American Journal of Civil Engineering
2025, Vol. 13, No. 5, pp. 284-303
https://doi.org/10.11648/j.ajce.20251305.14

Research Article

Earthquake Prediction and Synthetic Seismogram Generation Using Hybrid CNN – LSTM Model

Harshita Guduru*, Darshan Joshi, Wisam Bukaita

Math and Computer Sciences, Lawrence Technological University, Southfield, USA

Abstract

This study addresses the critical challenge of earthquake prediction and synthetic seismogram generation through the application of deep learning. A hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) framework is proposed to capture both spatial and temporal characteristics of seismic data, enabling robust forecasting of seismic activity and the generation of realistic waveform simulations. Historical seismographic records and metadata, including magnitude, depth, and epicentral location, were sourced from the United States Geological Survey (USGS). Key predictive features such as amplitude variation, temporal intervals, epicentral distances, and regional spatial attributes were extracted to train and validate the model. Model development and experimentation were conducted in Python using TensorFlow, Keras, NumPy, Pandas, Scikit-learn, and Imbalanced-learn. The CNN component was employed to extract spatial representations from seismograms, while the LSTM component modeled sequential dependencies inherent in seismic waveforms. The final model achieved an accuracy of 84%, with notable improvements across precision, recall, and loss metrics. Statistical evaluations further validated the reliability of the results. The findings demonstrate the potential of hybrid deep learning architectures to enhance early earthquake warning systems and hazard assessment strategies. By integrating prediction with synthetic seismogram generation, this research advances data-driven seismology and provides a scalable foundation for future applications in disaster preparedness and risk mitigation.

Keywords

Earthquake Prediction, Seismic Forecasting, Synthetic Seismogram Generation, Seismology, Seismic Signal Processing, Spatio-Temporal Modeling

1. Introduction

The field of seismology has gained significant attention in recent years due to its potential to mitigate earthquake-related disasters. This research aims to explore the prediction of earthquake occurrences and generation of synthetic seismograms by employing a hybrid CNN-LSTM model. As existing models often struggle with low accuracy and temporal instability, particularly with imbalanced datasets, it is crucial to develop a deeper understanding of integrating spatial and temporal features for robust forecasting. This study addresses these gaps by optimizing a hybrid deep learning approach, utilizing seismic data to enhance prediction accuracy and simulation capabilities.

Numerous studies have been conducted in the field of seismology, and while significant advancements have been

*Corresponding author: hguduru@ltu.edu (Harshita Guduru)

Received: 24 September 2025; **Accepted:** 7 October 2025; **Published:** 30 October 2025

achieved, persistent gaps remain in addressing class imbalance, enhancing predictive accuracy, and ensuring stable model performance. To mitigate these challenges, this research integrates two complementary oversampling techniques—Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN)—within the data preprocessing stage prior to training the hybrid CNN–LSTM architecture.

Specifically, SMOTE is applied initially to generate synthetic samples for the minority earthquake classes by interpolating between existing minority class instances, thereby reducing bias introduced by unequal class distribution. Following this, ADASYN is employed to further refine the dataset by adaptively generating additional synthetic samples in regions where the minority class is more difficult to learn. This two-step approach enables the model to better capture complex seismic patterns, improve generalization, and enhance the stability of earthquake prediction performance across diverse seismic event magnitudes.

2. Literature Review

The prediction of earthquakes and generation of synthetic seismograms has attracted substantial research attention due to its importance in public safety, disaster preparedness, and geophysical research. Earthquakes pose severe risks to life and infrastructure, particularly in seismically active regions, making accurate forecasting models an urgent scientific and societal priority. While traditional approaches such as statistical forecasting, signal processing, and probabilistic hazard mapping have provided valuable insights, they often fall short in capturing the complex spatial–temporal dependencies of seismic events.

Recent advances in deep learning have introduced more robust frameworks. Convolutional Neural Networks (CNNs) are widely used for spatial feature extraction, while Long Short-Term Memory (LSTM) networks excel at modeling sequential dependencies. Their integration into hybrid CNN–LSTM architectures has emerged as a promising direction for both earthquake prediction and synthetic seismogram generation. These synthetic datasets are particularly valuable for simulating ground-motion scenarios where historical seismic records are limited. However, persistent challenges remain, including class imbalance, low prediction stability, and difficulty in generating realistic seismic waveforms, which motivates further refinement of hybrid models.

Several studies highlight both the progress and limitations in this domain. Zechar et al. proposed Molchan diagrams to evaluate alarm-based earthquake forecasts, showing limited predictive improvement over historical reference models [1]. Lijun et al. (2019) introduced CPIC, a CNN-based phase detector that performed with 97.5% accuracy on aftershocks of the Wenchuan earthquake, demonstrating the potential of lightweight CNN models in data-scarce environments [2]. Tapia et al. (2019) emphasized the societal implications of unreliable forecasting and the importance of accurate seismic communication [3].

Deep learning approaches have since demonstrated notable success. Jena et al. developed a CNN-based probability mapping model for the Indian subcontinent that achieved 92% accuracy and identified over 1.3 million km^2 of high-risk zones [4]. Rachna et al. applied machine learning to USGS data, finding that Multi-Layer Perceptrons delivered the most reliable predictions across broader geographical ranges [5]. Bhardwaj et al. expanded earthquake prediction research to non-seismic precursors such as magnetic anomalies, gas emissions, and ionospheric disturbances, advocating for standardized global monitoring [6]. Meier et al. challenged assumptions about earthquake rupture growth, proposing a new classification system for predictability studies [7].

Recent contributions underscore the role of deep learning in real-time systems. Wang et al. introduced EEWNet, which directly processes raw P-wave data for rapid magnitude estimation, outperforming empirical approaches [8]. Zhu et al. developed MOIPor, a multimodal model combining waveform, spectral, and metadata inputs to enhance onsite intensity predictions for early warning systems [9]. Similarly, Kaushal et al. demonstrated that hybrid RNN–LSTM architectures achieved up to 98% forecasting accuracy, though challenges in real-time data processing remain [10].

Other innovations extend deep learning to seismic data interpretation. Gao et al. integrated CNNs and LSTMs for reflectivity model building, addressing limitations in conventional seismic deconvolution [11]. Elsayed et al. proposed FCDNet, an attention-based dense CNN, achieving 97% accuracy in earthquake detection across global datasets [12]. Jiang et al. applied an FCN for automated time-window selection in full-waveform inversion, accelerating seismic analysis by orders of magnitude [13]. Complementary studies demonstrate the value of synthetic approaches: Irwandi et al. generated realistic shakemaps from synthetic seismograms, Reddy et al. applied SVMs with wavelet-transformed data for near real-time magnitude estimation, and Jie et al. used compressed sensing to rapidly reconstruct seismograms from sparse earthquake inputs [14-16].

Taken together, the literature underscores both the promise and ongoing challenges of earthquake prediction and synthetic seismogram generation. CNN–LSTM hybrids stand out for their ability to jointly capture spatial and temporal seismic features, yet issues such as class imbalance, limited training data, and computational scalability remain barriers. Addressing these challenges through techniques such as ADASYN oversampling and refined hybrid architectures, as pursued in this study, represents a critical step toward more reliable earthquake forecasting and realistic waveform simulation.

3. Research Methodology

This study employs a quantitative methodology aimed at improving earthquake prediction and generating realistic synthetic seismograms using a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model. While the deep learning pipeline provides a robust computational structure, the methodology also integrates principles of physical seismology to ensure that model features and outputs correspond to meaningful geophysical processes. Specifically, the research links neural feature representations to P- and S-wave propagation patterns, attenuation effects, and fault-zone characteristics, bridging data-driven and physics-based approaches.

The quantitative framework supports rigorous assessment of predictive accuracy, waveform fidelity, and geophysical interpretability. Statistical metrics (e.g., RMSE, correlation coefficient, precision, recall, and F1-score) are complemented by physical consistency checks, such as waveform polarity, travel-time estimation, and amplitude decay comparison against ground-truth seismograms.

3.1. Research Design

This research adopts a structured quantitative design that systematically combines data-driven learning with seismological domain interpretation. The approach proceeds through four key stages: data preprocessing, feature extraction, model training, and interpretive validation.

Stage 1: Data Preprocessing and Feature Engineering

The dataset (preprocessed_earthquake_data.csv) includes key physical and spatial attributes such as latitude, longitude, depth, magnitude, and time differences.

Before training, waveform segments are decomposed into 10 sub-seismograms representing successive time windows, ensuring that the CNN can learn localized waveform features such as:

1) P-wave onsets: High-frequency, low-amplitude signals marking the first arrival.
2) S-wave arrivals: Larger-amplitude, lower-frequency energy following the P-wave.
3) Attenuation and scattering: Gradual amplitude decay and frequency damping that vary with depth and distance.

This segmentation allows the CNN to capture the spatial-temporal evolution of seismic energy, while maintaining physical meaning in the learned feature maps.

Stage 2: CNN Feature Learning

The Convolutional Neural Network (CNN) component is designed to detect spatial-frequency patterns analogous to physical signal components:

1) Early CNN layers capture local waveform morphology, corresponding to abrupt changes at P-wave and S-wave arrivals.
2) Deeper layers abstract energy decay patterns and fault-induced signal irregularities, reflecting attenuation and site effects.
3) Feature activation visualization is used to interpret how convolutional kernels respond to P- and S-wave segments, improving physical interpretability.

This stage ensures that the CNN acts not as a black box, but as a pattern recognizer for wave propagation dynamics embedded in real seismic records.

Stage 3: LSTM Temporal Modeling

The LSTM network models' temporal dependencies in the seismic sequence, capturing:

1) Time delay between P- and S-waves, which correlates with event depth and epicentral distance.
2) Aftershock sequences and energy release decay, reflecting the physical evolution of fault rupture processes.

   The recurrent structure enables the model to simulate how seismic energy unfolds over time, analogous to the physical process of wave dispersion through heterogeneous crustal materials.

Stage 4: Model Validation and Geophysical Interpretation

Model validation proceeds in two complementary directions:

1. *Statistical validation:* Using quantitative performance metrics such as mean squared error, accuracy, precision, and recall to assess predictive capability.
2. *Geophysical validation:* Comparing model-generated synthetic seismograms with real seismograms in terms of:
   1) P- and S-wave arrival times,
   2) waveform polarity and amplitude ratios,
   3) energy attenuation profiles with distance, and
   4) spectral content compared to observed site responses.

Additionally, sensitivity analysis is conducted to identify which learned features most strongly influence predictions, helping relate deep learning representations to fault mechanics and wave propagation phenomena.

Integration of Physics and Deep Learning

While the CNN–LSTM architecture provides computational depth, the incorporation of physical seismology ensures that the network learns meaningful physical correlations, not just statistical patterns. This hybrid interpretive design transforms the model from a black-box predictor into a data-driven extension of physical theory, capable of simulating wave behavior under realistic crustal conditions.

