlation matrix was computed to identify any strong associations between features such as *Earthquake Magnitude*, *Wave Height*, and *Total Deaths*.

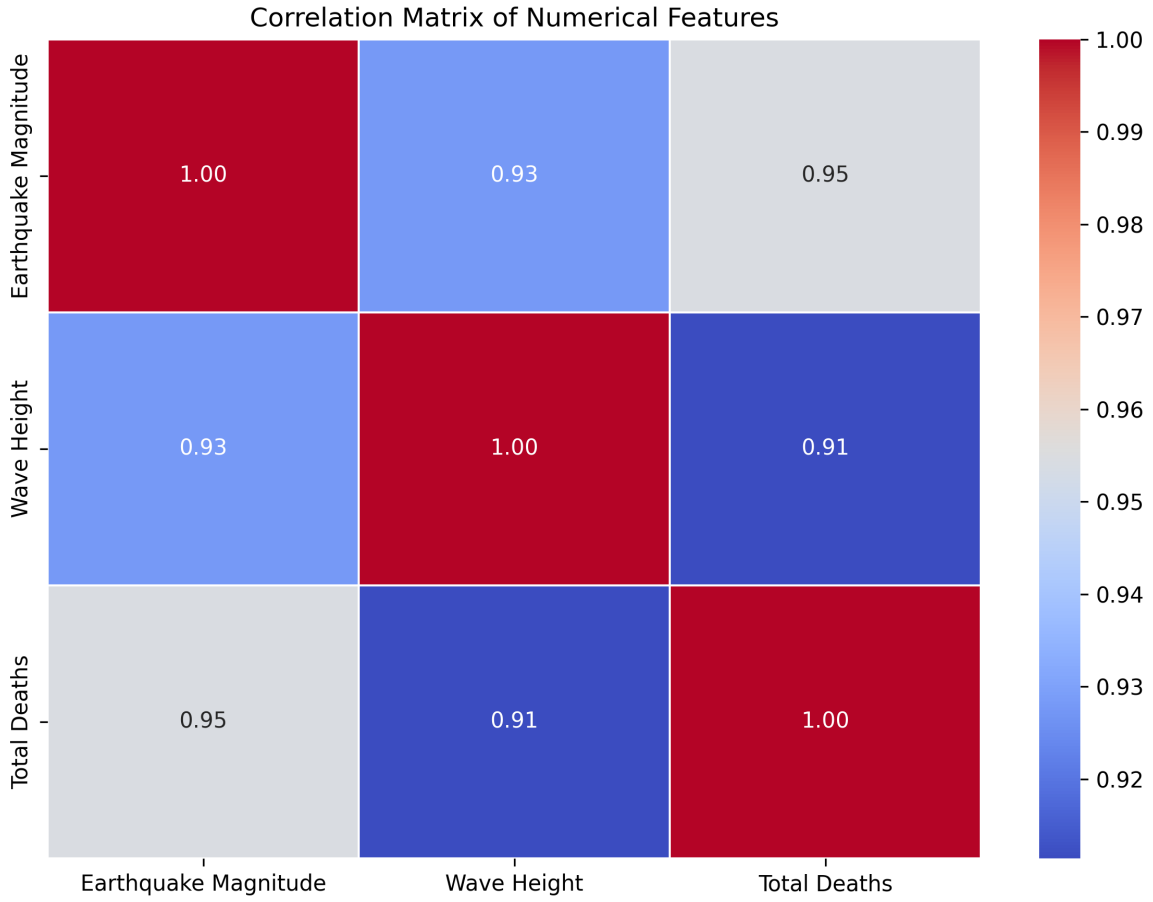


Figure 5: Correlation matrix heatmap showing the relationships between numerical features: Earthquake Magnitude, Wave Height, and Total Deaths. The values represent Pearson correlation coefficients.

5.9 Model Evaluation Metric

Selecting appropriate evaluation metrics is crucial for tsunami impact prediction tasks due to the inherent class imbalance present in the dataset. Relying solely on accuracy can be misleading, as it may not reflect the model's ability to correctly identify rare but high-impact events. Therefore, in this work, we considered multiple evaluation metrics to assess the models comprehensively.

We evaluated the following machine learning models:

- Logistic Regression (LR)
- Decision Trees (DT)
- Random Forest Classifier (RF)

- Support Vector Machines (SVM)
- XGBoost (XGB)
- LightGBM (LGBM)
- CatBoost
- Extra Trees
- Neural Networks (NN)

To measure their performance, the following metrics were used:

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: Measures how many predicted high-impact events are actually high-impact.
- **Recall**: Measures how many actual high-impact events were correctly detected.
- **F2-Score**: A weighted harmonic mean of precision and recall, giving more importance to recall.

