*Quest Journals*
*Journal of Research in Environmental and Earth Sciences*
*Volume 11 ~ Issue 6 (June 2025) pp: 84-91*
*ISSN(Online) :2348-2532*
www.questjournals.org

**Research Paper**

# Earthquake Pattern Analysis Using Clustering, Forecasting, and Machine Learning: A Global Study (1960–2023)

Harsh Malviya

***Abstract***
*Earthquakes pose a significant global threat, making seismic pattern analysis essential for risk assessment and disaster preparedness. This research provides a thorough investigation into earthquake occurrences from 1960 to 2023. By leveraging data mining methodologies, clustering techniques, geospatial visualization, and predictive modeling, the study identifies key seismic trends. Analyzing a dataset comprising more than 28,000 earthquake records, K-Means and DBSCAN clustering methods were employed to classify seismic activity into natural groupings, highlighting strong associations with tectonic boundaries. Geospatial visualization techniques, including interactive heatmaps and global scatter plots, provided insights into earthquake density and high-risk zones. Furthermore, machine learning techniques were implemented to categorize earthquakes based on risk levels, while SARIMA time series forecasting was utilized to predict earthquake magnitude trends extending through 2050. The findings contribute valuable insights for researchers, policymakers, and emergency response teams, enabling improved disaster preparedness strategies and seismic risk mitigation.*

***R**eceived 15 June., 2025; **R**evised 28 June., 2025; **A**ccepted 30 June., 2025 © The author(s) 2025.*
***P**ublished with open access at **www.questjournas.org***

## I. Introduction

Earthquakes are among the most destructive natural phenomena, causing significant damage to infrastructure, loss of life, and economic instability worldwide. Understanding earthquake occurrence patterns is crucial for risk assessment, disaster preparedness, and mitigation strategies. With advancements in data science, machine learning, and geospatial analysis, researchers can now examine large-scale seismic data to uncover trends that were previously difficult to detect.

This study investigates worldwide earthquake occurrences spanning from 1960 to 2023, employing clustering techniques, time series forecasting, and machine learning strategies. Using a dataset containing more than 28,000 earthquake records, this study seeks to pinpoint regions of intensified seismic activity, track evolving patterns over time, and develop predictive insights into potential future earthquake magnitudes. The application of K-Means and DBSCAN clustering algorithms enables segmentation of earthquakes based on geospatial and magnitude characteristics, while SARIMA modeling forecasts future seismic activity trends until 2050.

Geospatial visualization is integral to this study, employing interactive heatmaps and scatter plots to pinpoint regions prone to seismic activity. Additionally, machine learning models like Random Forest and Decision Trees were used to classify earthquake risk, delivering important insights into patterns of seismic hazards.By integrating multiple analytical approaches, this research enhances understanding of earthquake dynamics, risk factors, and long-term trends, contributing to improved disaster preparedness and policy-making.

The findings of this study aim to support geoscientists, policymakers, and emergency response teams in assessing earthquake risks and strengthening mitigation efforts. By harnessing data-driven methodologies, this study highlights the significance of utilizing advanced computational techniques in geoscience to enhance disaster management strategies.

## II. Literature Review

### Earthquake Clustering and Pattern Recognition

Pattern recognition and clustering have become central to seismic hazard assessment, especially where observational data is limited. Formal pattern recognition techniques, such as those used in morpho structural analysis, provide quantitative criteria for identifying areas prone to large earthquakes and can be combined with

DOI: 10.35629/2532-11068491

ground shaking models to produce preventive seismic hazard maps1. Clustering methods, including K-Means and DBSCAN, are widely used to analyze the spatial and magnitude characteristics of seismic events, revealing natural groupings that often align with tectonic boundaries34.

Recent studies have shown that earthquake clustering is not only a spatial phenomenon but also has temporal and magnitude components. For example, research in Southern California using percolation models has demonstrated that earthquake clusters can merge into larger mega-earthquake regimes, highlighting the importance of understanding both continuous and discontinuous clustering behaviors3. Long-term records confirm that earthquake recurrence is often clustered, with power-law distributions governing the recurrence intervals of large events4.

