Dear Editor and Reviewers

We are grateful for your constructive feedback and for answering the very thorough comments from both reviewers. We acknowledge the convoluted and long manuscript, which resulted from a deep exploration of multi-class discrimination of seismic events. We respond to the comments in blue.

Reviewer #1

We are deeply grateful for the amazing and thorough review. We thank you for your time and focused attention to our paper.

General Comments:

Thank you for sharing this interesting manuscript. The substantial effort required to assemble diverse training and benchmarking datasets, train multiple models, and plan for operational deployment is rare and valuable in seismic event discrimination research. This comprehensive scope is a clear strength. That said, the breadth of datasets and model variants can make the narrative dense. To improve focus and readability, consider moving datasets or model versions that did not contribute to the final refined datasets or top-performing models into the Supplement. This preserves transparency about the research path without distracting from the core results and conclusions.

A few sections revisit material already introduced. You can reduce repetition while maintaining clarity by:

- Defining key terms once, then using acronyms consistently thereafter.
- Referring back to earlier sections rather than restating definitions or findings.
- Consolidating repeated results into a single, clearly referenced location.
- Acronym usage: ensure each acronym is defined at first occurrence, then used consistently. Avoid redefining the same acronym later.
- Figure labeling: be consistent with “Figure” or “Fig.” throughout. Pick one style and apply it uniformly.

Thank you for these suggestions. We have revised the manuscript to reduce repetition by defining key terms once and using acronyms consistently thereafter, referencing earlier sections instead of restating definitions, and consolidating repeated results into single, clearly referenced locations. We also standardized acronym usage (defined at first mention only) and made figure references consistent throughout (using a single style uniformly).

I include a detailed list figure comments and of line-by-line questions and minor corrections below to support final polishing. Overall, the manuscript presents strong technical work and a thoughtful operational perspective; these refinements will further strengthen its impact and readability.

General Figure Comments:

- Many figures need more descriptive detail for interpretability. These additions will significantly improve figure readability, methodological clarity, and alignment between narrative and visuals.
- I recommend only including figures that are explicitly referenced in the text. If Figure 7 and Figure 10 are intended to support specific arguments, add explicit call-outs in the appropriate sections. Otherwise, remove from the main text.
- Reorder supplementary figures so they appear in the order of first mention in the manuscript. Suggested order: S1, S3, S4, S5, S6, S7, S8, S9, S11, S13, S14, S15. Place any non-referenced figures, including S2, S10, S12, at the end of the supplement or consider removing them.

Thank you for these helpful suggestions. We have revised the manuscript accordingly. Specifically, we (i) expanded the captions across the figure set to include clearer methodological context and definitions where needed, improving interpretability and alignment with the main text; (ii) ensured that all figures included in the main manuscript are explicitly referenced—Figures 7 and 10 are now called out in the relevant sections; and (iii) reorganized the Supplementary Figures to follow the order of first mention in the manuscript. In addition, we removed Supplementary Figures S2 and S12, which were not referenced and did not materially support the core results.

Specific Figure suggestions

- Figure 1
  - Add volcano locations to align with the text that emphasizes Mt. St. Helens, Mt. Rainier, and Mt. Baker.

    We have added the locations of volcanoes in the current figure as red triangles and updated the caption.
  - Showing location of your benchmark datasets could be helpful too.

We have added the location of datasets used in training and testing in the supplementary materials.

§ Consider encoding earthquake depth or magnitude.

We are opting out of this modification because depths are shown already in the Ni et al, 2023 paper, and the definition of magnitude across event types is not fully standard yet (e.g., magnitudes were defined with coda waves before 2014 and with local magnitude after that, but there are known effects of these two magnitude estimates between explosions and earthquakes (Kholer et al, 2024), and we do not have magnitude estimates for the surface events)

§ You only define surface event stations? What about other stations?

The surface event stations are shown solely to point the readers to where the events are, because there is no location yet.

o Figure 2-3

§ The figure is referenced in quick succession with Figure 3. Consider combining Figure 2 and Figure 3 into a single multi-panel figure to show waveform, spectrogram, and Fourier spectrum together for each class, or remove Figure 2 and retain Figure 3.

We have moved the current Figure 2 into the supplementary and retained the current Figure 3 as Figure 2

§ Add labels indicating event metadata such as epicentral distance and depth for the earthquake examples, to demonstrate that events spanning varied locations in the training dataset fit within the selected training windows.

We have added the magnitude, depth, and epicentral distance information for all of the current events.

o Figure 4

§ Consider combining Figure 4 with Supplementary Figure S5. This would better connect the conceptual workflow to the specific architectures used, and provide a more informative overview of "fat shallow" versus "skinny deep," with clear input dimensions for 1D time series and 2D spectrograms.

We have now combined Figure 4 with Supplementary Figure S5, showing the architectures of the main CNN models used in the study, and corrected the references wherever necessary.

o Figure 5 and Figure 6

§ Much of the interpretive detail currently embedded in the text would be more useful if moved into the figure captions.

We have revised the caption of these figures to incorporate more details.

o Figure 7

- § Not explicitly mentioned in the text.

  We have added the reference to this figure in the section "Overall performance," where the computational time and memory requirements of individual models are discussed.

- Figure 8
  - § Add actual descriptions or definitions of the features on the y-axis. Current labels appear to be internal variable names, which are hard to interpret.
  - § Replace or augment with meaningful terms, for example:
    - § "Kurt_3_10" to "Kurtosis in 3–10 Hz band"; "KurtoSig" to "Kurtosis of full signal, 0.5–15 Hz"; "EnvAvgAmp" to "Average envelope amplitude"

  Thank you for this suggestion—we agree that the y-axis labels in Figure 8 use internal feature/variable names and are not immediately interpretable. To address this, we added a comprehensive feature glossary table in the Supplementary Materials that lists **every feature** used in the analysis together with a clear description/definition (including frequency-band context where applicable). We also updated the Figure 8 caption to explicitly reference this Supplementary table so readers can map each label in the figure to its full definition. This provides a complete and unambiguous interpretation of all feature labels without overcrowding the y-axis.

