SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I7P106

Volume 13 Issue 7, 77-95, July 2026

Original Article

An Explainable Machine Learning Framework for Processing Heterogeneous Seismic Sensor Data

Deepak Kumar Sinha[1], Sujata Kulkarni[2]

[1]*Department of Electronics and Telecommunication Engineering,*
Manjara Charitable Trust's Rajiv Gandhi Institute of Technology, Andheri (West), Mumbai, India.
[2]*Department of Electronics and Telecommunication Engineering,*
Sardar Patel Institute of Technology Andheri (West), Mumbai, India.

[1]*Corresponding Author : deepak.sinha@mctrgit.ac.in*

Received: 22 March 2026 Revised: 07 May 2026 Accepted: 17 June 2026 Published: 29 July 2026

Abstract - *Predicting the severity of seismic events relies on a large amount of data collected from remote sensor networks. However, processing long-term telemetry records is computationally challenging. For example, the data are typically subject to an extremely unbalanced class distribution, and sensor type variations are extremely high over time, thus reducing the effectiveness of the standard classification methods. This study presents an explainable machine learning framework for classifying event severity using long-term data from mixed seismic probes. Event records of the Indian tectonic region from 1947 to 2022 were extracted from the USGS ComCat database. To process the different information of the sensors, a processing pipeline based on a class-weighted random forest was assembled, which was assisted by data harmonization and feature engineering procedures. This pipeline was combined with SHapley Additive exPlanations (SHAP) to trace where and when the model considers spatial constraints and measurement uncertainty, that is, the focal depth and azimuthal gaps. Even with the complete elimination of explicit magnitude scalar objects to prevent target leakage, the system attained an overall accuracy of 91 %, along with a macro F1-score of 0.43. SHAP diagnostics were applied to demonstrate the existence of a hard physical boundary. Although secondary sensor proxies can be successful in classifying normal baseline tremors, there is a mathematical failure during extreme events. Catastrophic hazard detection is performed using direct waveform measurements. This work demonstrates that XAI can be used as a powerful diagnostic tool to identify the actual predictive limitations of legacy telemetry networks.*

Keywords - *Explainable Artificial Intelligence, Machine Learning, Random Forest, SHAP, Seismic event classification, Seismic Hazard Assessment.*

1. Introduction

In tectonically active regions, emergency response frameworks rely on remote sensor signals to estimate earthquake severity. Historically, seismologists have mapped these hazards using rigid magnitude thresholds. In contrast, modern telemetry networks generate multidimensional data streams across distributed arrays, rendering isolated peak value readings insufficient for comprehensive hazard triage.

This study uses the contextual proxy severity classification limits to demonstrate the utility of contextual proxies in baseline seismicity triage and to support the necessity of explicit magnitude in high-severity occurrences in post-event response planning and long-term disaster management risk reduction plans. Reliably defined levels of severity are better for resource allocation, infrastructure resilience planning, and scientific understanding of the behavior of buildings and other infrastructure.

Classical methods used in seismology for the determination of severity are almost entirely based on deterministic thresholds of moment magnitude Mw or intensity scales. Although such criteria based on rules provide standardization and reproducibility, they are not necessarily in line with multidimensional contextual variables such as focal depth, network geometry, measurement uncertainty, and local tectonic variability. Over the past 75 years, seismic telemetry and instrumentation have changed significantly. Early analog monitoring equipment has been completely replaced by highly sensitive digital network equipment. Because of this evolution in hardware, historical seismic databases are incredibly heterogeneous. Classical rule-based classification algorithms are generally unable to deal with inconsistent multidimensional data. For these types of sensors, it is not easy to adapt to changes in network geometry and large changes in measurement errors in different generations of sensors. Interpreting complex hazard patterns using a single magnitude

reading is practically impossible because of changes in the underlying hardware.

Standard machine learning architectures excel at handling noisy telemetry, yet their opaque 'black box' nature limits their deployment in critical early-warning sensor networks. Decision-makers managing operational hazard pipelines require transparent proof that an algorithm detects actual geophysical anomalies rather than memorizing historical reporting constraints. Currently, most applied research on machine learning in seismology prioritizes high predictive accuracy while compromising interpretability. Existing studies report feature importance but mostly use simple heuristic measures that lack mathematical rigor. In addition, local event-by-event rigorous diagnostic testing is rare in the literature, especially with respect to Indian telemetry networks. Consequently, there is a major research gap in developing transparent frameworks for auditing multi-era historical data.

To address these limitations, the novelty of this study lies in the development of a fully explainable machine learning framework that is explicitly designed to process heterogeneous multi-era seismic sensor data. Unlike conventional approaches that struggle with inconsistent sensor types and temporal drift, the proposed framework systematically isolates genuine geophysical signals from administrative data artifacts across various instrumentation regimes.

In this study, SHAP was used to explain a machine-learning model for forecasting seismic severity. An integrated catalogue of Indian seismic events (1947–2022) obtained from the United States Geological Survey (USGS) was used to train high-performance models for the classification of the severity of seismic events. SHAP was subsequently employed to deconstruct the rationale of the trained model, such that the marginal contributions to the magnitude, focal depth, azimuthal gap, and magnitude error were among the contextual attributes. The relationship between the decisions made by the model and the ground rules of seismology was investigated based on global attribution analysis, visualization of feature interactions, and local event case studies.

The primary objectives of this study were as follows.

1. To develop a robust data harmonization architecture that can standardize 75 years of heterogeneous, multi-era seismic sensor records without developing an empirical conversion bias.
2. To establish an effective ensemble pipeline for imbalanced sensitivity, particularly in scenarios involving severe anomaly detection and a 1:580 minority-class imbalance.
3. To investigate the predictive power of spatial, geometric, and uncertainty proxies for the controlled loss of deterministic scalar measures.
4. To implement a dual-track measurement system to separate operational real-time predictive systems from old administrative data.
5. To operationalize SHAP as a diagnostic mechanism for detecting indirect metadata-driven target leakage.

In this study, seismic severity labels were defined deterministically using magnitude thresholds.

Let:

$$S = f(M) \tag{1}$$

denotes the true severity class, where M represents the reported scalar magnitude and f(.) maps the magnitude to categorical severity levels.

Because the magnitude is excluded from the predictive feature space to avoid target leakage, the classifier must estimate the severity using contextual and geometric proxy variables X_proxy. Therefore, the predicted class can be expressed as:

$$\hat{S} = g(X_{proxy}) \tag{2}$$

where g(.) represents the trained class-weighted Random Forest model.

The central hypothesis of this study is that contextual seismic proxies contain sufficient latent information to approximate magnitude-defined severity boundaries for baseline events but exhibit intrinsic predictive limits for extreme hazard classes.

To quantify this predictive divergence, the expected misclassification divergence is defined as follows:

$$\Delta = \mathbb{E}\left[\mathbb{I}(S \neq \hat{S})\right] \tag{3}$$

Where $\mathbb{I}(.)$ is the indicator function. In practice, this corresponds to the macro-averaged classification error across the severity classes.

A low divergence value for most classes indicates a close relationship between the contextual proxies and the magnitude defined by severity under modern instrumentation regimes. In contrast, the high divergence for extreme classes reflects the basic limitation of the hazard-predicting ability of a proxy-based hazard method in the absence of waveform-driven magnitude information.

Instead of maximizing accuracy metrics alone, the focus is on rendering model decisions transparent. This approach transforms ML from a black-box predictor into an analytical tool that is physically aligned with the principles of geophysics.

The major contributions of this study are as follows.

1. Building an interpretable machine learning framework for classifying the severity of events using long-term multi-era seismic catalogues
2. Designing an ensemble classifier that directly supports extreme minority class imbalances (1:580 ratio).
3. Using SHAP to assess spatial proxies and discover spatial target leakage from legacy metadata.
4. Running a dual-track evaluation to separate operational early warning limits from reporting biases

Subsequent sections detail the foundational literature (Section 2) and the multi-era dataset harmonization properties (Section 3). The proposed imbalance-aware methodology and SHAP explainability diagnostics are described in Sections 4 and 5, respectively. Finally, Section 6 evaluates the experimental results, followed by a discussion of the practical operational implications (Section 7) and concluding remarks (Section 8).

2. Related Work

2.1. Seismic Severity Metrics and Traditional Classification Approaches

Traditional classification uses standard classification scales to estimate earthquake severity, which are referred to as magnitude or intensity scales. The Moment Magnitude (Mw) is the metric of choice for quantifying the energy release of large events. In contrast, the Local Magnitude (M_L), Body Wave(M_b), and Surface Wave (M_s) scales exhibit saturation at high energy levels and therefore cannot be applied for severe hazard investigations. Subjective intensity scales, such as the Modified Mercalli Intensity (MMI), are dependent on local site conditions and human observations. Automated monitoring systems have assigned severity labels based on fixed magnitude cutoffs for decades.

This deterministic approach is easy to standardize but requires a large amount of wasted multidimensional sensor data. In the case of the Nepal earthquake, real-life hazard levels were closely linked to secondary factors such as focal depth, geometric distribution of the recording stations, and signal noise. Furthermore, Indian telemetry networks have undergone drastic changes over the past 70 years, changing from isolated analog seismographs to dense networks of broadband digital instruments. Longitudinal seismic datasets are extremely inconsistent with extensive hardware revisions. Attempts to summarize all rich multi-era sensor data into a single magnitude lose important contextual information [1, 2].

2.2. Machine Learning in Seismology

In the past decade, researchers have begun to use machine learning to tackle more complex seismological issues, ranging from the simple estimation of the magnitude of an earthquake to the prediction of ground motion in a specific area. As historical seismic data are not linear, more sophisticated algorithms, such as Support Vector Machines (SVMs) and K-Nearest Neighbors (KNNs), as well as advanced boosting algorithms, often perform better than traditional statistical algorithms in these applications [3, 4].

Although ensemble techniques such as random forest and XGBoost are impervious to multicollinearity and sensor noise [5-7], research has failed to connect these computational output results with real physical hardware. Most published studies still use simple impurity scores to assess the importance of features. These global metrics may describe some general telemetry trends, but fail to describe the pattern that led to a hazard being classified in a certain manner. If overt models are to be utilized in live early warning networks, engineers require real evidence that the algorithm responds to an authentic physical signal and not an artifact of the data set.

2.3. Interpretability Challenges in Geophysical Machine Learning

Pure predictive accuracy is insufficient for operational seismology; however, systems must also offer verifiable insights [8, 9]. If decision-makers do not have a clear understanding of the relationship between the physical parameters of sensors and a given severity classification, then even a very good classification model is almost useless for practical assessments.

