Prediction of Ground Motion using Data on Seismic Waves

Praanesh Balakrishnan Nair
Dept. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Bengaluru, India
bl.en.u4aie23123@bl.students.amrita.edu

Varun Adhitya G B
Dept. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Bengaluru, India
bl.en.u4aie23135@bl.students.amrita.edu

Sanjushree Rajan
Dept. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Bengaluru, India
bl.en.u4aie23130@bl.students.amrita.edu

***Abstract*—This paper presents a methodology for analyzing seismic data from significant earthquakes using the ObsPy Python library and various machine learning techniques. Our dataset encompasses four major seismic events: the 2011 Tohoku earthquake (Japan), the 2010 Chile earthquake, the 2016 Kumamoto earthquake (Japan), and the 2015 Nepal earthquake. We process seismic waveforms to extract relevant features, including peak ground acceleration and Arias intensity. A Short-Term Average/Long-Term Average (STA/LTA) algorithm is used for P- and S-wave identification. A k-Nearest Neighbours (kNN) regression model predicts seismic intensity measures, while K-means clustering reveals underlying patterns in ground motion behavior. Furthermore, a custom Decision Tree classifier was built from scratch using entropy, Gini index, and information gain criteria. This model was used to classify ground motion behavior based on binned seismic features and visualize decision boundaries.**

***Index Terms*—Seismic Waves, Ground Acceleration, Arias Intensity, STA/LTA, Decision Tree, Entropy, Information Gain, Classification**

I. Introduction

The analysis of seismic data provides crucial insights into earthquake dynamics and Earth's subsurface structure. Various seismic waveform data from several significant earthquakes were processed and analyzed by investigating the relationship between maximum amplitude and distance to the event. This was done by extracting features, using matrix operations and similarity measures. We focus on four major seismic events: the 2011 Tohoku earthquake (Japan), the 2010 Chile earthquake, the 2016 Kumamoto earthquake (Japan), and the 2015 Nepal earthquake. This approach facilitates access to a large volume of data from diverse sources.

II. Literature Survey

Seismic signal processing and analysis have garnered significant attention in recent years, with various machine learning methodologies being employed for classification, event detection, and predictive modeling. This section reviews key contributions in the field, highlighting different approaches and techniques applied to seismic data analysis.

Li et al. [**?**] investigated seismic data classification using supervised machine learning techniques. Their study focused on extracting features such as spectral content, amplitude variations, and waveform characteristics. They evaluated the performance of multiple machine learning models, including Support Vector Machines (SVM), Decision Trees, and Neural Networks, which were trained on labeled seismic event datasets to assess classification accuracy.

Ramirez and Meyer [**?**] explored seismic phase classification through manifold learning techniques. Their approach involved mapping high-dimensional seismic data onto a lower-dimensional manifold using Laplacian Eigenmaps, improving classification performance by preserving local waveform structures. Their method demonstrated enhanced differentiation between P-waves and S-waves using nearest-neighbor-based classifiers in the transformed space.

Chakraborty et al. [**?**] employed statistical feature extraction techniques for micro-seismic event detection. The extracted features included peak amplitude, energy, zero-crossing rate, and entropy. Machine learning models such as Random Forest, SVM, and k-Nearest Neighbors (KNN) were applied to classify seismic events, distinguishing natural seismic activity from noise with improved accuracy.

Shu et al. [**?**] conducted a comprehensive survey on machine learning applications in microseismic signal recognition and classification. They categorized methodologies into three major groups: feature-based models utilizing statistical descriptors, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and hybrid techniques that integrate statistical feature extraction with deep learning frameworks. Additionally, their survey highlighted key challenges and potential advancements in the domain.

Varshney et al. [**?**] developed a machine learning-based earthquake monitoring system leveraging real-time seismic data streams. Their approach involved preprocessing seismic data using Fourier and Wavelet Transforms before training classification models, such as Decision Trees and Neural Networks, to detect and monitor earthquake events dynamically.

Chin et al. [**?**] enhanced earthquake detection accuracy by implementing a hybrid deep learning model integrating CNNs and Long Short-Term Memory (LSTM) networks. Their framework, trained on labeled seismic waveform datasets, incorporated data augmentation techniques to improve model robustness. The proposed system demonstrated superior detection performance compared to traditional signal processing methods.