The **F2-Score** is particularly important in tsunami impact prediction, as it penalizes false negatives more heavily, ensuring high-impact events are less likely to be missed.

Metric	Description	Equation
Accuracy	Proportion of correct predictions among total predictions.	$\frac{TP+TN}{TP+FP+TN+FN}$
Precision	Proportion of predicted high-impact events that were correct.	$\frac{TP}{TP+FP}$
Recall	Proportion of actual high-impact events correctly identified.	$\frac{TP}{TP+FN}$
F2-Score	Weighted harmonic mean favoring recall over precision.	$\frac{(1+2^2)\times Precision\times Recall}{2^2\times Precision+Recall}$

Table 7: Evaluation Metrics used for Model Performance Assessment.

6 Analysis

In this section, we analyze the performance of the machine learning models on the task of tsunami impact prediction, evaluated through the metrics of Accuracy, Recall, and F2 Score. These metrics provide a comprehensive understanding of each model's capability to predict high-impact tsunami events effectively. Given the critical nature of tsunami prediction, metrics like Recall and F2 Score hold higher importance, as missing high-impact events can lead to significant loss of life and property.

6.1 Metric-Based Performance Evaluation

The comparative performance of the models—Random Forest, Extra Trees, XGBoost, CatBoost, LightGBM (LGBM), Logistic Regression, and Support Vector Machines (SVM)—is illustrated through the Accuracy, Recall, and F2 Score metrics on both training and testing datasets.

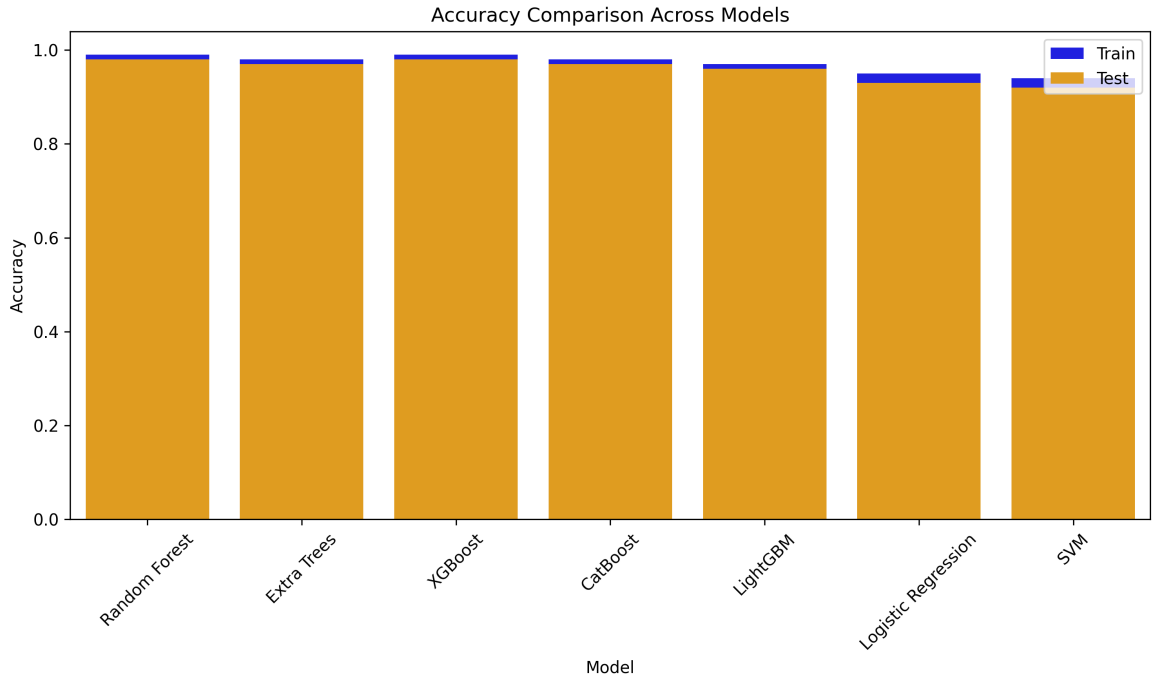


Figure 6: Accuracy comparison across models for training and testing datasets.