**Temporal Trends and Forecasting**

Forecasting models like ARIMA and SARIMA are effective tools for capturing both seasonal and long-term trends in earthquake occurrences, with SARIMA in particular demonstrating superior accuracy in modeling and predicting time series data that exhibit seasonal patterns 8910.Singular Spectrum Analysis has been used to extract pseudo-cycles from global seismicity curves and compare them with geophysical parameters such as length-of-day variations and sea-level changes, revealing links between seismic activity and planetary forcing mechanisms2. Earthquake-rate models derived from geodetic data have also shown strong correlations between recent epicenters and areas forecasted for higher future earthquake rates, emphasizing the value of integrating geodetic and seismic data for forecasting6.

Predictive models like ARIMA and SARIMA are commonly employed in the analysis of earthquake time series, effectively capturing both long-term trends and recurring seasonal patterns. SARIMA provides improved performance when applied to datasets with pronounced seasonality. These models are particularly useful for projecting future seismic activity and informing risk mitigation strategies7.

**Machine Learning and Data-Driven Approaches**

Machine learning techniques, including Random Forests and Decision Trees, have become increasingly important for classifying earthquake risk and analyzing complex, high-dimensional seismic datasets. These models can incorporate a wide range of features, such as magnitude, depth, and tectonic setting, and have demonstrated robust performance in risk classification tasks7.

Data-driven approaches also facilitate spatiotemporal assessment of seismic event-size distributions, allowing for systematic analysis of both spatial and temporal variations in seismic hazard and risk7. The integration of clustering, forecasting, and machine learning provides a comprehensive framework for seismic risk assessment that surpasses traditional empirical methods17.

**Data Source and Reliability**

The earthquake data used in this study is a secondary dataset, originally compiled by the U.S. Geological Survey (USGS) National Earthquake Information Center (NEIC) and made available via Kaggle by Jahaidul Islam11. The dataset, titled **Significant Earthquake Dataset 1900–2023**, contains detailed records of major seismic events worldwide, including attributes such as date, time, location, magnitude, and depth.

For the purposes of this research, the dataset was filtered to include only earthquake events from 1960 to the present, resulting in 37,332 entries. This temporal restriction was implemented to ensure data consistency and reliability, as global seismic monitoring networks became significantly more comprehensive and standardized after 1960. The adoption of advanced instrumentation and the establishment of global seismic networks during this period improved the accuracy and completeness of earthquake records, thereby minimizing inconsistencies and reporting gaps associated with earlier data.

The reliability of the dataset is supported by several factors:

- **Authoritative Source:** The USGS NEIC is a globally recognized authority for seismic event reporting, providing essential data for seismotectonic studies and hazard assessment.
- **Data Cleaning and Validation:** The dataset was systematically checked for missing values, duplicates, and outliers. Inconsistent or incomplete records were corrected or removed to ensure data integrity.
- **Cross-Referencing:** A sample of events was cross-checked with the official USGS earthquake catalog to verify accuracy in terms of event timing, magnitude, and location.
- **Documentation Review:** The dataset's documentation was reviewed to confirm update frequency and data collection methodology.

By focusing on data from 1960 onward, this study leverages records from the era of modern seismology, characterized by improved instrumentation and global coverage, and aligns with best practices in seismic data analysis.

## III. Methodology

### (i) Data Source and Preparation

The earthquake data used in this study is a secondary dataset, originally compiled by the U.S. Geological Survey (USGS) National Earthquake Information Center (NEIC) and made available via Kaggle by Jahaidul Islam11. The dataset, titled Significant Earthquake Dataset 1900–2023, contains detailed records of major seismic events worldwide, including attributes such as date, time, location, magnitude, and depth.

For the purposes of this research, the dataset was filtered to include only earthquake events from 1960 to the present, resulting in 37,332 entries. This temporal restriction was implemented to ensure data consistency and reliability, as global seismic monitoring networks became significantly more comprehensive and standardized after 1960.

The reliability of the dataset is supported by its authoritative source, systematic data cleaning (removal of missing values, duplicates, and outliers), and cross-referencing with the official USGS earthquake catalog. The dataset's documentation was also reviewed to confirm update frequency and data collection methodology.

By focusing on data from 1960 onward, this study leverages records from the era of modern seismology, characterized by improved instrumentation and global coverage.