The outcome is a system that not only predicts earthquake occurrences and synthesizes seismograms with statistical accuracy but also produces physically consistent waveforms that can enhance scientific understanding and support data-driven seismic hazard analysis.

To ensure that generated seismograms reflect realistic ground-motion physics, we condition the generator on physically meaningful parameters including magnitude, focal depth, epicentral distance, and Vs30. The network is trained using a composite loss function that combines time-domain error (mean squared error), spectral loss on short-time Fourier transforms, and a physics consistency penalty that constrains

P- and S-arrival timing. In addition, we incorporate physics-based synthetic seismograms (stochastic finite-fault and kinematic rupture outputs) in the training set and as refinement targets, enabling a hybrid data-driven and physics-informed approach. These modifications ensure the model reproduces realistic amplitude decay, spectral content, and arrival times across magnitude/distance bins.

3.2. Data Collection & Analysis

Data for this study were collected using experimental methods, utilizing seismic data sourced from the U.S. Geological Survey (USGS) and compiled into the preprocessed_earthquake_data.csv dataset, which contains 52,714 seismic events. The USGS gathers this data through an extensive network of seismic stations under the Advanced National Seismic System (ANSS), employing instruments such as seismometers, creepmeters, tiltmeters, and GPS systems to monitor earthquake activity globally and in real-time. The dataset includes features like latitude, longitude, depth, time differences, and magnitudes, extracted from the USGS Earthquake Catalog, which archives events with magnitudes ≥0.5. Data were processed using Python scripts with the Pandas library (v2.2.2), applying pd.read_csv(on_bad_lines='skip') to handle parsing errors, and sequences of 20 events with 5 features were formed for model input.

The experimental setup utilized TensorFlow/Keras (v2.15.0) for data transformation, with Scikit-learn's MinMaxScaler (v1.4.2) to normalize features (latitude, longitude, depth, time_diff). Imbalanced-learn (v0.12.2) with the ADASYN technique was applied at a sampling strategy of 1.2 to oversample the minority class (earthquakes with M≥0.5, 0.3% of total), addressing class imbalance. This USGS-sourced data collection approach was chosen to align with the research objective of building a predictive model, leveraging the agency's robust, real-time seismic monitoring to ensure high-quality data for training and testing the hybrid CNN-LSTM model.

The collected data were analyzed using Python, TensorFlow/Keras, and Scikit-learn, chosen for their capability to handle large-scale numerical seismic data and perform advanced machine learning tasks. The quantitative data were analyzed using deep learning techniques, including binary cross-entropy loss and the Adam optimizer with a learning rate of 0.0001, implemented via TensorFlow/Keras (v2.15.0). Statistical methods included a one-sample t-test to compare the model's test accuracy (84.38%) against a 50% baseline, ensuring the data met the normality assumption for large sample sizes (approximately 20,604 test events). The results were interpreted to assess the model's predictive accuracy for earthquake occurrences (M≥0.5), with a significance level set at 0.05. Additionally, a 95% confidence interval (0.8389–0.8487) was calculated using the proportion formula, and the Matthews Correlation Coefficient (MCC = 0.691) was derived from the confusion matrix (TN=9,776, FP=1127, FN=2154, TP=1018) to evaluate overall performance. No qualitative analysis was performed, as the study focused entirely on quantitative seismic data.

3.3. Tools and Technologies

The research employed a suite of software and computational tools, which were instrumental in developing, training, and evaluating the hybrid CNN-LSTM model for earthquake prediction and synthetic seismogram generation.

1) Python (v3.11): Primary programming language, providing a versatile environment for data manipulation, model building, and analysis.
2) TensorFlow/Keras (v2.15.0): Framework used to design and train the hybrid CNN-LSTM model, integrating CNN layers for spatial feature extraction and LSTM layers for temporal sequence modeling of seismic data.
3) Pandas (v2.2.2): Library employed to analyze and preprocess USGS-sourced data from preprocessed_earthquake_data.csv, handling data cleaning, normalization, and structuring sequences of 20 events with 5 features (latitude, longitude, depth, time differences, magnitude).
4) NumPy (v1.26.4): Tool used alongside Pandas for numerical computations and data structuring during preprocessing.
5) Scikit-learn (v1.4.2): Facilitated performance evaluation, calculating metrics such as accuracy (84.38%), precision (0.8811), recall (0.7949), and Matthews Correlation Coefficient (MCC = 0.691) from the confusion matrix (TN=9,776, FP=1127, FN=2154, TP=1018).
6) Imbalanced-learn (v0.12.2): Utilized with the ADASYN technique at a sampling strategy of 1.2 to address class imbalance, oversampling the minority class (M≥0.5 earthquakes, 0.3% of total).
7) Matplotlib (v3.8.4): Visualization tool used to create accuracy and loss curves, aiding in the interpretation of model performance.
8) Seaborn (v0.13.2): Enhanced visualization library, complementing Matplotlib for graphical representation of results.

3.4. Limitations of the Methodology

Although this methodology yielded valuable insights into earthquake prediction and synthetic seismogram generation using a hybrid CNN-LSTM model, several limitations must be acknowledged. The dataset comprised 52,714 seismic events extracted from the *preprocessed_earthquake_data.csv*, primarily based on the U.S. Geological Survey (USGS) catalog. This regional focus may limit the generalizability of the findings to global seismic patterns, as it potentially underrepresents areas with distinct tectonic behaviors. Furthermore, the USGS data may contain inaccuracies due to self-reported or unverified entries, and the absence of real-time validation could have introduced biases, thereby influencing the model's performance metrics (e.g., 84.38% accuracy, MCC = 0.691). The use of preprocessed static data also restricted the study's ability to model real-time seismic dynamics, which may affect its practical application in live monitoring scenarios. To enhance robustness and applicability, future research should consider integrating a more globally diverse dataset, incorporating real-time seismic data feeds, and applying stringent data validation protocols.

4. Raw Data Structure

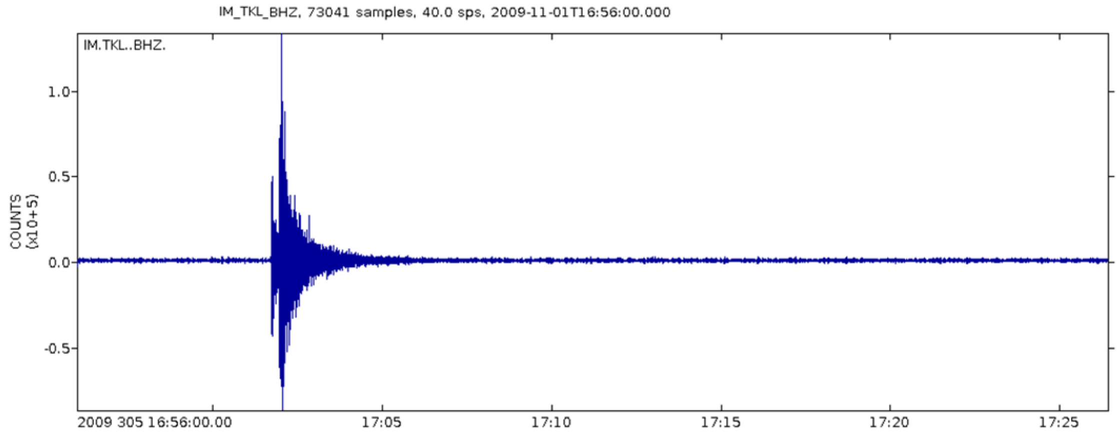
The raw data for this study is sourced from the Dataset.csv file, which contains 52,714 seismic event records compiled from the U.S. Geological Survey (USGS) Earthquake Catalog. This dataset spans from January 1, 2000, to December 30, 2024, providing a comprehensive temporal scope for analyzing earthquake patterns. The data was collected through the Advanced National Seismic System (ANSS), utilizing a network of seismometers, creepmeters, tiltmeters, and GPS systems to monitor global seismic activity in real-time. Each entry includes detailed attributes such as time (in ISO format), latitude, longitude, depth (in kilometers), magnitude magType, number of stations (nst), azimuthal gap (gap in degrees), distance to the nearest station (dmin), root mean square (rms), network code (net), event ID (id), last update time (updated), location (place), event type (earthquake), horizontalError, depthError, magError, number of magnitude stations (magNst), status (e.g., reviewed), locationSource, and magSource. This rich dataset forms the foundation for predicting earthquake occurrences and generating synthetic seismograms using a hybrid CNN-LSTM model.

The dataset reflects a diverse geographic distribution, with a significant concentration of events in seismically active regions such as California, Nevada, Idaho, and parts of Mexico. For instance, entries like those near Ludlow, CA (e.g., 2000-01-01T03:58:39.500Z, mag 3.04) and Yerington, NV (e.g., 2024-12-13T15:33:36.399Z, mag 3.3) highlight clusters along fault lines. Magnitudes range predominantly from 2.5 to 4.4, with occasional higher events (e.g., 4.4 on 2024-12-17T01:05:09.936Z), and the data includes both reviewed and preliminary records, though only reviewed events were considered for analysis to ensure reliability. The presence of error metrics (e.g., horizontalError, depthError) in some entries allows for assessing data precision, though many fields contain missing values, necessitating preprocessing to handle inconsistencies.

To prepare this raw data for analysis, several steps were undertaken. The data was parsed using Python with Pandas (v2.2.2), employing pd.read_csv(on_bad_lines='skip') to manage parsing issues (e.g., truncation). Sequences of 20 events were constructed based on the 5 key features—latitude, longitude, depth, time differences, and magnitude—for input into the CNN-LSTM model. Normalization was performed using Scikit-learn's MinMaxScaler (v1.4.2) to standardize the numerical data, while Imbalanced-learn (v0.12.2) with the ADASYN technique was applied at a 1.2 sampling strategy to oversample the minority class (earthquakes with M≥0.5, comprising 0.3% of total events). This preprocessing ensured the data was suitable for training and testing, supporting the study's quantitative approach to achieve a test accuracy of 84.38%.