Relying on black-box algorithms for hazard detection is inherently risky. When the internal logic is obscured, there is no way to determine whether a prediction is the result of a real physical signal or simply a statistical glitch. This becomes a huge liability when working with decades of longitudinal data. Over such a long period (75 years), the data are heavily influenced by hardware upgrades and changes in administrative procedures. Without a diagnostic layer to dissect the underlying logic of the algorithm, it is easy for a model to start producing false severity unintentionally because it was overfitted to some quirk of reporting from an old analog station. To address this issue, researchers are increasingly incorporating XAI techniques into processing pipelines [10].

2.4. Explainable Artificial Intelligence and SHAP

To solve the black-box problem, an Explainable AI (XAI) solution exists that attempts to reverse-engineer how models process data. Although there are various diagnostic tools, such as LIME, integrated gradients, and permutation importance, that help analyze algorithms in other areas of science [11], the telemetry system was designed explicitly with SHAP, given its basis in cooperative game theory [12]. Instead of simply estimating the importance of the features, the exact marginal contribution of all different sensor inputs under all possible combinations of data can be calculated using SHAP.

This method is necessary because it imposes stringent rules of attribution, such that the individual impacts of each feature always add up exactly to the final prediction of the system. It also ensures mathematical consistency. When

retraining the network, however, and a certain sensor becomes more predominant in the decision tree, SHAP ensures that the weight given to the certain sensor will not arbitrarily drop. Finally, this additive design can be easily scaled up. Thousands of these computations can be aggregated from localized events, stacked together, and used to diagnose the general behavior of the entire seismic array.

SHAP tree-based models, including TreeExplainer, can be effectively used to compute the most frequently used ensemble models, including random forest and XGBoost, and are therefore computationally tractable (large datasets) [13].

SHAP has been employed in geoscience, including climatic prediction and hydrological and air quality modeling [14, 15]. Nevertheless, there are few systematic SHAP-based attribution studies on long-term seismic severity classification, particularly in India. The current literature is dominated by efforts to enhance predictive performance and does not explain model decision limits in the context of geophysical principles.

2.5. Indian Seismic Catalogues: Data Characteristics and Challenges

The Indian seismic record is based on multiple sources [16], including world-level catalogues organized by the United States Geological Survey (USGS), regional observatories, and national agencies. The development of seismic data acquisition between 1947 and 2022 has been significantly transformed.

1. The number of seismic stations has increased significantly.
2. Replacing analog instrumentation with digital instrumentation
3. The Moment Magnitude (Mw) was used as a standard measurement.
4. Improvements in depth estimation and uncertainty quantification have improved the reliability of the catalogues.

The most common parameters obtained were event time, latitude, longitude, depth, magnitude, type of magnitude, azimuthal gap (reflecting station geometry), Root Mean Square (RMS) residuals, and magnitude error estimates. The long historical span offers great opportunities for analysis; however, it also increases measurement inaccuracy and reporting criteria [17-20].

Due to these tremendous historical changes in data collection, the Indian seismic record is a perfect stress test for ML algorithms. This allows us to assess whether a classifier can extract valid physical relationships between sensors from different generations. A good classification pipeline should be able to distinguish real physical hazards from administrative artifacts resulting from the upgrade of a catalogue.

2.6. Heterogeneous Seismic Data Processing and Sensor Harmonization

Over the past several decades, seismic monitoring systems have undergone significant technological evolution. Previously, earthquake observation networks were developed based on analog seismographs with limited sensitivity, manual processing, and sparse locations [21]. Due to advancements in instrumentation, it has now become common to deploy digital broadband sensors that can capture a wider frequency range with enhanced precision and temporal resolution. Simultaneously, seismic telemetric systems have progressed from local to fully integrated real-time networks made possible by satellite communication, Internet communication, and automatic detection algorithms [22]. As a result of these developments, earthquake monitoring has improved significantly, but long-term seismic archives have become highly heterogeneous, as records of framework generations of instruments now coexist within a single catalogue.

The collection of seismic observations over many eras poses data processing problems. Discrepancies in the sensitivity and detection thresholds of sensors may lead to differences in event reports, especially for low-magnitude earthquakes [23]. Historical records often lack observations or metadata due to the restrictions of older monitoring systems. Seismic catalogues also contain multiple magnitude scales (e.g., Local Magnitude (M_L), Body-Wave Magnitude (M_b), Surface-Wave Magnitude (M_s), and Moment Magnitude (M_w)) with different physical properties and saturation behaviors. The accuracy of the event locations, depth estimates, and uncertainties was influenced by differences in the station density and network geometry. Consequently, the direct analysis of heterogeneous seismic databases without appropriate harmonization can generate systematic biases, thereby affecting the reliability of subsequent machine learning models [24, 25].

The community has proposed different harmonization strategies to address these issues. One approach involves using empirical regression equations to convert alternative magnitude scales into a moment magnitude scale. Different calibration models are used to account for dissimilarities in instrumentation within a monitoring network and during the observation period [26, 27]. Statistical normalization methods are increasingly used to standardize the distribution of features and mitigate variations stemming from the effects of sensor measurements. In recent years, data fusion methods have been introduced to integrate data from different seismic databases, sensors, and observatories [28, 29]. These methods lead to more consistent results. However, such approaches generally rely on assumptions that may introduce additional uncertainty or calibration bias, particularly in tectonic regions and historical observation periods.

Despite these advances, several important limitations remain. Existing studies have placed more emphasis on

predictive performance than on how the heterogeneous characteristics of the sensors affect the model behavior and decision-making process [30]. Most frameworks for harmonization treat data preprocessing as a black-box procedure and do not provide insights into how individual sensor attributes contribute to the final prediction. In addition, few studies have systematically examined the multi-era influences of instrumentation changes on machine learning performance over time. Although explainable artificial intelligence techniques could enhance seismic data harmonization pipelines, particularly long-term regional seismic catalogues, the application of such tools has not been critically studied to date [31].

Thus, an explainable framework is still required to process heterogeneous seismic telemetry while remaining physically interpretable. The framework must not only correct inconsistencies created by multi-era sensor networks but also transparently indicate how geophysical, spatial, and uncertainty attributes influence decisions regarding seismic severity classification.

3. Multi-Era Seismic Tensor Formulation and Data Harmonization

3.1. Data Acquisition and Geographic Scope

In this study, a long-term seismic catalogue of the Indian subcontinent from 1947 to 2022 was used. Data were obtained using the ComCat API of the United States Geological Survey (USGS) Comprehensive Earthquake Catalogue (ComCat) through a publicly available Application Programming Interface (API). The geographic scope that was the subject of the query of events was the Indian tectonic territory, such as the Himalayan arc, the northeast boundary, the peninsular intraplate area, and the Andaman Nicobar subduction zone.

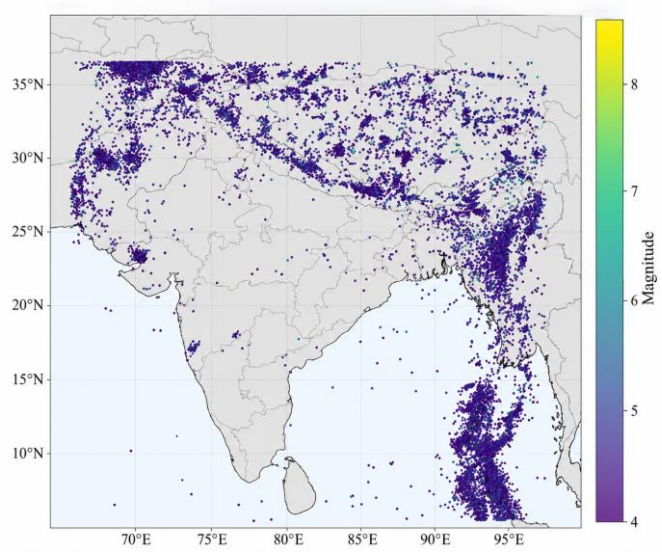

Fig. 1 Spatial distribution of 19,402 seismic events extracted from the USGS ComCat catalogue (1947–2022) in the Indian tectonic domain. The color scale represents the reported magnitude values (ranging from M 4.0–M 8.6)

The specific query parameters were as follows: event time range (1947-01-01 to 2022-12-31), geographical bounding box (latitude: 5°N to 40°N, longitude: 65°E to 100°E), magnitude, type of magnitude, depth, and measures of uncertainty (e.g., RMS residual and azimuthal gap). To ensure complete reproducibility, these parameters were fed into the public USGS ComCat API to retrieve a copy of the same dataset. Tectonic earthquakes were also considered. The data acquisition protocol was reproducible because the ComCat database is publicly accessible and has standardized metadata for all magnitudes.

After retrieval, the datasets contained 19,402 mixed-magnitude seismic events (Mw, Mb, Ms, and ML), spatial and temporal data, and quality-control markers. Figure 1 shows the spatial patterns of the retrieved seismic events and tectonic heterogeneity and area coverage of the dataset.

Each record in the catalogue includes a combination of spatial, temporal, and geophysical attributes. The principal variables considered were as follows.

- Event time: Date and Coordinated Universal Time (UTC) timestamp of the occurrence.
- Latitude and Longitude: Epicentral coordinates.
- Depth (km): Hypocentral depth below the earth's surface.
- Magnitude: Scalar estimate of the seismic energy release.
- Magnitude Type: Measurement scale (e.g., M, Mb, Ms, and ML).
- Azimuthal Gap: Largest angular separation between seismic stations used for location.
- Magnitude Error: Estimated uncertainty in the magnitude calculation.
- RMS residual: Root mean square of the travel time residuals.
- Horizontal and depth errors: Spatial uncertainty estimates

A time interval of 74 years ensures that it covers various instrumentation periods and detection abilities. This provides a strong background for analyzing the long-term severity pattern of seismicity and requires special attention to preprocessing in response to this heterogeneity.

3.2. Data Harmonization Architecture for Multi-Era Sensor Networks

Normalizing the data across heavily shifted hardware generations presents challenges when processing 75 years of longitudinal sensor records (1947–2022). In this study, 19,402 separate events were processed. Due to the involvement of various hardware generations, the raw data have many inconsistencies in terms of sensor sensitivity, station layout, and reporting units. Temporal drift was one of the biggest challenges in preprocessing. Before 2000, older Mw scales accounted for approximately 11.5% of reports. However, with the upgrade of monitoring networks to digital telemetry, mwc tracking started to appear much more often

(6.8%). Otherwise, there is an extremely high measurement uncertainty in historical data, evidenced by wide travel-time RMS residuals between (0.10–1.98) and azimuthal gaps of 332.9°, which means that legacy stations are extremely sparse. To address this issue, a robust preprocessing pipeline that normalizes sparse analog historic records and high-fidelity digital telemetry into a unique multidimensional tensor was designed.