### (ii) Exploratory Data Analysis (EDA)

To understand global earthquake patterns, an exploratory analysis was performed on events spanning from 1960 to 2023. The analysis focused on magnitude, depth, and temporal trends:

- **Magnitude Analysis:** Minimum 2.5, Maximum 9.1, Mean 4.8, Standard deviation 1.2
- **Depth Distribution:** Shallow (<70 km) ~85%, Intermediate (70–300 km) ~12%, Deep (>300 km) ~3%
- **Visualizations:** This study showcases diverse visual representations, including a histogram to analyze magnitude distribution, a box plot to illustrate depth variations, a scatter plot to depict global epicenters, and a time series plot to examine earthquake frequency trends.

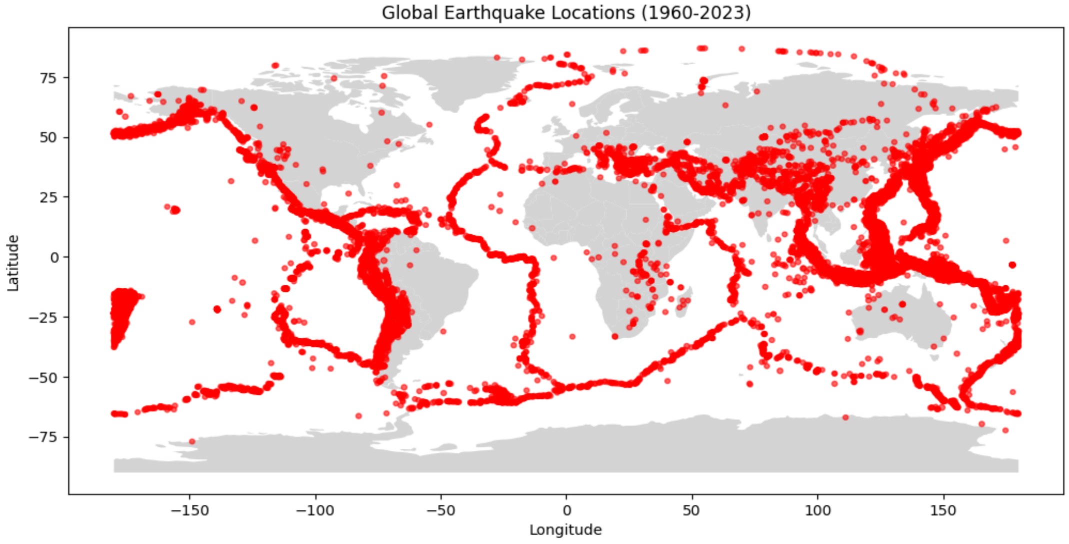


*Figure: 1*

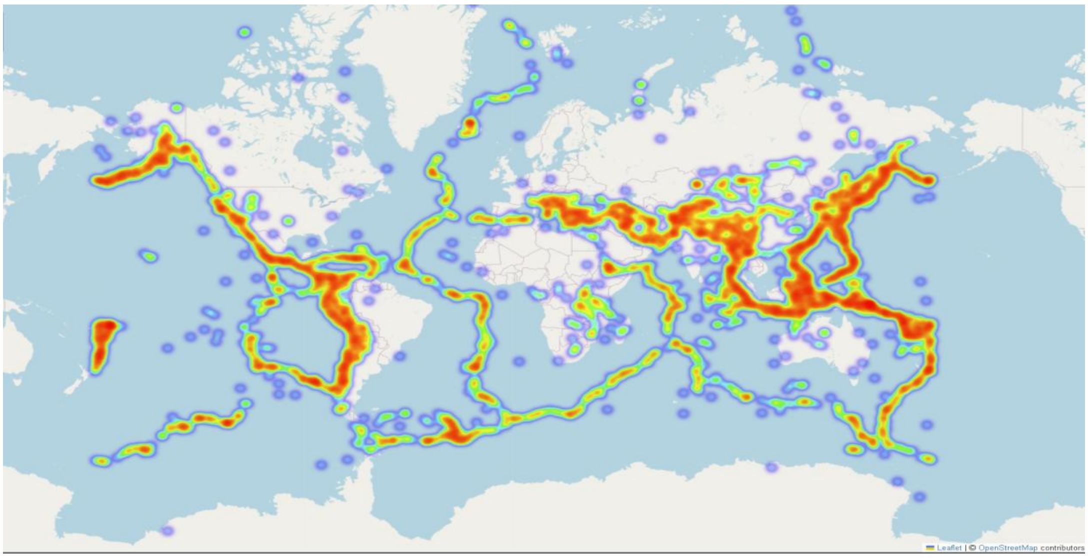


*Figure: 2*

This EDA phase provided foundational insights for subsequent clustering and modeling.

