doi:10.26443/seismica.v5i1.2068

Exploration of Machine Learning Methods to Seismic Event Discrimination in the Pacific Northwest

Akash Kharita * [1], **Marine Denolle** [1], **Alexander R. Hutko** [1,2], **J. Renate Hartog** [1,2], **Stephen D. Malone** [1,2]

[1]Earth and Space Sciences, University of Washington, Seattle, USA, [2]Pacific Northwest Seismic Network, University of Washington, Seattle, USA

Author contributions: *Conceptualization*: AK,MD. *Methodology*: AK,MD. *Software*: AK. *Validation*: AK,AH,SM. *Formal Analysis*: AK,MD. *Investigation*: AK,MD,AH. *Resources*: RH. *Writing - Original Draft*: AK. *Writing - Review & Editing*: AK,MD,AH,SM,RH. *Visualization*: AK. *Supervision*: AK,MD,RH. *Project Administration*: MD,RH. *Funding Acquisition*: MD,RH.

Abstract Accurately separating tectonic, anthropogenic, and geomorphologic seismic sources is essential for Pacific Northwest (PNW) monitoring but remains difficult as networks densify and signals overlap. Prior work largely treats binary discrimination and seldom compares classical machine learning (feature-engineered) and deep learning (end-to-end) approaches under a common, multi-class setting with operational constraints. We evaluate methods and features for four-way source discrimination – earthquakes, explosions, surface events, and noise – and identify models that are both accurate and deployable. Using ∼200k three-component waveforms from >70k events in an AI-curated PNW dataset, we test random-forest classifiers on TSFEL, physics-informed, and scattering features, and CNNs that ingest time series (1D) or spectrograms (2D); we benchmark on a balanced common test set, a 10k-event network dataset, and out-of-domain data (global surface events; near-field blasts). CNNs taking spectrograms lead with accuracy performance over 92% for within-domain (as a short-and-fat CNN SeismicCNN 2D) and out-of-domain (as a long and skinny CNN QuakeXNet 2D), versus 89% for the best random forest; performance remains strong at low signal-to-noise ratio (SNR) and longer distances, and generalizes to independent network and global datasets. QuakeXNet (2D) is lightweight (70k parameters; 1.2 MB) and integrated into SeisBench. On commodity hardware, it processes a full day of 100 Hz three-component data in 9 s. These results show spectrogram-based CNNs provide state-of-the-art accuracy, efficiency, and robustness for real-time PNW operations and transferable surface-event monitoring.

Production Editor: Gareth Funning
Handling Editor: Marlon Ramos
Copy & Layout Editor: Abhineet Gupta

Received: October 10, 2025
Accepted: January 18, 2026
Published: February 26, 2026

1 Introduction

The Pacific Northwest (PNW) region of the United States, situated at the dynamic boundary between the North American continental plate and the Juan de Fuca oceanic plate, presents unique challenges and opportunities in seismic monitoring in a multi-geohazard-prone landscape with a subduction-zone plate boundary. The PNW experiences diverse seismic sources (Fig. 1), including large megathrust earthquakes (e.g., Witter et al., 2003), intraslab (e.g., Ichinose et al., 2004) and crustal earthquakes (e.g., Gomberg and Bodin, 2021), slow repeating earthquakes (e.g., Bartlow, 2020; Rogers and Dragert, 2003; Wech and Bartlow, 2014), tectonic tremors (e.g., Wech et al., 2010), and low-frequency earthquakes (e.g., Royer and Bostock, 2014). Beyond earthquakes, the region includes more than a dozen potentially active Cascade volcanoes, and extensive mountain ranges that experience frequent landslides and debris flows (e.g., Luna and Korup, 2022). Anthropogenic activities, such as quarry blasts, generate ground motion intensities comparable to those of small-magnitude earthquakes, further complicating the source of this seismicity (Kramer et al., 2024). Such a variety of seismic sources necessitates robust classification methods to accurately label and catalog these events.

The PNW Seismic Network (PNSN) (Hellweg et al., 2020), a key component of the Advanced National Seismic System (ANSS), has been operating since 1969 and currently manages over 600 seismic stations in the states of Washington and Oregon, providing essential data for seismic event analysis. Current event detection relies on traditional techniques such as the Short-Time Average to Long-Time Average (STA/LTA) ratio algorithm (e.g., Allen, 1982). While effective for basic event detection, this approach has limited accuracy when discriminating between visually similar waveforms from different event types, such as earthquakes, controlled explosions, and mass wasting or slope failure events. These limitations have become more pressing with the expansion of seismic networks and the increasing volume of data (e.g., Carniel and Raquel Guzmán, 2021; Kong et al., 2018), particularly in regions like the PNW, where the simultaneous occurrence of multiple seismic sources adds to the complexity of waveform interpretation. Traditional discrimination techniques – such as P/S spectral (amplitude) ratios and differential magnitudes (e.g., ML–MC) – were largely developed for binary earthquake–explosion classification (e.g., Koper et al., 2016, 2024), but they also face several limitations. First, the required parameters (reliable phase measurements and stable magnitude estimates) are often un-

*Corresponding author: ak287@uw.edu

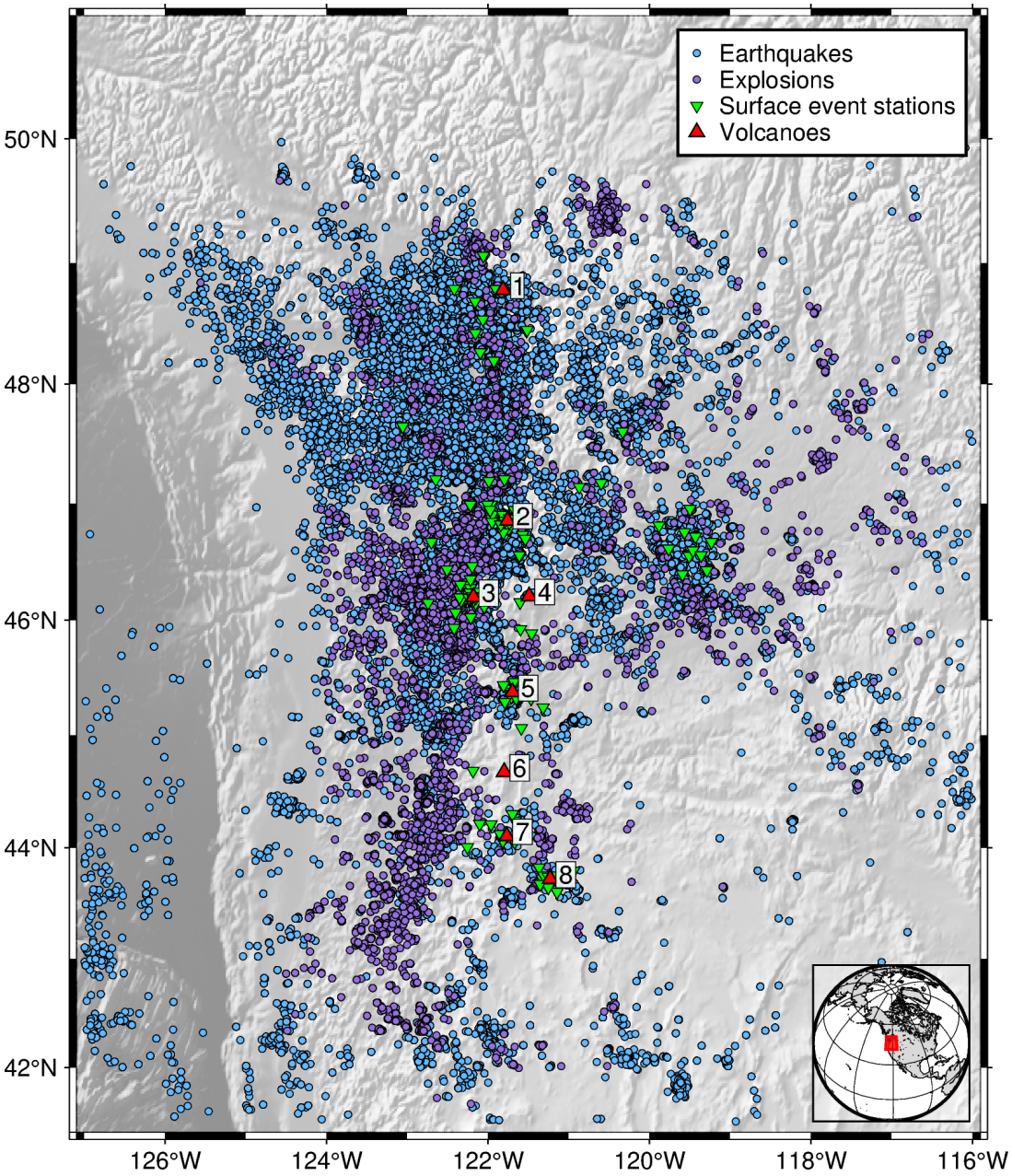

Figure 1 Map of seismic events in the curated catalog by Ni et al. (2023). Earthquakes (blue circles) and explosions (purple) are located by the PNSN. Surface events are only marked at the seismic stations where they are recorded (green triangles). Major volcanoes are shown as red triangles and are numbered as follows: 1. Mt. Baker, 2. Mt. Rainier, 3. Mt. St. Helens, 4. Mt. Adams, 5. Mt. Hood, 6. Mt. Jefferson, 7. Three Sisters, and 8. Newberry Volcano.

available during preliminary processing, which limits their utility for near–real-time classification. Second, these discriminants can exhibit increased variability and reduced robustness at local distances due to strong path and site effects, often requiring region-specific calibration (e.g., Pyle and Walter, 2019, 2021). Finally, some metrics – particularly ML-MC – can act primarily as a depth/shallow-source discriminant rather than a unique indicator of source type, which can complicate interpretation when different source types occur at similar depths (e.g., Koper et al., 2016, 2020, 2024).

Small magnitude earthquakes (M_L<3) have been difficult to distinguish from mining and single-shot explosions recorded at local distances (<50–150 km) due to similarities in their seismic waveforms and spectral characteristics (Fig. 2), and particularly at frequencies below 10 Hz (e.g., Tibi et al., 2018; Pyle and Walter, 2019; Wang et al., 2020; Koper et al., 2020; Pyle and Walter, 2021). Similarly, seismic signals generated by mass movements on volcanoes, such as landslides, debris flows, and lahars, closely resemble those associated with low-frequency volcanic seismicity and shallow volcano-tectonic earthquakes. These events are characterized by emergent waveforms, indistinguishable P and S phases, and dominant frequencies below 5 Hz (e.g., Wassermann, 2012; Allstadt et al., 2014, 2018). Furthermore, seismic signals from rockfalls with a significant free-fall component can exhibit similarities to small earthquakes when recorded locally (Hibert et al., 2011, 2014). In volcanic regions near populated areas, volcano-seismic signals, including those related to low-frequency volcanic events or shallow volcano-

tectonic earthquakes, can also be misinterpreted as anthropogenic noise (Wassermann, 2012).

Many studies have focused on the binary classification of seismic events, particularly distinguishing between earthquakes and explosions (e.g., Koper et al., 2020; Kong et al., 2022), or other sources such as slope failures (e.g., Wenner et al., 2020; Chmiel et al., 2021; Hibert et al., 2017) or icequakes (e.g., Pirot et al., 2023; Kharita et al., 2024). Binary classification is often simpler because it requires separating only two classes and learning a single decision boundary, and it is frequently studied in settings where the target classes have relatively distinct features. As a result, reported performance in binary earthquake–explosion problems is often high, with many studies reporting accuracies exceeding 95% (e.g., Wang et al., 2020; Koper et al., 2020).

By contrast, in complex environments such as volcanic regions, where seismic signals from landslides, pyroclastic flows, and low-frequency volcanic events share similar waveforms, multi-class classification is necessary and considerably more difficult (e.g., Wassermann, 2012; Allstadt et al., 2014, 2018). Discriminating among multiple event types is more challenging because the boundaries between classes can be less distinct, and performance is typically lower, often in the range of 75–90% (e.g., Hibert et al., 2014, 2017, 2019).

Classical Machine Learning (CML) and Deep Learning (DL) have shown promise in seismic event classification by enabling automated feature extraction and effective discrimination between event types, even in noisy environments. CML techniques, which use engineered features as input (see Fig. 3), may offer interpretability and have shown success in distinguishing between event classes such as earthquakes, explosions, and glacier seismicity (e.g., Koper et al., 2016; Zeiler and Velasco, 2009; Kharita et al., 2024; Hibert et al., 2017; Pirot et al., 2023; Domel et al., 2023; Wang et al., 2022). However, these approaches often require extensive and computationally costly feature engineering, which may limit adaptability to new event types. DL methods, on the other hand, automatically extract features from raw data through neural network optimization (see Fig. 3), resulting in good classification performance for more nuanced differences in seismic signals (e.g., Mousavi and Beroza, 2022; Bergen et al., 2019). Studies have typically chosen either CML or DL approaches and used varied data sets, which limits our ability to draw a general understanding of feature extraction and machine learning classification on multi-class discrimination.

Interpretability is crucial for understanding model behavior and generalizability. In CML, it is often assessed via feature importance, which identifies the waveform attributes most responsible for separating classes and can guide physical interpretation of the underlying source processes. The interpretability of deep learning models is less straightforward, mostly because isolating parts of the feature space that most contribute to the classification is buried in neural networks. There exist methods today to estimate feature importance in seismological applications. For example, Kong et al. (2021) and Kong et al. (2022) used a method called Grad-CAM to trace back regions (time and feature space) of the seismic waveforms that most contribute to classification, Linville et al. (2019) used attention mechanisms to identify key temporal patterns in seismic signals, and Clements et al. (2024) directly visualized a given feature value after activation to isolate the wave types that contribute to improving shaking intensity forecast for early warning. A comprehensive comparison between CML-based and DL-based methods is still lacking.

Accurate multi-class classification in the PNW is not only a technical challenge but also a scientific and operational priority. The region's overlapping seismic sources generate visually similar signals, making them difficult to distinguish in real time for the PNSN analysts. Improving classification reduces analyst workload, enhances the reliability of catalogs, and supports downstream applications such as hazard assessment, early warning, and tomography. Beyond operations, clean catalogs enable new scientific insights, for instance, clarifying the physical differences between explosions, earthquakes, and mass movements, or quantifying the seasonal and climatic controls on surface processes. In particular, building the first comprehensive catalog of PNSN "surface events" – a catch-all analyst label near volcanoes dominated by mass-movement signals (e.g., rockfalls/avalanches), and sometimes including glacier-related or low-frequency volcanic activity – would provide a baseline for future research on landslide frequency–magnitude statistics and volcano–geomorphic interactions.

The unique contribution of this work is a comprehensive evaluation of engineered waveform feature sets and model architectures (classical machine learning and deep learning) for multi-class seismic event classification. We compile and curate diverse datasets spanning tectonic, anthropogenic, and geomorphological sources (including curated data from Ni et al. (2023), an exotic event catalog from Bahavar et al. (2019), and additional data prepared for model development and testing), and evaluate model performance, generalization to unseen data, and interpretability through feature-importance analyses.