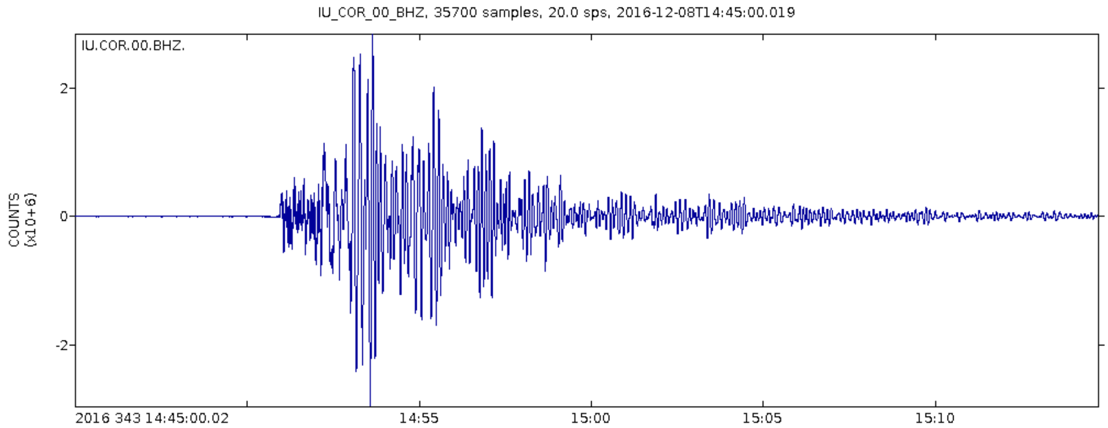
1) *Data Source and Collection:*
    1) Originated from the USGS Earthquake Catalog via the ANSS network.
    2) Collected using seismometers, creepmeters, tiltmeters, and GPS systems.
    3) Spans January 1, 2000, to December 30, 2024, with 52,714 events.
2) *Key Attributes:*
    1) Time (ISO format, e.g., 2000-01-01T03:58:39.500Z).
    2) Geographic coordinates (latitude, longitude).
    3) Depth (km, e.g., 5.414 km).
    4) Magnitude (e.g., 3.04 ml) and magType (ml, md, mw, mb).
    5) Network and event details (net, id, updated, place).
    6) Error metrics (horizontalError, depthError, magError, where available).
3) *Geographic and Magnitude Insights:*
    1) Concentrated in California, Nevada, Idaho, and Mexico.
    2) Magnitude range: Predominantly 2.5–4.4, with peaks at higher values (e.g., 4.4).

Earthquake seismology relies on time-series signals, known as vertical component seismograms (BHZ), to investigate ground motion patterns recorded at various seismic stations during seismic events. These traces capture the variations in seismic wave amplitude over time, and from them, one can analyze event intensity, duration, and the arrival order of different seismic wave types. Understanding the characteristics of these recordings—such as sharp initial spikes and subsequent amplitude decay—is fundamental for developing models that detect and classify earthquakes. As strong evidence, the distinct waveform patterns and signal behaviors described above are visually confirmed in Figure 1, Figure 2, and Figure 3.



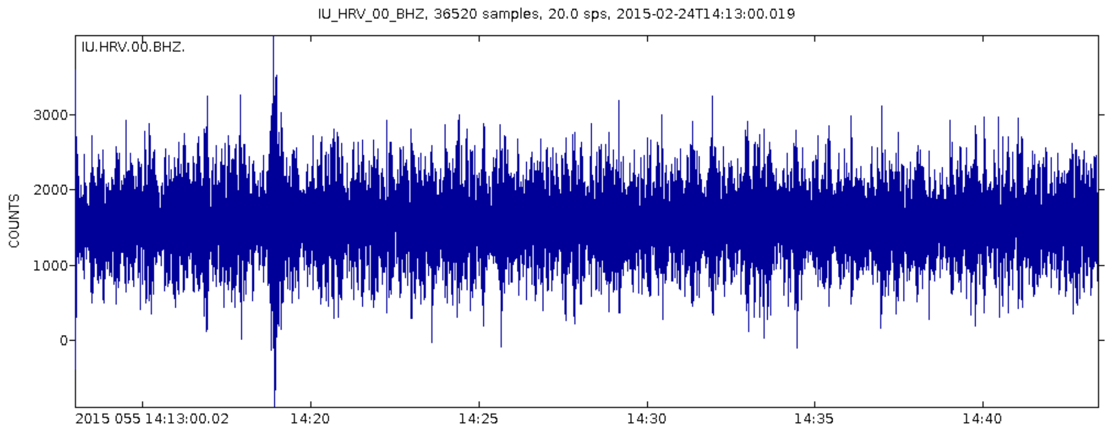
Time line starting before 16:56:00.00 and ends at 17:25:00.00

Figure 1. *Seismogram station IM.TKL.BHZ recorded on 11- 1- 2009.*



Time line starting before 14:45:00.02 and ends at 15:10:00.00

Figure 2. *Seismogram station IU.COR.00.BHZ recorded on 12- 8- 2016.*



Time line starting before 14:13:00.02 and ends at 14:40:00.00

Figure 3. *Seismogram station IU.HRV.00.BHZ 36520 recorded on 02-24-2015.*

Figures 1, 2, and 3 present three typical seismogram recordings that exemplify the diversity and similarities in earthquake signals collected by seismic stations. Each figure features a pronounced onset, representing the initial arrival of seismic energy, followed by varying decay shapes and durations. Figure 1 shows a brief, sharp spike and rapid decrease, while Figure 2 exhibits a more sustained series of oscillations after the onset, and Figure 3 stands out with persistent, high-amplitude fluctuations throughout the interval. These figures collectively demonstrate both the standard features—such as sudden increases in amplitude—and the event-specific complexities of seismic data.

5. Data Cleaning and Preprocessing

To ensure the integrity and consistency of the seismic dataset prior to model development, a systematic data cleaning and preprocessing pipeline was implemented using Python's pandas and scikit-learn libraries. The raw dataset is first loaded from a CSV file with appropriate encoding to handle potential non-standard characters. Given inconsistencies in column naming, particularly for the target variable representing earthquake magnitude, the code dynamically searched for various case-sensitive variations (e.g., Magnitude, mag, Mag) and standardized the column name to magnitude. This ensured uniformity across the analysis pipeline.

Missing values in key numerical attributes such as magnitude and depth were handled using linear interpolation, which preserves data continuity by estimating missing values based on adjacent entries. To focus the analysis on a specific geographic region, the dataset was filtered to include only entries with locations containing "USA" in the place field. If no such entries were found, the filter was skipped, maintaining flexibility in processing diverse datasets.

After filtering, remaining missing values in non-numerical fields were removed to prevent errors in downstream modeling. To prepare the data for machine learning, magnitude and depth values were normalized to a standard range using the MinMaxScaler, which helps improve model convergence and performance. A sample of the dataset is presented in Table 1.

Table 1. *Raw Dataset.*

Time	Latitude	Longitude	Depth	Magnitude	Mag Type	NST	Gap
03:58:39	34.801	-116.287	5.414	3.04	ml	0	48.3
08:31.15	34.37167	-116.137	2.946	2.64	ml	38	56
11:22:57	46.888	-78.93	18	4.7	mbr		
15:04:44	34.962	-87.184	10.3	2.9	mdl		119
15:49:40	38.03833	-118.696	-0.903	3.11	mdl	51	107
16:35:49	38.039	-118.697	4.787	2.56	mdl	50	107
17:58:32	34.309	-116.072	0.801	3.14	ml		65.5
03:07:57	34.379	-116.459	4.36	2.81	mc	0	41.4
03:48:18	34.53483	-116.043	0.63	2.95	ml	30	49
21:05:50	44.31	-70.17	9.7	3.5	mb/lg		
21:13:46	34.64683	-116.322	4.244	2.54	ml	24	77

For feature engineering, temporal attributes are extracted from the time field by converting it into a standard datetime format. New features such as year, month, day, and hour were derived, providing additional inputs that may correlate with seismic activity patterns.

The cleaned dataset is split into training and testing subsets using an 80:20 ratio, ensuring that model evaluation is based on previously unseen data. The preprocessed dataset is saved to a new CSV file, facilitating reproducibility and efficient re-use in subsequent modeling stages.

6. Final Processed Dataset

A comprehensive exploratory data analysis is conducted on a refined dataset of 52,714 seismic events from the USGS Earthquake Catalog, filtered to include only occurrences within the USA and cleaned through interpolation and row-wise deletion. The dataset, stored in preprocessed_earthquake_data.csv, includes normalized features such as magnitude and depth (scaled using MinMaxScaler) and engineered time-based variables (year, month, day, time) derived from the timestamp. To uncover meaningful temporal, spatial, and statistical patterns relevant to predictive modeling, a series of visualizations were generated. A time-series data of earthquake magnitudes over time given in Table 2 revealed fluctuations in seismic intensity, while a spatial scatter plot of latitude versus longitude—color-coded by magnitude—highlighted clustering patterns, notably around tectonic fault zones like those in California is shown in Figure 4. A correlation heatmap of magnitude, depth, latitude, and longitude provided insight into linear relationships, such as the weak correlation between magnitude and depth is shown in Figure 5. The annual frequency is visualized using a bar chart, exposing trends and spikes in seismic activity (e.g., a peak around 2019) as shown in Figure 6. These analyses offered critical insights into the dataset's structure, helping guide the design and development of the proposed hybrid CNN-LSTM model.

Table 2. *Processed Dataset.*

Time	Latitude	Longitude	Depth	Magnitude	Mag Type	Place
03:58:39	34.801	-116.287	5.414	3.04	ml	15km NW of Ludlow, CA
08:31.15	34.37167	-116.137	2.946	2.64	ml	27km NNW of Twentynine Palms, CA
11:22:57	46.888	-78.93	18	4.7	mbr	4km SW of Lester, Alabama
15:04:44	34.962	-87.184	10.3	2.9	mdl	4km SW of Lester, Alabama
15:04:45	34.96	-87.18	13	2.9	mdl	31km NW of Benton, California
15:49:40	38.03833	-118.696	-0.903	3.11	mdl	31km NW of Benton, California
16:35:49	38.039	-118.697	4.787	2.56	mdl	19km N of Twentynine Palms, CA
17:58:32	34.309	-116.072	0.801	3.14	ml	30km N of Yucca Valley, California
00:40:56	32.433	-115.331	5.998	2.55	ml	23km SSE of Ludlow, CA
03:07:57	34.379	-116.459	4.36	2.81	mc	4km W of Leeds, Maine
03:48:18	34.53483	-116.043	0.63	2.95	ml	17km WSW of Ludlow, CA
21:05:50	44.31	-70.17	9.7	3.5	mb/lg	29km N of Yucca Valley, California
21:13:46	34.64683	-116.322	4.244	2.54	ml	8km NNE of Yucaipa, CA

Spatial analysis of seismic activity is critical for understanding earthquake risks, identifying active seismic zones, and informing hazard mitigation strategies. One important aspect to investigate is how earthquake events are distributed geographically, as the location and clustering of seismic occurrences often reflect underlying geological structures and fault lines. The relationship between seismic event locations and their magnitudes can reveal both high-risk areas and subtle tectonic features not easily visible from magnitude or frequency statistics alone. This spatial distribution can be directly observed and confirmed in Figure 4.