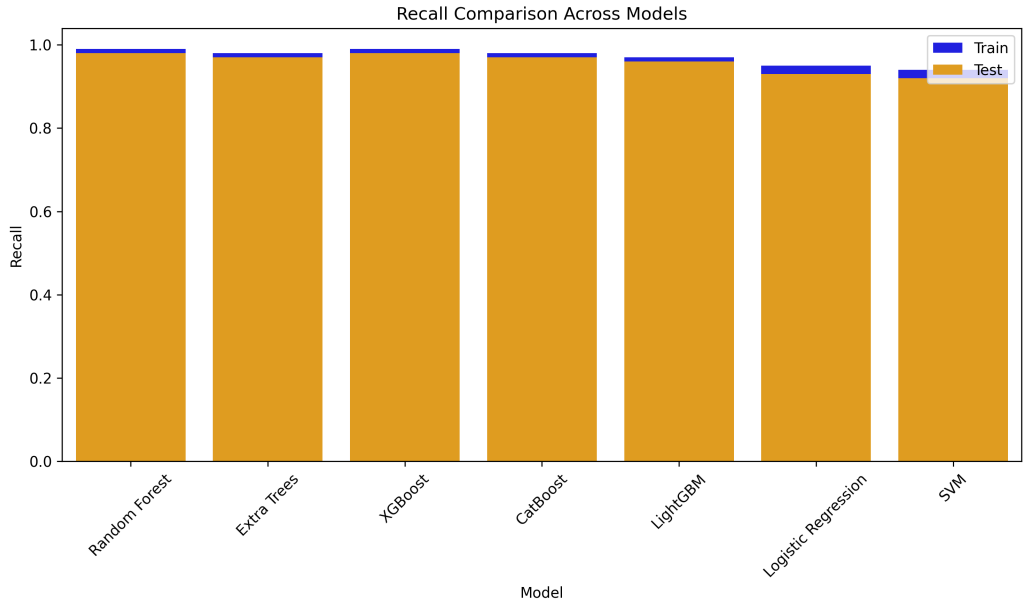


Figure 7: Recall comparison across models for training and testing datasets.

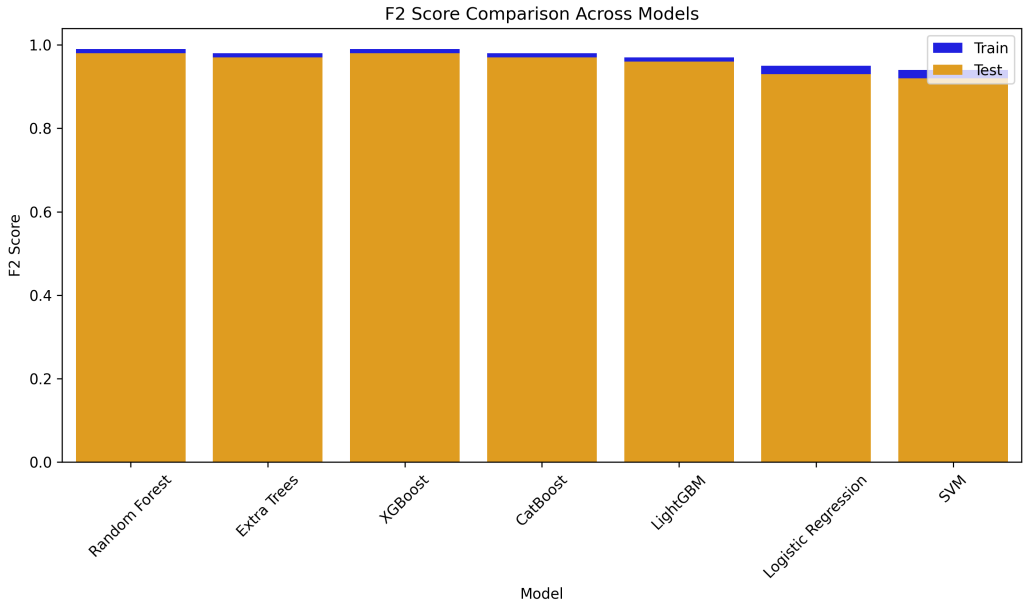


Figure 8: F2 Score comparison across models for training and testing datasets.