2 Data

This study utilizes multiple datasets to train and test the models. The first is a subset of the comprehensive dataset curated by Ni et al. (2023), which spans 21 years, from 2002 to 2022, and additional waveforms we collected (as described later in the section). The curated dataset comprises approximately 200 000 seismic waveforms and associated metadata, corresponding to approximately 70 000 events. The source types of these events are primarily classified into four distinct categories: earthquakes, explosions, surface events, and noise, with other categories only having a few samples. PNSN only locates the sources of earthquakes and explosions, which we show in Fig. 1. PNSN analysts only identify the surface events at one or two stations; thus, their locations are assumed to be close to the seismic station that records them, as shown in Fig. 1. The classes are not balanced (see Supp. Fig. S1): About 90% (163 064 out of the 185 081 labeled seismic records) are

earthquakes, which is expected given the PNSN's mandate to monitor earthquakes. The noise class is artificially generated by Ni et al. (2023) to provide sufficient examples and was verified using the transfer-learned earthquake transformer model (Mousavi et al., 2020) to ensure that it does not contain earthquake waveforms. There are over 8000 examples of surface events and 15 000 examples of explosions. Note that this dataset was manually classified by PNSN analysts. Through our preliminary analysis, we found that a small portion of the data may have been mislabeled (shown later in Section 6.1).

We refer to traces as fixed-length, three-component time series. We use the curated Pacific Northwest dataset of Ni et al. (2023), in which waveforms from mixed sampling rates (commonly 40 Hz for BHE/BHN/BHZ, and 100 Hz for EHE/EHN/EHZ and HHE/HHN/HHZ) are resampled to 100 Hz to support deep neural networks with fixed input sizes (including upsampling of BHE/BHN/BHZ channels). For stations with only a vertical component available (e.g., EHZ), the missing horizontal components are filled with zeros to maintain a consistent three-component input representation across stations. In the following, we describe the characteristics of each class.

2.1 Earthquakes

Earthquakes within the dataset are those whose information is sent to the ANSS Comprehensive Earthquake Catalog (ComCat) from the PNSN. Ni et al. (2023) collected the waveform data and associated attributes in a metadata table that encompasses event-related attributes, such as source location, depth, local or duration magnitudes, and station-specific terms, including P and S picks, which PNSN analysts have generated. Their magnitudes span from -0.3 to 5 (local magnitude for most events after 2014, duration/coda magnitude for events before 2014, Ni et al., 2023); with a peak around 1 to 3. The PNW subduction zone hosts a diverse range of earthquake sources, including shallow crustal events, deeper intraslab earthquakes that extend to depths of ∼100 km, and volcano-tectonic earthquakes. The depth distribution is bimodal: while the majority of earthquakes occur at shallow depths, a secondary concentration is found between 30–50 km, indicating the presence of intraslab seismicity. The corresponding seismic waveforms exhibit distinct P and S arrivals (Fig. 2), characterized by an impulsive onset and relatively higher frequencies exceeding 5 Hz (Fig. 2). The duration of these waveforms mostly varies between 10 and 30 seconds (Fig. 2) (see Supp. Fig. S2). For each event sent to the ANSS ComCat catalog, Ni et al. (2023) selected events with both P and S picks, which preselects for high-quality data. Each waveform in the curated dataset spans 150 seconds, sampled 50 seconds before and 100 seconds after the source origin time. The label for earthquakes is referred to as `eq`.

2.2 Explosions

PNSN analysts classify events similar to shallow quarry blasts and those occurring near recognized quarry blast sites as "probable explosions", "shots", or simply "explosions". While these mechanisms differ, they are collectively categorized as "explosions". In our dataset, it is dominated by probable quarry blasts (∼99%), which are typically routine industrial single-shot chemical explosions. We treat this as an operational explosion category and do not further subtype explosions here; a dedicated analysis of waveform similarity across explosion subtypes is left for future work. These events are characterized by a prolonged coda (Fig. 2) and relatively lower and monochromatic frequency content, with dominant frequencies typically falling within the range of 1 to 3 Hz (Fig. 2). The label for explosions is referred to as `px`.

2.3 Surface Events

Surface events are identified at the PNSN near volcanoes, seemingly as emergent ground motions, and categorized by PNSN analysts as "surface events". This category typically includes a variety of mass-movement sources likely associated with rockfalls and avalanches, although some may also resemble low-frequency volcanic earthquakes or glacier-related activity, depending on the setting. First arrivals are typically picked at one or two stations per event and stored in the ANSS Quake Monitoring System (AQMS) database (Renate Hartog et al., 2019). Given the unknown source origin time of the surface events, waveforms are sampled from 70 seconds before the first arrival pick, as designated by PNSN analysts, and extend to 110 seconds post-P-wave pick to accommodate potentially longer-duration events. The waveforms of these events exhibit a wide range of characteristics, lasting from 20 seconds to several minutes for the longest, but rare, debris flows (Fig. 2). The waveforms are less broadband than earthquake waveforms, typically falling within the range of 1 to 15 Hz, and they often feature emergent onset (Fig. 2) (see Supp. Fig. S1). While the origin time and location of these events are not confirmed, the PNSN analysts have mostly labeled these events at stations near volcanoes. The most active places for such labeling are Mt. St. Helens and Mt. Rainier. A thorough characterization of these events, including their origin and mechanism, is worthy of investigation and will not be addressed in this paper. In the curated dataset, surface events were initially underrepresented, with waveform data available for only about 5200 events (or 8912 traces). This imbalance has hindered model performance in our preliminary exploration, limiting its ability to learn distinctive characteristics of surface events. To overcome this, we expanded the dataset by incorporating three-component waveform data from additional stations located within 30 km of each event. Although phases were not always picked and reported on these nearby stations, they often recorded strong signals due to their proximity to surface sources. This augmentation added 6495 new three-component traces, increasing the total count to 15 407 traces and thereby enhancing the dataset's diversity, while improving the model's capacity to generalize surface event patterns. The label for surface events is referred to as `su`.

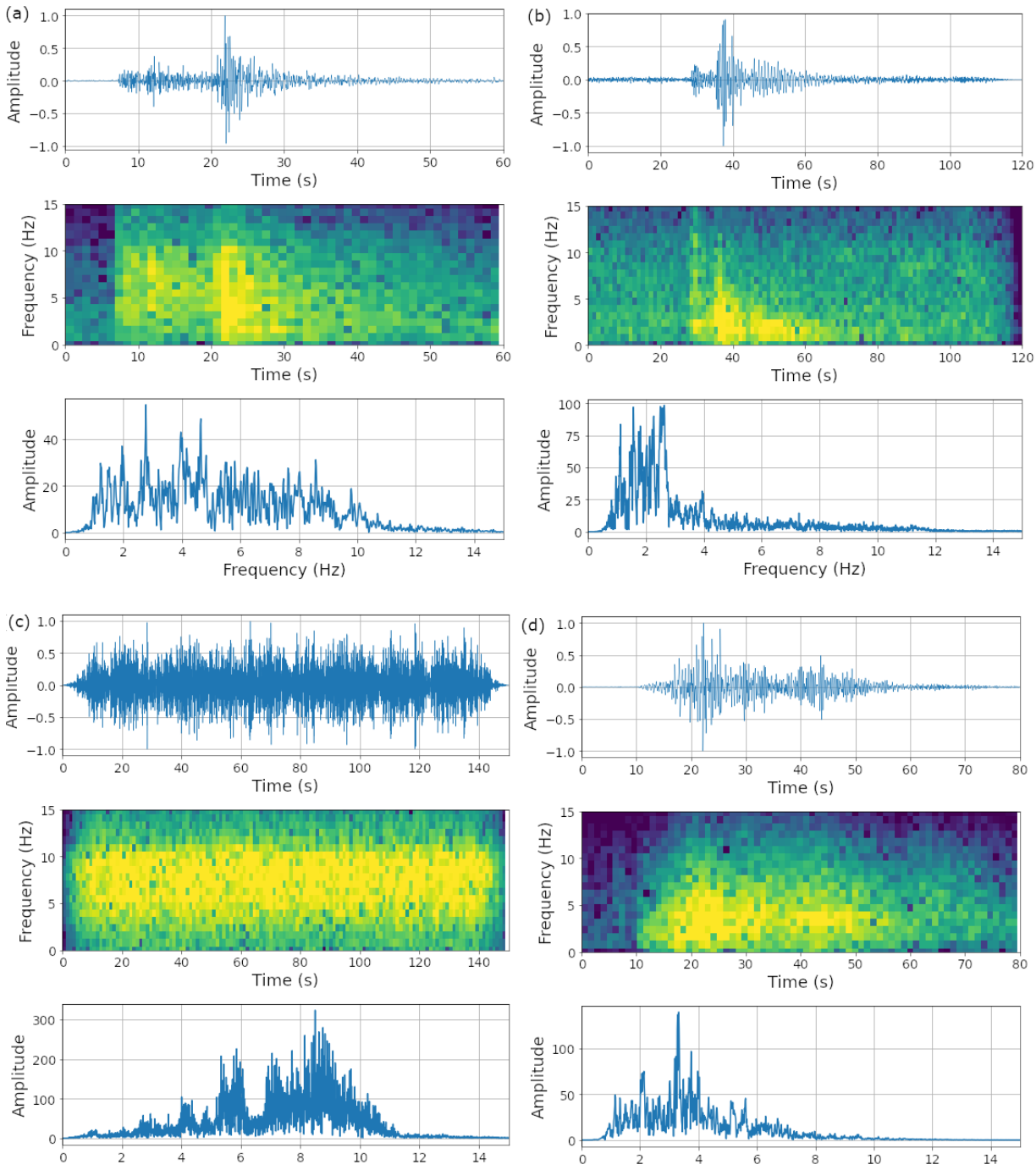

Figure 2 Example of a tapered single-component waveform, its corresponding spectrogram, and Fourier Amplitude spectrum of (a) earthquake, (b) explosion, (c) noise, and (d) surface event.

2.4 Noise

The "noise" class includes 150 s waveforms extracted immediately preceding the P-wave of a ComCat event. These recordings come from the curated dataset of Ni et al. (2023), where automated screening with a retrained Earthquake Transformer was applied to exclude hidden seismic events, resulting in very few cases identified. Given the dataset's large size (>50 000 traces), comprehensive visual inspection of all noise records is impractical, but the deep learning picker used in the curation has demonstrated near 100% accuracy on benchmark datasets, providing high confidence that the noise class contains minimal contamination. Nonetheless, we acknowledge the possibility that a small number of unusual or unpicked events may remain. Some of the noise recordings are characterized by numerous impulsive arrivals and non-typical impulsive earthquake signals, exhibiting a substantial amount of high-frequency content with peak frequencies typically falling within the range of 6 to 10 Hz. This distinctive waveform signature distinguishes them from typical seismic signals associated with earthquakes or explosions, further underlining their classification as ambient noise within the dataset. The label for noise is referred to as no.

2.5 Training, Validation, and Test Data

We designed separate training and validation datasets for classical machine learning (CML) and deep learn-

ing (DL) models, while using a shared testing dataset to ensure a fair comparison across approaches. In preliminary experiments, we found that CML models achieved their strongest performance using the vertical component alone, whereas DL models improved when trained on three-component (3C) inputs. We attribute this improvement to the additional information available in multi-component waveforms – particularly polarization and the relative distribution of P- and S-wave energy across components – which can be informative for discriminating earthquakes and explosions (e.g., Kong et al., 2022). Because the curated dataset contains relatively few fully 3C traces, we supplemented the surface-event class with additional 3C recordings as described in Section 2.3.

To evaluate models on an equal basis, we created a **common test dataset** composed entirely of three-component (3C) traces. We randomly selected 10 000 traces per class. For earthquakes, explosions, and noise, the curated dataset provided sufficient 3C traces. As mentioned in Section 2.3, for surface events, we augmented the curated dataset with 6500 additional 3C traces from nearby stations, yielding a sufficient pool for balanced sampling. From the 10 000 traces available for each class, we randomly split the data into training, validation, and testing subsets using an 80:20 ratio. This produced 2000 traces per class for testing (total of 8000 traces) and reserved the remaining 8000 per class (total of 32 000) for DL training and validation. We ensured that no event data was split between the training and testing datasets by verifying that no event identifiers were found in both the testing and the training/-validation datasets.

We generated the **CML training/validation datasets** using all events in the curated dataset, excluding the common test dataset. The training data set uses randomly sampled 6000 traces per class. The validation data set uses the remaining 2000 traces that do not share event ID from the training data set. If fewer than 2000 traces were available for a class, we sampled with replacement. To account for variability, we repeated the validation sampling process 50 times with different random seeds and averaged the results across iterations.

For the **DL training and validation data sets**, we used the remaining 8000 3C traces per class that remained after the common test data were set aside, and split them into 6000 (training) and 2000 (testing) sets, also ensuring that event ID did not leak between both subsets.

2.6 Network-Testing Dataset

To evaluate model performance in routine network operations, we generate a *network testing dataset* that differs in several key ways from the curated dataset. Unlike the curated dataset, which is balanced and based on carefully reviewed analyst picks, the network dataset reflects the realities of day-to-day operations: incomplete analyst picks, class imbalance, and heterogeneous station coverage. This dataset was designed to test the robustness of models beyond controlled conditions.

We selected the most recent events within the PNSN authoritative boundary, reviewed by the same analyst to ensure consistency across classes. A balanced set of 10 000 events was assembled, with 3333 earthquakes, 3333 explosions, and 3334 surface events; noise was excluded. For each event, waveform data from up to ten stations with the earliest picks were included. Because surface events often lack locations and multi-station coverage, only a small fraction (n=187) were located at more than one station. To address this, most surface events were supplemented with up to nine nearby stations that have historically recorded such events, typically within a 40 km radius. One more distant station, UW.JCW (>100 km, near Mt. Baker), was also included because of its frequent surface-event detections.

The resulting dataset captures the natural variability of routine operations. Earthquakes generally had magnitudes between 0 and 2, epicentral distances below 50 km, and SNR values ranging from 0 to over 20 (capped at 20 to avoid skewing). Explosions showed similar magnitudes and SNR ranges but occurred at slightly larger distances, clustering around 50 km. Surface events lacked magnitude estimates but were typically recorded at distances of less than 25 km, with a few extending to ∼150 km. These long-distance cases may reflect either misclassified artifacts or genuine surface events at Mt. Baker, where volcano–station separations are larger (see Supp. Fig. S12). By including these variations in class balance, SNR, and station coverage, the network testing dataset provides a realistic benchmark for assessing model performance in operational settings.

2.7 Generalization Datasets

An important property for models is to generalize beyond the training datasets. Because training, validation, and testing data sets were compiled for the PNW region, we generated three additional datasets: (1) the Exotic Seismic Event Catalog (ESEC), which provides verified global surface events (Bahavar et al., 2019), (2) a near-field explosion dataset, designed to test distance effects, and (3) incrementally expanded training datasets that incorporate additional surface events and near-field explosions. Together, these datasets allowed us to probe the limits of model generalization and to iteratively refine the training set when systematic misclassifications were observed.

2.7.1 Exotic Seismic Event Catalog Testing Dataset

The Exotic Seismic Event Catalog (ESEC; Bahavar et al., 2019) is an expanding global database of non-tectonic seismic events verified using independent, non-seismic observational evidence (e.g., field/visual reports and remote-sensing products). The catalog includes surface processes such as rock and debris avalanches, rock/debris/ice falls, debris flows/lahars, and snow avalanches, and while it is global in scope, many events are concentrated in North America and Europe. Each entry provides event metadata including origin time, location, and source type.