Shimshoni and Intrator [?] applied ensemble learning techniques for seismic signal classification. Their work utilized bagging and boosting strategies to enhance classification accuracy, demonstrating that ensemble-based neural network approaches could improve differentiation between seismic events.

Agliz and Atmani [?] employed multi-layer perceptron (MLP) neural networks for seismic signal classification. Their methodology involved extracting both frequency-domain and time-domain features from seismic waveforms, which were subsequently used as input to the neural network for classification and training.

Akhouayri et al. [?] introduced a fuzzy expert system for automatic seismic signal classification. Their method defined fuzzy sets based on key signal parameters such as amplitude, frequency content, and duration, allowing for a more flexible and interpretable classification framework compared to traditional rule-based systems.

Curilem et al. [?] investigated the application of genetic algorithms for optimizing neural network classifiers in the context of volcanic seismic signal classification. Their findings demonstrated that evolutionary computation techniques could enhance the performance of neural network models in differentiating between various types of volcanic seismic events.

III. Methodology

A. Data Extraction

The ObsPy Python library was used for obtaining seismic wave data within 20° of the epicenter. The epicenter of the earthquake is at 37.52° N and 143.04° E, and the seismic activity was observed from 60 seconds before the earthquake, to 600 seconds after the earthquake. Using `obspy.mass_downloader()`, waveform data was retrieved in miniSEED format corresponding station metadata in StationXML format from multiple data centers, including IRIS. The downloaded miniSEED files were processed to extract key features for analysis. The seismic channels prioritized are "Horizontal North-South" (HN), "Broadband North-South" (BN), "Horizontal Horizonal" (HH) and "Broadband Horizontal" (BH).

B. Data Preprocessing

Since different seismic instruments have different frequencies at which they are most sensitive, the actual ground motion had to be isolated from the instruments' behaviour. The raw waveform files were read and the instrument response was removed, by using the right acceleration units and apply pre-filtering of 0.01 Hz to 50 Hz to stabilize the data.

C. Feature Extraction

The features extracted were station latitude, longitude, station elevation, distance from earthquake epicenter, Peak Ground Acceleration (PGA), Root Mean Square (RMS) acceleration, Arias Intensity (measure of ground shaking severity), Significant duration (time between 5-95% cumulative energy), Predominant frequency, Mean frequency, Spectral centroid and Energy distribution across frequency bands.

D. Training

After feature extraction, K-means algorithm was used to divide the data into 3 clusters, with the aim of identifying similar ground motion behaviour. This was done after Principal Component Analysis (PCA) was used for reduction in dimensionality.

IV. Results

TABLE I
Results and their details

Aspect	Details
Entropy of `pga_binned`	0.2869
Gini Index of `pga_binned`	0.0750
Information Gain of `pga_binned`	0.287
Root node feature based on Information Gain	`sig_duration_cat`

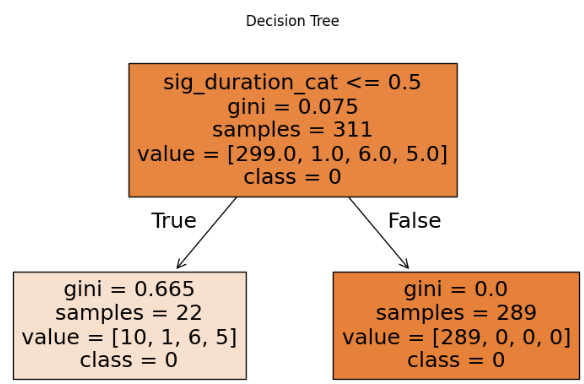

Fig. 1. Decision Tree Visualization

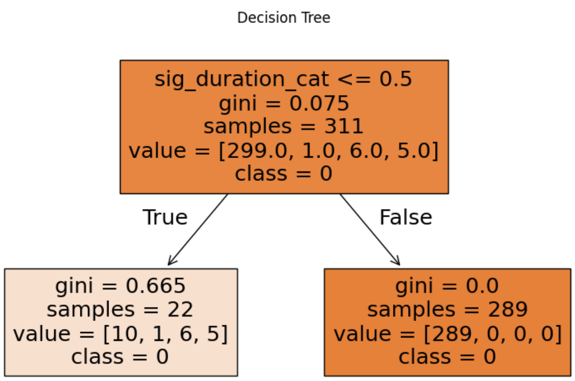

Fig. 2. Decision Tree Boundaries