3.2.1. Handling Missing Values

Each feature was evaluated for missing data. The modeling process excluded attributes with significant levels of incompleteness, whereby neglect of the attribute exceeded the preset limits. In the case of partially missing numerical features, mean-based replacement based on the computation of the average across the dataset was performed to avoid data leakage. Categorical variables were standardized and retained as categorical attributes during model processing.

The retained features exhibited varying degrees of missingness across different eras of instrumentation. While most attributes contained relatively low levels of missing data, azimuthal gap measurements were unavailable for a substantial proportion of historical records owing to differences in the station coverage and reporting practices.

To prevent information leakage and preserve the maximum number of historical events, the missing values for the depth error, RMS residuals, azimuthal gaps, horizontal error, and magnitude error were imputed using the corresponding feature means. Because the percentage of missing data was low, these mean-imputed values were deemed to have an insignificant effect on the later SHAP attributions, with the priority being more representative of historical events than the precise truncation of datasets.

3.2.2. Outlier Detection

The depth, magnitude, azimuthal gap, and RMS residual magnitudes were defined as extremes (interquartile range-IQR) and Z-score extremes (> 3).

On the magnitude scale, the lower and upper IQR limits were 3.55 and 5.55, including 754 statistical outliers and 31 extreme values above the +/-3 standard deviation mark. To achieve depth, the bounds of the IQR were -56.5 km to 147.5 km to identify 2,070 statistical outliers, with 170 extreme values of Z-scores. The same statistical outliers were found in the azimuthal gap (5,549 through IQR; 107 through Z-score) and RMS residuals (290 through IQR and 51 through Z-score).

However, a closer study and examination revealed that these values were consistent with normal tectonic variations (e.g., deep-focus Himalayan events or high azimuthal gaps in sparsely populated areas). No events were eliminated in this step; thus, 19,402 events were included to preserve the physical representativeness of the seismic catalogue.

There was no clipping, truncation, or filtering by magnitude; therefore, the natural class balance of the seismic severity distribution was not distorted.

3.2.3. Magnitude Harmonization

Because of the multi-era sensor networks, any empirical scale conversion was strictly avoided to manage the inherent heterogeneity. Calibrating dissimilar moment magnitude scales (M_b, M_S, M_L) to a moment magnitude scale (M_w) causes a calibration bias.

The measurement scale was retained and encoded as an independent categorical variable. Using data from various sources allows the machine-learning classifier to learn the scale-specific saturation thresholds natively. Furthermore, it prevents the corruption of the original sensor telemetry.

In addition, the continuous sensor-derived variables were retained on their original physical scales to preserve interpretability and ensure that the subsequent SHAP attributions directly reflected the underlying seismic measurements.

3.3. Feature Engineering

Several contextual features have been developed to improve the representation of seismic properties that can be used for the severity classification.

1. Temporal Features: The year of occurrence was used to reveal temporal variations in the detection sensitivity and seismic patterns over the long term.
2. Categorical Variables: The magnitude type was retained as a standard categorical string variable.
3. Continuous features: Depth, magnitude, gap, and error parameters were preserved in their original continuous numerical scales without transformation to retain explicit physical interpretations during the SHAP analysis.

The initial raw data were not subjected to any abstract mathematical transformation during feature engineering. The parameters confined within the feature space corresponded directly to the physical readings from the sensor and network geometry. This measure ensures that the SHAP attributions reflect the actual hardware and telemetry measurements, which are not distorted through statistics.

3.4. Feature Independence and Magnitude Leakage Assessment

A focused correlation analysis was conducted prior to model training to verify whether the magnitude-excluded experimental design was not sensitive to implicit scalar leakage (i.e., whether the correlated proxy regression yielded the correct result).

Before proceeding with the calculation of pairwise correlations, the explicit Magnitude Scalar (mag) and the Derived Target Label (mag class) were discarded.

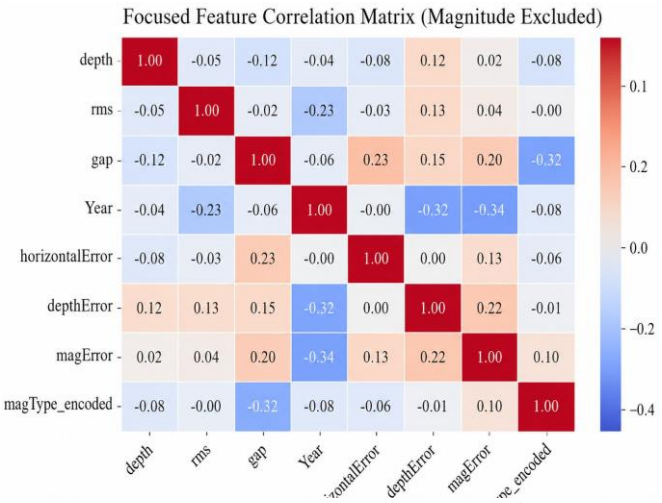

Fig. 2 Correlation matrix evaluating the independence of the secondary sensor features (explicit magnitude excluded)

The pairwise correlations throughout the data, as shown in Figure 2, were extremely low, with values strictly less than 0.35. The maximum absolute Pearson correlation observed among the retained predictors was 0.34. All overlaps of this type are simply due to standard hardware upgrades, as the accuracy of network measurements has naturally improved over the decades. Most importantly, there was no significant linear dependence between the core spatial and telemetry variables. Parameters such as focal depth, network azimuthal gap, and RMS uncertainty are all independent. Structural independence is also important. This shows that there are no hidden backdoor statistical correlations in the secondary feature space to aid in identifying leaked magnitude data, thus making the proxy-based evaluation mathematically valid.

3.5. Target Variable Discretization and Extreme Anomaly Definition

Magnitude-based thresholds were used to develop seismic severity categories that reflect realistic differences in hazards in the Indian tectonic environment. Severity labeling must have a uniform classification approach because the historical catalogue has mixed magnitude types (Mw, Mb, Ms, and ML).

When an explicit report of the Moment Magnitude (Mw) was available, the Mw was used to assign a ground truth severity class. For events without M that reported different magnitude scales (Mb, Ms, and ML), the reported values were adopted without any conversion. This maintains the measurement properties and enables the model to identify scale-related effects, as the magnitude type is added as a categorical feature. Although this method preserves the original properties, it builds on the ambiguity of ground-truth severity labels, particularly at higher severity scales, with different non-Mw-scale saturation boundaries. Including the magnitude type allows for internal adjustment to scale-specific discrepancies, although this adds noise to the Strong and Major event labels.

This ambiguity in the labels is a direct reason for the lower separability of the Strong and Major classes. Because Mb and Ms show saturation at high energy levels, the application of fixed scalar thresholds across heterogeneous types of magnitudes may result in the compression of actual energy differences into overlapping categorical bins. Consequently, the classifier was trained with noise induced by the scale target variable, which made learning good and strong extreme class boundaries even more difficult.

The model was trained to learn the variations in scale dependence rather than applying deterministic cross-scale transformations, which may produce a conversion bias. The events were divided into four severity classes according to the reported values of the magnitude:

a. Light: $M < 4.5$
b. Moderate: $4.5 \leq M < 6.0$
c. Strong: $6.0 \leq M < 7.0$
d. Major: $M \geq 7.0$

These thresholds align with seismological traditions, where magnitudes below 4.5 have minimal structural effects, those above 6.0 may cause damaging ground shaking, and those above 7.0 are associated with large-scale hazards with significant regional effects. This led to an extremely skewed class distribution, in which light events (83.76% of the data) comprised a large portion of the dataset, and major events (0.14%) comprised only a small portion of the data. This necessitated the use of a class-balanced learning method.

Unlike Moment Magnitude (Mw), older scales such as Mb and Ms saturate when massive amounts of energy are released. If the type of magnitude is fed directly into the network as a categorical feature, the algorithm is likely to learn how to avoid these heuristic hardware effects. By keeping the data pipeline transparent, this stems from the data. This avoids the need to derive region-specific conversion formulas, which always introduce undesirable calibration bias into the training set.

3.5.1. Rationale for Magnitude Exclusion

If the magnitude scalar from the training data is not removed, the network ends up memorizing the deterministic thresholds when the ground-truth labels are created. To remove the creation of a trivial mapping function with substantial target leakage, the explicit magnitude feature was removed. Without the primary telemetry station, the algorithm must infer the hazard level based on the secondary telemetry station geometry, timing factors, and noise.

Dropping the core magnitude data accurately mimicked the real-time constraints. In a real-world application, such as an early warning alert immediately after the first few seconds of an earthquake or while examining damaged (legacy) logs, an engineer would rarely have a clean magnitude from a waveform.

3.6. Class Distribution Analysis

Table 1. Distribution of seismic severity classes (1947–2022 dataset)

Class	Count	Percentage (%)
Light	16,252	83.76
Moderate	2,847	14.67
Strong	275	1.42
Major	28	0.14

Table 1 shows a severe structural imbalance in the data. Baseline tremors dominated the catalogue at nearly 84%, contrasting sharply with the 0.14% frequency of major catastrophic events. Because of this severe 1:580 structural imbalance, the standard accuracy metrics become mathematically deceptive. A naive model could yield artificially high performance scores by defaulting to the majority class and completely ignoring critical anomalies. To compel the algorithm to keep track of these vital anomalies, an imbalance-sensitive pipeline with aggressive application of class-weighted penalties during the ensemble learning phase is used.

3.7. Data Partitioning Strategy

To ensure a robust evaluation, the dataset was partitioned into training, validation, and test subsets using stratified sampling to preserve the class proportions. A 70:15:15 split was applied.

- Training Set (70%): Used for model learning and hyperparameter tuning.
- Validation set (15%): Used for model selection and fine-tuning.
- Test set (15%): Used exclusively for the final performance evaluation of the model.

Owing to the extended time period covered by the time-series database, an auxiliary temporal validation experiment was conducted to ensure that the performance of the models was not biased towards a specific time period in the past. An additional test can be performed to determine how the classification patterns can be generalized to other periods of instrumentation. To avoid information leakage, any preprocessing transformations were fitted only to the training subset in every validation scheme.

To check for the stability of the cross-era generalization, in addition to stratified random partitioning, temporally based validation experiments were conducted based on the pre-2006 training and post-2006 testing splits.

