Detailed Review Report

General Impressions

The authors clearly propose that traditional machine learning models for seismic event detection often carry biases due to training on specific, limited catalogs. Their approach, utilizing pseudo-labeling based on pre-trained systems to enhance model generalization, is compelling and well-motivated. The manuscript presents a highly relevant and interesting investigation into improving seismic-volcanic catalogs through weakly supervised machine learning techniques. After multiple rounds of review, significant improvements have been made. However, readability and methodological clarity remain core concerns. Below, I outline detailed recommendations to address these issues constructively.

- About Methodology Section.

- Section 3.1 (Methodology Clarity):

The manuscript currently states that the proposed method aligns with the open-set domain adaptation paradigm, explicitly designed to handle novel event categories. However, the authors subsequently note a significant limitation: the method only labels events within categories already present in the master database. This limitation appears to directly contradict the previously stated open-set capability. I recommend clarifying this contradiction explicitly. The authors should specify whether the approach is truly open-set (capable of detecting and handling unseen seismic categories) or acknowledge clearly that it is currently limited to closed-set scenarios.

Although the assumptions clearly indicate that label spaces may only partially overlap, and thus novel categories could be present, the authors later explicitly state their methodology can only label categories that exist in the master database. Therefore, the authors should clarify how their approach practically handles (or does not handle) the novel categories mentioned in their assumptions. If the method currently doesn't handle these novel categories, explicitly stating this limitation and distinguishing clearly between the theoretical scenario and the actual method implementation would strengthen the manuscript.

- Main issues in the Methodology (Experiment 3.2.1):

1) Insufficient methodological details to reproduce the experiment
    - Problem: Currently, the authors only briefly mention:
        - Three model architectures (RNN-LSTM, Dilated-LSTM, TCN),

- Pre-training on **MASTER-DEC**, then re-training with **POPO2002**

However, readers might ask:

- How were the models pre-trained initially (hyperparameters, training set sizes, epochs, loss functions, etc.)?
- What specific transfer learning strategies were applied (e.g., layer freezing, fine-tuning, learning rate adjustments)?
- How exactly were data split (train-validation-test)?
- What were the evaluation metrics or validation methods?
- Did authors address class imbalance or category distribution?

Without these details, readers cannot reproduce the experiments. It seems that some important information about this is in section 4 (results). We suggest to the authors that change the text to the methodology section.

2) Confusion caused by two alternatives, stating:

- Problem: The authors propose two alternatives, stating:
  - **Option A**: Only use the **5 categories** in common with the MASTER-DEC catalog.
  - **Option B**: Adapt the model output to accommodate all **7 categories** present in POPO2002 by updating the output layer only.

But, critically:

- Authors state vaguely: "these two approaches have no major implications from a ML perspective".
- This statement is confusing because, practically, these two options have **very different implications**:

Option A completely excludes new categories, simplifying the task significantly. **Option B** involves at least minor model changes (output layer modification), and crucially, implies retraining with novel data categories (a clearly significant ML implication). This confusion significantly weakens methodological clarity.

- Second Experiment (3.2.2):

Clearly distinguish the novelty of Experiment 2 by explicitly stating upfront that the primary difference from Experiment 1 is the source of labels (pseudo-labels rather than true annotations). Emphasizing this difference early in the description would enhance readability.

- **About Results Section.**
  - **Lines 334-357:** This methodological detail is beneficial and should be moved explicitly into the methodology section for clarity and improved reproducibility.

- Line 362: The statement regarding "two experiments" conducted at this stage is confusing and should be simplified in the methodology section.

- Should the reader be beneficed with more transfer learning details? (e.g., fine-tuning strategies, freezing layers explicitly, loss functions and training epochs?).