- Figure 9
  - § Define "Intensity" and "Attribution Intensity" colors explicitly. For example, state:
    - § "Intensity" refers to spectrogram power or normalized PSD in specified frequency bands. -"Attribution Intensity" indicates Integrated Gradients attribution magnitude per time–frequency pixel, showing regions most influential for the model's class decision..

    Thank you for this suggestion. We agree that the terms "Intensity" and "Attribution Intensity" should be defined explicitly. We updated the Figure 9 caption to clarify this.

- Figure 10
  - § Differences between station-level and event-level performance should be labeled directly on the plots for easier interpretability, for example, panel titles "Station-level" and "Event-level," or embedded annotations.

    Thank you for this suggestion. We have updated the figures with clear titles on each panel for easier interpretability

  - § If this figure remains in the main text, add an explicit call-out in the section where ESEC and near-field performance is discussed. Otherwise, move to the supplement.

    We have added an explicit reference to this figure in the mentioned sections for ease of interpretability

Supplementary Figures

- Supplementary Figure S9 - Label of the star marker missing.

Thank you for the comment. In Supplementary Figure S9, the yellow star markers denote computational time as a function of the number of features. This quantity was already shown on the secondary (twin) y-axis, and we have now also added an explicit legend/label entry for the star markers to make this mapping unambiguous.

- The manuscript includes magnitude and epicentral distance distributions for the testing dataset, but not for the training dataset. It would be useful to show comparable summaries for the training data.

The relevant information for both the training and testing datasets have been added in the form of Supplementary Figure 5

Acronyms Issues (not defined at first use):

§ SNR - Used before explicit expansion. Add "signal-to-noise ratio (SNR)"

Corrected

§ PSD - Used without expansion in spectrogram discussion. Add "power spectral density (PSD)"

Corrected

§ FFT-Used without expansion. Add "fast Fourier transform (FFT)" at first use.

Corrected

§ ScatNet- Shorthand used without explicit tie to "scattering convolutional neural networks." Add "ScatNet (scattering network)"

Corrected

§ NAS - Used without expansion. Add "neural architecture search (NAS)" at first use.

Corrected

§ AWS, S3 - "Amazon Web Services S3" appears, but AWS is not expanded. Add "Amazon Web Services (AWS)" and "Simple Storage Service (S3)" at first use.

Corrected

§ VT, LF, DLF - Volcano-tectonic (VT) and low-frequency (LF) are referenced without ensured first-use expansion. Deep low-frequency (DLF) appears later without prior expansion. Add expansions at first mention.
Corrected

§ 3C, 1C - Three-component is used, but "3C" should be expanded at first use.
Corrected

§ Z, E, N channel labels- Channel shorthand appears without expansion. Add "vertical (Z), east (E), north (N)" at first use.
Corrected

§ HOD - "Hour of day" appears in feature lists and figure context, but if "HOD" is used as a label, expand at first use in text or caption.
Corrected

§ Additional Notes:- Proper names like SeisBench, PhaseNet, QuakeScope, tsfresh, Captum, and model names (SeismicCNN, QuakeXNet) -> *ensure consistent capitalization.*
Corrected

Line By-Line Comments

· Line 19: "test set of a.." remove comma
Corrected

· Line 18–24: Consider splitting for clarity - "QuakeXNet-2D is lightweight…, implemented …, with released checkpoints."
Corrected

· Line 33: "over twenty" is vague, why not just say the number of volcanos?
Corrected

· Line 35 & 36: "REF" I'm assuming are reference placeholders that need to be replaced (multiple instances in manuscript)
Corrected

· Line 49–53: You mention limitations but only describe one – that these techniques are not proper for near-real time.. you can also mention that more than event type these discriminants can have decreased performance at local distances (Pyle and Walter 2019, 2021) and some such as ML-MC are more of a depth discriminant that source type (Tibi et al.,2018; Koper et al. 2021).

Thank you for the suggestion. We agree that the original text implied multiple limitations, but only discussed near-real-time availability. We revised the paragraph to (i) note that key inputs (reliable phase picks and magnitude estimates) are often unavailable early in processing, (ii) acknowledge reduced robustness at local distances due to path/site effects (Pyle and Walter, 2019, 2021), and (iii) clarify that some metrics (e.g., ML–MC) can act primarily as depth/shallow-source discriminants rather than uniquely indicating source type (Tibi et al., 2018; Koper et al., 2021).

· Line 54-56: This section notable missing references Pyle and Walter (2019, 2021) and Tibi et al. (2018)

We thank the reviewer for pointing out the missing references! They have been added.

§ https://doi.org/10.1785/0220210270
§ https://doi.org/10.1785/0120180256
§ https://doi.org/10.1785/0120180024

· Lines 67–68: “Binary classification is generally easier…” Too general a statement. I wouldn’t agree “easier”, perhaps simpler of a problem due to less data, and not needing to discriminate between more nuanced signals like this manuscript focus.

We agree that “easier” was overly general. We revised the text to describe binary classification as a simpler formulation (two classes, one decision boundary, often less nuanced class overlap) rather than inherently easier, while retaining the supporting performance context and citations.

· Line 69: These studies likely incorporated detection thresholds and were testing solely on clean data with those classifications.

Corrected

· Line 74: “..classes are less distinct..” *can be* less distinct?

Corrected

· Line 88-91: “In CML, this … seismic wavefield signatures).” This reads repetitive.

Corrected

· Line 100-101: "… difficult to distinguish..." for analysts or conventional methods?