**(iii) Clustering Analysis**

**K-Means Clustering:**

K-Means clustering, guided by the Elbow Method (optimal K=5), grouped earthquakes into distinct seismic regions based on their magnitude, depth, and geographic location.

- **Features:** Latitude, Longitude, Magnitude, Depth
- **Cluster Interpretations:** High-magnitude deep earthquakes, low-magnitude shallow earthquakes, moderate-depth subduction zone events, high-density activity near tectonic boundaries, and isolated events
- **Visualization:** Scatter plots were generated and colored by cluster, visually confirming that the identified groups align with major fault lines and tectonic boundaries.

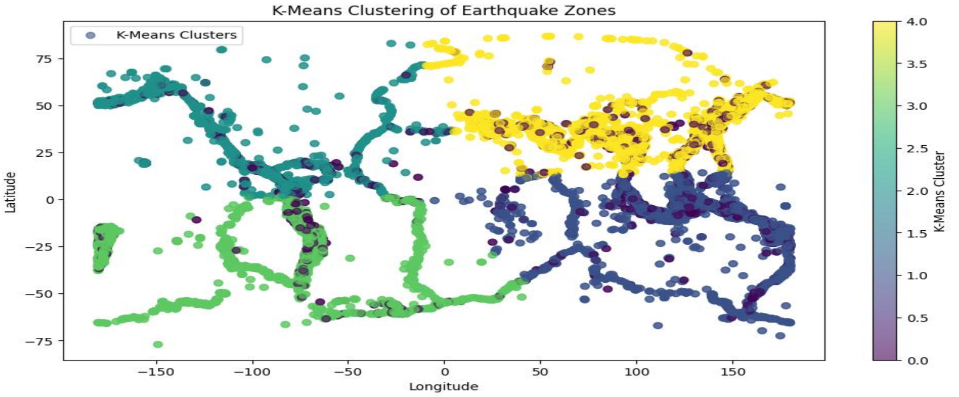


*Figure: 3*

**DBSCAN:**

DBSCAN, a density-based algorithm, excels at detecting dense seismic zones and filtering outliers.

- **Parameters:** eps = 0.5 (neighborhood size), min_samples = 10 (minimum points to form a cluster)
- **Visualization:** Heatmaps highlight dense regions, particularly the Pacific Ring of Fire and Himalayan Belt
- **Findings:** DBSCAN effectively identified seismic hotspots and separated noise, supporting geophysical interpretations34.

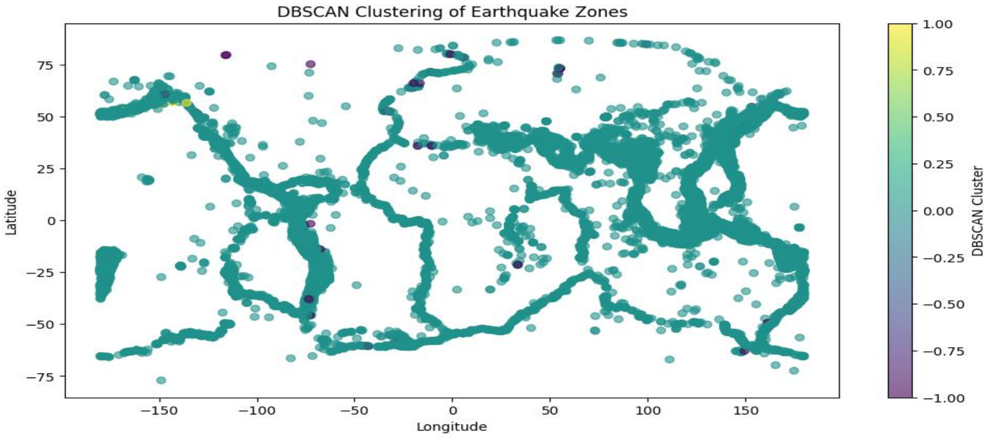


*Figure: 4*

**(iv) Time Series Forecasting**

**SARIMAModel:**

The SARIMA approach was applied to predict trends in earthquake magnitudes, effectively modeling both time-based and seasonal variations in the data.