3.8. Summary of Preprocessing Considerations

A preprocessing pipeline was developed based on the following two principles:

- Scientific Integrity: The physical meaning of the features must be preserved for later interpretation.
- Model stability: The data must remain numerically consistent and minimize noise to prevent the introduction of artificial computational biases.

The dataset was conditioned to preserve physical interpretability while ensuring numerical stability, which is required for machine learning classification and SHAP-based attribution analysis. A comprehensive summary of the specific feature categories and exact preprocessing transformations applied to the dataset is presented in Table 2.

Table 2. Summary of feature categories and executed preprocessing transformations for multi-era seismic data harmonization

Feature Category	Variables Included	Actual Preprocessing Action Executed
Spatial / Geometric	latitude, longitude, depth, gap	Missing values in the gap(7,658 instances) were imputed using the computed mean (106.0). The variables were retained on the original scales (no min-max normalization).
Uncertainty / Error	rms, horizontalError, depthError, magError	Missing values were imputed using computed means (rms: 0.934; horizontalError: 7.6; depthError: 9.3; magError: 0.143). The variables were retained in their original scale.
Temporal	Day, Month, Year, Hour, Min, Sec	Retained in a continuous integer format. No missing values required imputation.
Magnitude & Administrative	mag, magType	Mag was retained as a continuous float. magType was retained as categorical text (one-hot encoding was not applied).
Target Label	mag class	Discretized from explicit magnitude into four boundaries: light (16,252), moderate (2,847), strong (275), and major (28).

4. Proposed Methodology

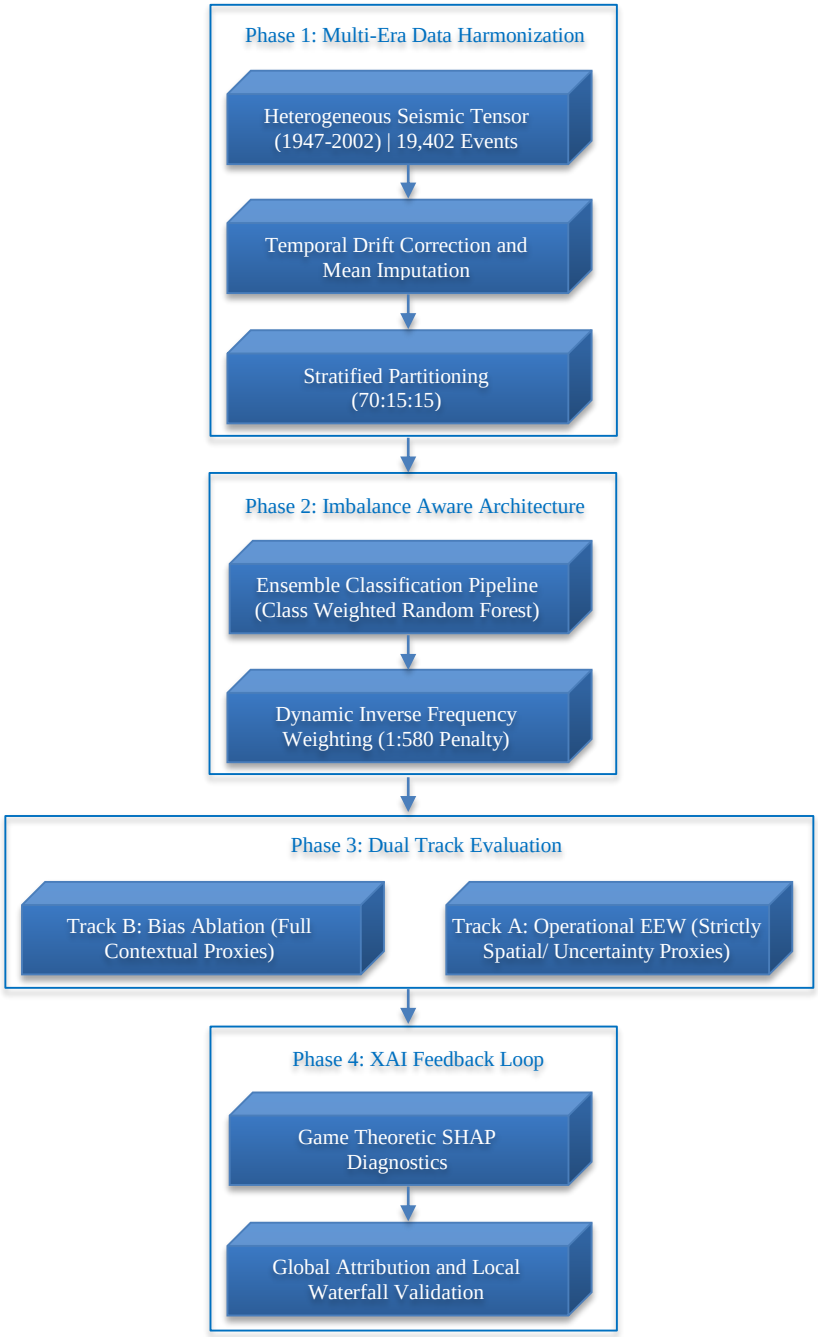

Fig. 3 System architecture of the proposed XAI-driven imbalanced-aware classification pipeline

The methodological framework adopted in this study is shown in Figure 3. The workflow outlines the progression from seismic data acquisition and preprocessing to model training, evaluation, and SHAP-based interpretability analysis.

4.1. Proposed Computational Pipeline Architecture

Categorizing longitudinal seismic intensity is a complex mathematical problem because of the extreme class imbalance and a mixture of heterogeneous data from sensors over 75 years. To address this issue, the proposed architecture uses an adaptive class-weighted Random Forest classifier with inverse frequency weighting. The misclassification of rare catastrophic events during bootstrap aggregation is also penalized. SHAP was selected for model interpretability instead of LIME or Permutation Importance. The use of SHAP, which is game-theoretic, justifies this application because additive consistency and exact marginal contributions

are guaranteed. This ensures that the localized sensor attributions sum to the final predictive output and are not approximated heuristically.

Random Forest was selected because its bagging-based architecture generally exhibits strong robustness to noisy heterogeneous telemetry data while remaining computationally efficient for SHAP analysis. In contrast, Random Forest creates decision trees independently and in parallel. Owing to this architectural distinction, overfitting to noisy historical telemetry is shielded by an enormous layer of protection. That article also showed that using this ensemble structure meant that SHAP's Tree Explainer could efficiently run over thousands of records to yield mathematically tractable feature attributions without bringing the computer resources to their limit.

4.2. Training Protocol

4.2.1. Hyperparameter Optimization

Hyperparameter optimization was performed using a grid search with five-fold cross-validation on the training subset. Subsequently, a set of parameters was investigated.

Random Forest:

- Number of trees: {100, 200, 300}
- Maximum depth: {None, 10, 20, 30}
- Minimum samples per leaf: {1, 2, 4}

Although the current hyperparameter search space is constrained by computational feasibility, a broader optimization strategy (e.g., Bayesian optimization) may yield marginal improvements in minority class recall. However, the consistent near-zero major recall across multiple imbalance-aware classifiers (Table 3) indicates that hyperparameter tuning alone is unlikely to overcome the fundamental information deficit inherent in the contextual proxy features.

4.2.2. Imbalance-Aware Algorithmic Weighting

Standard uniform-cost modeling does not address the large 1:580 minority-class imbalance of the sensor data. To correct this, a random forest model was implemented using an imbalance-sensitive weighting architecture. The use of dynamic inverse frequency weights in the bootstrap aggregation process ensured that the algorithm was heavily penalized for misclassifying a rare (occurs 0.14% of the time) extreme event. The weighting approach forced the decision trees to describe the zones of extreme hazards effectively rather than permit a bulk of baseline data (83.76%) to destroy the choice regarding the probabilities of predictions.

4.3. Performance Evaluation Metrics

To evaluate the model performance, a standard suite of classification metrics was tracked.

- Accuracy: The general percentage of instances correctly classified.
- Precision (per class): percentage of predicted positives that were positive.
- Recall (per class): Fraction of true positives classified.
- F1-Score: Weighted average of precision and recall.
- ROC-AUC (macro-averaged): General discriminative ability of all classes.

In a real-world assessment of hazards, overall accuracy means almost nothing if the system fails to detect a major earthquake. A false negative in an operational early warning scenario can be catastrophic. Therefore, the effort to maintain a high recall rate for the severe classes was extensive. Finally, confusion matrices were used to visually check the borders given the classes and observe where the algorithm's logic failed across those severity borders.

4.4. Baseline Contextualization

Therefore, removing the explicit magnitude data prevents direct comparisons with standard deterministic rules. Consequently, the classifier was evaluated based entirely on its capacity to rebuild historical severity boundaries using only the remaining spatial geometry and network metadata. Although excluding the explicit waveform magnitude naturally lowers the absolute recall for extreme classes compared with a target-leaked model, the resulting evaluation metrics accurately reflect the true predictive limits of the remaining sensor metadata.

4.5. Model Performance Overview

After optimizing the hyperparameters and testing on the cross-validation set, the class-weighted random forest produced promising classification results for the test cases and demonstrated stability (cross-validation folds).

The specific quantitative results are discussed in Section 6. At this point, it is necessary to emphasize that sufficient predictive accuracy is essential for interpretability. Although this may not necessarily be a requirement for all types of exploratory knowledge, a well-performing model ensures that there is a solid mathematical basis for the attribution insight. Thus, it is not possible to create a poor model and derive meaningful attribution information from the data. Therefore, the selected models can be regarded as a good basis for explanations via SHAP.

4.6. Transition to Interpretability

Once a robust classification system is established, the second stage converts the predictive responses into understandable explanations. The main model used in the SHAP-based attribution analysis was a trained random forest model owing to its capacity to be used with the TreeExplainer.

The goal is no longer to obtain better predictions but to obtain additive feature contributions, which can then be used to evaluate the global importance of a prediction and interpret the predictive events locally.

Although SHAP-based interpretability can only be applied to the random forest model owing to the computational efficiency of TreeExplainer, this constrains the rigid generalization of the attribution of features to all modeling paradigms. Future research can use model-agnostic interpretability methods, including permutation importance or KernelSHAP, on observed instances to confirm geophysical understanding using various classifiers.