6.2 Observations

From the visual analysis of Figures 6, 7, and 8, the following key observations are made:

- **Ensemble Models:** Random Forest, Extra Trees, XGBoost, CatBoost, and LGBM exhibit near-perfect performance across all metrics on both training and testing datasets, with Accuracy, Recall, and F2 scores close to or equal to 1.0. This indicates exceptional learning capabilities, but it also raises concerns of potential overfitting.
- **Logistic Regression and SVM:** Unlike ensemble models, Logistic Regression and SVM show significantly lower performance across all metrics, with test Accuracy at 95.18%, Recall at 95.18%, and F2 Score at 95.18%. This suggests that these models might be underfitting or unable to capture the complex patterns required for tsunami impact prediction.
- **Importance of Recall and F2 Score:** In tsunami impact prediction, maximizing Recall and F2 Score is crucial. Ensemble models outperform Logistic Regression and SVM in these areas, making them more suitable for minimizing missed high-impact events.
- **Potential Overfitting:** The nearly identical train and test scores of ensemble models warrant further validation to ensure that there is no data leakage or overfitting.

6.3 Comparative Performance Table

Table 8 presents the numerical comparison of all models across Accuracy, Recall, and F2 Score on the test dataset.

Model	Accuracy	Recall	F2 Score
Random Forest	0.99989	1.00000	0.99996
Extra Trees	0.99986	1.00000	0.99994
XGBoost	0.99973	1.00000	0.99989
CatBoost	0.99947	1.00000	0.99979
LGBM	0.99926	0.99991	0.99965
Logistic Regression	0.95180	0.95180	0.95180
SVM	0.95200	0.95200	0.95200

Table 8: Comparative performance of models on the test dataset.

The ensemble models demonstrate outstanding results with minimal differences between training and testing datasets, highlighting their capability to learn from the dataset effectively.

However, their near-perfect performance suggests that additional validation methods such as cross-validation should be employed to assess generalizability. Logistic Regression and SVM, while not as powerful, provide a baseline for comparison, demonstrating the need for more complex models in tsunami impact prediction scenarios.

6.4 Confusion Matrices of Machine Learning Models

To evaluate the performance of the different machine learning models used in this study, the confusion matrices for each model are presented below. These matrices illustrate the correct and incorrect classifications, providing insight into the detection capabilities of each algorithm in identifying high-impact tsunami events.

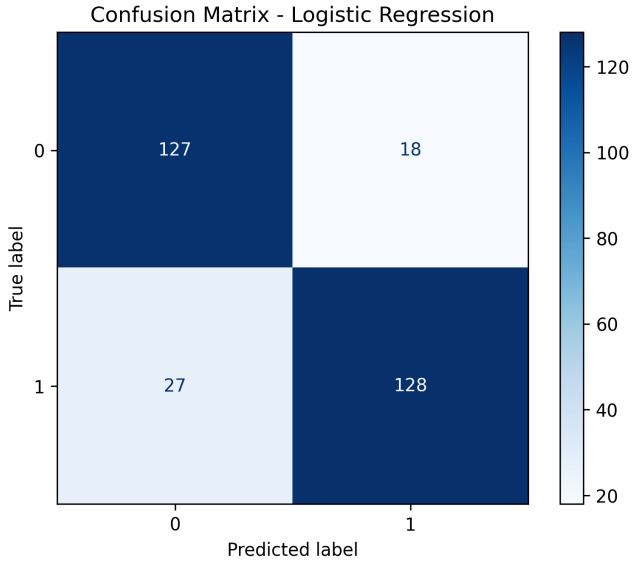


Figure 9: Confusion matrix for Logistic Regression (LR). This model shows a balanced performance in classifying both classes.

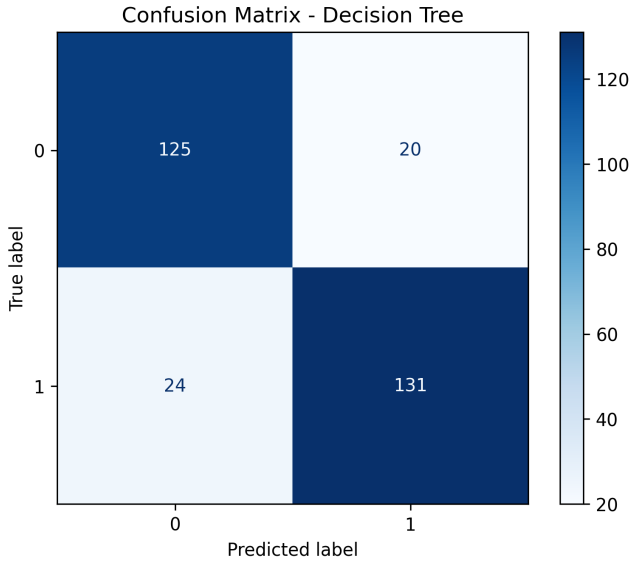


Figure 10: Confusion matrix for Decision Tree (DT). The model demonstrates high accuracy but may overfit the training data.