To assess generalization beyond the Pacific Northwest, we tested our classifier on 245 ESEC events.

For each event, we extracted 270 s of three-component waveform data (70 s before to 200 s after the catalog origin time) from stations within 100 km of the event location; larger radii led to poorer performance due to increased noise levels and waveform dispersion. We additionally required a signal-to-noise ratio (SNR)>10, computed as the ratio between the mean absolute amplitude in a signal window (start time +30 s) and the mean absolute amplitude in a pre-event noise window (start time −50 s to −20 s).

2.7.2 Near-Field Explosion Test Dataset

We report that model performance when testing models on the ESEC dataset indicated that surface events were often misclassified as explosions at larger epicentral distances. To test whether this confusion was due to distance-dependent signal characteristics, we constructed a dedicated near-field explosion test set (0–50 km). We downloaded approximately 10 000 waveforms from ∼1100 explosion events within the PNSN authoritative boundaries and saved in the ANSS catalog, restricting the data to 2023–2025 to ensure no overlap with the training set. This dataset allowed us to directly test whether models trained primarily on far-field explosion signals could generalize to near-source recordings.

2.7.3 Incrementally Improved Training Data Sets

The systematic misclassifications observed on the ESEC and near-field datasets motivated an iterative, data-centric expansion of the curated training set to improve generalization. We developed two successive versions. Version 2 improved robustness to background noise and variability in explosion signals by adding randomly sampled noise-only waveforms, effectively doubling the dataset to ∼12 000 traces per class. Version 3 further expanded the data with targeted additions: (1) 1866 surface-event waveforms from ESEC (excluding lahars, debris flows, and low-SNR events previously misclassified as earthquakes), and (2) 2502 near-field (<50 km) explosion waveforms from the ANSS catalog within the PNSN authoritative boundary. These additions were designed to address dominant confusions seen in preliminary tests – surface events being mistaken for explosions at larger distances and some true explosions being misclassified as surface events – by exposing the models to more representative surface-event and near-field explosion examples.

3 Methods

We explore the two main branches of machine learning in classification, which we summarize in Fig. 3. CML algorithms utilize engineered features (Jordan and Mitchell, 2015) that may be automatically generated using time series toolboxes such as tsfresh (Christ et al., 2018) and TSFEL (Barandas et al., 2020; Kharita et al., 2024), or chosen based on physical models (e.g., Chmiel et al., 2021). The DL approach learns feature representations as part of the optimization (Zheng and Casari, 2018; Kong et al., 2021). In addition to performance and robustness, we will consider the computational cost of deploying either strategy as part of our evaluation.

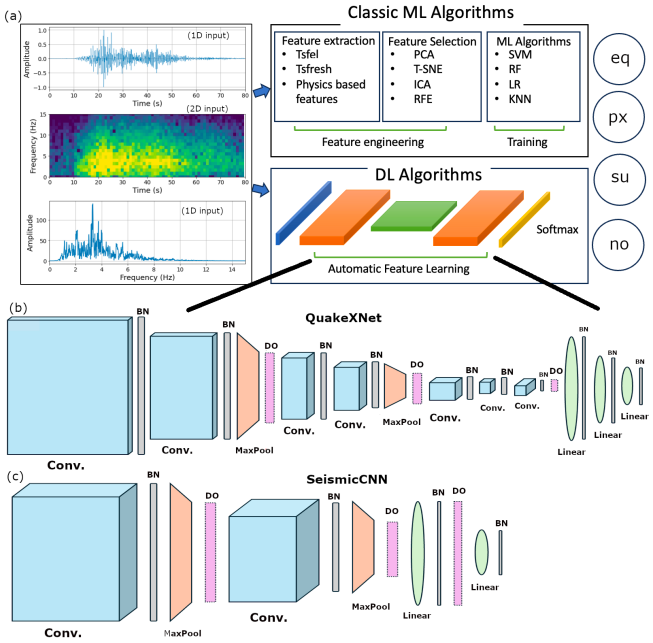

Figure 3 (a) Two main approaches to supervised machine learning in event discrimination: Classical Machine Learning (CML) requires feature engineering before classification. Deep Learning (DL) encompasses feature extraction and classification within a single optimization framework. Input data may be the raw time series, the Fourier Amplitude Spectrum, or the short-time Fourier amplitude spectrum (i.e., spectrograms). Features are either extracted and selected or transformed before CML classification, or they are solved in a single network using deep learning (DL). Finally, the four classes are predicted: eq (earthquake), px (explosion), no (noise), and su (surface events). **Neural network architectures:** (b) QuakeXNet and (c) SeismicCNN. The abbreviations and full forms are as follows: Conv. – Convolutional Layer, BN – Batch Normalization, DO – Dropout, and MaxPool – Maximum Pooling Layer.

3.1 Classical Machine Learning

3.1.1 Data Processing

Raw seismic data are typically high-dimensional, making them unsuitable for direct use in conventional machine learning (ML) algorithms. Feature engineering – the process of creating, transforming, and selecting new features from raw data – is therefore critical for reducing dimensionality and improving model performance of CML algorithms. The goal is to extract essential information that enhances the model's ability to identify patterns and make accurate predictions (Zheng and Casari, 2018).

In this study, we extracted a comprehensive set of features commonly employed in ML applications for seismological analysis. Drawing on both established time-series toolboxes and prior seismological studies, we grouped features into the following categories:

1. **TSFEL features**: Extracted using the Time Series Feature Extraction Library (TSFEL) in Python (Barandas et al., 2020), consisting of 390 features calculated from time, Fourier, and wavelet domains. TSFEL categorizes these features into statistical, temporal, and spectral domains; we did not use the recently introduced fractal domain. We selected TSFEL for its simplicity and broad coverage of features that have proven effective in prior seismological studies (e.g., Kharita et al., 2024; Malfante et al., 2018).

2. **Physical features**: Designed based on seismological observations and physical models of mass movements, these features capture characteristics that distinguish slope failures (surface events) from earthquakes (Hibert et al., 2017; Provost et al., 2017; Maggi et al., 2017; Hibert et al., 2014; Domel et al., 2023; Kharita et al., 2024; Huynh et al., 2024). Examples include the ratio of ascending to descending time, which differentiates impulsive fault sources (peaks at onset) from granular flows such as landslides (peaks mid-event) (e.g., Allstadt et al., 2018; Hibert et al., 2011). Other features include dominant and centroid frequencies, which provide information on source type, and kurtosis and skewness across frequency bands, which indicate impulsiveness. We do not include explicit P/S-ratio features (Kong et al., 2022) or magnitude-based estimates (Koper et al., 2024) because computing them consistently across all classes – particularly surface events and phase-poor/emergent signals – requires phase windowing or picks that are not uniformly defined. Instead, we use phase-agnostic proxies (e.g., frequency-band energy partitions (E1–E4 FFT), time-frequency features, and envelope-shape/timing features as listed in Supp. Table 1) that capture related discriminatory information.

3. **ScatNet features**: Derived from scattering convolutional neural networks (ScatNet) (Anden and Mallat, 2014), these higher-order wavelet transforms have recently proven effective in distinguishing seismic sources (Seydoux et al., 2020; Moreau et al., 2022; Köpfli et al., 2024; Steinmann et al., 2023). Scattering transforms provide translation-invariant, noise-robust features by convolving input waveforms with wavelets across scales, forming first- and second-order scattering coefficients. We selected wavelet parameters following the guidelines in Seydoux et al. (2020). Details of feature extraction are provided in the Supp. Section S1.

4. **Manual features**: To account for anthropogenic patterns relevant to explosions, we added temporal descriptors such as Hour of Day (HOD) (local time), Day of Week (DOW), and Month of Year (MOY). These features reflect human activity schedules and complement the physically motivated features.

For the CML approach, we extracted features from vertical-component waveforms only, thereby maximizing the dataset size, as many PNSN stations record only a single component. After feature extraction, we performed standard data cleaning, removing samples with NaNs, Inf values, or identical constant values. We further reduced dimensionality by removing highly correlated features (Pearson correlation coefficient>0.95). Finally, we standardized each feature using z-score normalization (zero mean, unit variance) and then applied a threshold-based filter to remove outliers, discarding samples with any feature value exceeding ±5 standard deviations.

To investigate the impact of time windowing and prefiltering on extracted features and classification performance, we constructed six feature sets (M1–M6) with varying window lengths and frequency bands. For earthquakes and explosions, windows were aligned relative to the P arrival, while for surface events, they were aligned to the analyst-defined first arrival in the curated dataset. Noise windows were extracted using the same time spans and frequency bands. The models are defined as follows:

- M1: 40-s window (P−10 to P+30 s), filtered 1–10 Hz
- M2: 40-s window (P−10 to P+30 s), filtered 0.5–15 Hz
- M3: 110-s window (P−10 to P+100 s), filtered 1–10 Hz
- M4: 110-s window (P−10 to P+100 s), filtered 0.5–15 Hz
- M5: 150-s window (P−50 to P+100 s), filtered1–10 Hz
- M6: 150-s window (P−50 to P+100 s), filtered 0.5–15 Hz

3.1.2 Architecture Descriptions

We systematically compared seven widely used machine learning algorithms for seismic event classification: Logistic Regression (LR, Hosmer et al., 2013), Multi-Layer Perceptron (MLP), Support Vector Classifier (SVC, Hearst et al., 1998), K-Nearest Neighbors (KNN, Cover and Hart, 1967), Random Forest (RF, Breiman, 2001), XGBoost (XGB, Chen and Guestrin, 2016), and LightGBM (LGBM, Ke et al., 2017). Our evaluation resembled an automated ML approach in scope but was implemented manually and transparently: we defined parameter ranges for each algorithm, performed initial hyperparameter searches, and directly compared their performance.

Each algorithm was assessed using five-fold cross-validation, with the macro F1 score as the optimization metric. This stage of analysis was intended to identify which algorithm families consistently performed well on our data. Results are shown in Supp. Figs. S3–S4, with tested parameter ranges and optimal values summarized in Supp. Section S2. Tree-based algorithms (RF, XGB, and LGBM) consistently outperformed others. LGBM achieved the highest average macro F1 score (∼89%), followed by RF (∼85%). However, training LGBM was at least 20 times slower than RF under comparable settings, and runtime scaled steeply with larger

grids. Because our study required hyperparameter tuning across 48 feature sets and multiple validation seeds, tuning was computationally exhaustive. RF offered a more practical balance: strong performance, lower computational costs in training, and the ability for interpretable feature importance estimation. In addition, RF is widely used in seismology because of its robustness, relatively simple hyperparameter space, and intuitive feature-importance measures (e.g., Hibert et al., 2014, 2017; Provost et al., 2017; Kharita et al., 2024). For these reasons, we selected RF as the primary CML algorithm for the rest of the study.

3.1.3 Training and Tuning

After selecting RF, we ran a second, more extensive hyperparameter search tailored to each of the 48 feature sets. For each set, we randomly sampled ∼300 RF configurations varying the number of trees (*n_estimators*), maximum tree depth (*max_depth*), and the minimum samples required to split and form a leaf (*min_samples_split, min_samples_leaf*). Each candidate configuration was evaluated using five-fold cross-validation, and the best-performing combination (macro F1) was selected. To control runtime, we tuned on a balanced subset of 3000 samples per class. Across feature sets, this deeper optimization typically yielded limited additional 1–2% improvement in macro F1 compared to the initial broad search. Detailed parameter ranges are available in the public software repository linked in the code availability section. (Supp. Section S2, Supp. Figs. S3,S4)

3.2 Deep Learning

3.2.1 Data Processing

For deep learning models, we prepared three-component waveform traces using a standardized preprocessing pipeline. For each trace, we extracted a 100-s window, with the window start chosen randomly between 5 and 20 seconds before the analyst's pick time. All waveform traces were preprocessed before model training and evaluation. Each trace was first linearly detrended to remove baseline offsets and then tapered with a 1% cosine taper to minimize edge effects introduced during filtering. We applied a band-pass filter between 1–20 Hz to suppress long-period noise as well as high-frequency artifacts, ensuring the retention of the frequency content most relevant for seismic event discrimination. Finally, each trace was normalized by its standard deviation, following common practice in deep learning applications for seismic data, so that all channels contribute comparably during training. To ensure data quality, we only retained traces with an SNR greater than 1, computed following the same procedure described by Ni et al. (2023). In addition to raw waveform inputs, we generated spectrogram representations of each three-component trace. Spectrograms provide a joint time–frequency characterization of the signals, which has been shown to improve classification performance in many deep learning applications. We implemented a PyTorch-based spectrogram computation function (`compute_spectrogram`), which transforms each input batch of waveforms (B, C, T) into power spectral density representations $(B, C, F, T_{\mathrm{spec}})$. Waveforms were segmented into overlapping windows using a Hann taper, with each segment set to 256 samples (`nperseg = 256`) and 50% overlap. Each segment was then mean-centered, tapered, and transformed into the frequency domain using the real-valued Fast Fourier Transform (FFT). Power Spectral Density (PSD) estimates were obtained by normalizing the squared magnitudes of the Fourier coefficients by the window power and sampling rate. To ensure energy conservation under the Parseval convention, one-sided spectra were produced by doubling all frequencies except the DC and Nyquist components. Finally, frequency and time axes were derived from the FFT length and hop size, providing consistent alignment across all traces. The resulting spectrograms preserve three channels per trace (E, N, Z), while adding frequency and time dimensions, producing an input well-suited for convolutional neural networks (Yeck et al., 2020). By combining both waveform- and spectrogram-based representations, we ensured that the deep learning models had access to both temporal and spectral features of the seismic signals.

3.2.2 Architecture Descriptions

In recent years, deep learning methods, particularly Convolutional Neural Networks (CNN, LeCun et al., 2015), have emerged as powerful algorithms for seismic event classification, leveraging their ability to extract hierarchical features from raw waveform data automatically. CNNs excel in learning complex patterns directly from the data, making them particularly effective for distinguishing between events such as earthquakes, explosions, surface events, and noise. In recent studies, CNNs have been successfully applied to process seismic signals, demonstrating superior performance over conventional methods in both accuracy and scalability (e.g., Mousavi and Beroza, 2022). The multi-layered architecture of CNNs allows them to capture both local and global features in seismic waveforms, making them well-suited for detecting subtle variations in seismic signatures. Moreover, CNNs can be extended to incorporate multi-channel inputs, such as combining signals from different seismic components, which further enhances their classification capability (e.g., Mousavi and Beroza, 2022; Ross et al., 2018; Perol et al., 2018; Linville et al., 2019). As a result, CNNs have become a valuable tool in seismic monitoring, aiding in the real-time detection of events and enhancing the accuracy of seismic discrimination systems.

Designing an optimal CNN architecture is a complex task, primarily due to the many hyperparameters that must be tuned for best performance. These hyperparameters include the number and size of filters, the configuration of the convolutional, fully connected, dropout, and pooling layers, as well as the choice of activation functions. The search for an ideal hyperparameter combination is computationally intensive and far more demanding than traditional machine learning ap-

proaches, often requiring heuristic optimization methods, such as random search or grid search, to explore the vast parameter space (e.g., Network Architecture Search (NAS), Elsken et al., 2019). Alternative methods may reuse the architecture of foundation models (e.g., VG16).