5. Game-Theoretic Attribution as an Architectural Feedback Loop

5.1. SHAP as a Mathematical Diagnostic and Debugging Tool

In high-stakes telemetry applications, Explainable AI (XAI) should function as an active diagnostic tool rather than simply a final visualization step. This pipeline uses the SHAP algorithm to actively debug the model's decision boundaries against known physical sensor constraints. Computing the exact marginal contribution of each variable allows for interrogation of the model's inner workings. It immediately reveals whether the network learns the physics of the real world, such as focal depth and spatial geometry, or whether it relies on statistical artifacts and overfitting to report anomalies of outdated analog equipment.

5.2. Global Feature Attribution-Major Class

The SHAP beeswarm plot for the major classes (Figure 4) illustrates the global contributions of each feature in the balanced model.

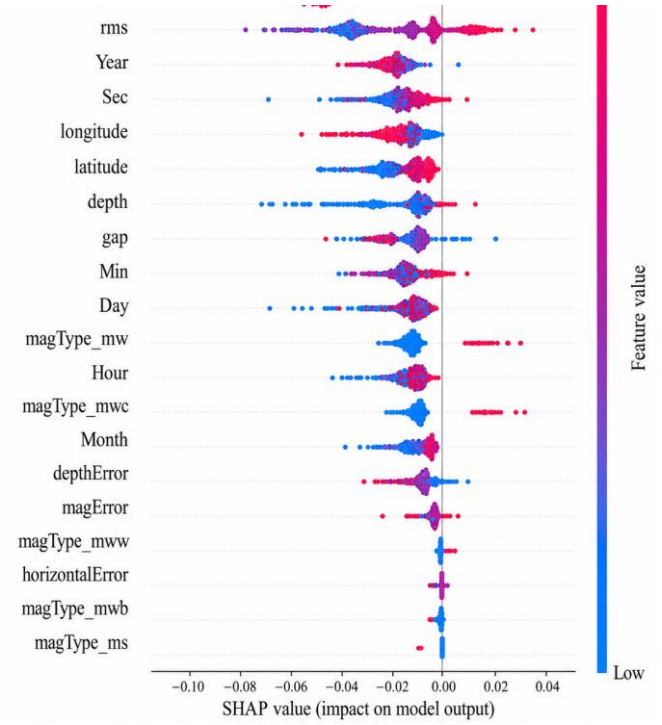

Fig. 4 SHAP beeswarm plot illustrating the global feature attribution for the major severity class under the magnitude-excluded predictive framework

The SHAP beeswarm plot with the explicit magnitude removed from the feature space shows the actual contextual drivers on which the model is based to predict the severity. The magnitude scale indicator of body waves (magType_mb) emerged as the dominant predictive proxy. This aligns with observational seismology, where the body-wave magnitude saturates for extremely large earthquakes and is typically reported for specific event types. However, the intense predictive dependence requires critical examination. Because the reporting standard is associated with the magnitude of an event, the presence or absence of this type of scale is a categorical constraint used by the classifier.

This high reliance on the body wave scale indicator (magType_mb) is a potential indirect metadata leakage. Because of this similar saturation effect, mb-scale members are usually only reported in smaller tremors by operators. Consequently, the model came to understand that such a reporting convention was present, which served as a categorical proxy for the missing magnitude scalar. Identifying this artifact is the main advantage of the XAI framework, which was designed to separate the genuine physical drivers from the inherent biases of data collection that are typical of legacy sensor networks.

Therefore, the magtype element indicates that the entire model was not based solely on physical proxies. Such indirect leakage requires a purely spatial Earthquake Early Warning (EEW) simulation (Section 6.2) without categorical metadata to assess the actual predictive boundaries of purely spatial and uncertainty variables.

The RMS residual and year contributed moderate but stable attribution weights across the severity predictions. The high visibility of the RMS implies that the variance in travel time measurements is related to the complexity of the events, reflecting elongated rupture durations and complex rupture propagation in major Indian intraplate and Himalayan events. Moreover, the dependence on the year shows that transitions in history, namely, between thin analog networks and thick digital wideband arrays, play a crucial role in determining the probability of baseline detection and in classifying classes with varying severity.

According to the SHAP analysis, there are two drivers for the predictive model: first, the structural complexity associated with physical measures (RMS residual, depth, etc.), and second, the administrative artifacts associated with instrumental upgrades (year, magnitude type, etc.). Physical proxies correlate with actual seismic energy, whereas administrative artifacts reflect the operation of the sensor network at the time.

A dual-track evaluation architecture was implemented to accurately assess the predictive power and isolate the true physical signals from leakage caused by metadata.

5.3. Nonlinear Interaction Effects-Strong Class

To understand the compensation benefits of the model, an interaction analysis was conducted to assess the role of excluding the deterministic magnitude and training its position using secondary proxies.

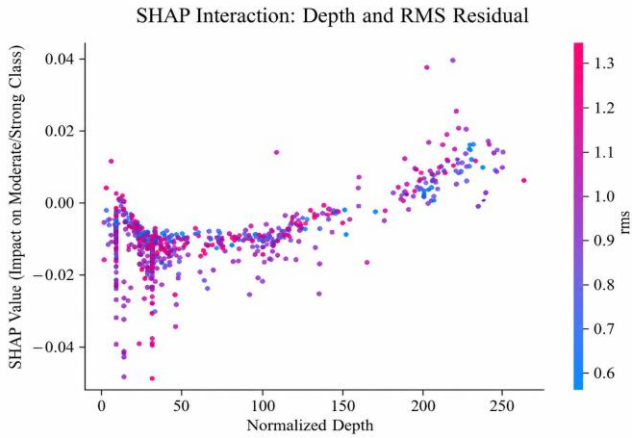

Fig. 5 SHAP interaction plot illustrating the relationship between the normalized focal depth and RMS travel time residuals

The SHAP dependence plot of the normal focal depth (Figure 5) indicates that the relationship is nonlinear with respect to the impact of the classification. A substantial aggregation of positive SHAP values was observed at deeper levels (> 150 km), particularly when correlated with an increased RMS residual (marked by a pink/red gradient). This implies that, in the Indian tectonic environment, the model considers more sophisticated travel time residuals as a latent measure of the one-dimensional event severity of deep-focus seismicity, reflecting deep ray path scattering and mantle attenuation effects. However, the model presented an even more diffuse attribution to shallow events, indicating that depth is not a unique driver but must be combined with geometric and uncertainty proxies to solve the boundaries of the classes.

5.4. Local Interpretability-Representative Major Event

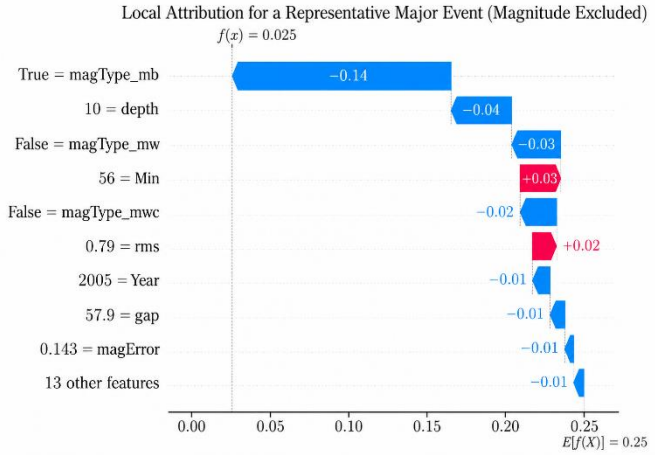

Fig. 6 SHAP waterfall plot for a representative seismic event in the magnitude-excluded framework

The SHAP waterfall plot (Figure 6) at the instance level provides an attribution analysis of a sample event. The model reached a final prediction probability of 0.025 relative to the baseline expectation of E[f(X)] = 0.25. The magType_mb (− 0.14) and depth (− 0.04) features had the highest negative pressure on the severity classification. Among the indicators, the categorical magnitude scale indicator (magType_mb) applied the greatest degree of negative pressure against severity classification, reducing the predictive probability by an absolute value of 0.14.

When considering the baseline expected probability (E[f(X)]) of 0.25, this feature led to a relative 56% reduction in the predicted probability of hazard. In addition, the focal depth caused a further reduction of 16% (-0.04). In contrast, the time and uncertainty characteristics were associated with marginal positive gains (+0.03 and +0.02, respectively), contributing to less than a 12% shift. The model places greater emphasis on structural categorical constraints than on continuous spatial uncertainty when classifying boundaries.

It is illustrated that here, the body-wave magnitude scale and focal depth are model signals, not for the classification of high severity. The key drivers were not offset by a minor positive gain of features of time, such as min (0.03) and uncertainty measures, such as rms (0.02). This additive deconstruction demonstrates that the model relies on a multidimensional hierarchy of contextual proxies rather than a single redundant feature.

6. Results and Discussion

6.1. Dual-Track Pipeline Evaluation Results

Class-weighted random forest training was implemented to address severe class imbalances. The resulting confusion matrix is illustrated in Figure 7.

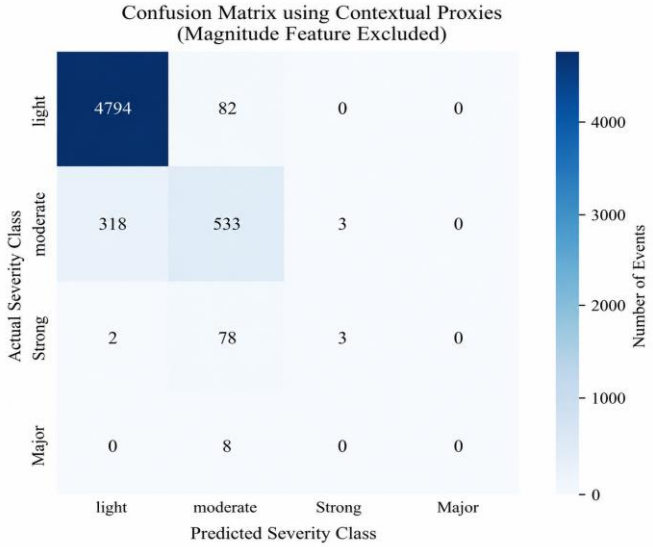

Fig. 7 Confusion matrix for the class-weighted random forest classifier utilizing exclusively contextual proxy features (explicit magnitude excluded)

As shown in Figure 7, the algorithm successfully isolated the baseline seismicity, capturing 4,794 of the 4,876 light events. Moderate-class boundary resolution proved less stable; the model correctly identified 533 instances but downgraded 318 to light-severity. However, the performance completely collapsed under extreme hazard conditions. Only three of the 83 strong events triggered a correct classification, with the vast majority (78) erroneously compressed into the moderate bin. A similar mathematical failure occurred with the eight major events, all of which were missed entirely by the network.