For the PNSN analysts

· Line 102–107: "Beyond operations, clean catalogs enable new scientific insights—for instance,…" Replace em dash with comma; maybe consider replacing em dashes (you have many) with commas or parentheses.

Corrected

· Line 105: "surface events" would be helpful to be a little more clear with what you classify as a surface event- it reads that it encompasses all landslide and volcanic related events, tremor bursts, or shallow events?..

Thank you for noting this ambiguity. We clarified the definition of " surface events", a catch-all analyst label near volcanoes dominated by mass-movement signals (e.g., rockfalls/avalanches)", while acknowledging that some events can resemble glacier-related or low-frequency volcanic signals depending on the setting.

· Line 109: "… curated by (e.g..." change to "curated by Ni et al. (2023)"

Corrected

· Line 108-117: It is not necessary to include what your sections will entail, this seems a little redundant.

Thank you for this suggestion. We agree that the detailed section-by-section outline in Lines 108–117 is redundant. We have revised this paragraph by removing the extended roadmap and replacing it with a concise statement of the paper's scope and organization.

· Line 120: "..and additional waveforms we included." From where?

We have specified that this is described later in the section

· Line 126: Instead of percentage perhaps use number like you do for other classes (for consistency)

Corrected

· Line 131: "Through.. mislabeled." Confusing on first read through, consider removing or mention the section you describe this later in the text.

We have mentioned the section where this is described

· Line 132-135: Confusing phrasing / comma usage

  o Consider clarifying why zero padding is used, and potential model impact..? Do you only down sample? (I can see problems if you up-sampled

the traces and I'm not sure why you re-sample to 100hz when the highest frequency you use is 15/20 Hz?)

Thank you for this comment. We clarify that the curated waveform dataset used here was produced following the preprocessing described in Ni et al. (2023), rather than being newly curated in this study. In that dataset, all waveforms are resampled to 100 Hz to support deep neural networks with fixed input sizes; this includes upsampling BH? channels that are commonly sampled at 40 Hz. Because our analyses focus on frequencies ≤15–20 Hz, the 100 Hz sampling rate is not required for bandwidth reasons, but is used to standardize the discrete representation (consistent sample counts, batching, and spectrogram computation) across instruments and channel types. Ni et al. (2023) also address missing components by filling absent channels (e.g., vertical-only stations such as EHZ) with zeros to maintain a consistent three-component input format. We have revised the manuscript text to (i) explicitly attribute these preprocessing steps to Ni et al. (2023), (ii) state that BH? channels are upsampled as part of this standardization, and (iii) briefly discuss that zero-filled channels contain no signal and primarily serve input-shape consistency.

Ni, Yiyu, Alexander Hutko, Francesca Skene, Marine Denolle, Stephen Malone, Paul Bodin, Renate Hartog, and Amy Wright. "Curated Pacific Northwest AI-ready seismic dataset." (2023).

· Line 142: "..earlier events.." Events before 2014? (please be specific)

Corrected

· Line 143: "magnitude ranges are similar to those from explosions." Ambiguous.. and concerning. I would be surprised that there are explosion magnitude higher than magnitude ~2.5 for mining and volcanic explosion sources. Please clarify.

Thanks for catching this. The magnitude of earthquakes range from -1 to 5 with most of these events having magnitudes from 1 to 3. We have added the actual magnitude range in the text to clear any confusion

· Line 148-149: Does the duration correlate with the magnitude? And selecting only traces with both P and S wave arrivals for your chosen window also limits the epicentral distance and depth. This is imposing a bias in your data. How do you balance this to surface events that do not have phase picks? What was your reasoning for picking a long window length? You may see better performance with shorter windows.

Thank you for raising this important point. The waveform curation and windowing choices for the curated ComCat traces follow the procedure of Ni et al. (2023), who

selected only station–event pairs with both P- and S-phase picks to ensure high-quality, reliably labeled training data. We agree that this selection criterion can introduce a bias toward higher-SNR records and may preferentially include nearer events (and, by extension, a subset of depths/distances), since smaller or older events may lack S picks. We have added explicit text in the manuscript acknowledging this potential bias and clarifying that it is an inherent property of the curated dataset rather than a new filtering step introduced in this work.
With respect to waveform duration and magnitude, we note that apparent signal duration can correlate with magnitude, but in practice it is strongly confounded by distance- and SNR-dependent detectability and by the fixed extraction window, making a direct duration–magnitude relationship difficult to interpret robustly for low-magnitude events. We now clarify this limitation in the text.

Regarding balance with surface events: surface-event waveforms come from the analyst-curated "exotic event" catalog in Ni et al. (2023), which is inherently high-quality (analyst-selected) and often lacks distinct phase picks; these traces are curated using a separate alignment strategy and longer windows to accommodate emergent, long-duration shaking typical of surface processes.

Finally, our motivation for using long traces is twofold: (1) it follows the AI-ready dataset design choice of using long windows to provide flexibility for trimming and time shifting, and (2) it helps capture the longer time scales observed in surface-event signals. Importantly, we also evaluated multiple window lengths in our own experiments (from short windows to longer windows) and found that an intermediate window length (100 s) provided the best overall performance; we now emphasize this sensitivity test and its outcome in the Results/Methods to directly address the concern that shorter windows might perform better.