Each study employs a specific model architecture, topology, and dataset, making inter-comparison challenging. To remediate this, without undergoing a full NAS, we propose representing diversity in architecture by utilizing two canonical architectures and two types of inputs: time series and spectrograms. We design a wide/fat shallow network (SeismicCNN) and a skinny deep network (QuakeXNet), which we illustrate in Fig. 3. For each architecture, we have a (1D) and a (2D) version, where (1D) refers to taking time series input data and (2D) refers to taking spectrograms as inputs. The time series has a dimension of 3x5001 points, and the spectrograms have a dimension of (3x129x38). Each architecture is designed to address the unique characteristics of seismic waveform data, aiming to improve classification performance.

The SeismicCNN architecture consists of two convolutional layers, each followed by a batch normalization layer and a max pooling layer. The architecture begins with a 1D convolutional layer that extracts features from the seismic waveforms. The first layer applies 32 filters, each with a kernel size of five, followed by a batch normalization and ReLU activation. The second convolutional layer expands to 64 filters, again with a kernel size of five, followed by batch normalization and ReLU activation. Both convolutional layers are followed by max-pooling layers with a kernel size of two, which reduces the temporal dimension and focuses on key features. Dropout layers with a rate of 0.2 are integrated after each pooling operation to prevent overfitting during training. The final layers are fully connected and output four class logits, which are converted to class probabilities using a Softmax activation (i.e., probabilities sum to 1 across classes for each input window). SeismicCNN (2D) has the same architecture but instead uses spectrograms as input and 2D convolutions. However, the size of the models has a significant difference: SeismicCNN (1D) has 10 227 340 parameters, whereas SeismicCNN (2D) has 1 986 572 parameters, primarily due to the difference in input size.

The QuakeXNet (1D) architecture consists of seven convolutional layers, each followed by batch normalization layers and two max pooling layers. The architecture begins with sequential 1D convolutional layers, each followed by batch normalization and ReLU activation. The first convolutional layer uses eight filters with a kernel size of nine and padding to preserve input dimensions, while subsequent layers gradually increase the number of filters to 64 and alternate between stride lengths of one and two. This progressive increase in the number of filters is thought to allow the model to capture more complex features at deeper layers. A max pooling operation is applied after every second convolutional layer to reduce the temporal resolution, enabling the network to focus on the most salient features, and is a form of blur pooling layer (Zhang, 2019) and has been utilized in denoising CNNs (e.g., Yin et al., 2022). The classifier takes the flattened output of features from the convolutional layers and passes it through a fully connected layer with 128 neurons, followed by batch normalization and ReLU activation, and another fully connected layer that outputs class logits. Dropout layers with a rate of 0.2 are inserted after pooling layers and fully connected layers to mitigate overfitting. The QuakeXNet (2D) architecture, adapted for 2D seismic spectrogram inputs, mirrors the QuakeXNet (1D) structure but uses 2D convolutional layers. QuakeXNet (1D) has 657 716 parameters, whereas QuakeXNet (2D) has 70 708 parameters.

For evaluation on curated/test datasets, we assign the predicted class by the maximum-probability (argmax) output; we do not apply a fixed probability threshold for classification.

3.2.3 Training and Tuning

For training, we used the Adam optimizer (Kingma and Ba, 2014), an initial learning rate of 0.001, and a cross-entropy loss. We train on a single NVIDIA RTX3090 24 GB RAM. We trained the models with a batch size of 128 for up to 100 epochs, implementing early stopping if validation performance did not improve within 30 epochs. This approach allowed us to fine-tune the model's ability to generalize while avoiding overfitting on the training set. We assess the performance of each model by comparing the training and validation losses, along with the validation accuracy, as the number of epochs increases.

The training loss of the SeismicCNN (1D) architecture showed a smooth and consistent decline over the epochs, indicating effective learning during training. In contrast, the validation loss exhibited a more irregular decay with the number of epochs, characterized by an initial increase followed by a gradual decrease. Training SeismicCNN achieved a peak validation accuracy of 89% (Supp. Fig. S6). In comparison, QuakeXNet (1D) demonstrated improved stability in the training, with the training loss steadily decreasing from 0.75 to 0.15 and the validation loss dropping from 0.9 to 0.45. The validation accuracy improved significantly, rising from 60% to 92%, indicating that this architecture was highly effective for the given task (Supp. Fig. S6).

Given the highly variable performance on the validation test while training a model on 1D waveforms (Supp. Fig. S6), we shifted our focus to 2D spectrogram-based architectures. We did not attribute this limitation specifically to window length or event magnitude; rather, we interpret it as an empirical advantage of the 2D representation, which makes discriminative time–frequency structure more explicit and facilitates learning with CNNs. The SeismicCNN (2D) architecture demonstrated that both training and validation losses decreased smoothly, with the training loss falling from 0.75 to 0.15 and the validation loss dropping from 0.75 to 0.25. The accuracy improved from 75% to 94%, highlighting the model's effectiveness in discriminating between classes. Similarly, the QuakeXNet (2D) architecture showed consistent improvement, with both train-

ing and validation losses decreasing from 0.9 to 0.3. The accuracy increased steadily from 78% to 92%, although the validation loss was slightly lower than the training loss. Overall, our findings suggest that transitioning from 1D to 2D architectures improves the model's ability to distinguish between various seismic event classes (Supp. Figs. S7, S8).

3.3 Model Deployment Workflows

To explore how our models could be applied in operational and research settings, we designed several deployment workflows. These workflows span real-time and retrospective use cases, ranging from integration with existing software ecosystems to large-scale cloud-ready pipelines. The following subsections describe the implementation details of each workflow.

3.3.1 Integration with SeisBench

We first implemented our deep learning classifiers in the SeisBench ecosystem (Woollam et al., 2022), enabling seamless integration into catalog-building workflows. The models are compatible with the `model.classify` and `model.annotate` functions, which return class labels, timestamps, and probability traces for use in real-time or retrospective analysis.

Listing 1 Read an obspy stream and classify or annotate with seisbench

```
# importing the dependencies
import seisbench.models as sbm
import obspy
# load the model
model = sbm.QuakeXNet.from_pretrained()
model.eval()  # Set to evaluation mode
# reading the data
z = obspy.read("file.mseed")
# classify the data
results = model.classify(z)
# anotate the data
annotations = model.annotate(z)
```

3.3.2 Deployment on the Network-Testing Dataset (Retrospective)

To evaluate model performance under controlled conditions, we applied our classifiers to the network testing dataset (Section 2.6). For each event, we retrieved waveforms from the 10 nearest stations and extracted 141-s windows (30 s before to 111 s after origin), resampled to 50 Hz. We then ran two DL models (QuakeXNet and SeismicCNN) and two ML models (M2_30 and M2_110) with a 5-s stride. Station-level probabilities were averaged and filtered using quality-based criteria (SNR thresholds, probability thresholds, probability–distance thresholds). Event-level predictions were obtained by majority vote across high-confidence stations. To assess robustness, we conducted a grid search over threshold values.

3.3.3 Deployment on Continuous Data (Cloud-ready Workflow)

Finally, to test how to run this continuously and deploy these on cloud data archives (e.g., Ni et al., 2025b,a), we developed a pipeline to scan daily waveform archives stored on Amazon Web Services (AWS) Simple Storage Service (S3). Day-long miniseed files were read, relevant channels (e.g., any HH or BH channels) extracted, and data preprocessed (filtering, trimming, spectrogram computation). We deployed the model using 100-s windows with a 20-s stride. Probability traces were smoothed with a moving average (window size=5). Detections were initiated when the smoothed probability exceeded 0.15 and ended when it dropped below, with valid detections requiring a maximum probability above 0.5. Each detection was assigned the class with the highest peak probability, and outputs included labels and timestamps. A quantitative comparison of continuous-data detections to the curated/test datasets is not included here. Such an evaluation would require catalog matching and/or analyst validation to characterize false positives/negatives and to tune detection and association thresholds, which is beyond the scope of this study.

4 Model Performance Results

Our performance evaluation proceeds in four stages.

1. We first benchmark all CML and DL models on a balanced test set derived from the curated dataset (Section 2.5) to enable a controlled comparison between feature-based and end-to-end approaches.

2. Based on this benchmark, we select the top candidates (CML: M2 and M4 with Physical + Manual features; DL: QuakeXNet and SeismicCNN) and evaluate them on the network-testing dataset (Section 2.6), which reflects more realistic conditions (class imbalance, variable SNR, and heterogeneous station coverage).

3. We then probe generalization of the best DL models using two independent datasets: the Exotic Seismic Event Catalog (ESEC) and a near-field explosion dataset. These tests reveal systematic confusions, most notably between surface events and explosions.

4. Finally, we address these failure modes through data-centric retraining with progressively expanded training sets (Versions 2 and 3). Version 3 provides the best overall trade-off between accuracy, robustness, and computational efficiency.

4.1 Evaluation Setup

We evaluated model performance using standard classification metrics. *Precision* is the proportion of true positive predictions out of all predicted instances for a given class, while *recall* (or sensitivity) is the proportion of true positives out of all actual instances of that

class. At a broader level, *accuracy* is the ratio of correctly predicted instances (true positives and true negatives) to the total number of instances, and the *F1 score* is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. F1 is particularly informative when datasets are imbalanced.

Model outputs were evaluated at two levels. At the *station level*, predictions were generated independently for each waveform trace, with the assigned class corresponding to the maximum predicted probability among the four classes (earthquake, explosion, surface event, or noise). At the *event level*, predictions from all available stations were aggregated by averaging class probabilities across stations, and the event was assigned to the class with the highest mean probability. For the balanced curated test dataset, station coverage was often limited – particularly for surface events, so we report only station-level performance. For the network testing dataset, where more complete station coverage was available, both station- and event-level performance were assessed (see Section 2.6). This evaluation framework provides the basis for the performance comparisons presented in the following sections.

4.2 Performance on the Curated Test Dataset

We first evaluated the performance of classical machine learning (CML) models trained on engineered features and deep learning (DL) models trained on waveform time series or spectrogram inputs. This controlled test set provided a benchmark for comparing the two approaches under balanced conditions.

4.2.1 CML Performance

We evaluated eight feature sets and their combinations to classify seismic traces using CML models (Supp. Sections S1 and S2): Scatnet features (110), Scatnet + Manual features (113), TSFEL (390), TSFEL + Manual (393), Physical (62), Physical + Manual (65), Physical + TSFEL (454), and Physical + TSFEL + Manual (457).

Precision Precision results shown in Fig. 4 highlight the importance of certain feature groups. Feature sets with Scatnet-derived features alone showed moderate precision, which improved when combined with Manual features. TSFEL-based features outperformed Scatnet. Combining Physical features with others consistently achieved the best precision, especially for earthquakes and explosions. Longer waveform windows (110–150 s; M3–M6) showed a slight improvement in precision for explosions and surface events compared to shorter windows (40 s; M1–M2), although the differences were generally small (on the order of 1–2%). For the best-performing feature set (Physical + Manual), shorter 40 s windows (M1–M2) performed marginally better than longer ones, underscoring that window length did not have a consistent effect across all feature sets. Broader frequency ranges (0.5–15 Hz; M2, M4, M6) tended to yield marginally higher precision than narrower bands (1–10 Hz; M1, M3, M5), but again the gains were minor. Noise precision remained stable across all configurations, with Physical + Manual features consistently achieving the highest values overall.

Recall Hybrid feature sets that included Physical features yielded higher recall across all classes (Fig. 4). Longer windows again improved recall for explosions and earthquakes, and broader frequency ranges (0.5–15 Hz) contributed to stronger performance compared to narrower filters. Noise recall was consistently high, reflecting its distinctive spectral content. For Physical features, a negligible difference in performance was observed.

Accuracy Accuracy values ranged from moderate (70–73% for Scatnet features) to strong (86–88% for physical-based features). TSFEL features significantly improved accuracy (83–86%), especially when combined with Manual features (up to 87%). Overall, feature type was the dominant factor for CML performance, followed by waveform window length, while preprocessing choices had smaller effects (Fig. 4). The best accuracy (88.5%) was achieved by Physical + Manual features, underscoring the value of physically interpretable descriptors.

4.2.2 DL Performance

We next evaluated four DL models: SeismicCNN (1D, 2D) and QuakeXNet (1D, 2D) with waveform time series (1D) or spectrograms (2D) as input.

Precision Precision results shown in Fig. 5(top) vary across models and classes. SeismicCNN (1D) achieved high precision for surface events (95%) and noise (99%), but struggled with earthquakes (75%) and explosions (74%). Using spectrograms improved results substantially: SeismicCNN (2D) reached 90–96% precision across all classes. QuakeXNet (1D) performed well for noise (99%) and earthquakes (88%) but showed lower precision for surface events (84%). QuakeXNet (2D) achieved the strongest and most balanced precision (90–98%) across all classes.

Recall Recall values shown in Fig. 5(middle) showed similar trends. SeismicCNN (1D) recalled earthquakes (94%) and noise (99%) well, but misclassified many surface events (58%). SeismicCNN (2D) delivered balanced recall across classes (90–97%), excelling in surface event detection (95%). QuakeXNet (1D) achieved high recall for surface events (94%) and noise (99%) but lagged on explosions (82%). QuakeXNet (2D) again provided the best balance (89–99%), with strong performance across all classes.

Accuracy Accuracy is shown in Fig. 5(bottom) and further highlights the advantage of spectrogram inputs. SeismicCNN (1D) achieved 84.2%, while SeismicCNN (2D) reached 93.7%. QuakeXNet (1D) reached 91.6%, and QuakeXNet (2D) 92.4%.