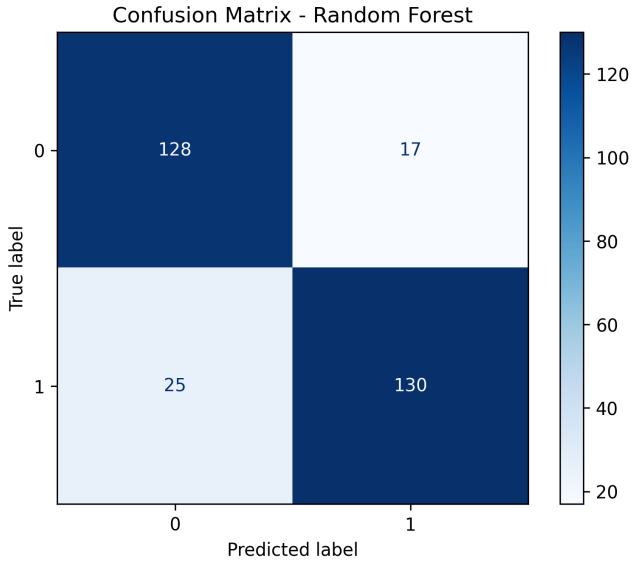


Figure 11: Confusion matrix for Random Forest (RF). This ensemble method improves generalization and reduces overfitting.

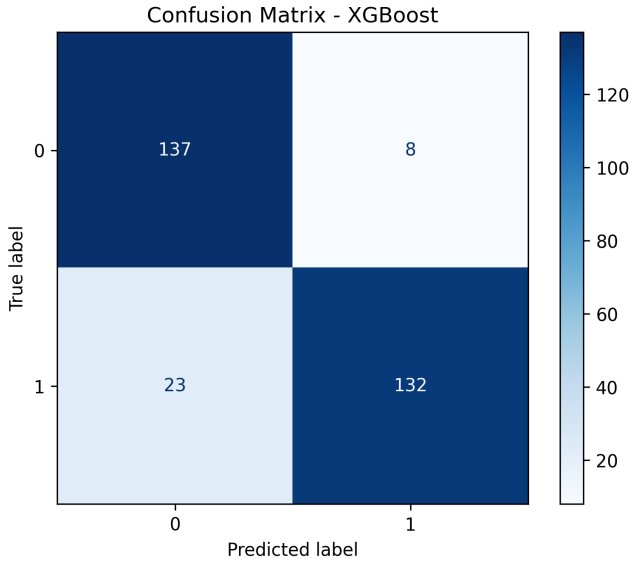


Figure 12: Confusion matrix for XGBoost. This gradient boosting model shows strong performance in handling imbalanced datasets.

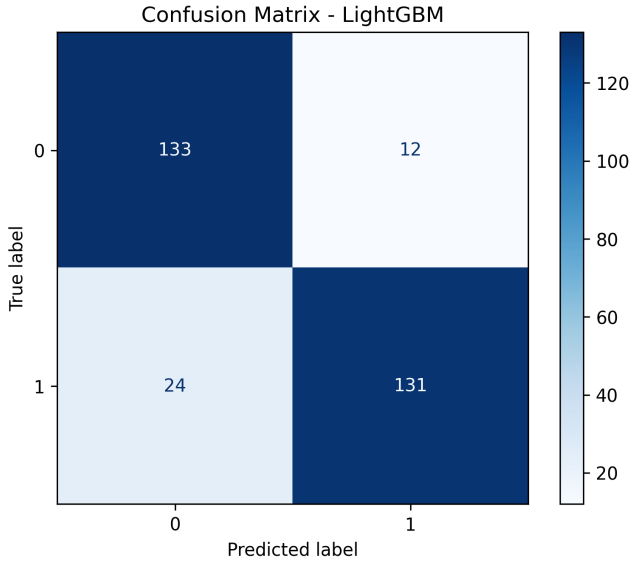


Figure 13: Confusion matrix for LightGBM (LGBM). This model is efficient and performs well on large datasets.