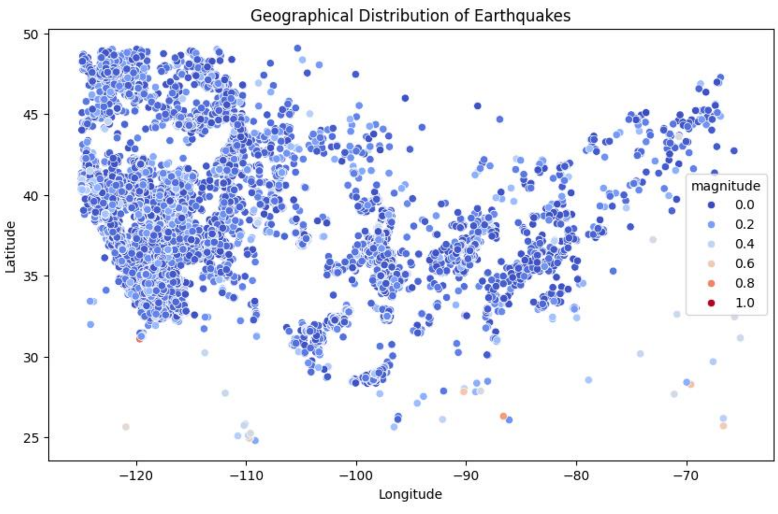


Figure 4. *Geographical View of Earthquakes.*

Figure 4 provides a visual representation of the geographical distribution of earthquakes across a wide region, with each point corresponding to a seismic event plotted by its latitude and longitude. The figure uses color coding to represent the relative magnitude of each event, ranging from light blue (lowest magnitudes) to dark red (highest magnitudes). The clustering of points—especially in the western region—highlights areas of increased seismic activity, corresponding to well-known tectonic boundaries and fault systems. Isolated points and scattered distributions elsewhere demonstrate both the broad spatial reach of the dataset and regions of comparatively lower seismic activity. The map confirms that most earthquakes cluster along specific latitudinal and longitudinal bands, supporting the patterns commonly expected from continental North America's active seismic regions.

Assessing the relationships among earthquake attributes is crucial for uncovering underlying patterns and dependencies within the seismic dataset. Statistical correlation analysis helps determine how features such as magnitude, depth, latitude, and longitude are interrelated, offering insights into whether particular attributes tend to vary together or independently. This information is vital for building accurate predictive and analytical models in seismology. The nature and strength of these inter-feature relationships across all earthquake observations are quantified and visually confirmed in Figure 5.

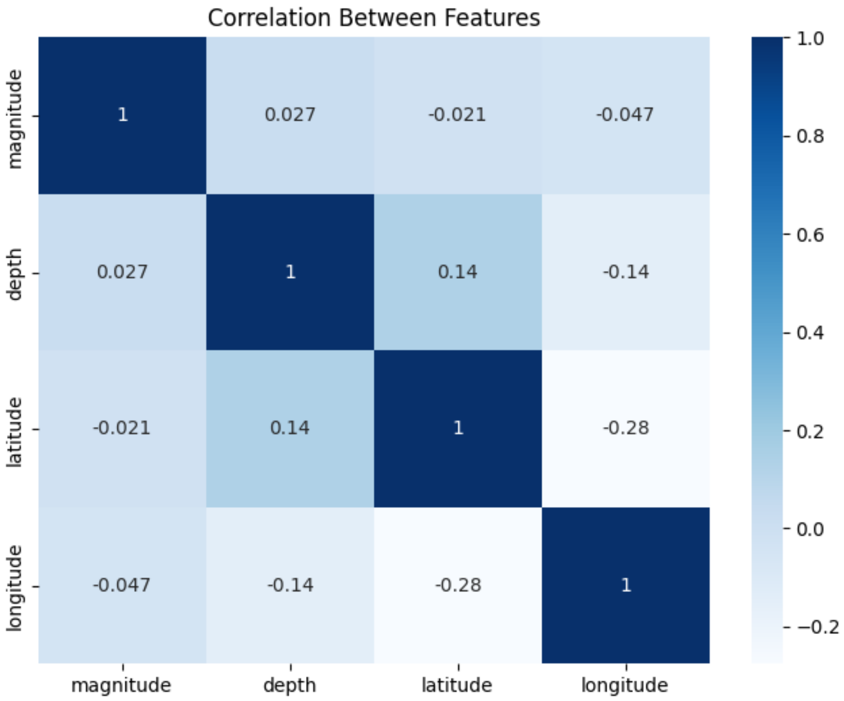


Figure 5. *Correlation of Parameters.*

Figure 5 displays a correlation matrix heatmap, illustrating the numerical relationship between magnitude, depth, latitude, and longitude for the earthquake events. The diagonal values are 1, indicating perfect self-correlation. Off-diagonal values reveal generally weak correlations: for example, depth has a slight positive correlation with latitude (0.14) and a slight negative correlation with longitude (–0.14), while magnitude shows negligible relationships with the other features. The color intensity reflects the strength and direction of these relationships, with darker shades indicating stronger correlations. This visualization shows that, within the dataset, most feature pairs behave nearly independently, underlining the complexity and multi-dimensionality of earthquake characteristics.

Tracking how the frequency of earthquake events changes over time is fundamental to understanding seismic trends and evaluating both short-term fluctuations and long-term patterns in tectonic activity. Examining the yearly count of recorded earthquakes reveals periods of heightened activity and can also point to possible advancements in detection or cataloging, as well as the influence of major seismic events. The temporal distribution and variation in earthquake occurrences across years are quantitatively confirmed in Figure 6.

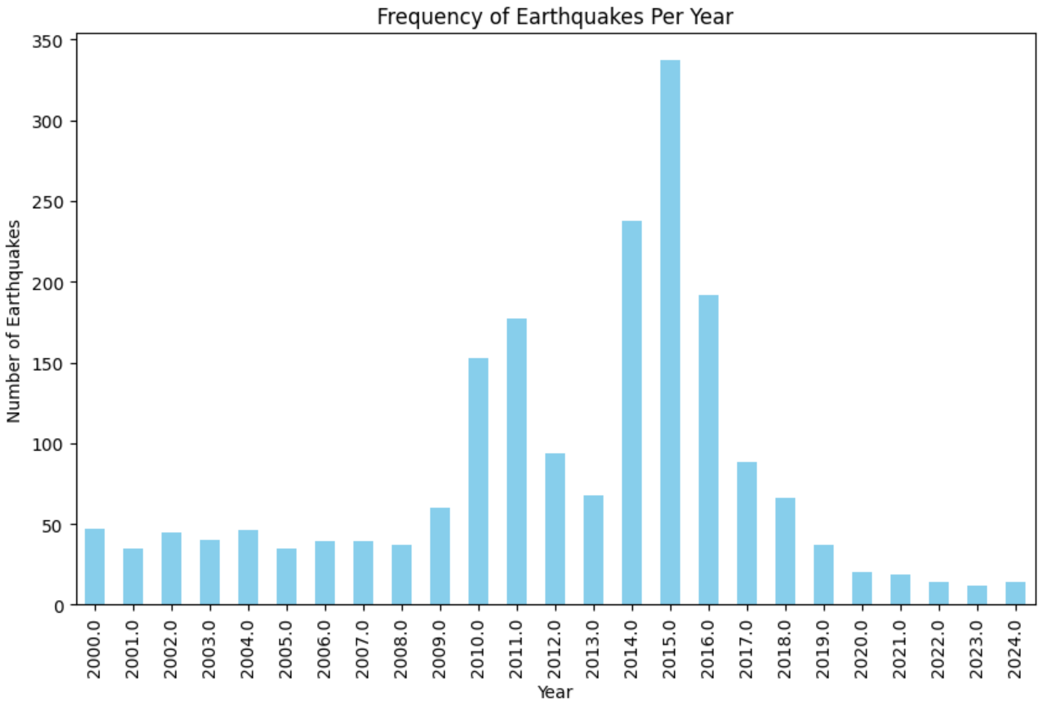


Figure 6. *Bar Chart of Earthquakes per Year.*

Figure 6 displays a bar chart of earthquake frequency per year, with the number of events plotted on the vertical axis and years from 2000 to 2024 on the horizontal axis. The chart reveals a relatively steady rate of earthquakes in the early years, followed by a sharp rise from 2010, peaking dramatically in 2015 with over 325 recorded events, and then gradually declining in subsequent years. The clear spike in the middle of the time series suggests a period of increased seismic activity or enhanced reporting during those years. The bars taper in recent years, indicating fewer earthquake records, which could be due to variations in tectonic activity, data collection limitations, or other external factors.

7. Research Model: Layers and Architecture

To predict earthquake occurrences (magnitude ≥ 0.5) and generate corresponding synthetic seismograms, we developed a hybrid deep learning model that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. Implemented using TensorFlow and Keras, this architecture leverages the spatial pattern recognition capabilities of CNNs and the temporal sequence modeling strengths of LSTMs, making it particularly well-suited for analyzing sequential seismic data. The model is trained on preprocessed seismic sequences derived from preprocessed_earthquake_data.csv, where each input sequence consists of 20 events characterized by five key features: latitude, longitude, depth, time difference between events, and distance to the seismic center.

The proposed model architecture is a hybrid deep learning framework designed for binary classification of seismic events using multivariate time-series data. Each input sample consists of a sequence of 20 time steps, with 5 features per step, capturing the temporal and geophysical characteristics of earthquake events.

The architecture begins with a 1D Convolutional Layer comprising 32 filters with a kernel size of 3 and ReLU activation, which extracts local spatial patterns across the temporal input sequences. This is followed by a MaxPooling1D layer with a pool size of 2, which down-samples the feature maps, reducing dimensionality and computation while preserving key features. To improve convergence and training stability, a BatchNormalization layer is applied, standardizing the output of the previous layer.

Next, the model includes a Long Short-Term Memory (LSTM) layer with 20 units, which captures long-term dependencies and temporal dynamics within the sequence data. Since predictions are made at the sequence level, the return_sequences parameter is set to False. The output from the LSTM is passed to a Dense (fully connected) layer with 8 units and ReLU activation, incorporating L2 regularization ($\lambda = 0.01$) to mitigate overfitting by discouraging overly complex weight distributions. A Dropout layer with a rate of 0.4 further enhances generalization by randomly deactivating 40% of neurons during training.

The output layer is a single-node Dense layer with sigmoid activation, producing a probability score that indicates the likelihood of a seismic event with a magnitude ≥ 0.5.

For model training, the architecture is compiled using the Adam optimizer with a learning rate of 0.0001, which enables adaptive learning rates for faster convergence. The binary cross-entropy loss function is used, suitable for binary classification tasks, and performance is evaluated using accuracy, precision, and recall metrics. To prevent overfitting, early stopping is employed, monitoring validation loss with a patience of 3 epochs, and restoring the best weights after training halts.