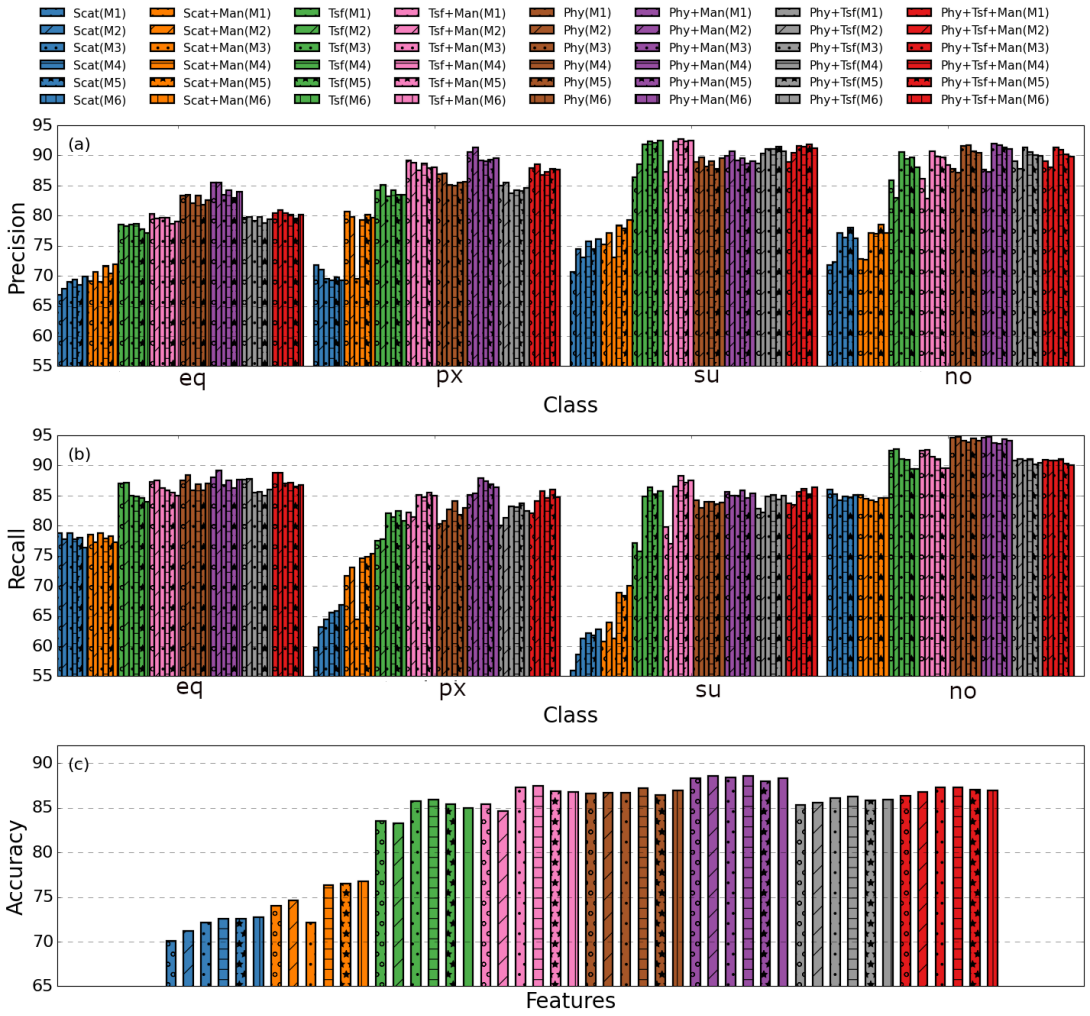

Figure 4 CML model performance on the curated test set. Precision (a), recall (b), and overall accuracy (c) for random-forest classifiers trained on eight engineered feature sets across six waveform configurations (M1–M6). Colors denote feature sets: Scat (ScatNet), Scat+Man (Manual), Tsf (TSFEL), Tsf+Man, Phy (Physical), Phy+Man, Phy+Tsf, and Phy+Tsf+Man; hatch patterns denote waveform configuration (window length and bandpass setting; M1–M6). Across metrics, feature type dominates performance: Scat-only features show the lowest scores, Tsf features improve substantially, and feature sets including Phy consistently yield the strongest precision/recall for earthquakes and explosions, and the highest overall accuracy. Man features provide modest gains when added to Scat or Tsf, while differences among waveform configurations are comparatively small (typically ∼1–2%) and most apparent for explosion/surface precision and recall. Noise is consistently classified with high precision and recall across all feature sets/configurations.

4.2.3 Comparison of CML and DL Approaches Within Domain

DL models, particularly spectrogram-based architectures (SeismicCNN 2D and QuakeXNet 2D), consistently outperformed CML models across all metrics (precision, recall, and accuracy). While CML models benefited from carefully engineered features, their performance plateaued below the best DL models. In contrast, DL approaches leveraged end-to-end learning to capture subtle waveform differences – especially between surface events and explosions – leading to superior generalization. Overall, these results establish DL spectrogram-based models as the strongest candidates for deployment, setting the stage for further evaluation on more challenging and realistic datasets (e.g., see Fig. 6).

Model	Accuracy	F1 score
SeismicCNN (1D)	84%	84%
QuakeXNet (1D)	92%	92%
SeismicCNN (2D)	**94**%	**94**%
QuakeXNet (2D)	92%	92%
Phy+Man (M2)	89%	87%
Phy+Man (M4)	88%	88%
Phy+Man (M6)	87%	89%

Table 1 Performance of different models on the common test sets from the curated catalogs described in Section 2.5.

4.3 Performance on the Network Testing Dataset

From the balanced curated test dataset, we identified three models that achieved the best overall perfor-

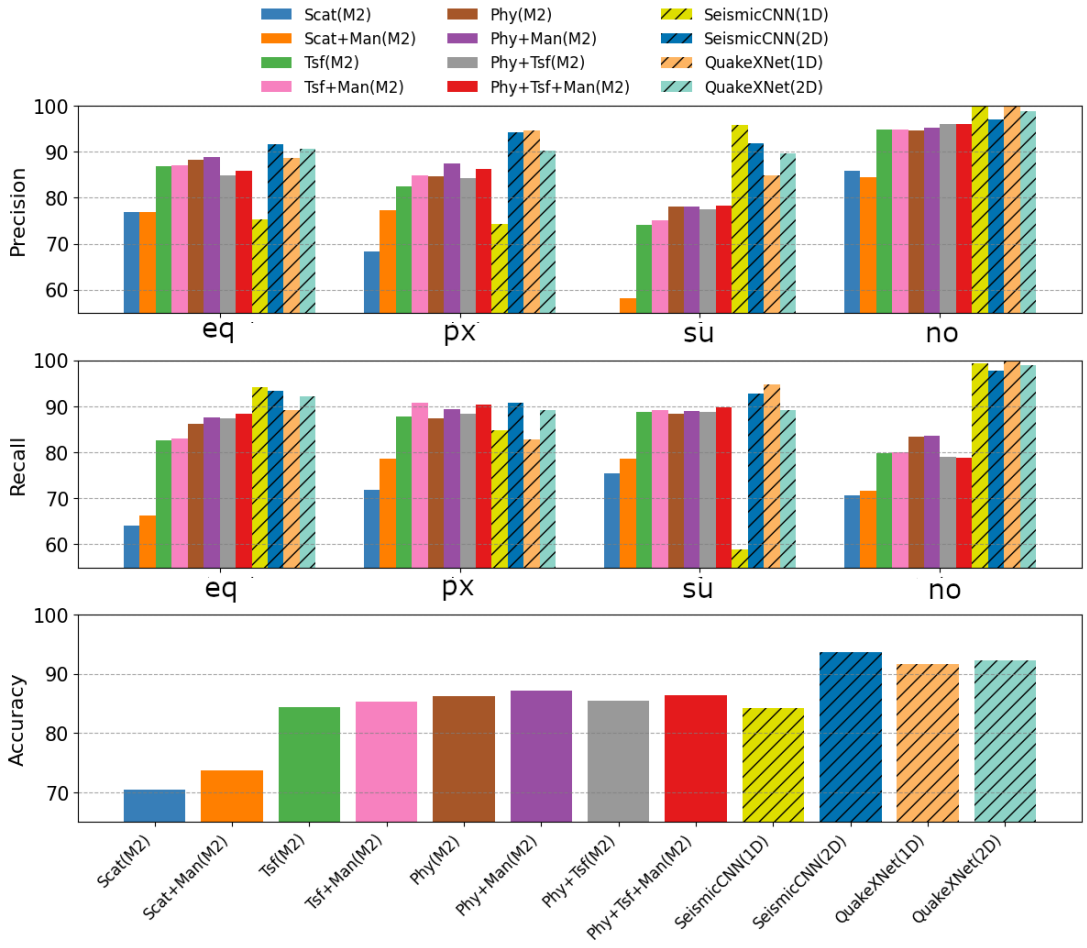

Figure 5 DL vs. engineered-feature baselines on the curated test set. Precision (top), recall (middle), and overall accuracy (bottom) for deep-learning models (hatched bars) and classical ML baselines trained on engineered features (solid bars) using the same waveform configuration (M2). DL models include SeismicCNN and QuakeXNet with 1D waveform inputs and 2D spectrogram inputs. Engineered-feature baselines include Scat (ScatNet), Tsf (TSFEL), and Phy (Physical) feature sets and their combinations with Manual features. Across classes, 2D spectrogram-based DL models provide the most consistent and highest performance, with SeismicCNN (2D) and QuakeXNet (2D) achieving uniformly strong precision and recall (typically ∼90–99%) and the highest accuracies.

mance: two deep learning classifiers (SeismicCNN 2D and QuakeXNet 2D) and one classical machine learning model (RF trained on Physical + Manual features with a 40 s, 0.5–15 Hz window, hereafter referred to as ML_40). These models capture both end-to-end feature learning (DL) and engineered feature-based approaches (CML), providing complementary perspectives on event classification. In this section, we evaluate these top-performing models on the network testing dataset to assess how well they generalize under realistic operational conditions.

The two deep learning models performed robustly, with accuracy and F1 scores only slightly lower than on the curated test set (Supp. Fig. S13). SeismicCNN (2D) achieved nearly equal performance across classes, with precision/recall/F1 of 96%/95%/95% for earthquakes, 98%/93%/95% for explosions, and 92%/97%/95% for surface events. QuakeXNet (2D) performed comparably overall but showed a noticeable drop in precision for surface events (87%), indicating greater confusion between surface events and explosions.

By contrast, the ML_40 model underperformed relative to the deep learning approaches, particularly for surface events. It correctly classified only about three-quarters of su cases (recall=77%), roughly 10% lower than the CNNs in mean F1 score. This performance gap underscores the advantage of end-to-end feature learning, which appears better suited to the heterogeneous and lower-quality records typical of routine operations.

To further investigate the causes of misclassification, we analyzed classification probability as a function of SNR and epicentral distance (Supp. Fig. S14). For both CNNs, high-confidence predictions ($P > 0.9$) were maintained for traces with SNR>5 out to distances of ∼20 km. Confidence declined sharply for low-SNR waveforms and for distances greater than ∼60 km. We did not stratify by magnitude because many surface events lack consistent magnitude estimates, and because station-level SNR is strongly influenced by distance and noise conditions in addition to event size; thus, SNR provides a more direct indicator of signal quality for classification confidence.

Surface events were most vulnerable to this decline in performance over SNR and distance, which we interpret

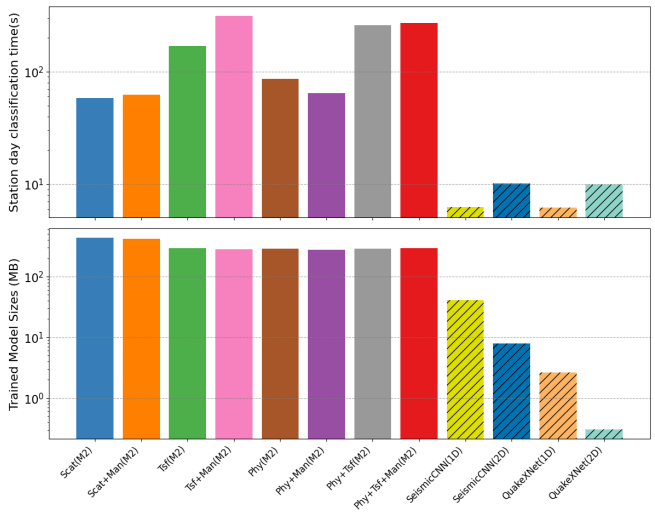

Figure 6 Performance comparisons of various ML and DL models in terms of (a) station-day classification time and (b) trained model sizes.

as a greater attenuation of signals in shallow wave propagation, explaining the precision shortfall observed for QuakeXNet (2D).

These results demonstrate that while DL models generalize well from curated to network data, their performance remains bounded by physical SNR–distance trade-offs inherent to the network geometry and noise environment, rather than by model bias.

4.4 Performance on the Generalization Datasets

Performance degraded on out-of-domain datasets, with systematic confusion between surface events and explosions (Fig. 7). On the ESEC dataset, SeismicCNN(v1) labeled most station-level records as noise (420) and only 100 as surface events (Fig. 7a). After noise augmentation, SeismicCNN(v2) increased surface-event assignments to 235 and reduced noise assignments to 270. At the event level, the number of ESEC events classified as surface events increased from 5 (v1) to 19 (v2) and 18 (v3) (Fig. 7c).

On the near-field explosion dataset, SeismicCNN(v1/v2) showed strong confusion between explosions and surface events at the station level (explosion 540 vs. surface 385), whereas QuakeXNet(v3) increased explosion assignments to 675 while reducing the number of explosions mislabeled as surface events to 225 (Fig. 7b). Event-level results showed the same trend: QuakeXNet(v3) increased explosions classified as explosions from 99 (v1/v2) to 130 and reduced explosions classified as surface events from 82 to 33 (Fig. 7d), albeit with a modest increase in explosion→noise assignments.

Closer inspection of probability curves revealed systematic patterns: surface event probabilities tended to peak early in the waveform, while explosion probabilities peaked later, consistent with similarities in their physical sources (surface waves, explosive or detachment phases). By aggregating probabilities across entire traces and assigning labels based on the dominant class, Version 3 achieved improved event-level accuracy on both datasets. QuakeXNet (2D), when analyzed with the same procedure, generalized better than SeismicCNN (2D), suggesting that architectural differences influence out-of-domain robustness.

Despite these improvements, some confusion between surface events and explosions persisted, particularly for cases with emergent arrivals, extended codas, or low signal quality. This ambiguity may also reflect similarities in shallow-source radiation and propagation effects, since both surface processes and many explosions are exceptionally shallow. These observations underscore the need for diverse training data and robust evaluation across global datasets.

4.5 Overall Performance

Across all stages of evaluation, deep learning models consistently outperformed classical machine learning approaches. On the balanced curated test dataset, spectrogram-based architectures (SeismicCNN 2D and QuakeXNet 2D) achieved the highest precision, recall, and accuracy, clearly surpassing feature-engineered CML models. On the network testing dataset, both CNNs maintained strong performance under realistic operational conditions, though QuakeXNet (2D) showed a modest loss of precision for surface events compared to SeismicCNN (2D).

Generalization tests revealed the limitations of training only on curated data sets: both CNNs struggled when applied to out-of-domain datasets, such as the Exotic Seismic Event Catalog (ESEC) and near-field explosions, frequently confusing surface events with explosions. These systematic errors motivated iterative augmentation of the training data. Version 2 incorporated noise-only records, yielding modest improvements, while Version 3 further expanded the dataset with 1866 ESEC surface event traces and 2502 near-field explosion traces (see Fig. 7).

QuakeXNet (2D) Version 3 emerged as the best-performing model overall. It provided the strongest balance of accuracy, robustness, and computational efficiency, outperforming SeismicCNN (2D) on out-of-domain tests while maintaining high performance on curated and network datasets. Importantly, QuakeXNet (2D) is lightweight, requiring only 70 708 parameters and ∼1.2 MB of memory (Fig. 6), with a full day of continuous data processed in ∼9 s (Table 2) at a stride of 10-s (Fig. 6). This combination of accuracy, robustness, and efficiency makes **QuakeXNet (2D) v3** the most reliable classifier developed in this study and the most suitable candidate for deployment in real-time network operations.