· Line 153-156: Interestingly low dominant frequency content. Explosions typically higher. The mixture of explosive types is a little concerning. Did you test how well correlated the waveforms are from these different explosive sources? Or were the features similar?
We thank the reviewer for this important point. In the curated catalog, the "explosion" class includes multiple sublabels (e.g., quarry blasts, confirmed explosions, probable explosions). In practice, however, the class is dominated by probable blasts (~99%), which are largely routine quarry/industrial single-shot chemical explosions, and we therefore treat the explosion label as representing this operational category.
We did not perform a dedicated waveform-correlation analysis across explosion subtypes in this study, and we now clarify this in the manuscript. Nevertheless, within the curated dataset the engineered feature distributions and the spectrogram-based

patterns for the explosion class are broadly consistent across events, and the observed lower-frequency importance likely reflects the dataset's specific mix of sources, distances/SNR, and our preprocessing band (e.g., 1–10 Hz vs. 0.5–15 Hz), rather than implying that explosions are universally lower-frequency than earthquakes. We have revised the text to (i) explicitly state the subtype composition of the explosion label and (ii) note that subtyping/correlation analysis across explosion categories is a worthwhile future extension.

- Line 165: 20 **seconds Corrected*
  - Do difference classes have different waveform lengths?

    Yes, different events may have different duration but the input window length to models are same for all classes.
- Line 167: "While the origin of these events…" origin time or location?

Both, Corrected

- Line 178–186: ESEC waveform definition, I would not include the website here. You mention "methods other than seismic" please clarify what methods. Additionally, this section is a confusing jumping to a benchmark dataset, maybe this should be a subsection. The detail is good, but please give similar or consistent detail to all datasets you describe in the manuscript for consistency.

We thank the reviewer for these suggestions. We agree that the ESEC benchmark was introduced too abruptly and that "methods other than seismic" was unclear. We have moved the ESEC description to a dedicated subsection (Section Exotic Seismic Event Catalog), removed the inline URL in favor of a formal citation to Bahavar et al. (2019) (with access date retained as appropriate), and clarified that "non-seismic methods" refers to independent observational evidence associated with ESEC entries (e.g., event-specific imagery/maps and supporting materials). We also standardized the level of detail across all dataset descriptions (event counts, window length, station selection radius, and SNR thresholds where applicable).

- Line 200-206: "The superiority of 3C data for DL…" I see where you're going but this is

vague and not really backing up this assertion.

We agree that the statement as written was too vague. We have revised the text to (i) explicitly frame the 3C observation as an empirical result from our preliminary experiments, (ii) clarify the physical motivation (availability of polarization and relative P/S energy information across components), and (iii) avoid overclaiming causality. We

also added a citation to prior work noting the value of multi-component information for discriminating between earthquakes and explosions.

· Lines 207–208: Common test dataset, "10,000 traces per class," then "From the 10,000 traces available for each class, we randomly split… 2,000 for testing" This is confusing, based on the bar graph in the supplement, 10,000 traces is more than some classes had available? Also, if you're training per-trace do you ensure that you are training and testing on data that come from separate and unique sources/events. (To avoid bias you don't want to be training on traces that come from the same event and test on traces that come from that same event but are just recorded on different stations).

We would like to clarify that these points are already stated in the manuscript. Specifically, we note that the surface-event class initially contains fewer than 10,000 three-component traces and is therefore supplemented with additional downloaded 3C recordings (Section 2.3) to enable balanced sampling. We also explicitly state that the train/validation/test split is performed at the **event level** (using event identifiers), so that no event contributes traces to more than one split (avoiding leakage across stations for the same event). To reduce any remaining ambiguity, we have slightly revised the wording in this paragraph to make the sequence (pool construction → event-level split → trace sampling) more explicit.

· Lines 285: Section 2.5 to 2.7.3 - Version 2 and 3 training expansions well described for your project/process but I'm not sure this level of detail is essential for the main body of the text. It can be distracting to the reader describing all the iterations and datasets. I recommend focusing these sections.

Thank you for this feedback. We agree that the original description of Versions 2 and 3 was overly detailed and could distract from the main narrative. We have condensed Sections 2.5–2.7.3 to focus on the key point: the final model was obtained through an iterative data-centric refinement process, where training data were expanded in response to systematic failure modes observed on external/near-field tests.

· Lines 292: Why not include more seismology-led features?

  o "We do not include other physics-based features used in the explosion P/S ratio (Kong et al., 2022) and various magnitude estimates (Koper et al., 2024) because these are not calculated for surface events." I understand that, but consider touching on limitations you expect this caused on model performance and capability?

  We have added a brief statement in the manuscript acknowledging this limitation and noting that our approach prioritizes cross-class consistency over class-specific features.

· Line 319: “?; ?” Formatting issues with your citations.

Corrected

· Line 331: Was any scaling done?

Yes. We applied z-score normalization prior to outlier removal, and then removed samples with any feature exceeding ±5σ. We clarified this in the revised manuscript.

· Line 343: Section 3: Architecture descriptions-

Corrected

· Line 364: This looks like your variables just copy-pasted, I suggest re-writing in a readable format.

Corrected

· Line 372: How did you handle balancing when many aforementioned stations had only vertical component?

Thank you for this comment. For the common 3C training/testing datasets, we restricted the analysis to stations that provide three-component recordings. Earthquakes, explosions, and noise in the curated dataset already contained a sufficient number of 3C traces for balanced sampling. The only class that was initially limited in 3C coverage was the surface-event class; to address this, we augmented surface events by downloading additional 3C waveforms from nearby stations as described in Section 2.3. As a result, all classes were balanced using three-component traces, and vertical-only stations did not constrain the final balanced dataset sizes.

· Line 379: “SNR greater than 1…” This is perhaps low for training? Did you pick SNR of 1 because of the small magnitude events or to include as much data as possible?

We used a permissive threshold (SNR >1) to retain small-magnitude and lower-SNR events and maximize diversity in the training set. A higher cutoff would bias the data toward larger, nearby events. This threshold serves as a minimal quality filter while maintaining broad coverage of recording conditions.

· Line 393: This needs a citation. I would suggest Yeck et al. (2020)

Citation added

§ https://doi.org/10.1785/0220200178

· Line 419: Repetitive. You already defined and described how you generated spectrograms earlier in the text.