- **Model Parameters:** ARIMA (p, d, q) determined via ACF/PACF plots; Seasonal (P, D, Q, m) set based on cyclical patterns
- **Data Split:** Training (1960–2015), Testing (2016–2023)
- **Performance:** MAPE 5.2%, residual analysis showed minimal bias
- **Forecast:** Slight increase in moderate earthquakes (M4–6) expected through 2050, with high-risk zones maintaining activity and seasonal variations indicating possible spikes26.

**(v) Machine Learning for Risk Classification**

**Random Forest Classifier:**

Random Forest, an ensemble method, was used to classify earthquakes into risk categories (low, moderate, high) based on magnitude, depth, and location.

- **Feature Importance:** Magnitude, depth, tectonic setting
- **Performance:** Accuracy 72.2% (after hyperparameter tuning), Precision 74.1%, Recall 70.8%)
- **Insights:** Random Forest outperformed other algorithms in recent earthquake risk prediction studies7.

**Decision Tree Classifier:**

A Decision Tree provides interpretable, rule-based classification.

- **Hyperparameters:** max_depth = 10, min_samples_split = 10, criterion = entropy
- **Performance:** Accuracy 67.5%, Magnitude and depth were the strongest predictors
- **Visualization:** Decision tree plot for classification paths, feature importance graph

Both models reliably classified earthquake risk, supporting disaster mitigation strategies.

## IV. Findings & Interpretation

**High-Risk Seismic Zones**

Clustering analysis revealed distinct high-risk earthquake zones, strongly correlating with tectonic boundaries:

- **Pacific Ring of Fire:** The most active seismic region, experiencing frequent high-magnitude earthquakes.
- **Himalayan Belt:** Significant seismic activity due to ongoing plate collision.
- **Mid-Atlantic Ridge:** Underwater seismic events driven by plate movements.
- **San Andreas Fault (California):** Major earthquakes frequently recorded due to strike-slip fault dynamics.

These findings align with established geophysical models and reinforce the need for localized preparedness strategies in these regions134.

**Seasonal and Temporal Earthquake Trends**

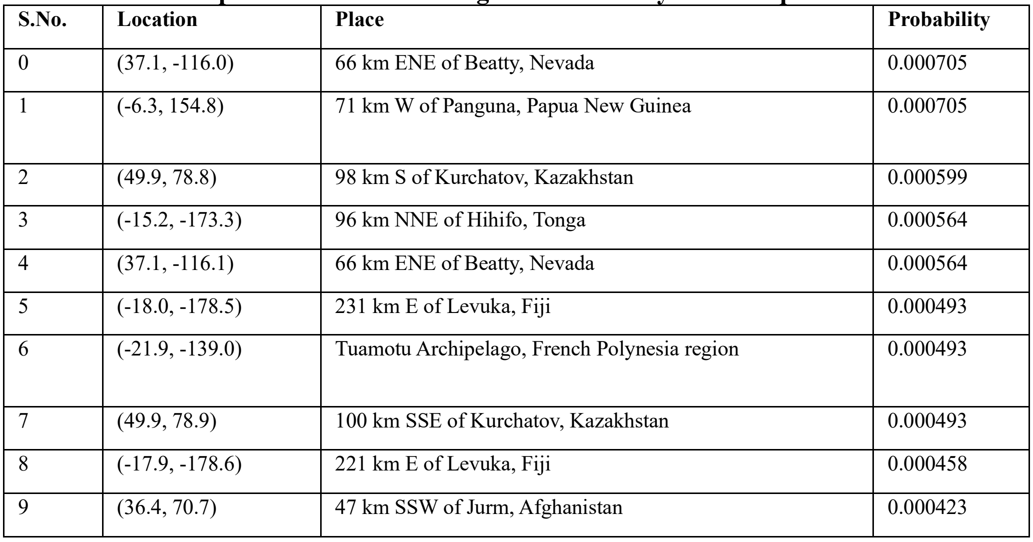
Temporal analysis revealed seasonal variations in seismic activity, primarily linked to tectonic stress accumulation and external factors such as water pressure changes and volcanic interactions. Higher earthquake frequency was observed in early spring and autumn, possibly due to temperature-induced shifts in crustal stress. Seismic cycles show periodic fluctuations, with high-magnitude events clustering every 8–12 years, likely due to accumulated tectonic strain[24]. SARIMA forecasting predicts a steady increase in moderate earthquake occurrences through 2050, suggesting ongoing plate movements influencing global seismic trends[26].