5 Feature Importance

5.1 Feature Importance from CML

Focusing on the CML model that includes TSFEL, Manual, and Physical features, we now explore the impor-

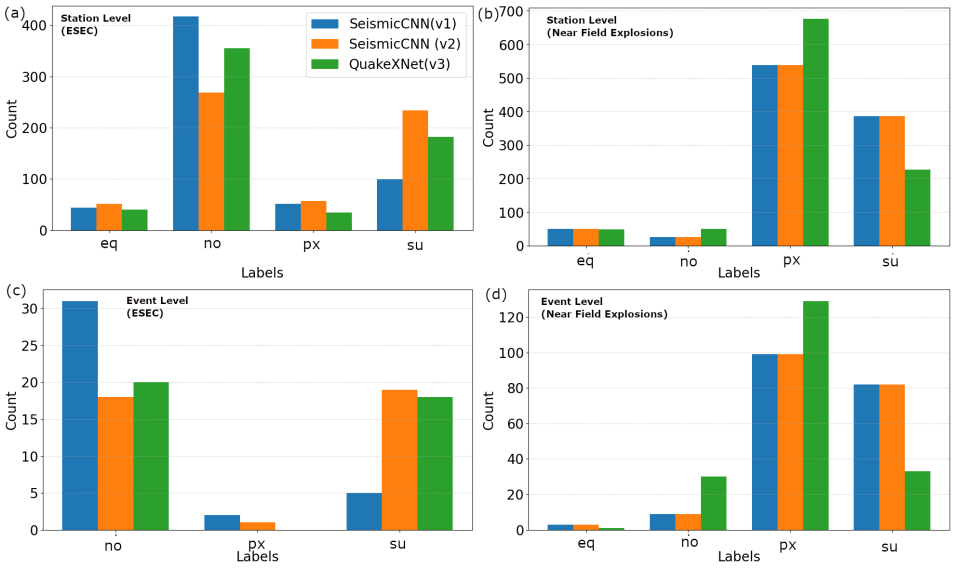

Figure 7 Performance of different models on the ESEC dataset (a, c) and the near-field explosion dataset (b, d), shown at both the station level (a, b) and the event level (c, d).

Model	Number of parameters	Memory usage (MB)	Deploying on 1 day of 100-Hz data (s)
QuakeXNet (1D)	657 716	4.55	6.17
SeismicCNN (1D)	10 227 340	46.39	6.22
SeismicCNN (2D)	1 986 572	11.61	10.07
QuakeXNet (2D)	**70 708**	**1.22**	**9.13**

Table 2 Computational performance of selected deep learning models.

tance of the feature calculated by the Random Forest algorithm and show the feature importance in Fig. 7. In Random Forest models, the importance of a feature is typically measured by the decrease in a performance metric, such as Gini impurity or accuracy, when the feature is used to split the data in a tree. To estimate feature importance, RF models aggregate the impact of each feature across all trees. The more a feature reduces uncertainty in predictions, the higher its importance score will be. This approach has several advantages: it handles large datasets and high-dimensional spaces easily and provides a measure of feature importance without the need for feature scaling or normalization.

Fig. 8 shows the feature importance of running model M2 with a combination of Physical + Manual features. The analysis of the feature set revealed that kurtosis-based features were the most important (Fig. 8). Kurtosis is a statistical measure that indicates the flatness of the signal amplitude distribution compared to a normal distribution. Signals with lower kurtosis values have flatter distributions and shorter tails, indicating fewer extreme values. The feature *Kurt_3_10* (kurtosis in the 3–10 Hz frequency band) was found to be the most important among all time-series features, followed by *KurtoSig* (kurtosis of the entire signal, filtered between 0.5–15 Hz), *Kurto_10_20* (kurtosis in the 10–20 Hz

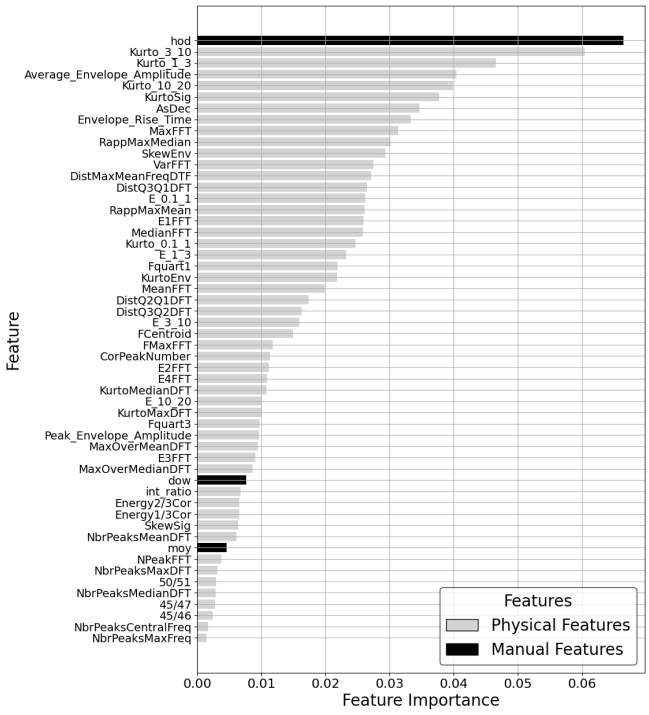

Figure 8 Feature importance for model configuration M2 using the Physical + Manual feature set, averaged across 10 iterations. Feature names correspond to the engineered-feature glossary provided in Supp. Table S1.

band, though our data was pre-filtered to 15 Hz), and *Kurto_1_3* (kurtosis in the 1–3 Hz band) (Supp. Fig. S9). These kurtosis-related features highlight the significance of amplitude distribution in different frequency bands for classifying seismic events. When looking at histograms of the distribution of the kurtosis values over the four classes, we see that noise is well separated from the event classes, while earthquakes, explosions, and surface events show partial overlap in the marginal distributions of individual kurtosis-based features; class discrimination emerges when

these kurtosis-based features are combined with other complementary features in the multivariate model: the noise kurtosis ranges between -0.01 and 1, the surface-event kurtosis between 1 and 20, the explosion kurtosis varies between 2 and 25, and the earthquake kurtosis varies between 6 and 60. These kurtosis-based features provided a strong separation between each class (Supp. Fig. S9). In addition to kurtosis-based features, the manual feature "hod" (hour of day) was found to be the most important feature. This was expected, as explosions typically occur during the day, making this feature highly informative for discriminating explosions from other events (Supp. Fig. S11). Although "hod" ranks highly in feature importance, it is a contextual, region-specific proxy for quarry blast timing. Its standalone predictive power is limited (Supp. Fig. S10), and adding Manual features yields only a modest gain over Physical features (Fig. 4), indicating that classification is primarily driven by waveform-derived attributes. Other notable features included the "Average envelope amplitude", which effectively shows the shaking duration and provides a good way to distinguish long-duration surface events from shorter-duration earthquakes and explosions, and "Envelope rise time", where surface events demonstrate a slower growth.

While Random Forest is a powerful algorithm for classification and explainability, it has some limitations. There is a bias toward high cardinality features as decision trees tend to assign higher importance to features with many unique values (high cardinality), regardless of whether they are truly informative. When features are correlated, the model tends to distribute importance across them. This can result in underestimating the importance of the more influential feature, but it is addressed here by removing one feature from highly correlated pairs (Supp. Table S2), retaining Feature 1 and removing Feature 2. Finally, the importance scores may vary significantly across different runs, especially if the dataset contains noise or irrelevant features. We mitigated this by averaging the importance calculated over ten iterations.

Feature Selection To identify the minimum number of features required for comparable performance to the full set of features, we computed the performance of our model with a progressively increasing number of the most important features. We find that the 20 most important features suffice to predict with an F1 score of 89%, which is as much as a model that includes 62 features (Supp. Fig. S10). Further, using just a single feature provides an F1-score of 60%, whereas using just the 20 most important features provides an F1-score of 90%. This results in a reduction of computational time by approximately 1.5, while providing similar performance (Supp. Fig. S10).

5.2 Feature Importance from DL

Difficulties in the interpretability of neural networks have been a major limitation for the broad adoption of deep learning methods in seismic discrimination. Kong et al. (2022) was among the first to utilize a gradient-based method to explore feature importance in deep learning feature extraction for event classification between earthquakes and explosions. Recently, Clements et al. (2024) showed activation feature maps to reveal the parts of the seismograms that were most contributing to predicting shaking intensity.

We chose to use the Integrated Gradients (IG) attribution method provided by the `Captum` Python library, offering a robust way to interpret model predictions (Sundararajan et al., 2017; Alzubaidi et al., 2021). IG attribution assigns importance scores to input features by computing the path integral of gradients along a straight line between a baseline input and the actual input. This method ensures that the attributions satisfy key properties such as completeness and sensitivity, making it particularly suitable for complex models like neural networks (Sundararajan et al., 2017; Alzubaidi et al., 2021). We apply IG on QuakeXnet (2D) to four representative cases for each of the four classes, presenting the seismograms, spectrograms, and IG maps in Fig. 9. For earthquakes, the most critical features are concentrated in the 5–15 Hz frequency band. They are primarily associated with the arrival of S-waves, consistent with the dominance of these wave phases in earthquake signal detection and classification. In contrast, the model's attributions for explosions concentrate more strongly in the 1–5 Hz band and are largely time-locked to P-wave onsets. We emphasize that this reflects the most informative time–frequency regions for the classifier, not necessarily the peak spectral energy of explosions; other studies (e.g., Kong et al., 2022) report dominant explosion bands at higher frequencies depending on region, distance, and preprocessing. Noise shows diffuse attribution across the wide range of frequencies, reflecting the broad and low-frequency nature of background or anthropogenic noise sources.

For surface events, the model highlights features primarily in the lower frequency range (below 5 Hz). Unlike earthquakes and explosions, surface events do not exhibit a clear S-wave phase, as analysts typically only pick the onset of such events (e.g., Ekström and Stark, 2013). Instead, the identified attribution intensity may correspond to emergent or prolonged energy release patterns, which are characteristic of surface processes. Volcanic seismicity often consists of high-frequency P- and S-waves from deeper sources or low-frequency (LF) events, where the frequency content is primarily controlled by the source mechanism rather than surface wave excitation (e.g., Chouet, 1996; Chouet and Matoza, 2013; Allstadt et al., 2014; Hürlimann et al., 2019). It is important to note that the "surface event" classification used by the PNSN is broad, encompassing a diverse set of mass-movement and volcanic processes with correspondingly varied waveforms. This diversity likely contributes to the difficulty of training machine learning models on a single unified `su` class, and should be recognized as an inherent limitation of this label.

6 Discussion

This study provides a comprehensive analysis of seismic event classification through the application of CML

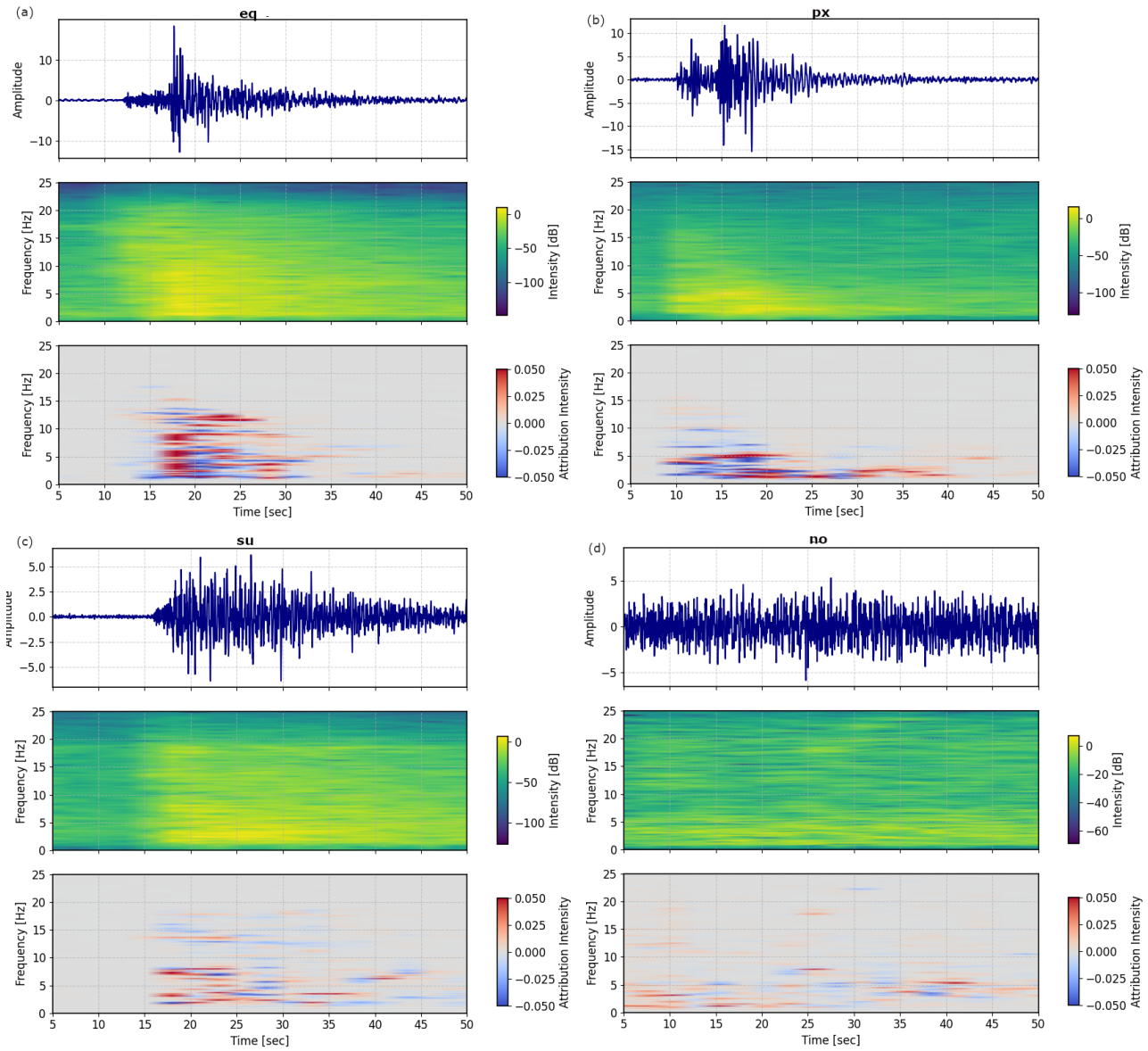

Figure 9 QuakeXNet (2D) time–frequency attributions. Representative examples for each class: (a) earthquake, (b) explosion, (c) surface event, and (d) noise. For each class, the top panel shows the input waveform amplitude, the middle panel shows the corresponding spectrogram, and the bottom panel shows the model attribution map. *Intensity* (middle panels) denotes the log-scaled spectrogram power (in dB) of the input signal as a function of time and frequency. *Attribution Intensity* (bottom panels) denotes the Integrated Gradients attribution value per time–frequency pixel with respect to the predicted class; larger absolute values indicate regions of the spectrogram that most strongly influence the model's decision (red/blue indicate positive/negative contribution, respectively).

and DL approaches, leveraging a diverse feature set derived from seismic waveforms. Our findings underscore the importance of feature selection in CML models and reveal the nuanced strengths and weaknesses of DL architectures in real-world seismic monitoring applications.

6.1 Analysis of Misclassified Events

We evaluate our best-performing model, QuakeXNet (2D) v3, on the held-out portion of the curated three-component dataset (excluded from training; 33 719 earthquakes with 109 687 traces; 3490 explosions with 5075 traces; 769 surface events with 902 traces; and 28 074 noise traces). We report both trace-level and event-level accuracy. To identify likely catalog issues, we additionally flag *high-confidence disagreements*, defined as cases where the predicted non-catalog class attains $P > 0.9$ (and, when available, remains high-confidence after averaging across multiple stations).

