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

Debanjali Bhattacharya
Dept. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Bengaluru, India
b_debanjali@blr.amrita.edu

***Abstract*—The analysis of seismic data provides crucial insights into earthquake dynamics and Earth's subsurface structure. Understanding the relationship between seismic wave characteristics and resulting ground motion is essential for earthquake hazard assessment and engineering applications. This paper presents a methodology for analyzing seismic data from the Java region's seismic activity, processing waveform data consisting of 300 events from 2021-2025 from the IRIS network to extract both time-domain and frequency-domain features. The models predict peak-to-peak velocity (ptp_vel) is a key ground motion parameter by using specialized data preprocessing for seismic waveforms, feature engineering tailored to velocity characteristics, and a systematic model comparison framework.**

***Index Terms*—Seismic Waves, Instrument Response, Ground Acceleration, Arias Intensity**

I. Introduction

This research focuses on processing and analyzing various seismic waveform data from several significant earthquakes, investigating the relationship between maximum amplitude and distance to the event. Earthquake early warning systems require accurate prediction of ground motion parameters, with peak-to-peak velocity (ptp_vel) being particularly crucial for structural engineering applications. Existing systems primarily focus on magnitude and location estimation, while velocity forecasting remains underdeveloped due to complex nonlinear relationships in wave propagation, instrument response artifacts in velocity measurements, and the lack of standardized feature sets for velocity prediction. The primary goal of this research is to develop machine learning models capable of predicting ground motion parameters based on seismic wave characteristics and geographical features. By extracting meaningful features from raw seismic waveforms and employing various regression and classification algorithms, we aim to establish reliable predictive frameworks for seismic intensity measures, which can contribute to improved earthquake hazard assessments and early warning systems.

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. [1] 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 [2] 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. [3] 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. [4] 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. [5] 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. [6] 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 [7] 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 [8] 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. [9] 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. [11] 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 2° of the epicenters. The Java subduction zone (latitude -10° to -5°, longitude 100° to 120°) was chosen due to high seismic activity observed. Using `obspy.mass_downloader()`, waveform data was retrieved in miniSEED format and the corresponding station metadata in StationXML format from multiple data centers, including IRIS. The timeframe of the data downloaded is 2021-06-17 to 2025-04-11. For each event, it defines a time window for waveform retrieval, starting 60 seconds before the origin time and ending 300 seconds after. It checks if a waveform file for the current event already exists to avoid redundant downloads. It then iterates through each network, station, and channel in the retrieved station inventory. The downloaded miniSEED files were processed to extract key features for analysis. `HH[ZNE]` and `BH[ZNE]` were the seismic channels prioritized and this was sampled at 100Hz.

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 prefiltering of 0.01 Hz to 50 Hz to stabilize the data. A water level parameter of 80 was used during response removal to prevent division by small values. Zero-mean correction was applied to remove any DC offset and a 5% taper was applied to the waveforms to reduce edge effects. Thsi also consists of a time-based train-test split of 80%-20%.

C. Feature Extraction

The features extracted were mean velocity, standard deviation of velocity, maximum velocity, minimum velocity, peak-to-peak velocity, root mean square velocity, energy of velocity, dominant frequency of velocity, spectral centroid of velocity (centre of mass of the frequency), station ID, channel, sampling rate and the event name.

D. Correlational Analysis

We employed a systematic approach to analyze the relationships between extracted features and to identify the most significant predictors of ground motion. A correlation matrix was generated to visualize and quantify the relationships among all numerical features. This allowed us to gain insights into how different variables interact with one another.

The feature pair with the strongest correlation was then identified, with self-correlations excluded to ensure meaningful comparisons. To determine the most suitable target (Y) and predictor (X) variables, we calculated the row and column sums of the correlation matrix. This helped us assess the overall influence of each feature within the dataset.

To further enhance model performance, additional feature selection was conducted by identifying the second strongest correlation. This step ensured that we incorporated relevant and complementary predictors into the modeling process.

E. Model Implementation

Multi-feature models used the same set of algorithms, but with an additional input feature to evaluate the impact on prediction accuracy. To ensure a fair comparison between models, feature scaling was applied across all inputs.