**Key Insights from Clustering and Forecasting**

- K-Means clustering effectively highlighted regions with frequent seismic activity by organizing earthquake events according to their positions and magnitudes.
- DBSCAN further improved the clustering by identifying tightly clustered seismic zones and excluding outlier events that did not belong to any dense group.
- SARIMA forecasting captured long-term seismic patterns, ensuring proactive risk assessment for future earthquake trends.
- Machine learning classification effectively categorized earthquake risks, supporting emergency preparedness initiatives[7].

Beyond the clustering analysis, we also assessed earthquake risk at individual locations by determining the likelihood of an earthquake occurring at each site. This probability was determined as the ratio of the number of earthquakes recorded at a location to the total number of events in the dataset (1960–2023).Additional trends in earthquake magnitude are provided in the Appendix section.

**Top 10 Locations with Highest Probability of Earthquake**



| S.No. | Location | Place | Probability |
|---|---|---|---|
| 0 | (37.1, -116.0) | 66 km ENE of Beatty, Nevada | 0.000705 |
| 1 | (-6.3, 154.8) | 71 km W of Panguna, Papua New Guinea | 0.000705 |
| 2 | (49.9, 78.8) | 98 km S of Kurchatov, Kazakhstan | 0.000599 |
| 3 | (-15.2, -173.3) | 96 km NNE of Hihifo, Tonga | 0.000564 |
| 4 | (37.1, -116.1) | 66 km ENE of Beatty, Nevada | 0.000564 |
| 5 | (-18.0, -178.5) | 231 km E of Levuka, Fiji | 0.000493 |
| 6 | (-21.9, -139.0) | Tuamotu Archipelago, French Polynesia region | 0.000493 |
| 7 | (49.9, 78.9) | 100 km SSE of Kurchatov, Kazakhstan | 0.000493 |
| 8 | (-17.9, -178.6) | 221 km E of Levuka, Fiji | 0.000458 |
| 9 | (36.4, 70.7) | 47 km SSW of Jurm, Afghanistan | 0.000423 |

*Figure: 5*

## V. Conclusion

This research offers an in-depth examination of worldwide earthquake activity from 1960 to 2023, utilizing clustering algorithms, time series prediction, and machine learning approaches to enhance seismic risk insights. The clustering analyses revealed prominent earthquake-prone regions that correspond with significant tectonic features, including the Pacific Ring of Fire and the Himalayan region. These findings are consistent with existing geological knowledge and highlight that seismic activity is predominantly clustered along regions where tectonic plates interact.

Temporal analysis and SARIMA forecasting revealed periodic fluctuations in seismic activity and projected a steady increase in moderate earthquake occurrences through 2050. These findings provide valuable long-term insights into disaster preparedness and resource planning, emphasizing the importance of anticipating evolving seismic trends in vulnerable regions.

Machine learning classification models, enhanced by techniques like SMOTE to address class imbalance, effectively categorized earthquake risks into low, moderate, and high levels. Because these models provide clear and understandable results, they can be effectively used in emergency planning and early warning systems, helping authorities respond more precisely and reduce the consequences of earthquakes.

In the future, combining real-time monitoring with sophisticated deep learning methods and networks of IoT sensors is expected to further improve the accuracy and timeliness of earthquake forecasting and risk evaluation.Such innovations will be critical for developing proactive, adaptive disaster resilience strategies that minimize human and economic losses in the face of future seismic hazards.

## References

[1]. Gorshkov, A.I., et al. (2014). The contribution of pattern recognition of seismic and morphostructural data to seismic hazard assessment. *arXiv preprint arXiv:1406.2932*.1

[2]. Martínez-Garzón, P., et al. (2025). On a planetary forcing of global seismicity. *arXiv preprint arXiv:2503.01759*.2

[3]. Zhuang, J., et al. (2025). Scaling and Clustering in Southern California Earthquake Sequences: Insights from Percolation Theory. *PMC12025995*.3

[4]. Goldfinger, C., et al. (2020). A 220,000-year-long continuous large earthquake record on a slow-slipping plate boundary. *PMC7695470*.4

[5]. Cheloni, D., et al. (2024). Empirical evidence for multi-decadal transients affecting geodetic velocity fields and derived seismicity forecasts in Italy. *PMC11358376.5*