Surface events The model achieves 81% trace-level and 84% event-level accuracy. Among 120 misclassified surface events, 65 (54%) are high-confidence disagreements, and 22 (18%) remain high-confidence after two-station averaging. These 22 candidates were reviewed by PNSN analysts: 8 were confirmed as earthquakes mislabeled as surface events, and 4 additional cases were flagged as likely shallow volcano–tectonic (VT) events (sharp onsets, distinct phases, short dura-

tions, often at low SNR). This suggests a non-negligible surface-event label error rate in the curated catalog.

Explosions The model achieves 80% trace-level and 84% event-level accuracy. Of the misclassified explosion events, 233 are high-confidence disagreements; 17 persist across more than two stations and were reviewed by six PNSN analysts. Five were unanimously judged to be earthquakes mislabeled as explosions; only four were unanimously confirmed as true explosions, and several others resembled volcanic deep low-frequency (DLF) events not represented in the training taxonomy, indicating potential label noise and/or missing classes.

Earthquakes The model achieves 84% trace-level and 87% event-level accuracy. Most errors are predicted as explosions; 1652 (36%) are high-confidence disagreements and 288 (6%) remain high-confidence across >2 stations. Three senior PNSN analysts reviewed up to 50 such cases, unanimously confirming 9 as true explosions, identifying 5 as possible DLF events from Mount Baker, and validating the remainder as earthquakes (often teleseismic or near known quarries).

Overall, the dominant error patterns are consistent with *label leakage* in the curated catalog (e.g., shallow VT-like events within the surface-event label; explosions labeled as earthquakes and vice versa; and DLF-like signals absent from the taxonomy), rather than purely model failure. These audits indicate that modest targeted relabeling (on the order of ∼0.2–8% by class) could meaningfully affect downstream analyses, and that ML-based screening can support iterative catalog refinement. Recent work suggests similar mislabeling in phase picking data sets (Suarez and Beroza, 2025).

6.2 Model Performance Relative to Other Published Models

To the best of our knowledge, this study is the first to address classification across the three event classes most encountered at seismic networks from tectonic, anthropogenic, and geomorphologic events. Model performance in the literature is highly sensitive to choices such as the number of classes, the balance of training and testing datasets, and the architecture employed. Direct comparisons are therefore not straightforward, but placing our results in context highlights both the progress and the challenges in seismic event classification.

Most prior work has focused on binary classification; reported performance is often higher, although difficulty depends on class overlap, variability, and dataset complexity. For example, Perol et al. (2018) reported 94.9% accuracy for discriminating earthquakes from noise, and Meier et al. (2019) achieved precision and recall above 99% for the same task. Quarry blast versus earthquake discrimination has also reached near-human accuracy, with Linville et al. (2019) obtaining accuracies above 99% and Kong et al. (2022) obtaining 95.2% accuracy by combining physics-based and learned features to discriminate earthquakes and explosions. Other binary classifiers targeting noise versus earthquakes or specific volcanic signals similarly report accuracies exceeding 95–98% (Wu et al., 2018; Chakraborty et al., 2022).

Extending to multi-class problems is more challenging. Canário et al. (2020) achieved 96–98% accuracy across multiple volcanic seismicity classes. More recently, Maguire et al. (2024) trained CNNs across diverse U.S. regions and achieved station-level accuracies of ∼90% on previously unseen areas, underscoring the difficulty of robust generalization.

Our CNN-based models achieve accuracy and F1 scores of 92–94% on a balanced curated dataset spanning four classes, and retain high performance (F1≈0.95) under more realistic network conditions. This places our work at the high end of reported accuracies for multi-class classification, while tackling a more complex data and problem than most binary approaches. Importantly, our results highlight not only the feasibility of four-class discrimination but also the limits of choosing a region-specific training dataset, motivating the need for iterative dataset augmentation and generalization testing.

6.3 Deployment on Continuous Data

A key goal of this study is operational deployment: applying classifiers not just to curated test sets but to continuous waveform archives and real-time streams. We integrated our best deep learning model, QuakeXNet (2D) v3, into the open-source `SeisBench` ecosystem (Woollam et al., 2022), extending its API to support multi-class classification in parallel with existing phase pickers.

Our classifier takes raw three-component seismic waveforms as input, performs preprocessing internally, and outputs four probability traces – one for each event class. The temporal resolution of these traces depends on the stride: for example, a 400 s input with a 5-s stride produces 61 probability values per class, each representing the likelihood that the following 100 s belong to that class. To convert probability traces into discrete detections, we smooth the outputs with a five-sample moving average to suppress spurious peaks. A detection is triggered when the smoothed probability exceeds 0.15, terminated when it falls below, and validated if the within-window maximum exceeds 0.5. If multiple classes peak in the same window, the class with the highest maximum probability is assigned. The 0.15 onset/offset threshold was chosen empirically by sweeping candidate values on development data and selecting a value that balanced missed detections and false positives in continuous/network-style scans. We further integrated the classifier into QuakeScope (https://github.com/SeisSCOPED/QuakeScope) , enabling joint operation with phase pickers on cloud-hosted data (e.g., SCEDC (Yu et al., 2021), NCEDC, EarthScope archives), expanding to the earlier data mining work from Ni et al. (2025a). This parallel design allows detection and discrimination to inform each other, an important feature since phase pickers are not trained on surface events.

We do not report a quantitative continuous-detection benchmark within QuakeScope here; evaluating detection performance across different onset characteristics (including less-emergent events) requires event association and catalog/analyst validation and is left for future work. Finally, we benchmarked computational performance. QuakeXNet (2D) v3 requires only ∼5 s to process a full day of three-component data at 100 Hz with a 20-s stride. This cost is comparable to PhaseNet, underscoring that integrating multiple DL models into routine workflows is computationally feasible. Together, these results highlight QuakeXNet (2D) v3 as not only the most accurate but also an operationally practical solution for large-scale cataloging and real-time monitoring.

6.4 Implications and Recommendations

Our experiments show that while classical machine learning approaches can provide seismologically interpretable insights, they are ultimately limited compared to deep learning models. CML performance depends heavily on feature design, with Physical + Manual features emerging as the most important. Shorter windows improved earthquake recall, whereas longer windows benefited explosion and surface event detection. Broader frequency ranges consistently improved performance across classes. These findings align with seismological expectations and validate the utility of feature-based approaches for exploring signal characteristics. However, even with extensive hyperparameter tuning and feature engineering, CML models plateaued below the performance of DL classifiers and struggled to generalize to the network testing dataset.

In contrast, DL models bypass the need for intermediate feature extraction, learning directly from waveform or spectrogram representations. They consistently achieved higher accuracy, precision, and recall, while also being more computationally efficient at inference (or deployment) time. This efficiency makes DL approaches better suited for operational deployment, where throughput and robustness are critical.

Among the DL architectures tested, QuakeXNet (2D) emerged as the most reliable model across all evaluation stages. It generalized better than SeismicCNN (2D) to network and out-of-domain datasets, while maintaining high performance on curated test sets. Crucially, QuakeXNet (2D) is also lightweight and fast, requiring only 70k parameters (∼1.2 MB memory) and processing a full day of continuous data in ∼9 s. This combination of accuracy, robustness, and efficiency makes QuakeXNet (2D) the recommended model for all use cases considered in this study: earthquake monitoring, explosion detection, and surface event cataloging.

Overall, these results highlight that while CML models remain valuable for interpreting seismic features, DL models – and QuakeXNet (2D) in particular – provide the most practical and scalable solution for modern seismic event classification.

7 Conclusions

In this study, we developed and evaluated classical machine learning and deep learning models for classifying dominant seismic events in the Pacific Northwest. Our results indicate that while deep learning models, such as QuakeXNet (2D) and SeismicCNN (2D), perform well on test datasets and offer higher classification confidence for the Pacific Northwest data, classical machine learning models have also demonstrated relatively good performance and can be used in an ensemble fashion.

Our analysis highlighted the key features that distinguish different seismic event types. Classical machine learning models emphasized kurtosis-based features, which provided clear separation among noise, surface events, explosions, and earthquakes. In particular, kurtosis in specific frequency bands (e.g., 3–10 Hz) proved highly informative, while contextual features such as time-of-day helped discriminate explosions from other classes. We also note that time-of-day is a region- and source-dependent contextual proxy (e.g., quarry/mining activity) and may not transfer to other explosion types (e.g., volcanic explosions) or to regions without diurnal blasting patterns. Deep learning attribution maps offered complementary insights, showing that earthquakes are characterized by energy concentrated in the 5–15 Hz band and linked to S-wave arrivals, whereas explosions emphasize lower-frequency energy (1–5 Hz) associated with P-wave onsets and extended codas. Noise showed broad, diffuse patterns, while surface events exhibited energy concentrated at low frequencies without distinct S-wave phases, reflecting emergent or prolonged release processes. Together, these findings demonstrate that both engineered and learned features converge on physically meaningful signal properties, and that integrating multiple perspectives improves our ability to discriminate between event types.

Our analysis also shows that combining data from multiple stations improves classification performance by averaging out noise and reducing the impact of individual station biases – another common form of network seismology and association. Deep learning models performed better than classical machine learning models both in terms of classification performance as well as computational costs on the curated PNW dataset. Further, we enhanced the generalizability of the original deep learning models by training them with an incrementally enriched out-of-domain dataset. Overall, we found that QuakeXNet (2D) v3 is the best model in terms of performance, computational costs, and generalizability. This model performed well on the curated PNW datasets and out-of-domain surface events and explosions, and should be utilized for large-scale detection of surface events.

Acknowledgements

We acknowledge support from the US Geological Survey Earthquake Science Center through Cooperative Agreement G23AC00278 and the partial funding provided by the PNSN (USGS cooperative agreement G20AC00035).

We thank seismic analyst volunteers Amy Wright, Paul Bodin, and Barrett Johnson for their help validating event categories. We thank the PNSN analysts who helped review the misclassified events. We thank the editor Marlon Ramos and the two anonymous reviewers for their thorough and constructive comments, which substantially improved the clarity and quality of this manuscript.

Data and Code Availability

The seismic waveform dataset used in this study is publicly accessible at https://github.com/EarthML4PNW/PNW-ML(e.g., Ni et al., 2023). All models evaluated in this study, including feature extraction scripts and hyperparameter tuning configurations, are available in the repository https://github.com/Akashkharita/PNW_Seismic_Event_Classification. Additionally, the trained random forest models and scaler parameters (Physical + Manual feature set; CML artifacts used in this study) are archived on Zenodo (https://zenodo.org/records/13334838). Deep-learning model architectures, training scripts, and inference or deployment notebooks are provided in https://github.com/Akashkharita/PNW_Seismic_Event_Classification. Real-time testing and deployment codes can be found at https://github.com/Akashkharita/Surface_Event_Detection.

Competing Interests

The authors declare no competing interests.

References

Allen, R. Automatic phase pickers: Their present use and future prospects. *Bulletin of the Seismological Society of America*, 72 (6B):S225–S242, Dec. 1982. doi: 10.1785/bssa07206b0225.

Allstadt, K., Malone, S., Vidale, J., Bodin, P., and Steele, B. Seismic signals generated by the Oso landslide. *Posted by Pacific Northwest Seismic Network. Accessed March*, 26:2014, 2014. https://wa.water.usgs.gov/data/SeismicReport2-OsoLandslide.pdf.

Allstadt, K. E., Matoza, R. S., Lockhart, A. B., Moran, S. C., Caplan-Auerbach, J., Haney, M. M., Thelen, W. A., and Malone, S. D. Seismic and acoustic signatures of surficial mass movements at volcanoes. *Journal of Volcanology and Geothermal Research*, 364: 76–106, Sept. 2018. doi: 10.1016/j.jvolgeores.2018.09.007.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), Mar. 2021. doi: 10.1186/s40537-021-00444-8.

Anden, J. and Mallat, S. Deep Scattering Spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, Aug. 2014. doi: 10.1109/tsp.2014.2326991.

Bahavar, M., Allstadt, K. E., Van Fossen, M., Malone, S. D., and Trabant, C. Exotic Seismic Events Catalog (ESEC) Data Product. *Seismological Research Letters*, 90(3):1355–1363, Apr. 2019. doi: 10.1785/0220180402.

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., Liu, H., Schultz, T., and Gamboa, H. TSFEL: Time Series Feature Extraction Library. *SoftwareX*, 11:100456, Jan. 2020. doi: 10.1016/j.softx.2020.100456.

Bartlow, N. M. A Long-Term View of Episodic Tremor and Slip in Cascadia. *Geophysical Research Letters*, 47(3), Feb. 2020. doi: 10.1029/2019gl085303.

Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C. Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433), Mar. 2019. doi: 10.1126/science.aau0323.

Breiman, L. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. doi: 10.1023/a:1010933404324.

Canário, J. P., Mello, R., Curilem, M., Huenupan, F., and Rios, R. In-depth comparison of deep artificial neural network architectures on seismic events classification. *Journal of Volcanology and Geothermal Research*, 401:106881, Sept. 2020. doi: 10.1016/j.jvolgeores.2020.106881.

Carniel, R. and Raquel Guzmán, S. Machine Learning in Volcanology: A Review, june 2021. doi: 10.5772/intechopen.94217.

Chakraborty, M., Fenner, D., Li, W., Faber, J., Zhou, K., Rümpker, G., Stöcker, H., and Srivastava, N. CREIME – A Convolutional Recurrent model for Earthquake Identification and Magnitude Estimation, Apr. 2022. doi: 10.1002/essoar.10511140.1.

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, Aug. 2016. doi: 10.1145/2939672.2939785.

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B. W., and Hibert, C. Machine Learning Improves Debris Flow Warning. *Geophysical Research Letters*, 48(3), Feb. 2021. doi: 10.1029/2020gl090874.

Chouet, B. A. Long-period volcano seismicity: its source and use in eruption forecasting. *Nature*, 380(6572):309–316, Mar. 1996. doi: 10.1038/380309a0.

Chouet, B. A. and Matoza, R. S. A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. *Journal of Volcanology and Geothermal Research*, 252: 108–175, Feb. 2013. doi: 10.1016/j.jvolgeores.2012.11.013.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307:72–77, Sept. 2018. doi: 10.1016/j.neucom.2018.03.067.

Clements, T., Cochran, E. S., Baltay, A., Minson, S. E., and Yoon, C. E. GRAPES: Earthquake Early Warning by Passing Seismic Vectors Through the Grapevine. *Geophysical Research Letters*, 51 (9), May 2024. doi: 10.1029/2023gl107389.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, Jan. 1967. doi: 10.1109/tit.1967.1053964.

Domel, P., Hibert, C., Schlindwein, V., and Plaza-Faverola, A. Event recognition in marine seismological data using Random Forest machine learning classifier. *Geophysical Journal International*, 235(1):589–609, May 2023. doi: 10.1093/gji/ggad244.

Ekström, G. and Stark, C. P. Simple Scaling of Catastrophic Landslide Dynamics. *Science*, 339(6126):1416–1419, Mar. 2013. doi: 10.1126/science.1232887.

Elsken, T., Metzen, J. H., and Hutter, F. *Neural Architecture Search*, page 63–77. Springer International Publishing, 2019. doi: 10.1007/978-3-030-05318-5_3.

Gomberg, J. and Bodin, P. The Productivity of Cascadia Aftershock Sequences. *Bulletin of the Seismological Society of America*, Apr. 2021. doi: 10.1785/01/20200344.

Hearst, M., Dumais, S., Osuna, E., Platt, J., and Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, July 1998. doi: 10.1109/5254.708428.