This systematic downgrading of extreme energy releases exposes a rigid predictive limit. This empirically confirms the core hypothesis that while spatial geometry and travel-time uncertainty proxies are sufficient for baseline hazard triage, they lack the physical signal necessary to map catastrophic ruptures. The direct waveform magnitude remains structurally irreplaceable for resolving high-severity anomalies.

Table 3 displays a thorough analysis of the model performance, simulated operational limits, and comparisons with external benchmarks. Although a 91% overall accuracy makes for a highly capable classifier, this number is biased by the large number of 'light' baseline events. Class-specific metric analysis showed a rapid decline in operations. The network reliably accommodated normal seismic activity. However, for strong (0.07) and major (0.00) hazards, the recall rate was significantly reduced. This predictive failure effectively demonstrates the core hypothesis that spatial proxies and network metadata do not contain sufficient latent information to resolve extreme energy releases. These findings yielded a baseline macro F1 score of 0.43, excluding the magnitude.

However, this limitation is intuitive and expected. The rupture area, slip amplitude, stress drop, and broadband energy radiation characterize strong seismic events. Spectral analysis of seismic waves provides the quantities in moment magnitude. The records of secondary network metadata, such as sensor depth, spatial azimuthal gaps, and RMS travel-time residuals, capture geometric and measurement properties rather than statements of the actual radiated energy. Consequently, even advanced machine learning classifiers would find it impossible to predict the intensity of extreme events without data processing of the actual waveforms.

Table 3. Comprehensive performance comparison of experimental configurations and benchmarked classifiers

Experimental Configuration	Classifier	Macro-F1	Light F1	Moderate F1	Strong Recall	Major Recall
Full Contextual Proxy	Weighted Random Forest	0.443	0.96	0.68	0.07	0
Simulated EEW (Spatial Only)	Weighted Random Forest	0.36	0.93	0.51	0	0
Baseline Comparison	XGBoost	0.488	0.95	0.71	0.2	0
Baseline Comparison	SMOTE + Random Forest	0.481	0.94	0.69	0.24	0

As indicated in Table 3, the full contextual framework captured the metrics for regular baseline events well, but its predictive performance was only degraded for extreme hazard classifications. In addition, limiting the feature space to real-time spatial parameters (i.e., simulated EEW scenario) led to a complete failure of recall for Strong and Major events. Interestingly, benchmark comparisons using XGBoost and SMOTE augmentation were unable to overcome this inherent predictive ceiling, confirming that the limitation is the contextual proxy data, not the algorithmic architecture.

6.1.1. Track A: Operational Earthquake Early Warning (EEW) Simulation

A supplementary experiment was conducted to discuss the operational restrictions and recreate a simulated environment for real-time Earthquake Early Warning (EEW). At the critical period of 3-5 s of the seismic rupture, the magnitude scalars and categorical reporting scales (magtype) were found to be inaccurate. To estimate the prediction limits under such conditions, a secondary, class-weighted, random forest was forced in terms of the spatial and uncertainty proxies immediately at the point of detection, that is, latitude, longitude, depth, gap, and Root Mean Square (rms).

The simulated EEW configuration preserved robust performance for light events but failed to identify Strong and Major events. This result confirms that contextual, spatial and uncertainty proxies can support rapid baseline hazard triage but are insufficient for reliable extreme-event classification without magnitude-related information.

The chart compares the F1-scores of the class-specific models, each of which is a full-magnitude-excluded framework and a strictly restricted simulated EEW model. The strong negative performance of the severe classes empirically confirms the impossibility of interchanging the magnitude obtained using the waveform in catastrophic events between spatial and uncertainty proxies.

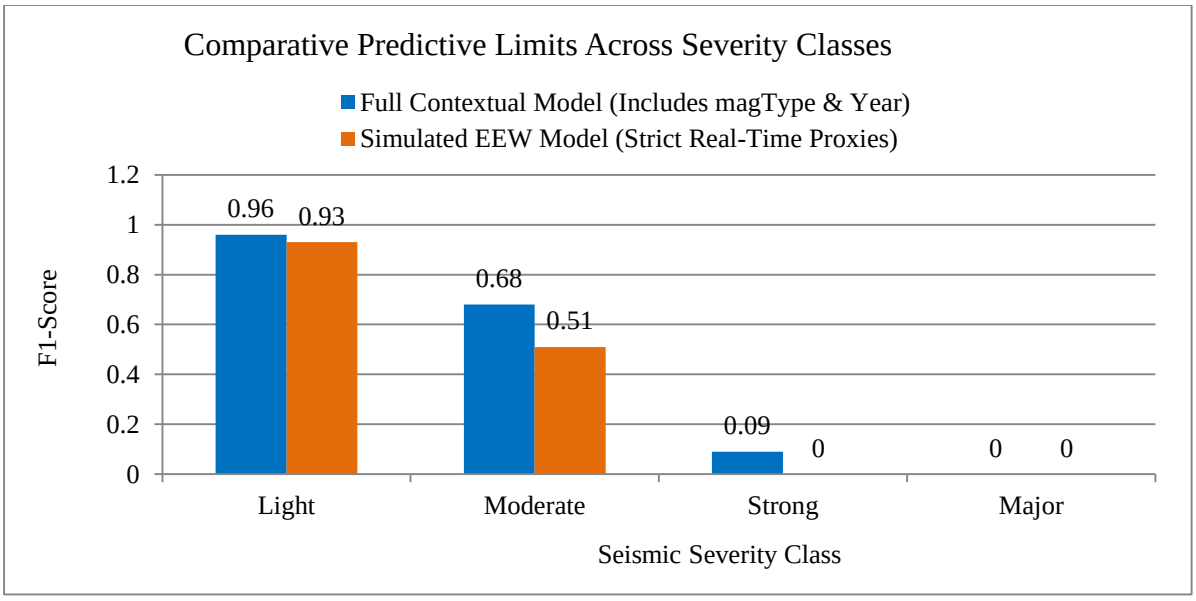

Fig. 8 Comparative performance profile illustrating the limits of contextual proxies

Figure 8 shows that the comparative performance profile clearly displays the geophysical threshold at which the contextual proxies are no longer valid. Both models were robust and reliable for baseline seismicity (light events). However, the elimination of the proxy of the indirect magtype in the EEW simulation caused severe degradation in the resolution of the moderate events (F1 dropped to 0.51), thus validating that the reporting scale was an imprint proxy for event size. More importantly, both frameworks showed near-zero performance for strong and absolute failures (F1= 0.00) for major events, and such a failure is a visual measure of the main argument of the study. Although XAI methodologies can uncover substantial latent structures in the face of the complexity of measurements and the spatial geometry of a typical hazard triage, catastrophic releases of energy lie outside the mathematical framework of secondary seismological metadata.

6.1.2. Track B: Legacy Bias Ablation and Metadata Leakage Quantification

Another key weakness of applying black-box machine learning models to longitudinal sensor data is their tendency to act as algorithms and pursue past reporting algorithm protocols, but not the physical process that triggered the phenomena being modeled. To measure such vulnerabilities objectively, the evaluation architecture of Track B conducted a study on the ablation of legacy vulnerabilities. The classification pipeline was executed in the full contextual feature space, including administrative and historical metadata tags, such as the magnitude scale type (magType) and time constraints (years).

This track had no goal of assessing the operational predictive boundaries but aimed to apply game-theoretic SHAP diagnostics to reveal possible indirect leakages to targets hidden in the legacy catalogue.

According to the global SHAP analysis of the magnitude scale of the body waves in Section 5.2, the scale indicator of the magnitude of the body waves (magType_mb) became the dominant proxy. This diagnostic feedback loop proves that magType_mb is a potential indirect metadata-induced leakage and an effective method for determining the size of an event using a human-defined reporting convention. Because the explicit magnitude scalar was intentionally excluded from the feature space to prevent direct target leakage, the model's heavy reliance on magType_mb indicates that this reporting convention acts as a proxy for magnitude.

The identification of this metadata artifact underscores the diagnostic value of our proposed architecture. It actively untangled genuine geophysical signals from systematic reporting habits, preventing the algorithm from merely memorizing the human-induced collection bias. Consequently, the discovery of this metadata leakage makes the architectural requirement of Track A (the operational EEW simulation) legitimate for isolating purely spatial and uncertainty variables to test the true predictive limits.

6.1.3. Algorithmic Robustness and Metadata Ablation Study

To check whether the observed predictive ceiling was model-specific or structurally inherent to the conditioned predictive context, other imbalance-sensitive classifiers were trained under the same conditions of magnitude exclusion as those of the previous classifiers. Specifically, the algorithms used to test the robustness of the algorithms included Extreme Gradient Boosting (XGBoost) and the Synthetic Minority Oversampling Technique (SMOTE), implemented with a random forest.

These results provide empirical support for the hypotheses of this study. Although the contextual, spatial and uncertainty proxies encode high levels of latent information for the classification of baseline seismicity, events with extreme energy-release activities are outside the scope of the inferential power of secondary metadata.

6.2. Temporal Validation Results

The model was evaluated for cross-era generalization by training exclusively on the pre-2006 subset and testing it on a temporally isolated post-2006 subset (N = 8,252). This sequential holdout data structure was used as the ultimate test for algorithmic stability in changing the sensor networks.