Corrected

· Lines 422-437: I would add this detail to the figure of the architecture schematic (and move figure S5 to main text). Also, I don’t see a soft max

activation layer in your description or in your schematic/figures. Does your model output so that all of the predicted probabilities of each class adds to 1? (For a discrimination of class, the predicted probability needs to be > 0.25? Or do you define a prediction threshold?)

Thank you for this suggestion. In the revision, we moved the architecture schematic (previously Fig. S5) into the main text and merged it with the existing figure to create a new Figure~3. We also updated the schematic to explicitly include the final Softmax layer and added the key architectural details (filters, kernel sizes, pooling, dropout) in the figure for clarity.

Regarding outputs: the model produces logits that are passed through Softmax, so the predicted probabilities across the four classes sum to 1 for each input window. For curated/test evaluations, we assign the predicted label by the maximum-probability (argmax) class (i.e., we do not require a fixed threshold such as >0.25). Probability thresholds are used only in the continuous-data detection workflow, where probability traces are converted into discrete detections (Section~3.3.3).

· Line 448: Adam Optimizer needs citation (Kingma and Ba, 2014)

Citation added

· Line 461: Is this limitation due to the long windows (and short/small magnitude events)?

We did not isolate window length or small-magnitude events as the cause of the 1D limitations. Using the same data, 2D spectrogram models trained more stably, likely because time–frequency patterns are made explicit and easier for CNNs to learn than raw 1D waveforms.

· Section 3.3: Not sure if this fits here or could be either condensed or moved to the supplement. It's taking focus from the story of your paper.

Thank you for this feedback. We agree that Section 3.3 can become implementation-heavy if presented in full detail. However, we believe a brief deployment/workflow description is important to the paper's contribution because our goal is not only to benchmark models, but also to demonstrate practical pathways for using them in operational and research settings (integration with SeisBench, retrospective network testing, and continuous/cloud-ready scanning).

· Lines 469–507: Deployment workflows; Listing 1 code shows "# anotate the data" typo; should be "# annotate the data". However, I don't think this adds anything substantial to the paper. I suggest moving to the supplement.

Thank you for catching the typo—we corrected "# anotate" to "# annotate" in Listing 1. Regarding placement, we agree that extended implementation details can be moved out of the main narrative; however, we believe it is important to retain a minimal SeisBench usage example in the main text because SeisBench is widely adopted by the seismological community and is a primary pathway for practical reuse of our models. The short snippet demonstrates, in a concrete and reproducible way, how the trained classifiers interface with standard SeisBench workflows (e.g., `classify` and `annotate`) for both retrospective and near-real-time use.

· Line 494: "Section ??" Missing section.

Corrected

· Lines 500-507: How did performance on continuous differ from test datasets?

Thank you for this question. In this study, we present the continuous-data workflow to demonstrate the feasibility and reproducibility of deployment, but we do not report a quantitative performance comparison on continuous archives. A rigorous evaluation on continuous data would require additional steps that are outside the current scope—e.g., defining event-matching rules against authoritative catalogs, validating detection/association thresholds, and performing analyst review to characterize false positives/negatives. We have clarified this limitation in the text and explicitly framed the continuous-data section as a deployment demonstration rather than a benchmark evaluation.

· Lines 513-514: …"two CML models and two DL models.." please clarify which models for clarity.

Models are specified

· Lines 519-523: I suggest restructuring. This section is a little confusing.

Thank you for the suggestion. We agree the evaluation overview was difficult to follow in its original form. We have restructured this paragraph to present the evaluation as a clear stepwise sequence (curated benchmark → network-testing → external generalization datasets → iterative retraining), with the model selection and motivation for dataset expansion stated explicitly.

· Line 543: Section 4.2.1 – Reference the supplement?

We have added a reference to the supplement.

· Line 569: Stray character after "with waveform time series (1D) or spectrograms (2D) as input.r" Remove stray "r".

Corrected

· Lines 591–605: Network dataset performance summary; OK.

Corrected

· Line 609: "… distances greater than 60 km." What about the magnitude of these events?

Thank you for the suggestion. We did not include magnitude in this analysis because (i) many surface events do not have a consistent magnitude estimate, and (ii) for the curated earthquake/explosion subset, the magnitude range is relatively narrow (predominantly small events), so stratifying by magnitude is less informative than SNR.

- Line 626: …"weaknesses." Such as? Please give example.

Thank you for pointing this out. We have clarified what we mean by "new weaknesses" by adding an explicit example in the text. Specifically, we now state that Version~3 - while improving ESEC surface-event classification - showed increased confusion for near-field explosions (more explosion traces receiving higher surface-event probabilities and being misclassified at the trace level), illustrating the trade-off in generalization.

- Lines 634–635: "This underscores the importance …across global datasets." Could also be due to the surface and explosion events being sources that are exceptionally shallow.

Thank you for this suggestion. We agree that the residual confusion may not be solely due to signal quality, and could also reflect the fact that both surface events and many explosions are exceptionally shallow sources with similar propagation characteristics. We have revised the text to explicitly acknowledge this as a plausible contributing factor while retaining our main point about the importance of dataset diversity and global evaluation.

- Line 669: Line 375 says data was filtered 1-20Hz?

Here we are discussing the feature importance for M2 configuration, where the waveforms were filtered between 0.5-15 Hz

- Line 653: "5Feature Importance" heading missing space after number or consistent formatting.

Corrected

- Lines 671–679: Kurtosis distributions: In figure S10 I do observe that noise and the other classes are clearly distinguished but other classes overlap with each other for all kurtosis distributions. Is this a concern?