Hellweg, M., Bodin, P., Bormann, J. M., Haddadi, H., Hauksson, E., and Smith, K. D. Regional Seismic Networks Operating along the West Coast of the United States of America. *Seismological Research Letters*, 91(2A):695–706, Feb. 2020. doi: 10.1785/0220190282.

Hibert, C., Mangeney, A., Grandjean, G., and Shapiro, N. M. Slope instabilities in Dolomieu crater, Réunion Island: From seismic signals to rockfall characteristics. *Journal of Geophysical Research*, 116(F4), Dec. 2011. doi: 10.1029/2011jf002038.

Hibert, C., Mangeney, A., Grandjean, G., Baillard, C., Rivet, D., Shapiro, N. M., Satriano, C., Maggi, A., Boissier, P., Ferrazzini, V., and Crawford, W. Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano. *Journal of Geophysical Research: Earth Surface*, 119(5): 1082–1105, May 2014. doi: 10.1002/2013jf002970.

Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A., and Ferrazzini, V. Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm. *Journal of Volcanology and Geothermal Research*, 340:130–142, June 2017. doi: 10.1016/j.jvolgeores.2017.04.015.

Hibert, C., Michéa, D., Provost, F., Malet, J.-P., and Geertsema, M. Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska. *Geophysical Journal International*, 219(2):1138–1147, July 2019. doi: 10.1093/gji/ggz354.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. *Applied Logistic Regression*. Wiley, Mar. 2013. doi: 10.1002/9781118548387.

Huynh, C., Hibert, C., Jestin, C., Malet, J. P., and Lanticq, V. A real scale application of a novel set of spatial and similarity features for detection and classification of natural seismic sources from distributed acoustic sensing data. *Geophysical Journal International*, 240(1):462–482, Oct. 2024. doi: 10.1093/gji/ggae382.

Hürlimann, M., Coviello, V., Bel, C., Guo, X., Berti, M., Graf, C., Hübl, J., Miyata, S., Smith, J. B., and Yin, H.-Y. Debris-flow monitoring and warning: Review and examples. *Earth-Science Reviews*, 199: 102981, Dec. 2019. doi: 10.1016/j.earscirev.2019.102981.

Ichinose, G. A., Thio, H. K., and Somerville, P. G. Rupture process and near-source shaking of the 1965 Seattle-Tacoma and 2001 Nisqually, intraslab earthquakes. *Geophysical Research Letters*, 31(10), May 2004. doi: 10.1029/2004gl019668.

Jordan, M. I. and Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, July 2015. doi: 10.1126/science.aaa8415.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. https://dl.acm.org/doi/10.5555/3294996.3295074.

Kharita, A., Denolle, M. A., and West, M. E. Discrimination between icequakes and earthquakes in southern Alaska: an exploration of waveform features using Random Forest algorithm. *Geophysical Journal International*, 237(2):1189–1207, Mar. 2024. doi: 10.1093/gji/ggae106.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint*, 2014. doi: 10.48550/ARXIV.1412.6980.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P. Machine Learning in Seismology: Turning Data into Insights. *Seismological Research Letters*, 90(1):3–14, Nov. 2018. doi: 10.1785/0220180259.

Kong, Q., Chiang, A., Aguiar, A. C., Fernández-Godino, M. G., Myers, S. C., and Lucas, D. D. Deep convolutional autoencoders as generic feature extractors in seismological applications. *Artificial Intelligence in Geosciences*, 2:96–106, Dec. 2021. doi: 10.1016/j.aiig.2021.12.002.

Kong, Q., Wang, R., Walter, W. R., Pyle, M., Koper, K., and Schmandt, B. Combining Deep Learning With Physics Based Features in Explosion-Earthquake Discrimination. *Geophysical Research Letters*, 49(13), July 2022. doi: 10.1029/2022gl098645.

Koper, K. D., Pechmann, J. C., Burlacu, R., Pankow, K. L., Stein, J., Hale, J. M., Roberson, P., and McCarter, M. K. Magnitude-based discrimination of man-made seismic events from naturally occurring earthquakes in Utah, USA. *Geophysical Research Letters*, 43(20), Oct. 2016. doi: 10.1002/2016gl070742.

Koper, K. D., Holt, M. M., Voyles, J. R., Burlacu, R., Pyle, M. L., Wang, R., and Schmandt, B. Discrimination of Small Earthquakes and Buried Single-Fired Chemical Explosions at Local Distances (<150 km) in the Western United States from Comparison of Local Magnitude (ML) and Coda Duration Magnitude (MC). *Bulletin of the Seismological Society of America*, 111(1): 558–570, Oct. 2020. doi: 10.1785/0120200188.

Koper, K. D., Burlacu, R., Armstrong, A. D., and Herrmann, R. B. Classifying small earthquakes, explosions and collapses in the western United States using physics-based features and machine learning. *Geophysical Journal International*, 239(2): 1257–1270, Sept. 2024. doi: 10.1093/gji/ggae316.

Kramer, R. L., Thelen, W. A., Iezzi, A. M., Moran, S. C., and Pauk, B. A. Recent Expansion of the Cascades Volcano Observatory Geophysical Network at Mount Rainier for Improved Volcano and Lahar Monitoring. *Seismological Research Letters*, 95(5): 2707–2721, July 2024. doi: 10.1785/0220240112.

Köpfli, M., Denolle, M. A., Thelen, W. A., Makus, P., and Malone, S. D. Examining 22 Years of Ambient Seismic Wavefield at Mount St. Helens. *Seismological Research Letters*, 95(5):2622–2636, June 2024. doi: 10.1785/0220240079.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521 (7553):436–444, May 2015. doi: 10.1038/nature14539.

Linville, L., Pankow, K., and Draelos, T. Deep Learning Models Augment Analyst Decisions for Event Discrimination. *Geophysical Research Letters*, 46(7):3643–3651, Apr. 2019. doi: 10.1029/2018gl081119.

Luna, L. V. and Korup, O. Seasonal Landslide Activity Lags Annual Precipitation Pattern in the Pacific Northwest. *Geophysical Research Letters*, 49(18), Sept. 2022. doi: 10.1029/2022gl098506.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., and Amemoutou, A. Implementation of a Multistation Approach for Automated Event Classification at Piton de la Fournaise Volcano. *Seismological Research Letters*, 88(3):878–891, Mar. 2017. doi: 10.1785/0220160189.

Maguire, R., Schmandt, B., Wang, R., Kong, Q., and Sanchez, P. Generalization of Deep-Learning Models for Classification of Local Distance Earthquakes and Explosions across Various Geologic Settings. *Seismological Research Letters*, 95(4):2229–2238, Feb. 2024. doi: 10.1785/0220230267.

Malfante, M., Dalla Mura, M., Mars, J. I., Métaxian, J., Macedo, O., and Inza, A. Automatic Classification of Volcano Seismic Signatures. *Journal of Geophysical Research: Solid Earth*, 123(12), Dec. 2018. doi: 10.1029/2018jb015470.

Meier, M., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., Li, Z., Andrews, J., Hauksson, E., and Yue, Y. Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning. *Journal of Geophysical Research: Solid Earth*, 124(1):788–800, Jan. 2019. doi: 10.1029/2018jb016661.

Moreau, L., Seydoux, L., Weiss, J., and Campillo, M. Analysis of micro-seismicity in sea ice with deep learning and Bayesian inference: application to high-resolution thickness monitoring, Nov. 2022. doi: 10.5194/tc-2022-212.

Mousavi, S. M. and Beroza, G. C. Deep-learning seismology. *Science*, 377(6607), Aug. 2022. doi: 10.1126/science.abm4470.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. Earthquake transformer–an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1), Aug. 2020. doi: 10.1038/s41467-020-17591-w.

Ni, Y., Hutko, A., Skene, F., Denolle, M., Malone, S., Bodin, P., Hartog, R., and Wright, A. Curated Pacific Northwest AI-ready Seismic Dataset, Feb. 2023. doi: 10.31223/x53w9q.

Ni, Y., Denolle, M., Thomas, A., Hamilton, A., Münchmeyer, J., Wang, Y., Bachelot, L., Trabant, C., and Mencin, D. A Global-scale Database of Seismic Phases from Cloud-based Picking at Petabyte Scale. *Seismica*, 4(2), Sept. 2025a. doi: 10.26443/seismica.v4i2.1738.

Ni, Y., Denolle, M. A., Münchmeyer, J., Wang, Y., Feng, K.-F., Garcia Jurado Suarez, C., Thomas, A. M., Trabant, C., Hamilton, A., and Mencin, D. A review of cloud computing and storage in seismology. *Geophysical Journal International*, 243(1), Aug. 2025b. doi: 10.1093/gji/ggaf322.

Perol, T., Gharbi, M., and Denolle, M. Convolutional neural network for earthquake detection and location. *Science Advances*, 4(2), Feb. 2018. doi: 10.1126/sciadv.1700578.

Pirot, E., Hibert, C., and Mangeney, A. Enhanced glacial earthquake catalogues with supervised machine learning for more comprehensive analysis. *Geophysical Journal International*, 236 (2):849–871, Oct. 2023. doi: 10.1093/gji/ggad402.

Provost, F., Hibert, C., and Malet, J. Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier. *Geophysical Research Letters*, 44(1):113–120, Jan. 2017. doi: 10.1002/2016gl070709.

Pyle, M. L. and Walter, W. R. Investigating the Effectiveness of P/S Amplitude Ratios for Local Distance Event Discrimination. *Bulletin of the Seismological Society of America*, 109(3), Mar. 2019. doi: 10.1785/0120180256.

Pyle, M. L. and Walter, W. R. Exploring the Effects of Emplacement Conditions on ExplosionP/SRatios across Local to Regional Distances. *Seismological Research Letters*, 93(2A):866–879, Dec. 2021. doi: 10.1785/0220210270.

Renate Hartog, J., Friberg, P. A., Kress, V. C., Bodin, P., and Bhadha, R. Open-Source ANSS Quake Monitoring System Software. *Seismological Research Letters*, 91(2A):677–686, Nov. 2019. doi: 10.1785/0220190219.

Rogers, G. and Dragert, H. Episodic Tremor and Slip on the Cascadia Subduction Zone: The Chatter of Silent Slip. *Science*, 300 (5627):1942–1943, June 2003. doi: 10.1126/science.1084783.

Ross, Z. E., Meier, M., and Hauksson, E. P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning. *Journal of Geophysical Research: Solid Earth*, 123(6):5120–5129, June 2018. doi: 10.1029/2017jb015251.

Royer, A. and Bostock, M. A comparative study of low frequency earthquake templates in northern Cascadia. *Earth and Planetary Science Letters*, 402:247–256, Sept. 2014. doi: 10.1016/j.epsl.2013.08.040.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. d., Campillo, M., and Baraniuk, R. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, 11(1), Aug. 2020. doi: 10.1038/s41467-020-17841-x.

Steinmann, R., Seydoux, L., Journeau, C., Shapiro, N. M., and Campillo, M. Machine learning analysis of seismograms reveals a continuous plumbing system evolution beneath the Klyuchevskoy volcano in Kamchatka, Russia, june 2023. doi: 10.22541/essoar.168614505.54607219/v1.

Suarez, A. L. A. and Beroza, G. Pervasive Label Errors in Seismological Machine Learning Datasets, 2025. doi: 10.48550/ARXIV.2511.09805.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv preprint*, 2017. doi: 10.48550/ARXIV.1703.01365.

Tibi, R., Koper, K. D., Pankow, K. L., and Young, C. J. Discrimination of Anthropogenic Events and Tectonic Earthquakes in Utah Using a Quadratic Discriminant Function Approach with Local Distance Amplitude Ratios. *Bulletin of the Seismological Society of America*, 108(5A):2788–2800, July 2018. doi: 10.1785/0120180024.

Wang, R., Schmandt, B., and Kiser, E. Seismic discrimination of controlled explosions and earthquakes near Mount St. Helens using P/S ratios, june 2020. doi: 10.1002/essoar.10503320.1.

Wang, T., Bian, Y., Zhang, Y., and Hou, X. Using Artificial Intelligence Methods to Classify Different Seismic Events. *Seismological Research Letters*, 94(1):1–16, Nov. 2022. doi: 10.1785/0220220055.

Wassermann, J. Volcano Seismology. In Bormann, P., editor, *New Manual of Seismological Observatory Practice 2 (NMSOP2)*. Deutsches GeoForschungsZentrum GFZ, 2012. doi: 10.2312/GFZ.NMSOP-2_CH13.

Wech, A. G. and Bartlow, N. M. Slip rate and tremor genesis in Cascadia. *Geophysical Research Letters*, 41(2):392–398, Jan. 2014. doi: 10.1002/2013gl058607.

Wech, A. G., Creager, K. C., Houston, H., and Vidale, J. E. An earthquake-like magnitude-frequency distribution of slow slip in northern Cascadia. *Geophysical Research Letters*, 37(22), Nov. 2010. doi: 10.1029/2010gl044881.

Wenner, M., Hibert, C., Meier, L., and Walter, F. Near Real-Time Automated Classification of Seismic Signals of Slope Failures with Continuous Random Forests, july 2020. doi: 10.5194/nhess-2020-200.

Witter, R. C., Kelsey, H. M., and Hemphill-Haley, E. Great Cascadia earthquakes and tsunamis of the past 6700 years, Coquille River estuary, southern coastal Oregon. *Geological Society of America Bulletin*, 115(10):1289, 2003. doi: 10.1130/b25189.1.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., Michelini, A., Saul, J., and Soto, H. SeisBench–A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*, 93(3):1695–1709, Mar. 2022. doi: 10.1785/0220210324.

Wu, Y., Lin, Y., Zhou, Z., and Delorey, A. Seismic-Net: A Deep Densely Connected Neural Network to Detect Seismic Events. *arXiv preprint*, 2018. doi: 10.48550/ARXIV.1802.02241.

Yeck, W. L., Patton, J. M., Ross, Z. E., Hayes, G. P., Guy, M. R., Ambruz, N. B., Shelly, D. R., Benz, H. M., and Earle, P. S. Leveraging Deep Learning in Global 24/7 Real-Time Earthquake Monitoring at the National Earthquake Information Center. *Seismological Research Letters*, 92(1):469–480, Sept. 2020. doi: 10.1785/0220200178.

Yin, J., Denolle, M. A., and He, B. A multitask encoder–decoder to separate earthquake and ambient noise signal in seismograms. *Geophysical Journal International*, 231(3):1806–1822, July 2022. doi: 10.1093/gji/ggac290.

Yu, E., Bhaskaran, A., Chen, S.-L., Ross, Z. E., Hauksson, E., and Clayton, R. W. Southern California Earthquake Data Now Available in the AWS Cloud. *Seismological Research Letters*, 92(5): 3238–3247, June 2021. doi: 10.1785/0220210039.

Zeiler, C. and Velasco, A. A. Developing Local to Near-Regional Explosion and Earthquake Discriminants. *Bulletin of the Seismological Society of America*, 99(1):24–35, Feb. 2009. doi:

10.1785/0120080045.

Zhang, R. Making Convolutional Networks Shift-Invariant Again. *arXiv preprint*, 2019. doi: 10.48550/ARXIV.1904.11486.

Zheng, A. and Casari, A. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018. https://dl.acm.org/doi/book/10.5555/3239815.