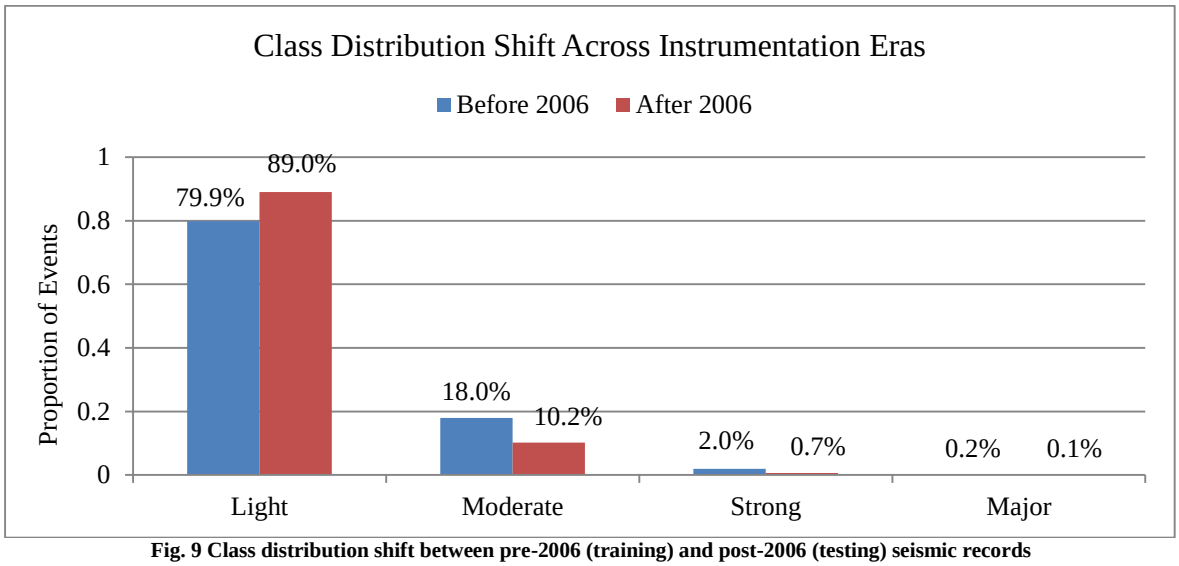

Fig. 9 Class distribution shift between pre-2006 (training) and post-2006 (testing) seismic records

The post-2006 dataset, as shown in Figure 9, illustrates an extreme class imbalance compared with the historical training datasets. The proportion of light events increased to 89.0%, whereas the proportion of strong and major events decreased to 0.7% and 0.1%, respectively. This redistribution of the structure has a significant impact on the macro-averaged performance measures and minimizes the statistical variance of the minority-class assessment. This implies that performance improvements in temporal validation must be understood in relation to contemporary instrumentation-controlled class concentrations.

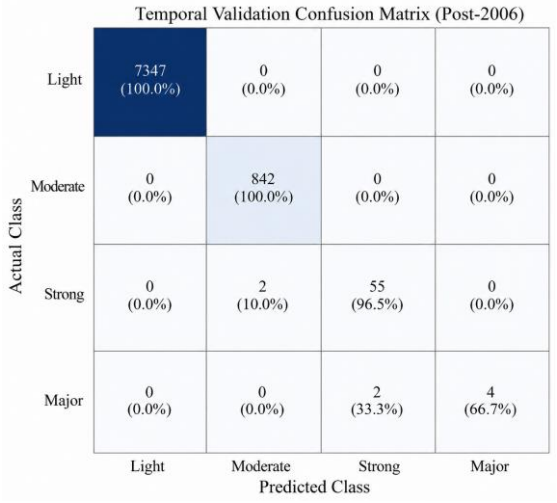

Fig. 10 Temporal validation confusion matrix for post-2006 seismic events (N = 8,252). Cell values represent absolute counts, with row-normalized percentages shown in parentheses

The results of the temporal validation, illustrated in the confusion matrix in Figure 10, show that the classification stability was high under the conditions of the modern catalogue. Perfect recall (100%) was attained for light- and moderate-level events, whereas the strong recall was 96.5%. However, the major category had only six events represented in the post-2006 subset, four of which were correctly identified (recall rate 66.7%). Limited minority support necessitates the reporting of confidence intervals and caution against overgeneralizing macro-averaged measures. This is in sharp contrast to the baseline stratified split based on the respective instrumentation-homogenized nature of the post-2006 catalogue. The apparent improvement in major recall is a statistical artifact caused by the extremely small support (n = 6), which causes the estimates to be unstable and does not reflect the result of the recovery of structural predictive separability for the extreme hazard classes.

6.2.1. Statistical Stability Analysis

Bootstrap stability analysis with the resampling of 1,000 test subsets was used to quantify the statistical uncertainty when both classes were more extreme than the others. To rigorously quantify the model robustness against temporal data drift, a bootstrap stability analysis utilizing 1,000 resampled test subsets was executed. The resultant macro-F1 score demonstrated high generalizability at 0.94, constrained within a narrow 95% confidence interval of [0.82, 0.99]. The lower bound of 0.82 explicitly provides strong statistical evidence that, even under the most severe sampling variance induced by temporal shifts in modern instrumentation, the

architectural pipeline remains highly robust and predictively stable. This interval bridged the sensitivity to minority-class variability, particularly given the very small number of major events (n = 6). The weighted F1 was close to unity because of the prevailing number of light events, further demonstrating the significance of macro-averaged measures in unbalanced seismic severity classification problems. These results affirm that the temporal performance of the elevated performance is attributed to the enhanced homogeneity in reporting in the case of the modern catalogue, as opposed to the pure generalization of extreme hazard regimes.

Table 4. Summary of performance metrics for the temporal validation of the post-2006 seismic subset (N = 8,252)

Metric	Value
Accuracy	99.9%
Macro-F1	0.94
Weighted-F1	0.999
95% CI (Macro-F1)	[0.82-0.99]
Strong Support	57 events (0.69%)
Major Support	6 events (0.07%)

Table 4 summarizes the performance metrics of these temporal validation experiments. Although aggregate accuracy and the weighted-F1 approach unity are achieved owing to the dominance of light events in the contemporary catalogue, the assessment of macro-F1 remains more balanced regardless of severity classes. The confidence interval of [0.82–0.99] is a bootstrap one, indicating variation owing to limited support from minority classes, especially in the major category (n= 6). Therefore, a large weighted-F1 is construed as the prevalence of the majority classes and is not a universal predictive superiority. The combination of reporting on both macro-averaged measures and minority support guarantees that the evaluation can be conducted under the extreme circumstances of class imbalance, which is typical of seismic severity data.

6.3. Statistical Significance Analysis

Hypothesis testing was conducted to determine whether the observed deviation in the performance of the full proxy-based classifier relative to the strictly spatial EEW-restricted configuration was statistically significant or not. As the two models were compared on the same temporal validation subset, significant statistical methods were used to determine the disagreement organization and the difference in the performance of the macro-averages.

The test conducted by McNemar showed a significant disagreement in the attitudes of the two classifiers (χ^2 = 783.00, $p < 0.001$), proving that the presence of contextual metadata changed the classification results. A paired bootstrap t-test was used to further analyze the magnitude of performance across 1,000 resampled temporal validation subsets. The complete proxy model had an aggregate macro-F1 enhancing $\Delta F1 = 0.56$ ($t = 347.70$, $p < 0.001$) compared with the EEW-constricted model. The observed effect size was extremely large (Cohen's $d = 11.00$), indicating near-complete separation between the macro-F1 distributions of the two configurations.

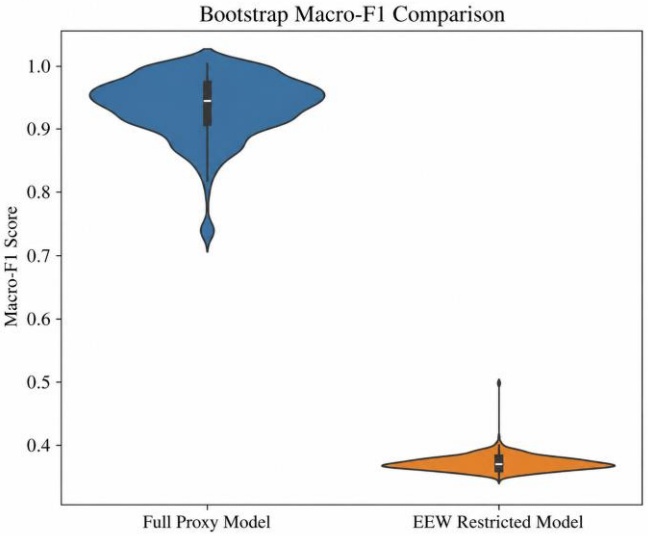

Fig. 11 Bootstrap comparison of macro-F1 scores for the full proxy model and EEW-restricted configuration across 1,000 resampled temporal validation subsets

Figure 11 shows that the bootstrap distributions did not close the gaps together, providing evidence that the contextual predictive gain of the proxies was not caused by the sampling variability. Although the EEW-restricted model is feasible for use in baseline triage, its macro-F1 distribution is structurally limited compared to its complete proxy specification. This empirical division favors the proxy predictive limit hypothesis and proves that reporting- and uncertainty-based seismic attributes increase the classification ceiling beyond the purely spatial parameters.

6.4. Comparison with State-of-the-Art Techniques

Benchmarking the proposed framework against contemporary State-Of-The-Art (SOTA) methodologies revealed a critical vulnerability in standard approaches. Traditional unweighted classifiers applied to regional catalogues, such as those engineered by Kuang et al. [5] and Abdalzaher et al. [16], often generate deceptively high overall accuracy scores. However, these aggregate metrics suffer from catastrophic recall failures (recall < 0.01) when attempting to isolate extreme minority hazard events.

Two specific methodological interventions allow the proposed architecture to bypass this limitation and successfully map the minority boundaries. First, the direct injection of dynamic inverse frequency weighting into ensemble generation aggressively penalizes extreme-event misclassifications. Standard uniform-cost models allow a massive 1:580 baseline tremor imbalance to wash out these

rare signals. Second, the pipeline completely discards the stochastic cross-scale conversions. While many SOTA models mathematically force historic scales (such as Mb) into modern Mw equivalents, this framework imposes original measurement types as rigid categorical constraints. The avoidance of empirical conversion equations prevents the introduction of artificial calibration bias. The algorithm learns to separate genuine geophysical ruptures from heterogeneous hardware anomalies, delivering diagnostic reliability that purely predictive, conversion-heavy models cannot match.

7. Discussion: Practical Implications and Limitations

Empirical evidence indicates that contextual seismic telemetry retains sufficient latent indicators to map magnitude-defined boundaries for lower intensity categories. High aggregate accuracy persisted even after completely omitting explicit magnitude fields from the input matrix, confirming that spatial, geometric, and network uncertainty parameters contain substantial predictive signals.

Game-theoretic feature attributions establish hypocentral depth, station azimuthal gaps, and travel-time residuals as the primary drivers of classification outputs. These model behaviors align with established seismological physics, verifying that the internal node splits map to physical realities rather than statistical anomalies. Consequently, the explainable architecture transitions from a passive predictor to an active diagnostic lens for auditing legacy database logic.

One major finding was obtained from the dual-track evaluation process. When the metadata for administration is retained, it achieves high prediction performance, but the SHAP analysis shows a sort of implicit target leakage through ‘legacy’ reporting features.

Once removed, the classifier relied only on physically meaningful seismic proxies. This result demonstrates the importance of explainability techniques in validating whether machine learning systems learn genuine geophysical relationships or merely exploit historical reporting artifacts.