Thank you for the observation. Yes, Figure S10 (now Figure S9) shows that kurtosis features cleanly separate **noise** from event classes, while **earthquakes, explosions, and surface events partially overlap** in any single kurtosis distribution. This is expected because these classes can share similar impulsiveness and amplitude-distribution characteristics in certain frequency bands. Our classifier does not rely on any single kurtosis feature; instead, it combines many complementary attributes (including multiple kurtosis bands plus additional spectral/time-domain

and manual features). As shown by the cumulative feature analysis (Fig. S9, now figure S10), performance improves substantially when using the top 20 features, indicating that class separation is achieved through multivariate combinations rather than one-dimensional thresholds. We have revised the text to clarify this point.

· Line 675: Hour of Day most important feature- This is concerning. I would actually consider removing this feature since it is not a physical feature of the source mechanism of the event nor of the waveform. It's just a correlation with mining operation times.

Thank you for raising this point. We agree that "hour of day" is not a physical source or waveform attribute and largely reflects operational timing of quarry/mining explosions. We included it as an optional contextual feature for the PNW setting, but it does not drive the classifier by itself. In fact, HOD alone yields limited skill (F1 ~0.5 - Supplementary Figure S10), and Figure~4 in the main manuscript shows that adding Manual features (including HOD) provides only a **modest** improvement for explosions relative to Physical features alone, with little change in overall performance between Physical and Physical+Manual feature sets. We have revised the text to describe HOD explicitly as a region-specific contextual proxy, note its limited standalone utility, and emphasize that the model's performance primarily derives from waveform-based features.

· Line 685: "…removing highly correlated features?" which features were removed?

Thank you for the comment. We have clarified which features were removed due to high correlation and now provide the complete list of correlated feature pairs and the removal choice in a Supplementary table (Table~S2).

· Line 709: "… explosions exhibit significant importance in the lower frequency band of 1-5Hz" This is a little surprising since many studies have shown dominant frequency for local and regional explosions in the range of 6-8 hz and higher, especially in volcanic regions. This dominant frequency band does match Barama et al. (2023) who used 1-5Hz successfully for regional to teleseismic P waveform discrimination (CNN multiclass discrimination of explosions-earthquakes-noise). However, Kong et al. (2022) saw success at < 15Hz and 10-18Hz for local explosion discrimination at similar distances and magnitudes to this study.

§ https://doi.org/10.1029/2022GL101528

Thank you for this comment. We agree that our wording could be misinterpreted as a statement about the dominant spectral energy of explosions. In our analysis, the 1–5 Hz band refers to the model attribution/feature-importance concentration (Integrated Gradients in the time–frequency domain), i.e., the frequency range that most

influenced the classifier's decision for our dataset—not necessarily the peak frequency content of explosion signals. We have revised the text to make this distinction explicit and to avoid implying a universal dominant-frequency band. We also acknowledge that prior studies report discrimination power at higher frequencies (e.g., 6–8 Hz and up to ~10–18 Hz) depending on region, distance, and preprocessing (e.g., Kong et al., 2022). Our model's emphasis on lower frequencies likely reflects dataset- and processing-specific factors (e.g., attenuation with distance, SNR variability, and the band-limited inputs used here). We now cite these studies and clarify that the attribution patterns are context-dependent.

· Lines 710-711: "… extended duration of coda waves." Needs citation. Mining explosions have increased coda duration and decreased S generation, but this strong or long coda is more a feature of shallow depth -> an explosion screening tool for unusually shallow sources.

Thank you for this point. We agree that the statement about "extended coda duration" requires support and clearer framing. We have added citations and revised the wording to emphasize that the longer-duration coda we observe is consistent with shallow-source effects (e.g., enhanced surface-wave energy and prolonged scattering/propagation) rather than a universal property of mining explosions alone. This is also consistent with explosion screening approaches that leverage unusually shallow sources and reduced S-wave generation. We have updated the text accordingly and added appropriate references.

· Lines 722: "… inherent limitation of this label." I agree, and your results are pretty good considering these limitations. However, why not make another event label, some thing that's a signal but not an earthquake or explosion? Some induced seismicity ML projects are focused on this problem because the datasets are small and the events so variable.

Thank you for this thoughtful suggestion. We agree that a finer-grained labeling scheme (e.g., an intermediate "non-EQ/non-EXP signal" class or multiple SU subclasses) could be valuable, and related work in induced-seismicity settings has shown that label refinement can improve performance when events are heterogeneous. In this study, however, we intentionally used the operational PNSN labeling scheme to (i) remain consistent with the authoritative catalog used in practice and (ii) enable direct benchmarking and deployment without introducing additional analyst-dependent relabeling. Creating a new label would require substantial manual review and clear criteria to ensure consistency, and the resulting class would likely still mix multiple physical source processes and recording conditions, making it non-trivial to define robustly.

· Line 728: I mentioned above, but where your holdout datasets randomly selected, or selected across a variety of distances / magnitudes / spatial distributions?

Thank you for the question. The curated 3C holdout set was created by random event-level splitting (i.e., events were held out at random, and all associated traces for a given event were assigned to the same split to avoid leakage). We did not stratify the split by magnitude, distance, or geography; instead, these properties are inherited from the underlying catalog and are therefore represented naturally in the holdout set. The resulting spatial distributions for the holdout test dataset are shown in Supplementary Figure~S5

· Lines 731-745: This wording is confusing. Please clarify/focus this section.

Thank you for this comment. We agree that the original presentation was too dense. We have rewritten Lines 731–745 to (i) define trace-level vs event-level metrics and "high-confidence disagreements" once, (ii) present surface-event/explosion/earthquake results in a consistent structure, and (iii) separate the analyst-audit findings from the broader interpretation. This restructuring improves readability while preserving the quantitative results and conclusions.

· Line 738: "These findings imply that approximately 1.5-8.4%..." How did you calculate this percentage range? Do you think these mis-labels are significantly contributing to performance? If so, would cleaning the data with an earthquake discriminator/classifier aid performance?