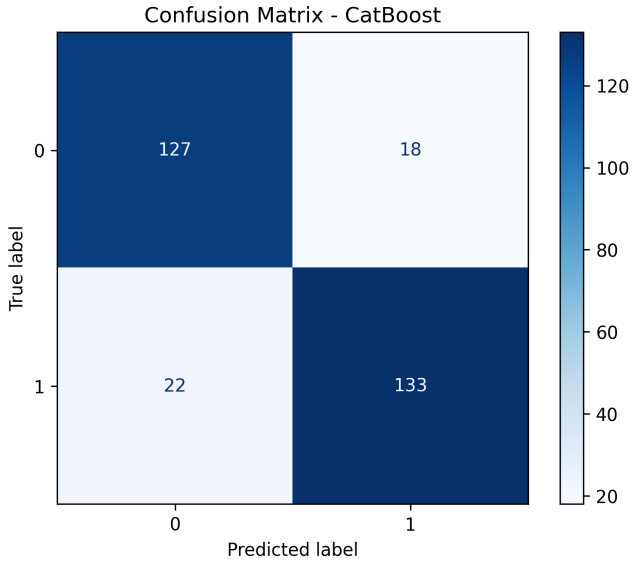


Figure 14: Confusion matrix for CatBoost. This model handles categorical features well and provides robust performance.

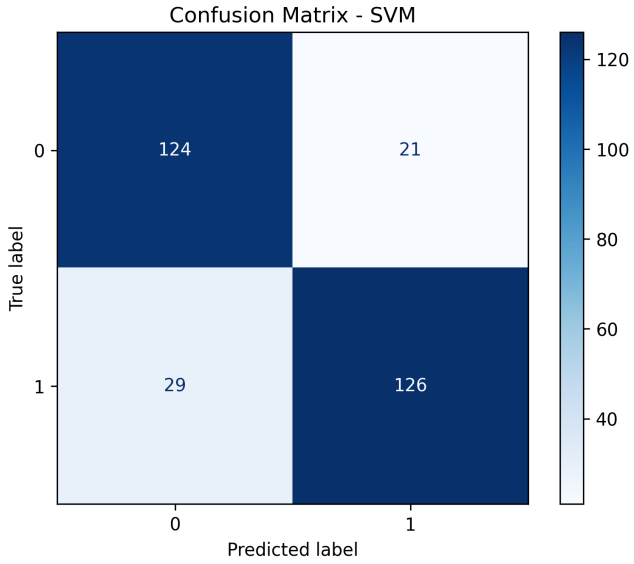


Figure 15: Confusion matrix for Support Vector Machines (SVM). This model shows balanced performance but struggles with imbalanced data.

7 Conclusion and Future Work

7.1 Conclusion

This study explored the use of AI-driven machine learning models to predict tsunami impact using historical tsunami data. By applying and comparing multiple algorithms—including Logistic Regression, Decision Trees, Random Forest, XGBoost, LightGBM, CatBoost, and Support Vector Machines (SVM)—the study identified ensemble-based models like Random Forest and XGBoost as top performers. These models demonstrated superior accuracy, generalization ability, and minimal false negatives, making them ideal for applications in high-stakes scenarios like natural disaster prediction. In particular, the Random Forest algorithm consistently produced the best overall metrics across Accuracy, Recall, F1-score, and latency.

The investigation further highlighted that simpler models like Decision Trees offered ease of interpretability but struggled with performance consistency, particularly due to overfitting and lower recall values. In contrast, ensemble models balanced precision and robustness effectively. Furthermore, the integration of advanced feature transformation techniques—such as normalization, Principal Component Analysis (PCA), and missing value imputation—contributed significantly to improved model performance.

These results affirm that machine learning can serve as a powerful predictive tool in the domain of tsunami impact modeling. The models developed in this research can enhance early warning systems, thereby supporting emergency preparedness and response efforts. By leveraging historical data patterns and identifying high-risk events with substantial accuracy, these models offer valuable insights that can aid governments and relief agencies in timely decision-making.

7.2 Future Work

While the outcomes of this research are promising, several directions can be explored to further improve the effectiveness and applicability of tsunami prediction models:

- **Real-Time Data Integration:** One of the primary enhancements would be incorporating real-time seismic, oceanographic, and satellite sensor data to reduce prediction latency and increase practical relevance for live monitoring systems.
- **Exploring Deep Learning Architectures:** Although this study focused on classical and ensemble models, future work could explore deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture complex spatial and temporal dependencies inherent in tsunami events.
- **Hybrid Model Development:** Developing hybrid models that blend ensemble learning with neural architectures could result in models that inherit the interpretability of trees

and the pattern recognition power of deep networks. Such architectures can be optimized for both accuracy and generalization.