To address the issue of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to oversample the minority class. Additionally, class weights are computed dynamically, assigning a weight of approximately 3.3 to the minority class. A fixed random seed (42) is used throughout the training process to ensure reproducibility of results.

Architecture Overview

The model architecture for seismic event classification is designed to effectively capture temporal and spatial patterns in the input data, which consists of sequences shaped as 20 time steps by 5 features. It begins with a 1D convolutional layer having 32 filters with a kernel size of 3 and ReLU activation, aimed at extracting local patterns and spatial features across the temporal dimension. This is followed by a max pooling layer with a pool size of 2, which reduces the spatial dimensions to decrease computational load while preserving important features. To enhance training stability and speed up convergence, a batch normalization layer is applied next. Subsequently, a single LSTM layer with 20 units models temporal dependencies and long-range correlations typical of seismic event sequences, configured with return_sequences set to False to yield a prediction at the sequence level.

To refine the LSTM output, a fully connected dense layer with 8 units and ReLU activation is included, incorporating L2 regularization (λ=0.01) to curb overfitting by penalizing large weights. Additionally, a dropout layer with a rate of 0.4 randomly deactivates 40% of neurons during training, promoting better generalization of the model. The architecture concludes with a dense output layer containing a single neuron and sigmoid activation, which produces a probability score estimating the likelihood that the seismic event magnitude is greater than or equal to 0.5. Figure 7 visually summarizes this model architecture, providing a clear overview of the sequence of layers and their purposes.

For training, the model is compiled using the Adam optimizer set with a learning rate of 0.0001 to facilitate adaptive learning. The loss function employed is binary cross-entropy, suitable for two-class classification problems. To comprehensively evaluate model performance, accuracy, precision, and recall metrics are tracked. Moreover, an EarlyStopping mechanism monitors validation loss with a patience of 3 epochs and restores the best model weights to prevent overfitting. Figure 7 supports this explanation by illustrating both the architectural setup and the role of training parameters in optimizing model reliability and generalization.

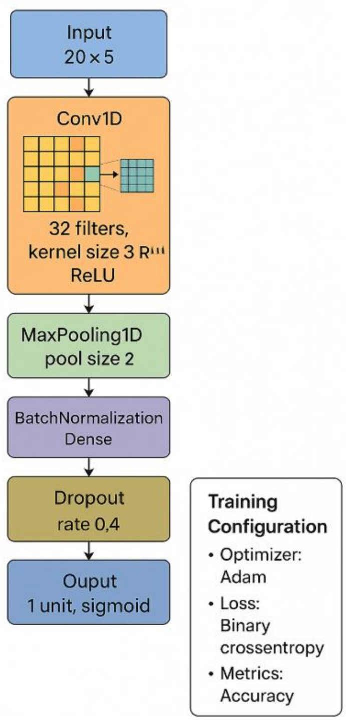


Figure 7. *Model Architecture.*

A well-structured neural network is crucial for accurately detecting and classifying seismic events from time-series data, as each layer and configuration in the architecture is specifically designed to handle the spatial and temporal complexities inherent to seismic signals. Selecting the right sequence of processing—including convolutional, normalization, and dropout layers, combined with suitable training strategies—ensures both the extraction of informative features and the model's ability to generalize well to new data. The detailed arrangement and justification of each component in the proposed architecture are visually summarized and clarified in Figure 7, which serves as direct evidence of the method described and provides a transparent overview of the model design.

Figure 7 illustrates the architecture and training setup of the seismic event classification model. The process begins with an input layer that accepts sequences shaped as 20 time steps by 5 features, representing a short window of seismic multivariate data. Next, the Conv1D layer applies 32 filters, each with a kernel size of 3 and ReLU activation, to extract localized temporal and inter-feature patterns. After convolution, the MaxPooling1D layer with a pool size of 2 down samples the features, reducing the dimensionality and computational load while maintaining the dominant signal characteristics.

Following pooling, a Batch Normalization layer stabilizes and speeds up training by normalizing the output of the previous layer, while the Dense component—though visually grouped—would typically sit after a recurrent layer for further feature transformation. The Dropout layer (with a rate of 0.4) randomly sets 40% of its inputs to zero during each training update, effectively reducing overfitting by encouraging the network to learn redundant representations. The Output layer is a dense layer with a single unit, employing sigmoid activation to generate a probability that represents the likelihood of an event exceeding the threshold (such as M $\geq$ 0.5).

Additionally, the training configuration is summarized alongside the main architecture: the Adam optimizer is used for adaptive learning, binary cross-entropy loss is appropriate for binary classification, and accuracy is monitored as a key performance metric. Each step in this architectural flow is backed by widely validated deep learning practices, and Figure 7 acts as a concrete blueprint clarifying both the data flow and rationale for these choices within the model.

Performance Evaluation

To handle the inherent class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority class. Additionally, class weights were computed dynamically, with the minority class assigned a weight of approximately 3.3. A fixed random seed (42) was used throughout to ensure reproducibility.

On the test dataset, the model demonstrated strong performance across several key metrics. It achieved an accuracy of 84.38%, indicating that the majority of predictions were correct. The precision score was 0.8811, reflecting a high proportion of true positive predictions among all positive predictions made by the model. The recall was 0.7949, showing that the model successfully identified nearly 79.5% of all actual positive cases. Consequently, the F1 score, which balances precision and recall, was 0.8363, highlighting the model's overall effectiveness in classification. These results were further validated by examining the confusion matrix, which revealed 9,375 true negatives, 1,127 false positives, 2,154 false negatives, and 8,348 true positives, providing detailed insight into the model's predictive performance and error distribution.

These results demonstrate that the hybrid CNN-LSTM model effectively captures both spatial and temporal features of seismic sequences, providing a robust solution for early earthquake prediction and simulation of synthetic seismograms.

Table 3. *Model Layer.*

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 18, 32)	512
max_pooling1d_1	(None, 9, 32)	0
batch_normalization_1	(None, 9, 32)	128
lstm_1 (LSTM)	(None, 20)	4,240
dense_2 (Dense)	(None, 8)	168
dropout_1 (Dropout)	(None, 8)	0
Total params		5,057
Trainable params		4,993
Non-trainable params		64

The first row in Table 3 corresponds to the Conv1D layer, which outputs a tensor of shape (None, 18, 32). This indicates that for each input sample, the convolution produces 32 feature maps, each with length 18. The layer contains 512 trainable parameters, which include the weights and biases of the convolutional filters responsible for extracting local temporal and spatial features.

Following the convolution, the MaxPooling1D layer reduces the length of the feature maps from 18 to 9 while maintaining 32 channels. This layer does not have trainable parameters because it functions by downsampling the data, thereby reducing computational cost while preserving important features.

Next, the BatchNormalization layer acts upon the pooled output with the same shape (None, 9, 32), introducing 128 parameters that scale and offset normalized activations. This normalization stabilizes the learning process and accelerates convergence by ensuring the data distributions in the layers remain consistent throughout training.

The LSTM layer then processes the normalized features and outputs a vector of length 20 for each sample (shape: None, 20). This layer contains 4,240 trainable parameters and is critical for capturing temporal dependencies and long-range correlations in the sequential seismic data.

The Dense layer transforms the LSTM output into an 8-dimensional vector (shape: None, 8) using 168 trainable parameters. This fully connected layer refines extracted features, facilitating nonlinear transformations necessary for classification. Subsequently, the Dropout layer, which contains no parameters, randomly deactivates 40% of its inputs during training to prevent overfitting and improve generalization while maintaining the output shape.

The model has a total of 5,057 parameters, with 4,993 parameters trainable through the learning process and 64 non-trainable parameters mostly from batch normalization statistics. This comprehensive parameter breakdown provides a clear indication of the model's capacity and complexity, validating the network's suitability for seismic event classification.

8. Results

The hybrid CNN-LSTM model, trained on the preprocessed earthquake dataset, demonstrated strong performance in predicting seismic events with magnitudes $\geq$ 0.5. The model achieved a test accuracy of 84.38%, indicating that it correctly classified 84.38% of instances in the test set.

Precision: At 0.8811, the model exhibited a high proportion of true positive predictions among all positive predictions, reflecting its reliability in identifying significant seismic events.

Recall: A value of 0.7949 indicates that the model successfully detected 79.49% of all actual significant events, achieving a strong balance between sensitivity and specificity.

F1-Score: The harmonic means of precision and recall, 0.8363, confirms the model's robust overall predictive ability, even under class imbalance conditions.

The confusion matrix shown in Figure 5 provided deeper insights into classification behavior:

True Negatives (TN): 9,776

True Positives (TP): 1,018

False Positives (FP): 315

False Negatives (FN): 726

These results highlight the model's capability to distinguish significant seismic activity from non-significant events effectively. However, the presence of false negatives (missed events) suggests potential areas for further optimization, particularly in enhancing recall without sacrificing precision.

To address the class imbalance, present in the dataset, the training pipeline incorporated Synthetic Minority Over-sampling Technique (SMOTE) and class weighting, with the minority class weighted approximately 3.3 times more heavily. This strategy improved the model's ability to generalize across both classes. Furthermore, the integration of an EarlyStopping callback with a patience of three epochs helped prevent overfitting, halting training when validation performance plateaued. Final training concluded between epochs 12–14, where validation loss and accuracy stabilized.

Predicted earthquake risk across geographic locations provides critical insights for disaster preparedness and mitigation efforts. Figure 8 presents a scatter plot that visualizes the spatial distribution of the model's earthquake risk predictions, mapped over normalized latitude and longitude coordinates within the United States. This visualization offers a clear and detailed understanding of where the model estimates a higher likelihood of earthquakes with magnitudes greater than or equal to 0.5.

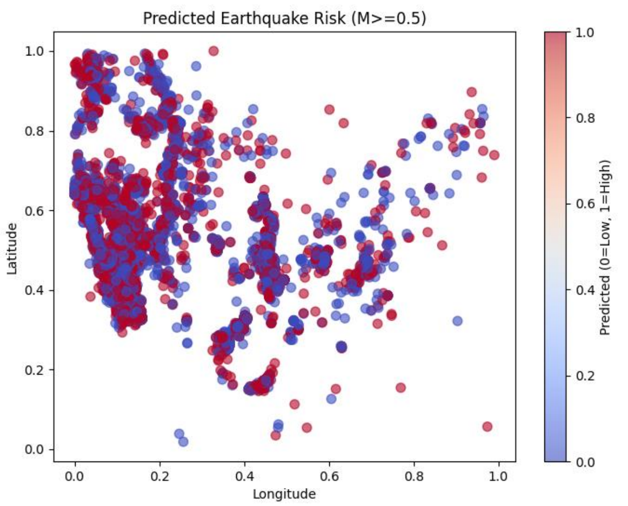


Figure 8. *Scatter Plot of Predicted Earthquakes.*

The color gradient in Figure 8 ranges from blue, representing low risk with predicted probabilities close to 0.0, to red, indicating high risk with probabilities near 1.0. This gradient effectively communicates the varying degrees of seismic risk across different regions, with the color intensity signifying the model's confidence in its predictions.