Thank you for the question. The 1.5–8.4% range was a back-of-the-envelope estimate based on the fraction of surface-event labels flagged as high-confidence disagreements in the held-out set. Specifically, among 768 surface events, 65 were high-confidence mismatches at a single-station level (65/768 ≈ 8.4%), and 22 remained high-confidence after requiring agreement across more than one station (22/768 ≈ 2.9%). We reported these values as an approximate range, but we agree that this inference depends on the assumption that all flagged cases correspond to true catalog mislabels. To avoid over-interpreting, we have removed the percentage range in the revised manuscript and instead report the raw counts (number flagged; number confirmed by analyst review).

We do expect catalog mislabels to contribute to apparent performance limits, particularly for the surface-event class, because they introduce irreducible label noise in both training and evaluation. Cleaning the training data—e.g., using an ML-assisted screening step followed by analyst validation—would likely improve performance, although the magnitude of improvement would depend on the prevalence and types of label errors and on the presence of missing classes (e.g.,

DLF-like signals). We now frame this explicitly as a future data-curation direction rather than a quantified claim.

· Line 747: "… many of which are teleseismic…" Do you think that only selecting local earthquakes as part of your training data biases your data, and reduces capability of identifying earthquake phases greater than local distances?

Thank you for this question. Yes—training primarily on local PNSN earthquakes likely limits the model's ability to generalize to teleseismic waveforms, whose phase content and time–frequency characteristics can differ substantially. This is largely an inherited constraint of the curated dataset we use (Ni et al., 2023), which is dominated by local/regional events. Our intent here is an operational classifier for small, local events in the PNSN, where discrimination among earthquakes, explosions, surface events, and noise is most challenging; teleseismic events are typically identifiable with standard network/location-based procedures and were not a target use case.

· Line 775: "...complex problem.." and complex data.

Corrected

· Line 762: "…which is inherently easier…" Repeated from earlier in the text and again, not sure if that is an entirely accurate statement. I would argue data complexity/correlation in each class is what makes classifying "easier" or "more difficult"

Thank you for this comment. We agree that "easier" is an oversimplification and that difficulty depends strongly on class variability/overlap and data complexity. We have revised this sentence to avoid repetition and to clarify that binary tasks often report higher performance largely because they involve fewer classes and are frequently posed with more separable class pairs, but this is not universally true.

· Line 789: "… probability exceeds 0.15…" How did you determine this threshold?

Thank you for the question. We selected the 0.15 onset/offset threshold empirically using a development workflow on continuous/network-testing data, sweeping candidate thresholds and inspecting the trade-off between missed detections and false positives. The chosen value provided a practical balance: it captured the majority of visually evident events while limiting spurious detections from transient noise after smoothing. We have clarified in the text that this threshold is a heuristic tuned for our deployment setting rather than a theoretically derived constant.

· Line 791: Add parenthesis around link

Corrected

· Line 794: Can you further comment on detection/classification performance on the less emergent events?

Thank you for this question. We have not yet completed a quantitative evaluation of end-to-end detection performance in QuakeScope on continuous archives (including less emergent, phase-like events), because this would require additional catalog matching/association logic and analyst validation to characterize false positives/negatives, which is outside the scope of this study (as noted above). In this manuscript, our evidence for robustness to more emergent signals comes from the curated/network-testing and ESEC-style evaluations, where surface events often have emergent onsets and extended codas; the model maintains reasonable performance under these conditions using long windows (e.g., 100–270 s) and probability aggregation. We have revised the text to clarify that the QuakeScope integration is presented as a deployment demonstration rather than a fully benchmarked continuous-detection study.

· Line 830: "… time of day helped discriminate explosions from other classes." Makes sense but it's important to note too that this would not be helpful for non-mining operation sourced explosions. Like volcanic explosions, etc..

Thank you for this point. We agree that time-of-day is not a physical source attribute and is primarily informative for anthropogenic (e.g., quarry/mining) explosions with strong diurnal operating schedules. We have added a sentence clarifying that this feature is context-dependent and may not generalize to other explosion types such as volcanic explosions or to regions without diurnal blasting patterns.

· Line 855: Zenodo repository link looks like it only links to different versions of Random Forest model?

Thank you for catching this. You are correct: the referenced Zenodo record archives the trained Random Forest models and scaler parameters (Physical+Manual feature set) (and associated preprocessing artifacts) to enable reproducibility of the CML results, since feature extraction can be time-consuming. We have revised the Data and Code Availability section to clarify that this Zenodo deposit corresponds to the CML (RF) artifacts only, while the deep-learning model architectures and training/inference notebooks are provided in our GitHub repository.

· Lines 897–898, 974–975: Duplicate Pirot citations with 2023 and 2024 entries that are identical; remove duplication or correct.

Corrected

Reviewer #2

Overview In this study, the authors present a thorough investigation of various machine learning methods applied to seismic signal classification in the Pacific Northwest (PNW), including deep learning, and classical machine learning based on feature engineering and, for example, random forests. The classification problem divides signals into four groups; earthquakes, explosions, surface events (e.g., landslides), and noise. Overall, they find that convolutional neural networks (CNNs) trained on spectrograms perform the best, with classification accuracy of >92%. Furthermore, they present a model that can aid in near-real time event classification and be used operationally in seismic networks in the PNW (e.g., through integration into the SeisBench machine learning toolbox). To date, this is the most comprehensive investigation I have seen regarding machine learning-based seismic signal classification in the PNW and could be useful for signal classification in other regions. I have few substantive comments regarding manuscript improvement. My comments below are mostly related to presentation of results and/or pointing out minor typos. I think that after minor modifications, the manuscript will be suitable for publication in Seismica.