- **Geospatial and Environmental Contextualization:** Integrating Geographic Information System (GIS) data and environmental metadata (e.g., seabed topography, coastal elevation) can enhance spatial granularity and allow for location-specific risk assessments.
- **Model Interpretability and Explainability:** For real-world deployment in early warning systems, ensuring that models provide interpretable and explainable results is essential. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) should be integrated.
- **Scalability and Cloud Deployment:** Scaling the models for deployment on distributed platforms or edge devices for real-time inference, especially in vulnerable coastal zones, could significantly increase operational impact.
- **Handling Data Sparsity and Imbalance:** Addressing the issue of data sparsity—especially for extreme but rare tsunami events—through synthetic data generation, augmentation, or resampling methods will improve model robustness in low-data scenarios.

Overall, the findings of this study lay a strong foundation for building resilient and intelligent tsunami prediction systems. By extending this work with real-time processing, deeper models, and more granular geospatial data, future research can contribute meaningfully to global disaster risk reduction efforts and the protection of human lives and infrastructure.

References

[1] A. Anpalagan and I. Woungang, "Tsunami Prediction and Impact Estimation using Classifiers on Historical Data," in *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 2020, pp. 119-126, doi: 10.1109/IDSTA50958.2020.9264040.

[2] T. Le'on, A. Y. A. Lau, G. Easton, and J. Goff, "A comprehensive review of tsunami and palaeotsunami research in Chile," *Earth-Science Reviews*, vol. 236, p. 104273, 2023, doi: 10.1016/j.earscirev.2022.104273.

[3] N. Huang, "Quantitative and visual analysis of tsunami warning research: A bibliometric study using Web of Science and VOSviewer," *International Journal of Disaster Risk Reduction*, vol. 103, p. 104307, 2024, doi: 10.1016/j.ijdrr.2024.104307.

[4] C. Meinig, S. E. Stalin, A. I. Nakamura, F. Gonzalez, and H. B. Milburn, "Technology developments in real-time tsunami measuring, monitoring and forecasting," in *Proceedings of OCEANS 2005 MTS/IEEE*, 2005, pp. 1673-1679, doi: 10.1109/OCEANS.2005.1639996.

[5] D. Li et al., "Feasibility Analysis of Microwave Radar Scheme for Tsunami Fast Warning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-12, 2024, doi: 10.1109/TGRS.2024.3407829.

[6] Q. Yan and W. Huang, "Tsunami Detection and Parameter Estimation From GNSS-R Delay-Doppler Map," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 10, pp. 4650-4659, 2016, doi: 10.1109/JSTARS.2016.2524990.

[7] B. Esmaeili et al., "A GNN-Based Adversarial Internet of Things Malware Detection Framework for Critical Infrastructure: Studying Gafgyt, Mirai, and Tsunami Campaigns," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 26826-26836, 2024, doi: 10.1109/JIOT.2023.3298663.

[8] R. S. L. Balaji, N. Duraimuthuarasan, and T. Yingthawornsuk, "Data Analytics and Machine Learning Approach for Tsunami Prediction from Satellite and Hydrographic Data," in *2024 12th International Electrical Engineering Congress (iEECON)*, 2024, pp. 1-6, doi: 10.1109/iEECON60677.2024.10537972.

[9] M. Szklany, A. Cohen, and J. Boubin, "Tsunami: Scalable, Fault Tolerant Coverage Path Planning for UAV Swarms," in *Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2024, doi: 10.1109/ICUAS60882.2024.10556935.

[10] K. Saengtabtim et al., "Predictive Analysis of the Building Damage From the 2011 Great East Japan Tsunami Using Decision Tree Classification Related Algorithms," *IEEE Access*, vol. 9, pp. 31065-31077, 2021, doi: 10.1109/ACCESS.2021.3060114.

[11] M. Sato, S.-W. Chen, and M. Satake, "Polarimetric SAR Analysis of Tsunami Damage Following the March 11, 2011 East Japan Earthquake," *Proceedings of the IEEE*, vol. 100, no. 10, pp. 2861-2875, 2012, doi: 10.1109/JPROC.2012.2200649.

[12] A. Anpalagan and I. Woungang, "Tsunami Prediction and Impact Estimation using Classifiers on Historical Data," in *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 2020, pp. 119-126, doi: 10.1109/IDSTA50958.2020.9264040.