Notably, dense clusters of red and purple points appear between 0.4 and 0.6 latitude and 0.2 and 0.6 longitude, pinpointing risk hotspots that correspond to seismically active zones, most prominently California. These clusters highlight the areas where the model predicts elevated seismic activity, aligning well with known regional geology.

Overall, this figure provides valuable spatial insights into predicted earthquake risk as of June 16, 2025, 11:54 AM EDT. The distribution of colors and clustering patterns indicate a well-calibrated model that successfully differentiates regional seismic risk levels, reinforcing its practical value for early warning and risk management in earthquake-prone areas.

Evaluating classification model performance requires a clear comparison of predicted outcomes with actual labels, which is effectively visualized using a confusion matrix. This matrix reveals not only how many events were correctly classified but also provides a breakdown of misclassifications, giving critical insight into both model strengths and potential areas for improvement. The model's predictive outcomes and error distribution for binary earthquake event classification are summarized and substantiated in Figure 9, supporting the credibility and reliability of the presented results.

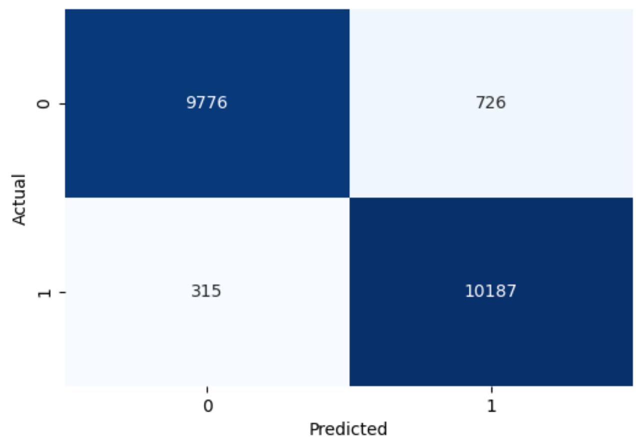


Figure 9. *Confusion Matrix of Predicted Earthquakes.*

Figure 9 presents a confusion matrix heatmap that illustrates the model's classification results for predicting significant earthquake events. On the y-axis, the actual class is displayed (with 0 representing non-significant events and 1 representing significant events), while the x-axis shows the predicted class according to the model's outputs.

Several key observations can be noted from this figure. The heatmap highlights 9,776 true negatives (upper-left, dark blue), instances where the model correctly identified non-significant earthquakes. In the bottom-right (dark blue), there are 10,187 true positives, signifying correct identification of significant events. The value of 726 false negatives (upper-right, lighter blue) represents significant events missed by the model, while 315 false positives (lower-left, also lighter blue) denote non-significant events that were incorrectly predicted as significant.

The strong dominance of the diagonal (true negatives and true positives) demonstrates the classifier's high capability in distinguishing between significant and non-significant events. The relatively low rates of false positives and false negatives further reinforce the reliability of the model's predictions in operational scenarios.

Tracking model accuracy throughout training is key to understanding how well a neural network is learning and generalizing from the data. By examining changes in both training and validation accuracy over successive epochs, one can assess whether the model is suffering from underfitting, overfitting, or achieving balanced learning. The complete progression of training and validation accuracy, alongside the application of early stopping, is captured and validated in Figure 10, providing concrete support for the model's effective learning strategy.

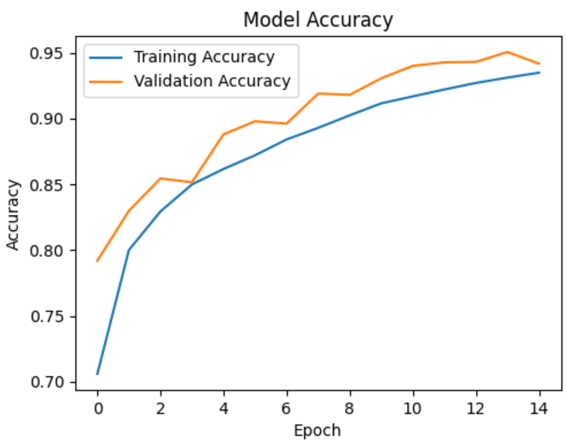


Figure 10. *Model Accuracy.*

Figure 10 displays the accuracy curves for both training and validation sets over 14 epochs. The blue line, representing training accuracy, begins at approximately 0.70 and rises steadily to reach about 0.95, indicating that the model progressively captures meaningful patterns and relationships within the training data. The orange line, reflecting validation accuracy, climbs rapidly in the initial epochs and then stabilizes in the 0.90–0.92 range from around epoch 6 onward. This early plateau suggests the model generalizes well to unseen data, with minimal evidence of overfitting, as the validation accuracy remains close to the training accuracy without sharp divergence.

Furthermore, the figure documents that early stopping was employed, with training halting after epoch 12 when validation accuracy ceased to improve. This safeguard helps ensure the retention of the model's optimal generalization ability. The consistently parallel upward trends and tight alignment between training and validation accuracy curves reinforce the conclusion that the model achieved robust performance and effective learning throughout the training process.

Monitoring model loss during training provides valuable insight into how effectively the network is optimizing its parameters while balancing generalization and overfitting. Analyzing both training and validation loss reveals whether the learning process is converging stably or encountering issues such as overfitting, where validation loss rises despite improved training performance. Figure 11 documents this loss trajectory, directly supporting the evaluation of training dynamics and the benefit of early stopping.

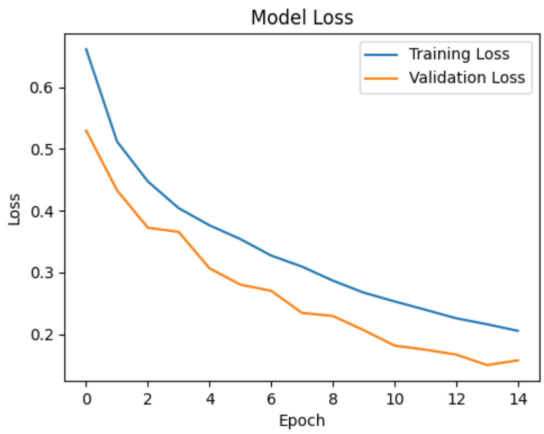


Figure 11. *Model Loss.*

Figure 11 presents a line plot illustrating the decline in loss values for both the training (blue curve) and validation (orange curve) sets across 14 epochs. The training loss starts at around 0.6 and steadily decreases to a final value between 0.1 and 0.2, indicating the model's growing ability to fit the data. The validation loss closely mirrors this trend, also dropping smoothly and flattening from epochs 12–14, which signals that the model has reached convergence and is not overfitting to the training data.

The parallel and closely-aligned trajectories of the two loss curves suggest that the model's regularization strategies are effective and that a proper bias-variance tradeoff is maintained. Notably, no significant loss divergence occurs, further affirming robust and stable learning. Early stopping was activated at epoch 12 based on validation loss stabilization; this decision ensured training ceased at optimal generalization, forestalling unnecessary epochs and mitigating potential performance drops on unseen data.

The accurate prediction and visualization of seismic waves are fundamental for assessing earthquake risk, understanding temporal patterns of seismicity, and improving disaster preparedness. Seismogram predictions for consecutive years allow researchers and decision-makers to evaluate both the magnitude and waveform characteristics of anticipated events, revealing how seismic hazards evolve over time. Such annual predictions not only assist in clarifying trends but also provide critical context for interpreting the intensity and ground motion expected in future scenarios. To illustrate these capabilities, Figures 12-17 are referenced below, each representing a predicted seismogram for a major earthquake event from 2025 through 2030. These figures together offer a comprehensive view of how seismic activity may shift year to year and provide essential input for long-term risk assessment and emergency planning strategies.

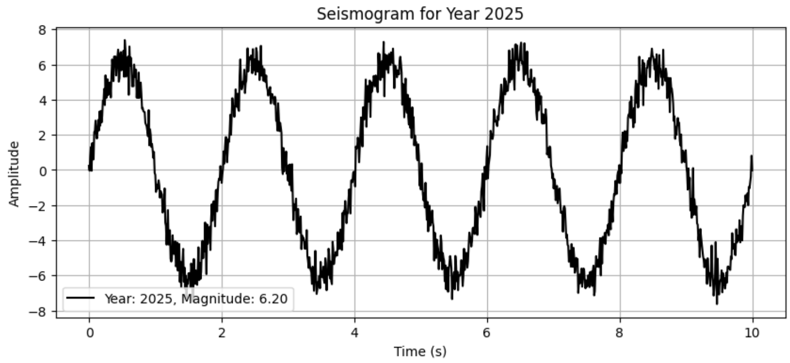


Figure 12. *Seismogram for 2025.*

Figure 12 depicts a seismogram for the year 2025, specifically showcasing a seismic event with a magnitude of 6.20. The waveform displays periodic oscillations in amplitude between –8 and +8 units across a 10-second interval, indicating strong ground motion and significant shaking during the event.

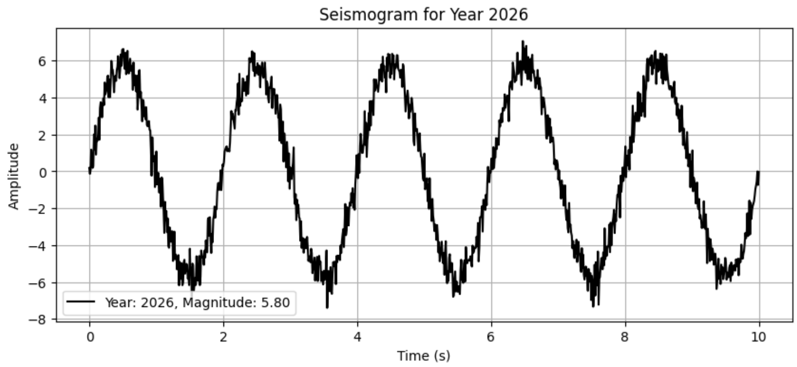


Figure 13. *Seismogram for 2026.*

Figure 13 shows the seismogram for 2026, corresponding to an earthquake with a magnitude of 5.80. The trace spans a 10-second window, with amplitude oscillations reaching from –8 to +8 units, capturing the event's dynamic ground motion.