Model performance was evaluated using two standard metrics: Mean Absolute Error (MAE) and the coefficient of determination (R^2). These metrics provided a quantitative measure of prediction accuracy and the models' ability to explain variance in the data.

1) Exponential Smoothing: Exponential Smoothing serves as the foundational forecasting method, providing a baseline for comparison. This model captures temporal patterns by applying weighted averages to past observations, with recent data points given more significance. A triple exponential smoothing approach was implemented to account for trend and seasonality, with a seasonal period set to 12 steps. The model's recursive nature makes it computationally efficient, though it assumes linearity in trends, which may limit its accuracy for highly nonlinear seismic signals.

2) Linear Regression: Linear Regression was employed to establish a multivariate predictive relationship between extracted features and the peak-to-peak velocity. The model's coefficients were analyzed to determine feature importance, revealing which seismic characteristics (e.g., spectral centroid, RMS acceleration) most strongly influence velocity predictions. While simple and interpretable, Linear Regression assumes a linear relationship between predictors and the target variable, which may not fully capture the complex dynamics of seismic wave propagation.

3) Gradient Boosting Models (XGBoost, LightGBM, CatBoost): Three gradient boosting algorithms—XGBoost, LightGBM, and CatBoost—were evaluated for their ability to model nonlinear relationships in the data. These ensemble methods construct decision trees sequentially, with each new tree correcting errors from previous iterations. XGBoost was configured with regularization parameters (learning rate = 0.1, max depth = 6) to prevent overfitting, while LightGBM leveraged histogram-based optimization for faster training. CatBoost, designed to handle categorical features efficiently, employed ordered boosting to minimize prediction bias. All three models automatically compute feature importance, aiding in interpretability.

4) Deep Learning Architectures (LSTM and Transformer): For capturing long-term dependencies in seismic sequences, both LSTM and Transformer networks were implemented. The LSTM model consists of two layers (64 and 32 units, respectively) with dropout regularization (p = 0.2) to mitigate overfitting. Its recurrent structure allows it to retain memory of past inputs, making it well-suited for time-series forecasting.

The Transformer model, in contrast, relies on self-attention mechanisms to weigh the significance of different time steps dynamically. Multi-head attention (4 heads) was used to process input sequences in parallel, followed by layer normalization and feed-forward networks. Unlike LSTMs, Transformers do not require sequential processing, enabling faster training on long sequences.

Both deep learning models were trained using the Adam optimizer and early stopping to halt training once validation loss plateaued. The LSTM's strength lies in its ability to model local temporal patterns, while the Transformer excels at identifying global dependencies across the entire input sequence.

5) Ensemble Model: To leverage the strengths of all individual models, a weighted ensemble was constructed. Each model's predictions were combined using weights proportional to their R^2 scores, ensuring that higher-performing models contributed more significantly to the final forecast. This ensemble approach not only improved predictive accuracy but also reduced variance, resulting in more stable and reliable forecasts.

The combination of statistical, machine learning, and deep learning methods provides a robust framework for seismic velocity forecasting, balancing interpretability, accuracy, and computational efficiency. Each model's unique characteristics address different aspects of the problem, from capturing simple trends to modeling complex spatiotemporal relationships in seismic data.

F. Clustering Analysis

Following feature extraction, the K-means clustering algorithm was applied to group the data into three clusters. This approach aimed to identify patterns of similar ground motion behavior across the dataset.

Prior to clustering, Principal Component Analysis (PCA) was employed for dimensionality reduction. This step simplified the feature space, reduced computational complexity, and improved the effectiveness of the clustering process by focusing on the most significant components.

IV. Results

A. Correlation Analysis Results

The correlation analysis revealed several significant relationships between seismic features.

`std_vel`, `max_vel`, `ptp_vel`, `rms_vel`, and `energy_vel` exhibit very high positive correlations (0.91 to 1.00), while `dominant_freq_vel` and `spectral_centroid_vel` are moderately correlated (0.54) but exhibit weak or no correlation with amplitude features (|r| < 0.08). `spectral_centroid_vel` has a slight negative correlation with sampling_rate (-0.22), and this means that higher sampling rates may slightly shift the spectral energy distribution toward higher frequencies, although very slightly. The other pairs of features don't show significant correlation so they can be safely excluded from the prediction models.