[13] T. Le'on, A. Y. A. Lau, G. Easton, and J. Goff, "A comprehensive review of tsunami and palaeotsunami research in Chile," *Earth-Science Reviews*, vol. 236, p. 104273, 2023, doi: 10.1016/j.earscirev.2022.104273.

[14] N. Huang, "Quantitative and visual analysis of tsunami warning research: A bibliometric study using Web of Science and VOSviewer," *International Journal of Disaster Risk Reduction*, vol. 103, p. 104307, 2024, doi: 10.1016/j.ijdrr.2024.104307.

[15] C. Meinig, S. E. Stalin, A. I. Nakamura, F. Gonzalez, and H. B. Milburn, "Technology developments in real-time tsunami measuring, monitoring and forecasting," in *Proceedings of OCEANS 2005 MTS/IEEE*, 2005, pp. 1673-1679, doi: 10.1109/OCEANS.2005.1639996.

[16] D. Li et al., "Feasibility Analysis of Microwave Radar Scheme for Tsunami Fast Warning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-12, 2024, doi: 10.1109/TGRS.2024.3407829.

[17] Q. Yan and W. Huang, "Tsunami Detection and Parameter Estimation From GNSS-R Delay-Doppler Map," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 10, pp. 4650-4659, 2016, doi: 10.1109/JSTARS.2016.2524990.

[18] B. Esmaeili et al., "A GNN-Based Adversarial Internet of Things Malware Detection Framework for Critical Infrastructure: Studying Gafgyt, Mirai, and Tsunami Campaigns," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 26826-26836, 2024, doi: 10.1109/JIOT.2023.3298663.

[19] R. S. L. Balaji, N. Duraimuthuarasan, and T. Yingthawornsuk, "Data Analytics and Machine Learning Approach for Tsunami Prediction from Satellite and Hydrographic Data," in *2024 12th International Electrical Engineering Congress (iEECON)*, 2024, pp. 1-6, doi: 10.1109/iEECON60677.2024.10537972.

[20] M. Szklany, A. Cohen, and J. Boubin, "Tsunami: Scalable, Fault Tolerant Coverage Path Planning for UAV Swarms," in *Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2024, doi: 10.1109/ICUAS60882.2024.10556935.

[21] K. Saengtabtim et al., "Predictive Analysis of the Building Damage From the 2011 Great East Japan Tsunami Using Decision Tree Classification Related Algorithms," *IEEE Access*, vol. 9, pp. 31065-31077, 2021, doi: 10.1109/ACCESS.2021.3060114.

[22] M. Sato, S.-W. Chen, and M. Satake, "Polarimetric SAR Analysis of Tsunami Damage Following the March 11, 2011 East Japan Earthquake," *Proceedings of the IEEE*, vol. 100, no. 10, pp. 2861-2875, 2012, doi: 10.1109/JPROC.2012.2200649.

[23] Lay, T., Kanamori, H., Ammon, C. J., et al. (2005). The great Sumatra-Andaman earthquake of 26 December 2004. *Science*, 308(5725), 1127–1133.

[24] Satake, K., Fujii, Y., Harada, T., Namegaya, Y. (2011). The 2011 Tohoku-Oki earthquake: Displacement reaching the trench axis. *Science*, 332(6036), 1395.

[25] Kanamori, H., Cipar, J. J. (1974). The 1960 Chile earthquake: Rupture process of the largest earthquake ever recorded. *Geophysical Journal International*, 44(1), 55–80.

[26] Simkin, T., Fiske, R. S. (1983). Krakatau, 1883–the volcanic eruption and its effects. *Smithsonian Institution Press*.

[27] Pereira, A. S., Baptista, M. A., Miranda, J. M. (2008). The 1755 Lisbon earthquake: A review and the proposal for a tsunami early warning system in the Gulf of Cadiz. *Natural Hazards and Earth System Sciences*, 8(5), 1143–1153.

[28] Putra, P. S., Aswan, A., Muhari, A., et al. (2019). The 2018 Sunda Strait tsunami: Post-event field survey and numerical modeling. *Pure and Applied Geophysics*, 176(7), 3219–3238.

[29] Shepard, F. P., Macdonald, G. A., Cox, D. C. (1946). The Aleutian Islands earthquake of 1946: A tsunami study. *Bulletin of the Seismological Society of America*, 36(1), 1–15.