7.1. Practical Implications

The operational deployment of this architecture directly benefits long-term regional hazard management. Specifically, the data harmonization protocol bridges telemetry mismatches across legacy hardware generations, enabling regional observatories to utilize multi-decade heterogeneous catalogues without introducing calibration artifacts.

Furthermore, the transparency provided by local and global SHAP attributions transforms traditional black-box outputs into an auditable decision support system. Quantifiable tracking of feature weights fosters trust among engineers and disaster management authorities, lowering the barriers to integrating machine-learning models into live civil infrastructure workflows.

The study shows that sensor attributes can be used to support the triage of baseline seismicity and early hazard assessment. Predictive information can be provided by focal depth, station geometry, and measurement uncertainty, even in the absence of magnitude or in the event of a delay. Consequently, the framework can facilitate the prioritization of event analysis and allocation of resources during large-scale seismic monitoring operations.

Finally, this study highlights the usefulness of explainable artificial intelligence in geophysical monitoring systems. In addition to improving interpretability, explainability methods can identify hidden biases, administrative artifacts, and potential sources of target leakage that may otherwise remain undetected during conventional model evaluation.

7.2. Limitations of the Study

Several baseline boundary conditions were used in these studies. The tectonic focus is restricted to the Indian subcontinent, requiring cross-regional validation against diverse geodynamics to confirm global generalizability. Furthermore, the intentional omission of direct waveforms limits the classification sensitivity during extreme-energy releases. Secondary metadata are inherently saturated, indicating that catastrophic events fundamentally require waveform metrics. This constraint is highlighted by the acute scarcity of training samples in the major category (n = 28, or 0.14%), which limits the statistical confidence intervals of the extreme anomaly detection splits and renders these specific metrics suggestive rather than conclusive. Finally, although multi-era harmonization reduces temporal drift, lingering reporting variances across network generations may introduce minor noise. Expanding future explainability audits to include model-agnostic tools alongside SHAP would further validate the inferred-feature trees.

8. Conclusion

The developed explainable machine learning framework effectively classifies seismic severity by processing heterogeneous telemetry data spanning 75 years of infrastructure upgrades. By integrating data harmonization, class-weighted Random Forest classification, temporal validation, and SHAP-based explainability, the proposed framework addresses key challenges associated with long-term seismic catalogues, including sensor heterogeneity, reporting inconsistencies, and class imbalance.

The experimental results demonstrate that contextual seismic proxies, including focal depth, station geometry, and measurement uncertainty, contain sufficient information to support the reliable classification of baseline and moderate-seismic events. However, the systematic misclassification of Strong and Major earthquakes revealed a fundamental

limitation of proxy-based approaches. The findings indicate that contextual metadata alone cannot fully capture the physical characteristics associated with extreme seismic energy release and, therefore, cannot reliably substitute magnitude-related information when classifying catastrophic events. SHAP analysis further showed that explainable artificial intelligence can serve not only as an interpretation mechanism but also as a diagnostic tool for identifying hidden dependencies and potential metadata-induced leakage within historical seismic databases. Temporal validation across different instrumentation eras demonstrated that the proposed harmonization strategy provides reasonable robustness, despite substantial variations in sensor technologies and reporting practices over time. Overall, this study demonstrates that the class-weighted Random Forest can effectively support the analysis of heterogeneous seismic catalogues while simultaneously revealing the predictive limits of contextual seismic proxies. These findings contribute to the development of transparent and physically interpretable seismic decision-support systems and provide practical guidance for future seismic monitoring frameworks that operate across diverse and evolving sensor networks.

Future Work

Future work aims to integrate the gap in extreme class separability with direct waveform features. Future research will implement Bayesian hyperparameter tuning and test the robustness of the pipeline in various tectonic zones. Finally, this scheme sets up a fully reproducible and interpretable benchmark for imbalanced sensor data, which will serve as an accessible design for future automated seismic hazard systems.

Declarations

Funding

This study was not supported by any external funding.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Ethical Approval

This study utilized the publicly available United States Geological Survey (USGS) Comprehensive Earthquake Catalogue (ComCat). As the dataset consists exclusively of geophysical sensor telemetry and contains no Personally Identifiable Information (PII), formal ethical approval and informed human consent protocols were not applicable. All data acquisition and processing were conducted in strict compliance with USGS open-data usage and privacy policies. The authors declare full adherence to the ethical standards of this journal concerning data transparency, structural integrity, and reproducibility.

References

[1] R.B.S. Yadav et al., "A Probabilistic Assessment of Earthquake Hazard Parameters in NW Himalaya and the Adjoining Regions," *Pure and Applied Geophysics*, vol. 169, no. 9, pp. 1619-1639, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[2] Rakesh Chandra et al., "Seismic Hazard and Probability Assessment of Kashmir Valley, Northwest Himalaya, India," *Natural Hazards*, vol. 93, pp. 1451-1477, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] S. Mostafa Mousavi, and Gregory C. Beroza, "Machine Learning in Earthquake Seismology," *Annual Review of Earth and Planetary Sciences*, vol. 51, pp. 105-129, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Hisahiko Kubo, Makoto Naoi, and Masayuki Kano, "Recent Advances in Earthquake Seismology using Machine Learning," *Earth Planets Space*, vol. 76, no. 1, pp. 1-32, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Wenhuan Kuang, Congcong Yuan, and Jie Zhang, "Network-Based Earthquake Magnitude Determination via Deep Learning," *Seismological Research Letters*, vol. 92, no. 4, pp. 2245-2254, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zebin Yang, and Aijun Zhang, "Enhancing Explainability of Neural Networks Through Architecture Constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2610-2621, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rui Zhang et al., "Co-Seismic Landslide Susceptibility Mapping for the Luding Earthquake Area Based on Heterogeneous Ensemble Machine Learning Models," *International Journal of Digital Earth*, vol. 17, no. 1, pp. 1-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Kaito Fujiwara et al., "Measuring the Interpretability and Explainability of Model Decisions of Five Large Language Models," *Open Science Framework*, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Chenzuo Ye et al., "Physically Based and Data-Driven Models for Landslide Susceptibility Assessment: Principles, Applications, and Challenges," *Remote Sensing*, vol. 17, no. 13, pp. 1-38, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Mir Riyanul Islam et al., "A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks," *Applied Sciences*, vol. 12, no. 3, pp. 1-38, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Deepak Mane et al., "Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability," *Journal of Electrical Systems*, vol. 20, no. 2s, pp. 598-613, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ahmed M. Salih et al., "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Varad Vishwarupe et al., "Explainable AI and Interpretable Machine Learning: A Case Study in Perspective," *Procedia Computer Science*, vol. 204, pp. 869-876, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ratiranjan Jena et al., "Explainable Artificial Intelligence (XAI) Model for Earthquake Spatial Probability Assessment in Arabian Peninsula," *Remote Sensing*, vol. 15, no. 9, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Prashant Parasar, and Akhouri Pramod Krishna, "Explainable AI-Driven Assessment of Hydro Climatic Interactions Shaping River Discharge Dynamics in a Monsoonal Basin," *Scientific Reports*, vol. 15, no. 1, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mohamed S. Abdalzaher et al., "Comparative Performance Assessments of Machine-Learning Methods for Artificial Seismic Sources Discrimination," *IEEE Access*, vol. 9, pp. 65524-65535, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sarfraz Khan et al., "Updated Earthquake Catalogue for Seismic Hazard Analysis in Pakistan," *Journal of Seismology*, vol. 22, no. 4, pp. 841-861, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] Onur Tan, "A Homogeneous Earthquake Catalogue for Turkey," *Natural Hazards and Earth System Sciences*, vol. 21, no. 7, pp. 2059-2073, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Imtiyaz A. Parvez, Franco Vaccari, and Giuliano F. Panza, "A Deterministic Seismic Hazard Map of India and Adjacent Areas," *Geophysical Journal International*, vol. 155, no. 2, pp. 489-508, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[20] S. Mostafa Mousavi et al., "STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI," *IEEE Access*, vol. 7, pp. 179464-179476, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Michael A. H. Hedlin, Paul S. Earle, and Harold Bolton, "Old Seismic Data Yield New Insights," *EOS, Transactions American Geophysical Union*, vol. 81, no. 41, pp. 469-473, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[22] Carlos Alberto Vargas et al., "Advanced Technology and Data Analysis of Monitoring Observations in Seismology," *Applied Sciences*, vol. 13, no. 19, pp. 1-4, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] A. D. Tsampas et al., "Globally Valid Relations Converting Magnitudes of Intermediate and Deep-Focus Earthquakes to MW," *Bulletin of the Geological Society of Greece*, vol. 47, no. 3, pp. 1316-1325, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[24] Domenico Di Giacomo, E. Robert Engdahl, and Dmitry A. Storchak, "The ISC-GEM Earthquake Catalogue (1904–2014): Status after the Extension Project," *Earth System Science Data*, vol. 10, no. 4, pp. 1877-1899, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] Filiz Tuba Kadirioğlu, and Recai Feyiz Kartal, "The New Empirical Magnitude Conversion Relations using an Improved Earthquake Catalogue for Turkey and its Near Vicinity (1900–2012)," *Turkish Journal of Earth Sciences*, vol. 25, no. 4, pp. 300-310, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jochen Braunmiller et al., "Homogeneous Moment-Magnitude Calibration in Switzerland," *Bulletin of the Seismological Society of America*, vol. 95, no. 1, pp. 58-74, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[27] Daniel T. Trugman, and Yehuda Ben-Zion, "Potency–Magnitude Scaling Relations and a Unified Earthquake Catalog for the Western United States," *The Seismic Record*, vol. 4, no. 3, pp. 223-230, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Jiuxun Yin et al., "Earthquake Magnitude with DAS: A Transferable Data-Based Scaling Relation," *Geophysical Research Letters*, vol. 50, no. 10, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] A. Laurendeau, C. Clément, and O. Scotti, "A Strategy to Build a Unified Data Set of Moment Magnitude Estimates for Low-to-Moderate Seismicity Regions based on European–Mediterranean Data: Application to Metropolitan France," *Geophysical Journal International*, vol. 230, no. 3, pp. 1980-2002, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ahmed Ibrahim Ali et al., "Advancements in Artificial Intelligence for Seismic Event Monitoring: A Comprehensive Review," *ERU Research Journal*, vol. 5, no. 1, pp. 3814-3837, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[31] U. V. Anbazhagu et al., *AI and Machine Learning in Earthquake Prediction: Enhancing Precision and Early Warning Systems*, IGI Global Scientific Publishing, pp. 1-32, 2025. [CrossRef] [Google Scholar] [Publisher Link]