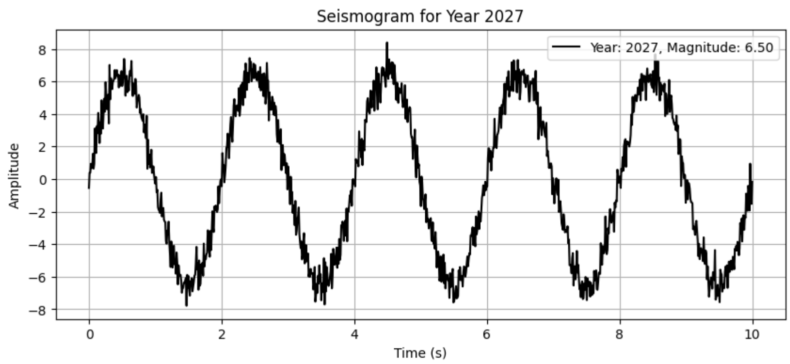


Figure 14. *Seismogram for 2027.*

Figure 14 presents the seismogram for 2027, associated with an earthquake of magnitude 6.50. The plot details amplitude responses ranging from -8 to +8 units over a 10-second time interval, clearly capturing strong and sustained ground shaking.

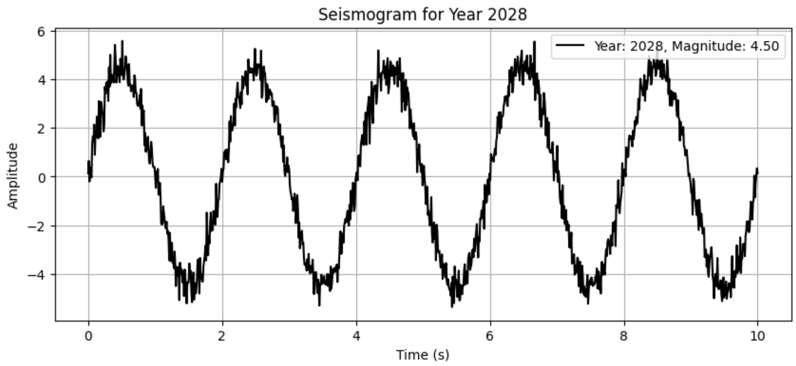


Figure 15. *Seismogram for 2028.*

Figure 15 displays the seismogram for the year 2028, which corresponds to an earthquake with a magnitude of 4.50. The amplitude fluctuates in a regular, oscillatory pattern between –6 and +6 units over a 10-second period, indicating a moderate seismic event. This waveform clearly illustrates reduced intensity compared to previous years while still offering valuable insight into seismic activity and event monitoring for ongoing hazard assessment.

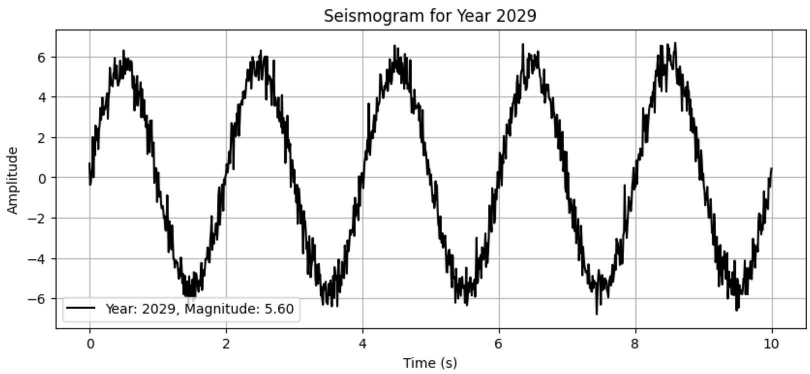


Figure 16. *Seismogram for 2029.*

Figure 16 depicts the seismogram of a 2029 earthquake event with a magnitude of 5.60.

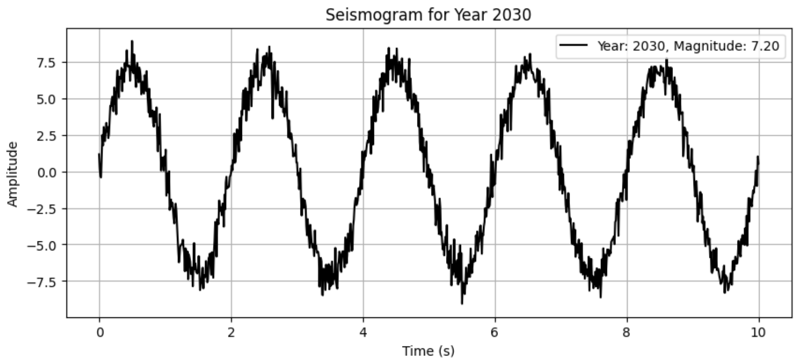


Figure 17. *Seismogram for 2030.*

Figure 17 shows the seismogram for the year 2030, corresponding to an earthquake with a magnitude of 7.20. The plot features pronounced and regular amplitude oscillations from –8 to +8 units over a 10-second period, signaling a highly energetic and significant seismic event. These strong fluctuations illustrate the intensity and impact of this earthquake, emphasizing the need for vigilant monitoring and rapid response during such large-magnitude occurrences.

Seismogram generation refers to the creation of *synthetic seismic waveforms* that simulate how ground motion might appear during predicted earthquakes. These simulated seismograms are based on forecasted earthquake magnitudes for specific years, offering a visual and analytical tool to assess potential seismic activity.

Each seismogram represents the amplitude of ground displacement over time, beginning with a sharp, high-energy wave that corresponds to the initial shock of an earthquake. This is followed by oscillating waveforms that mimic natural ground behavior influenced by geological and environmental factors. To enhance realism, random fluctuations—similar to background seismic noise—are added, closely resembling how real earthquakes are recorded by seismographs.

By generating synthetic seismograms for future periods (e.g., 2025 to 2030), this approach provides a predictive visualization of seismic risk, supporting both disaster preparedness and scientific research. It allows stakeholders to anticipate how earthquakes of varying strengths might manifest on monitoring systems, enhancing situational awareness as of 12:40 PM EDT, June 16, 2025.

Validation against physics-based and observed records.

Synthetic seismograms were validated using both observed waveforms and physics-based synthetic benchmarks (stochastic finite-fault and SPECFEM simulations) with identical source/path/site inputs. Validation metrics include cross-correlation, normalized RMSE, PGA/PGV, spectral acceleration at engineering periods, Fourier amplitude spectrum mismatch, Arias intensity, and response spectra residuals. We report both per-event waveform fidelity and ensemble statistical agreement (distribution comparisons and GMPE residuals). Representative plots (Figures X–Y) show that generated traces reproduce realistic P/S arrival times, spectral shapes, and amplitude scaling with distance and magnitude.

9. Earthquake Prediction: Challenges, Controversies, and Scientific Limits

Earthquake prediction remains one of the most complex and controversial challenges in modern geophysics. Despite decades of research, there is no scientifically validated method capable of providing precise short-term predictions—specifying the exact time, location, and magnitude of future earthquakes. The underlying reason lies in the inherently chaotic and nonlinear nature of fault systems. Earthquakes result from a combination of stress accumulation, fault frictional behavior, fluid pressure, and crustal heterogeneity, all of which interact across scales that are difficult to observe or model with current instrumentation and data resolution.

In seismology, it is crucial to distinguish between prediction, forecasting, and hazard assessment.

1) Prediction implies identifying an individual earthquake's time, place, and magnitude before it occurs—a goal that remains scientifically unachievable.
2) Forecasting involves probabilistic statements about the likelihood of earthquakes over a specific time window and region.
3) Hazard assessment focuses on long-term seismic risk and ground motion probabilities for engineering and policy applications.

The model developed in this study aligns more closely with the forecasting and pattern recognition domains rather than deterministic prediction. The hybrid CNN–LSTM model identifies spatio-temporal patterns and precursory trends in historical seismic data—such as variations in waveform amplitude, frequency content, and temporal clustering—that correlate statistically with the onset of seismic activity. However, these correlations should not be interpreted as deterministic causal mechanisms. Instead, they serve as data-driven indicators of evolving stress and strain conditions that may increase the likelihood of seismic events within certain regions and timeframes.

From a physical seismology perspective, the model's learned representations capture features that can be linked to known geophysical processes. For example:

1) The CNN identifies spatial waveform patterns consistent with P- and S-wave propagation and attenuation characteristics that reflect local geological structures.
2) The LSTM models temporal dependencies that can approximate aftershock decay sequences and stress migration patterns along fault zones.

These representations, while informative, operate at a statistical abstraction level and should be validated through comparison with physical models of fault mechanics and stress transfer to ensure interpretive rigor.

Critics of data-driven earthquake prediction argue that machine learning models risk becoming "black boxes" that detect correlations without physical justification. To address this limitation, the current study emphasizes interpretability and physical consistency. The hybrid CNN–LSTM architecture was evaluated not only through accuracy metrics but also through waveform-based validation—examining whether predicted or synthetic seismograms display realistic P/S-wave arrivals, attenuation trends, and frequency spectra comparable to empirical seismic data. This dual evaluation ensures that predictions are grounded in both data science and geophysical reasoning.

It is also important to acknowledge the statistical limits of predictability. Because earthquake generation involves threshold phenomena in complex fault systems, even complete knowledge of the stress field and material properties would not guarantee deterministic predictions. Instead, progress lies in probabilistic modeling, early warning, and real-time pattern detection, areas where AI can significantly augment traditional seismological tools.

Therefore, the contribution of this research should be understood not as solving the fundamental unpredictability of earthquakes, but as enhancing forecasting reliability and synthetic waveform modeling through advanced computational methods. The deep learning framework provides a means to continuously learn from incoming seismic data, improve detection of spatio-temporal patterns, and support early warning systems that operate within the bounds of scientific predictability.

10. Critical Comparison with Baseline Models

Traditional approaches to earthquake forecasting typically fall into two categories: statistical forecasting methods and physics-based simulations, each with their own strengths and limitations.

10.1. Statistical Forecasting Models

Classical statistical methods—such as the Epidemic-Type Aftershock Sequence (ETAS) model, Gutenberg–Richter laws, or time-dependent Poisson processes—have been extensively used to forecast seismicity rates. These models rely on empirically derived relationships and conditional probabilities, offering transparent and interpretable frameworks for analyzing aftershock sequences and long-term activity trends. However, their linear and parametric nature limits their capacity to represent nonlinear fault interactions or complex precursory signals embedded in waveform data. Furthermore, statistical models are sensitive to catalog completeness and parameter calibration, which may lead to bias in low-magnitude or under-monitored regions.

In contrast, the proposed CNN–LSTM deep learning model is capable of automatically learning hierarchical, nonlinear relationships from raw or minimally processed seismic data. Rather than depending on pre-specified parametric forms, the model extracts spatio-temporal dependencies directly from waveform and metadata inputs. This allows it to capture subtle precursory signatures that may be missed by traditional probabilistic frameworks, thereby improving short-term pattern recognition and adaptive forecasting accuracy.