Comments

1. In the introduction, the authors point out the diversity of different earthquake signals (e.g., megathrust, intraslab, crustal, slow repeating earthquakes, low frequency), and volcanic tremor signals. It seems that the most useful regional PNW classifier would also discriminate between different classes of earthquakes caused by different mechanisms (e.g., tectonic vs volcanic). However, the models are trained on 4 classes, with "earthquake" presumably encompassing many distinct types of events. The inclusion of 4 classes is an improvement over most source discrimination models that only consider binary earthquake/blast classification, but I wonder if the model could be pushed further to discriminate between different types earthquake signals.
   Thank you for this thoughtful comment. We agree that, in principle, a more detailed regional classifier that distinguishes between different earthquake source processes (e.g., tectonic vs volcanic, low-frequency/tremor, intraslab vs crustal) would be valuable, especially in a geologically diverse region like the PNW. In this study, however, we intentionally framed the task around the operational PNSN taxonomy and the most common discrimination need for routine cataloging: separating earthquakes from explosions, surface events, and noise. The curated dataset we build on provides a robust and well-vetted set of these four labels, whereas subdividing the "earthquake" class into

mechanism-based subclasses would require additional labeling effort, clear and consistently applied criteria, and likely supplementary information (e.g., event locations relative to volcanic edifices, focal mechanisms, spectral characteristics) that are not uniformly available across all events. As a result, introducing earthquake subclasses within the scope of this paper would risk creating small, heterogeneous, and inconsistently labeled categories that could reduce reproducibility and complicate fair benchmarking.

2. Scalograms based on, for example the Morlet wavelet transform, can in some cases be better suited for characterizing diverse seismic signals (e.g., volcano/seismic signals) (see for example Lapins et al., 2020). It would be interesting to see if models trained on scalograms can yield improved performance. Lapins, S., Roman, D. C., Rougier, J., De Angelis, S., Cashman, K. V., & Kendall, J. M. (2020). An examination of the continuous wavelet transform for volcano-seismic spectral analysis. Journal of Volcanology and Geothermal Research, 389, 106728.
   Thank you for this suggestion and for pointing us to Lapins et al. (2020). We agree that Morlet-wavelet scalograms can be a useful representation for complex and diverse volcano-seismic signals. In our study, we partially explored this direction through the ScatNet (wavelet scattering) feature set, which is explicitly derived from Morlet wavelets and closely related to the continuous wavelet transform (CWT): the scattering network computes stable, multi-scale coefficients from successive wavelet modulus operations and time-averaging, providing a CWT-based time–frequency representation that is robust to small time shifts. We have clarified this connection in the manuscript and added Lapins et al. (2020) as a relevant reference.

3. Lines 34/35. (REF) appears to have been used as a placeholder for references but not removed.
   Corrected
4. Line 319: There are two missing references, shown as question marks (Seydoux et al., 2020; Moreau et al., 2022; ?; ?; Steinmann et al., 2023). It looks like perhaps there are some bibtex entries missing the in the bibliography file.
   Corrected
5. Line 303: I don't think that "TSFEL" is a common enough acronym to not define it. Even if the TSFEL python package has widespread use, it would be helpful to explain more details. I had never heard of it.
   Thank you for pointing this out. We agree that TSFEL is not a standard acronym in seismology. We have revised the text to define TSFEL on first use and briefly describe what it provides. Specifically, we now spell out Time Series Feature Extraction Library (TSFEL), note that it is a Python library that computes a large set of standardized time-, frequency-, and wavelet-domain descriptors, and clarify which TSFEL domains we used

(statistical/temporal/spectral; excluding the fractal domain). We also retain the primary TSFEL reference (Barandas et al., 2020) and citations to prior seismological applications.

6. Line 314: "We do not include other physics-based features used in the explosion P/S ratio (Kong et al., 2022) and various magnitude estimates (Koper et al., 2024) because these are not calculated for surface events." Although I could see how ML – MC could be tricky to implement for the variety of events you examine, I don't follow why a P/S ratio couldn't be calculated in a straightforward way. Even for body wave signals that are emergent, or seemingly absent, the P/S ratio can be calculated by windowing based on expected P and S arrival times. It could be useful to include given how commonly it is used as a discrimination metric.
   Thank you for this suggestion. We agree that a P/S ratio can often be computed in a straightforward way when reliable phase windows can be defined. In our setting, however, implementing a consistent P/S ratio across *all four classes* (especially surface events and emergent/phase-poor signals) would require introducing additional assumptions: predicting or picking P and S arrival times, choosing window lengths, and handling cases where S is weak/absent or where energy is dominated by surface waves. These choices can become a major source of variability and would introduce a phase-dependent feature that is not uniformly defined for the surface-event class. Instead, we intentionally used phase-agnostic features that capture similar information in a more general way, including multiple frequency-band energy partitions (e.g., quartile band energies and centroids), time–frequency summaries from the spectrogram, and envelope-based timing/shape measures. Together these features provide proxies for relative early/late energy and frequency-dependent radiation without requiring explicit phase windowing. We have revised the text to clarify that our feature set contains several phase-agnostic proxies for P/S-type discrimination, and we note explicit P/S-ratio features as a promising extension for future work when consistent phase-window definitions (or robust pickers) are available.

7. Line 494: There is a section label missing, which again appears to be a LaTex issue. (Section ??).
   Corrected
8. In Section 4.4, the description of model performance on Generalization Datasets is fairly qualitative. For example, the authors mention that surface events and often misclassified as explosions (and vice versa), but did not provide the numbers, nor did they mention by how much additional training improved the misclassification of these events
   Thank you—this is a fair point. We have revised Section 4.4 to include quantitative station-level and event-level results on both generalization datasets and explicitly reference Figure 7, which summarizes how errors shift across training versions.