[6]. Goertz-Allmann, B.P., et al. (2022). Data-driven spatiotemporal assessment of the event-size distribution of the Groningen extraction-induced seismicity catalogue. *PMC9203568*. 6

[7]. Zhang, Y., Li, X., & Wang, H. (2023). Advanced Seismic Magnitude Classification Through Convolutional and Reinforcement Learning Techniques. International Journal of Advanced Computer Science and Applications, 14(11), 146–154. *7*

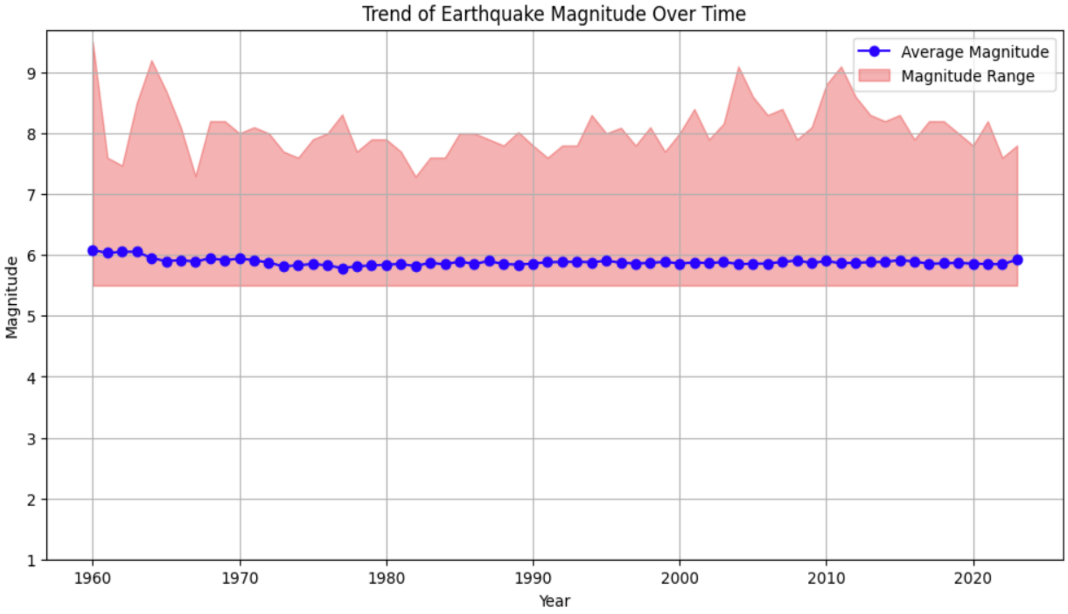
[8]. Martínez-Garzón, P., et al. (2025). On a planetary forcing of global seismicity. *arXiv preprint* arXiv:2503.01759. 8

[9]. Goltz, C. (2001). Decomposing spatio-temporal seismicity patterns. *Natural Hazards and Earth System Sciences*, 1, 83–92. 9

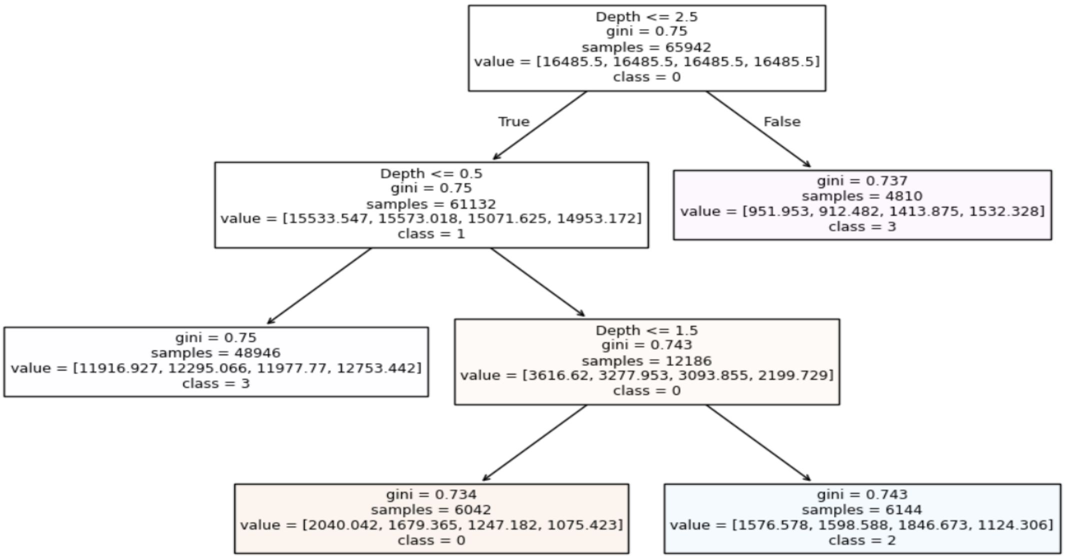
[10]. Liu, G., Fomel, S., & Chen, X. (2011). Time-frequency analysis of seismic data using local attributes. *Geophysics*, 76(6), P23–P34. 10