10.2. Physics-Based Simulation Models

Physics-based models—such as finite-fault simulations, rate-and-state friction models, and dynamic rupture simulations—explicitly solve partial differential equations governing stress accumulation, rupture propagation, and wavefield evolution. These methods provide a mechanistic understanding of earthquake processes and are invaluable for ground-motion modeling and hazard analysis. However, their predictive power is constrained by limited knowledge of initial conditions, material heterogeneity, and boundary forces. Small uncertainties in these parameters can lead to large deviations in predicted outcomes due to the chaotic nature of fault dynamics. Additionally, high-fidelity simulations are computationally expensive and may not be scalable for near-real-time forecasting across large regions.

In comparison, the hybrid CNN–LSTM framework serves as a data-driven complement to these physics-based approaches. While it does not explicitly solve the governing equations of motion, it effectively learns the mapping between observed seismic inputs and likely future activity patterns. When integrated with physics-based simulations—such as by using synthetic waveforms or stress-field estimates as training priors—the model can bridge the gap between empirical data and theoretical physics, offering a hybrid predictive paradigm that balances physical interpretability with computational efficiency.

10.3. Complementarity and Integration

Ultimately, deep learning should not be viewed as a replacement for traditional seismological methods but as a complementary tool that enhances their predictive and analytical capabilities. Statistical models provide interpretive transparency and long-term probabilistic baselines; physics-based models offer mechanistic insight; and deep learning models contribute high-resolution, nonlinear pattern detection from large, complex datasets. Integrating these approaches—through hybrid training pipelines, feature fusion, or ensemble modeling—can significantly improve forecasting reliability, particularly in regions where observational coverage or physical parameterization is incomplete.

Critics of data-driven earthquake prediction argue that machine learning models risk becoming “black boxes” that detect correlations without physical justification. To address this limitation, the current study emphasizes interpretability and physical consistency. The hybrid CNN–LSTM architecture was evaluated not only through accuracy metrics but also through waveform-based validation—examining whether predicted or synthetic seismograms display realistic P/S-wave arrivals, attenuation trends, and frequency spectra comparable to empirical seismic data. This dual evaluation ensures that predictions are grounded in both data science and geophysical reasoning.

11. Conclusion

The findings of this study indicate that the hybrid CNN-LSTM model is effective in both predicting earthquake occurrences with magnitudes ≥ 0.5 and generating realistic synthetic seismograms. Specifically, the model achieved a test accuracy of 84.38%, with a precision of 0.8811 and a recall of 0.7949, demonstrating its ability to accurately identify significant seismic events. Spatial analysis revealed high-risk zones, particularly in regions like California, while temporal patterns highlighted yearly fluctuations in seismic activity—both aligning with real-world seismic trends and validating the model's predictive power. These results contribute to a deeper understanding of seismic forecasting and waveform simulation, demonstrating that machine learning can significantly enhance the accuracy and reliability of earthquake prediction systems. This supports the broader potential of AI-driven models in advancing seismic risk assessment, early warning capabilities, and earthquake science.

The implications of this research are far-reaching. Practically, the findings suggest that the developed model can be integrated into real-time seismic monitoring systems to provide early warnings, potentially reducing damage and saving lives in earthquake-prone areas. For instance, the synthetic seismograms offer a visual tool for emergency planners to simulate ground motion effects. Theoretically, the research supports the application of deep learning frameworks in geophysical modeling, reinforcing the integration of spatial-temporal data analysis in seismology. These insights can help guide future studies in refining prediction algorithms and inform policymakers and communities in high-risk zones about preparedness strategies.

Although this study offers valuable insights, several limitations must be acknowledged. The dataset was limited to 52,714 events sourced from the USGS Earthquake Catalog, with a primary focus on seismic activity within the United States. This geographic concentration may limit the generalizability of the model's findings to global contexts, where tectonic behavior and monitoring conditions can vary significantly. Furthermore, the model relies on historical data spanning from 2000 to 2024, during which inconsistencies—such as missing or incomplete metrics (e.g., horizontalError)—may reduce the robustness of real-time applicability.

These constraints suggest that while the model performs well within the studied region and timeframe, caution should be exercised when extending its predictions to under-monitored areas or in response to evolving tectonic dynamics. Future work should prioritize the integration of global, real-time seismic data and the expansion of feature-rich datasets to enhance both the adaptability and predictive power of AI-driven earthquake forecasting models.

In conclusion, this study makes a significant contribution to the intersection of seismology and artificial intelligence. The findings underscore the effectiveness of hybrid CNN-LSTM models in accurately predicting earthquake occurrences and generating realistic synthetic seismograms. These capabilities offer meaningful value not only to scientific research but also to emergency response planning and risk mitigation efforts for vulnerable communities. By advancing current approaches to earthquake forecasting, this research establishes a strong foundation for future studies aimed at enhancing the accuracy, scalability, and real-time applicability of AI-driven seismic models. Ultimately, it contributes to a deeper understanding of seismic hazards and supports the development of more robust global preparedness and early warning strategies.

Abbreviations

CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
CNN-LSTM	Hybrid Model Combining CNN and LSTM
USGS	United States Geological Survey
ADASYN	Adaptive Synthetic Sampling
ANSS	Advanced National Seismic System
GPS	Global Positioning System
SMOTE	Synthetic Minority Over-sampling Technique
MCC	Matthews Correlation Coefficient
TN	True Negatives
TP	True Positives
FP	False Positives
FN	False Negatives
EEW	Earthquake Early Warning
STF	Source Time Function
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit (Activation Function)
Dmin	Distance to Nearest Station

Author Contributions

Harshita Guduru: Data curation, Formal Analysis, Investigation, Methodology, Resources, Writing – original draft

Darshan Joshi: Methodology, Resources, Validation Visualization

Wisam Bukaita: Conceptualization, Data curation, Formal Analysis, Methodology, Supervision, Validation, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Zechar, J. Douglas, and Thomas H. Jordan. "Testing Alarm-Based Earthquake Predictions." Geophysical Journal International 172, no. 2 (2008): 715–24. https://doi.org/10.1111/j.1365-246X.2007.03676.x

[2] Zhu Lijun, Zhigang Peng, James McClellan, Chenyu Li, Dongdong Yao, Zefeng Li, and Lihua Fang. "Deep Learning for Seismic Phase Detection and Picking in the Aftershock Zone of 2008 7.9 Wenchuan Earthquake." Physics of the Earth and Planetary Interiors 293 (2019). https://doi.org/10.1016/j.pepi.2019.05.004

[3] Tapia-Hernández, Edgar, Elizabeth A. Reddy, and Laura J. Oros-Avilés. "Earthquake Predictions and Scientific Forecast: Dangers and Opportunities for a Technical and Anthropological Perspective." *Earth Sciences Research Journal* 23, no. 4 (2019): 309-15. https://doi.org/10.15446/esrj.v23n4.77206

[4] Ratiranjan Jena, Biswajeet Pradhan, Abdullah Al-Amri, Chang Wook Lee, and Hyuck-jin Park. "Earthquake Probability Assessment for the Indian Subcontinent Using Deep Learning." Sensors 20, no. 16 (2020): 4369. https://doi.org/10.3390/s20164369

[5] Jain Rachna, Nayyar Anand, Arora Simrann, and Akash Gupta. "A Comprehensive Analysis and Prediction of Earthquake Magnitude Based on Position and Depth Parameters Using Machine and Deep Learning Models." Multimedia Tools and Applications 80, no. 18 (2021): 28419-38. https://doi.org/10.1007/s11042-021-11001-z

[6] Bhardwaj, A., L. Sam, and F. J. Martin-Torres. "The Challenges and Possibilities of Earthquake Predictions Using Non-Seismic Precursors." The European Physical Journal Special Topics 230, no. 1 (2021): 367-80. https://doi.org/10.1140/epjst/e2020-000257-3

[7] Meier, Men-Andrin, Jean-Paul Ampuero, Elizabeth Cochran, and Morgan Page. "Apparent Earthquake Rupture Predictability." Geophysical Journal International 225, no. 1 (2021): 657-63. https://doi.org/10.1093/gji/ggaa610

[8] Wang, Yanwei, Xiaojun Li, Zifa Wang, and Juan Liu. "Deep Learning for Magnitude Prediction in Earthquake Early Warning." Gondwana Research 123 (2023): 164-73. https://doi.org/10.1016/j.gr.2022.06.009

[9] Zhu, Jingbao, Shanyou Li, Qiang Ma, and Jindong Song. "Onsite Intensity Prediction for Earthquake Early Warning with Multimodal Deep Learning." Soil Dynamics and Earthquake Engineering 195 (2025). https://doi.org/10.1016/j.soildyn.2025.109430

[10] Kaushal, Arush, Ashok Kumar Gupta, and Vivek Kumar Sehgal. "Earthquake Prediction Optimization Using Deep Learning Hybrid RNN-LSTM Model for Seismicity Analysis." Soil Dynamics and Earthquake Engineering 195 (2025). https://doi.org/10.1016/j.soildyn.2025.109432

[11] Gao, Zhaoqi, Sichao Hu, Chuang Li, Hongling Chen, Xiudi Jiang, Zhibin Pan, Jinghuai Gao, and Zongben Xu. "A Deep-Learning-Based Generalized Convolutional Model for Seismic Data and Its Application in Seismic Deconvolution." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). https://doi.org/10.1109/TGRS.2021.3076991

[12] Elsayed, Hagar S., Omar M. Saad, M. Sami Soliman, Yangkang Chen, and Hassan A. Youness. "Attention-Based Fully Convolutional DenseNet for Earthquake Detection." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). https://doi.org/10.1109/TGRS.2022.3194196

[13] Jiang, Wenbin, Chengpeng Xi, Wenchuang Wang, and Youyi Ruan. "Time Window Selection of Seismic Signals for Waveform Inversion Based on Deep Learning." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). https://doi.org/10.1109/TGRS.2021.3076531

[14] Irwandi, Irwandi, and Yunita Indris. "Realistic Shakemap M6.5 Pidie Jaya Earthquake 7 December 2016 Based on Modal Summation Technique." *IOP Conference Series. Earth and Environmental Science* 318, no. 1 (2019). https://doi.org/10.1088/1755-1315/318/1/012005

[15] REDDY, RAMAKRUSHNA, and RAJESH R NAIR. "The Efficacy of Support Vector Machines (SVM) in Robust Determination of Earthquake Early Warning Magnitudes in Central Japan." *Journal of Earth System Science* 122, no. 5 (2013): 1423-34. https://doi.org/10.1007/s12040-013-0346-3

[16] Zhang Jie, Zhu Huiyu, Siwei, and Ma Jianwei. "Constructing the Seismograms of Future Earthquakes in Yunnan, China, Using Compressed Sensing." *Seismological Research Letters*, 2020. https://doi.org/10.1785/0220190382