Machine Learning–Driven Mainshock Identification Using Seismic Spatiotemporal Patterns and Enhanced Earthquake Feature Engineering

Kazi Sakib Hasan, Most. Afia Anjum Borsha, Faiyaz Bin Zaman,
Azra Humayra Alam Prova, Mahir Hasan Chowdhury
Epoch One AI Research Unit, Department of Computer Science and Engineering
BRAC University
Dhaka 1212, Bangladesh
kazi.sakib.hasan@g.bracu.ac.bd, afia.anjum.borsha@g.bracu.ac.bd, faiyazinan@gmail.com,
azrahumayraprova@gmail.com, mahir.hasan.chowdhury@g.bracu.ac.bd

***Abstract*—Rapidly distinguishing whether a seismic event is a mainshock or a foreshock is essential for minimizing public panic and enabling timely disaster-response in earthquake-prone regions. This study proposes a data-driven framework that employs an 88-year USGS seismic catalog, a bidirectional spatiotemporal labeling algorithm, physics-informed feature engineering, and optimized ensemble learning to classify mainshocks using only metadata features. Models including XGBoost, LightGBM, CatBoost, and AdaBoost were tuned using Bayesian optimization. The best-performing model, CatBoost, achieved an AUC of 0.965 and a recall of 0.97 for identifying mainshocks—crucial for safety-critical scenarios. SHAP interpretability reveals that magnitude differential relative to recent local seismicity is the dominant predictor. The proposed methodology offers a scalable foundation for real-time seismic role classification and can be integrated into operational early-warning systems.**

***Index Terms*—Mainshock Detection, Seismic Classification, Ensemble Learning, Spatiotemporal Labeling, SHAP, Bayesian Optimization, Early Warning Systems**

I. Introduction

Moderate earthquakes in densely populated developing regions often trigger widespread panic and uncertainty regarding whether the event is a mainshock or an early precursor to a larger event. Distinguishing mainshocks from foreshocks is essential for appropriate disaster-response strategies, where false negatives can result in catastrophic human and infrastructural loss. However, because foreshocks and mainshocks originate from similar tectonic processes, real-time discrimination remains a major challenge.

Traditional stochastic models such as ETAS provide long-term probabilistic estimates but lack precision for individual event classification. Existing ML efforts have primarily focused on structural damage prediction or aftershock hazard modeling, leaving a substantial research gap in rapid mainshock identification. This work fills that gap by proposing a fully data-driven mainshock classification framework incorporating spatiotemporal labeling, physics-informed features, and optimized ensemble models.

II. Literature Review

Recent ML research in seismology has predominantly explored structural damage prediction, aftershock hazard modeling, and ground motion forecasting. Gradient-boosting methods have shown strong performance in predicting structural fragility, while deep learning models provide accurate damage-state classification under MS–AS sequences.

More relevant studies, such as the NESTORE algorithm, attempt to assess the likelihood of subsequent strong earthquakes, but they begin analysis *after* a mainshock has occurred. Other works focus on predicting future magnitudes using seismicity patterns but do not classify whether an event *is* a mainshock. Thus, current literature lacks methods for real-time mainshock role classification.

This study positions itself uniquely by classifying mainshocks using only historical spatiotemporal context and seismic metadata, addressing an unmet operational need.

III. Methodology

The proposed framework consists of:

1) Dataset preprocessing and magnitude normalization,
2) Bidirectional spatiotemporal labeling,
3) Physics-informed feature engineering,
4) Chronological train–test split with MICE imputation,
5) Bayesian-optimized ensemble modeling, and
6) SHAP-based interpretability analysis.

A. Dataset Collection and Preprocessing

The dataset originates from a consolidated USGS global earthquake catalog (1930–2018). Magnitudes were normalized to the moment magnitude scale (Mw) using standard conversion functions. Irrelevant fields were removed, and geospatial coordinates were transformed to 3D Cartesian vectors for KD-Tree neighbor queries.

B. Spatiotemporal Labeling and Feature Engineering

An event is labeled a *mainshock* if its magnitude is the largest local event within a 50 km radius and a $\pm$30-day window. Only backward-looking features were used to prevent leakage. Key features include:

- Local seismicity rate (7-day),
- Cumulative energy release (30-day),
- Magnitude differential (relative to 30-day local history),
- Inter-event time gap, and
- Seismicity Z-score.

C. Data Splitting and Imputation

A strict chronological 80/20 split was applied to ensure realistic future prediction. MICE imputation was performed using statistics learned only from the training period.

D. Model Selection and Optimization

Four ensemble classifiers were tested: XGBoost, LightGBM, CatBoost, and AdaBoost. Hyperparameters were tuned using Optuna's Bayesian optimization (TPE) with F1-score as the objective due to moderate class imbalance.

E. Model Interpretability

SHAP values were computed for global and local interpretability. Magnitude differential emerged as the dominant feature, with geospatial coordinates contributing minimally—indicating strong generalizability.

IV. Results and Discussion

All models achieved approximately 91% accuracy, with CatBoost yielding:

- Recall = 0.97,
- AUC = 0.965,
- F1-score = 0.88.

High recall is critical for disaster management, as missing a mainshock could lead to severe consequences. ROC–AUC curves show nearly identical performance across boosting models, demonstrating strong feature robustness.

SHAP analysis confirms that the key predictor is magnitude contrast against recent local activity, aligning with physical seismology expectations that mainshocks exhibit abrupt energetic dominance.

V. Conclusion

This work introduces a fully interpretable, physics-informed, data-driven framework for mainshock classification. Ensemble models tuned via Bayesian optimization provide high accuracy and sensitivity. Magnitude differential is identified as the primary predictor, enabling transparent geophysical interpretation.

Future work will extend this framework to:

- real-time operational earthquake early-warning,
- integration with ETAS-based probabilistic models, and
- regional threshold adaptation across global fault systems.

References