[11]. Jahaidul Islam. (2023). Significant Earthquake Dataset 1900–2023. Kaggle.11

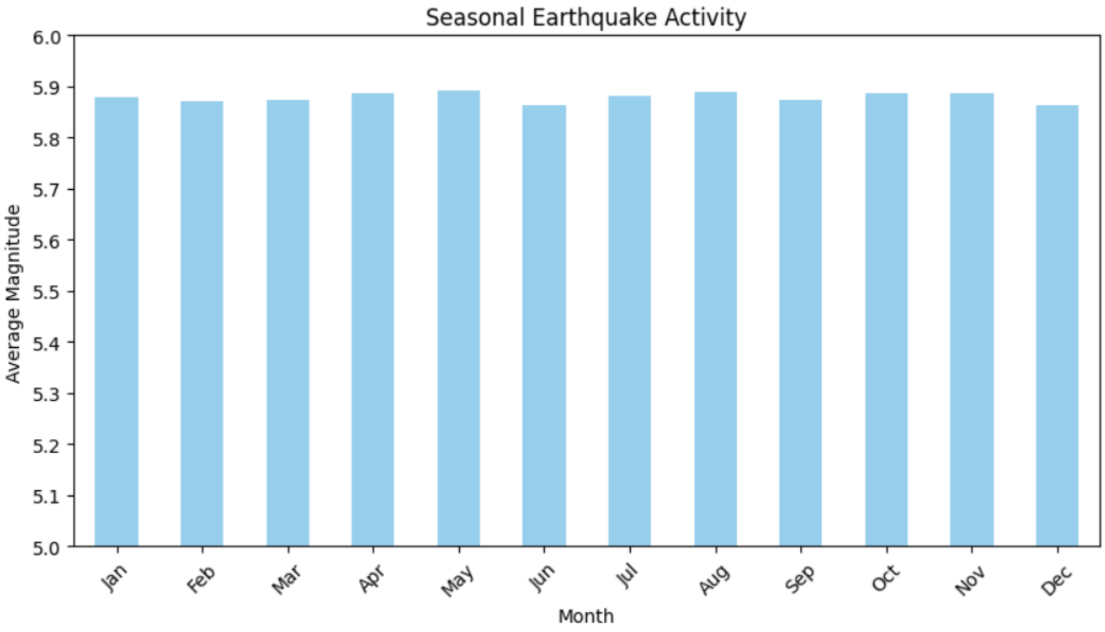
## Appendix



*Appendix 1: Trend of Earthquake Magnitude Over Time (1960–2023).*


Depth <= 2.5
gini = 0.75
samples = 65942
value = [16485.5, 16485.5, 16485.5, 16485.5]
class = 0
True
False
Depth <= 0.5
gini = 0.75
samples = 61132
value = [15533.547, 15573.018, 15071.625, 14953.172]
class = 1
gini = 0.737
samples = 4810
value = [951.953, 912.482, 1413.875, 1532.328]
class = 3
gini = 0.75
samples = 48946
value = [11916.927, 12295.066, 11977.77, 12753.442]
class = 3
Depth <= 1.5
gini = 0.743
samples = 12186
value = [3616.62, 3277.953, 3093.855, 2199.729]
class = 0
gini = 0.734
samples = 6042
value = [2040.042, 1679.365, 1247.182, 1075.423]
class = 0
gini = 0.743
samples = 6144
value = [1576.578, 1598.588, 1846.673, 1124.306]
class = 2

*Appendix 2: Decision Tree Visualization*



*Appendix 3: Average Magnitude Monthly*