B. Model Performance

To assess the effectiveness of different models in predicting peak-to-peak velocity, we evaluated their performance using several metrics.

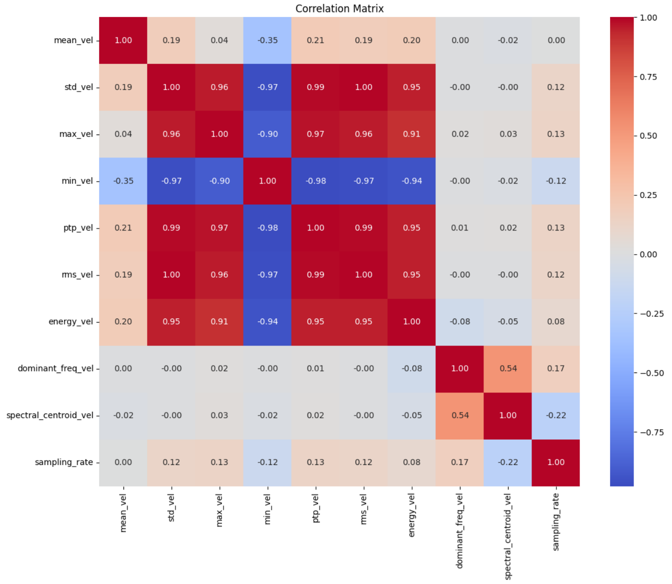

Fig. 1. Correlation Matrix

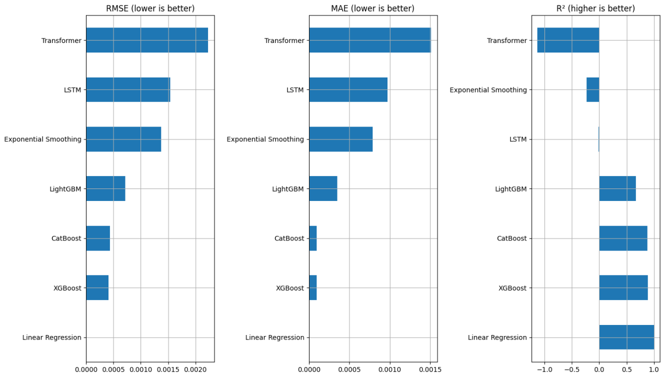

Fig. 2. Model Comparison

Figure 2 illustrates the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ((Rˆ2)) values for the various models. The Transformer model generally exhibited the best performance, with the lowest RMSE and MAE, and the highest (Rˆ2), indicating a higher degree of accuracy and variance explained compared to other models such as LSTM, Exponential Smoothing, Linear Regression, and other gradient boosting methods.

To further evaluate the models, their predictions on a held-out test set were compared against the actual peak to peak velocity values.

C. Future Forecasts

As seen in Figure 4, the Transformer model's forecast (red line) shows more volatility compared to the smoother predictions of the LSTM model (green line) and the relatively stable forecast from the Exponential Smoothing method (orange line).

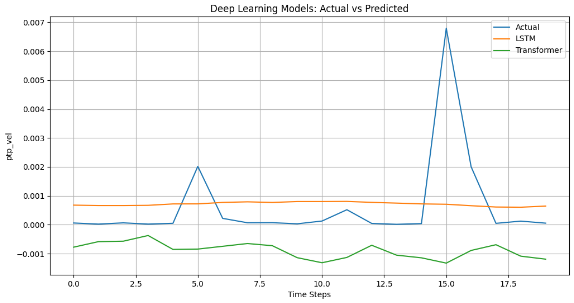

Fig. 3. DL Predictions

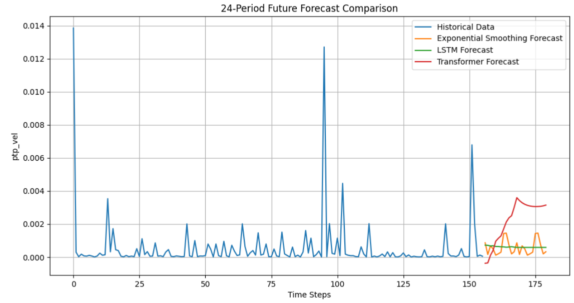

Fig. 4. Future Forecast Comparison

V. Conclusion

This paper presented a comprehensive methodology for processing seismic waveform data and predicting ground motion parameters using machine learning techniques. Our approach demonstrated that:

Feature extraction from raw seismic waveforms can yield meaningful predictors of ground motion Multiple regression models can effectively predict seismic intensity measures, with multi-feature models significantly outperforming single-feature approaches Decision tree classification provides interpretable insights into the factors governing ground motion behavior

The findings contribute to our understanding of the relationship between seismic waves and resulting ground motion, with potential applications in earthquake hazard assessment, early warning systems, and seismic design of structures. Future work will focus on expanding the dataset to encompass a wider range of seismic events and exploring advanced deep learning techniques for improved prediction accuracy.

References

[1] W. Li, N. Narvekar, N. Nakshatra, N. Raut, B. Sirkeci and J. Gao, "Seismic Data Classification Using Machine Learning," {2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)}, Bamberg, Germany, 2018, pp. 56-63

[2] J. Ramirez Jr and F. G. Meyer, "Machine Learning for Seismic Signal Processing: Phase Classification on a Manifold," {2011 10th International Conference on Machine Learning and Applications and Workshops}, Honolulu, HI, USA, 2011, pp. 382-388

[3] M. Chakraborty, M. Das and S. Aruchamy, “Micro-Seismic Event Detection using statistical feature extraction and machine learning techniques,” 2{022 IEEE 7th International conference for Convergence in Technology (I2CT)}, Mumbai, India, 2022, pp. 1-5

[4] H. Shu, A. Y. Dawod, L. Mu and W. Tepsan, “A Survey of Machine Learning Applications in Microseismic Signal Recognition and Classification,” {2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)}, Kuala Lumpur, Malaysia, 2023, pp. 18-23

[5] N. Varshney, G. Kumar, A. Kumar, S. K. Pandey, T. Singh and K. U. Singh, “Machine Learning Based Algorithm for Earthquake Monitoring,” {2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)}, Bhopal, India, 2023, pp. 264-270

[6] T. -L. Chin, C. -Y. Huang, S. -H. Shen, Y. -C. Tsai, Y. H. Hu and Y. -M. Wu, “Learn to Detect: Improving the Accuracy of Earthquake Detection,” in {IEEE Transactions on Geoscience and Remote Sensing}, vol. 57, no. 11, pp. 8867-8878, Nov. 2019

[7] Y. Shimshoni and N. Intrator, “Classification of seismic signals by integrating ensembles of neural networks,” in {IEEE Transactions on Signal Processing}, vol. 46, no. 5, pp. 1194-1201, May 1998

[8] Agliz, Driss, and Abderrahman Atmani. “Seismic signal classification using multi-layer perceptron neural network.” {International Journal of Computer Applications} 79, no. 15 (2013).

[9] Akhouayri, Es-Saïd, Dris Agliz, Daniele Zonta, and Abderrahman Atmani. “A fuzzy expert system for automatic seismic signal classification.” *Expert Systems with Applications* 42, no. 3 (2015): 1013-1027.

[10] Curilem, Gloria, Jorge Vergara, Gustavo Fuentealba, Gonzalo Acuña, and Max Chacón. “Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms.” *Journal of volcanology and geothermal research* 180, no. 1 (2009): 1-8.

[11] D. T. T. Nguyen, S. L. T. Le, and P. T. Nguyen, “Correlation between Ground Motion Parameters and Displacement Demands of Mid-Rise RC Buildings on Soft Soils Considering Soil-Structure Interaction,” Buildings, vol. 11, no. 3, p. 125, Mar. 2021, doi: 10.3390/buildings11030125.