- First Experiment Results:

While the overall self-consistency result (e.g., 77.38% accuracy for the RNN-LSTM model) provides a general sense of model performance, the confusion matrix reveals important class-specific differences—most notably, the relatively low recall for VT events (0.51) compared to much higher values for noise (0.97) and other event types. This suggests that the model may be biased toward the dominant class (likely noise), potentially inflating the global performance metric. I recommend that the authors include additional evaluation metrics, such as precision, recall, and F1-score for each class, as well as macro-averaged or balanced accuracy scores. These would provide a more nuanced understanding of how well the model generalizes across all event types, especially the underrepresented or more challenging classes like VT. Including this information would strengthen the assessment of the model's real-world applicability in diverse seismic scenarios.

- Second Experiment Results:

While the weakly supervised fine-tuning improved global accuracy, the model's ability to detect meaningful seismic events—especially VT and LP types—remains limited, with VT nearly absent in the confusion matrix. The dominance of the noise class likely inflates the global metric. Additionally, the model's detection rate far exceeds the label count, which may reflect over-sensitivity rather than true discovery. More rigorous evaluation, including precision-recall analysis, event-level validation, or expert review of excess detections, would strengthen confidence in the weak supervision pipeline.

- Third Experiment Results:

The use case of applying weakly supervised models during a pre-eruptive crisis is compelling and highlights the practical value of such approaches. However, the presentation of results—particularly the so-called "recognition results" table—is unclear. It is not evident whether the numbers reflect validated detections, raw counts, or comparisons to any ground truth. The sudden introduction of PhaseNet, while relevant, is also only partially integrated, with no evaluation metrics provided to contextualize its outputs or compare them to the proposed models. A more transparent and consistent presentation of results, including quantitative

comparisons, ground truth validation, and clearer labeling of what each table or number represents, would greatly improve the interpretability and impact of this section.
While the discussion highlights VT confusion rates exceeding 60% in some cases, this appears to reference only the worst-performing model (Dilated-LSTM). The other models achieve higher recall (e.g., 59% for TCN), and the average across all three models is closer to 47%, not 40%. A more balanced summary would acknowledge this range to accurately reflect performance variability across architectures.

Summary about results:

Throughout the results and discussion sections, the manuscript refers to "confusion matrices" and reports numerical values (e.g., 0.51, 0.31, 0.59 for VT events across models) without clearly stating whether these represent recall or confusion rates. However, the structure of the matrices—particularly the fact that each row sums to one—strongly suggests that the values correspond to **per-class recall**, i.e., the proportion of correctly classified instances for each true class. This is the standard interpretation for row-normalized confusion matrices in the machine learning literature. The ambiguity around this point makes the discussion difficult to follow and may contribute to the impression of poor presentation. For instance, the statement that "confusion rates exceed 60%" appears to refer to only the worst-performing model and does not align with the higher recall values seen in other models unless the reader assumes a confusion rate = 1 - recall. For the sake of clarity and consistency, it is essential that the manuscript explicitly define how these matrices are computed and what the reported values represent. This will not only improve readability but also help readers interpret the results accurately.

While the qualitative example shown in Figure 4 is compelling and suggests the model is capable of discovering events missed during the initial labeling, these anecdotal demonstrations are not sufficient to validate the effectiveness of the weakly supervised system. To move beyond suggestive visuals and convincingly argue for the scientific value of these new detections, the study would benefit from a more rigorous validation approach—such as expert review, waveform similarity analysis, or cross-comparison with independent models like PhaseNet. Without such steps, the claim that these new detections are not false positives remains speculative and limits the broader impact of the proposed method.

- About the Discussion Section:
    - **Line 416:** Verify if percentages presented in the discussion exactly match the results section; discrepancies would confuse readers.

While the qualitative example shown in Figure 4 is compelling and suggests the model is capable of discovering events missed during the initial labeling, these anecdotal demonstrations are not sufficient to validate the effectiveness of the weakly supervised system. To move beyond suggestive visuals and convincingly argue for the scientific value of these new detections, the

study would benefit from a more rigorous validation approach—such as: expert review, waveform similarity analysis, or cross-comparison with independent models like PhaseNet (we'll talk about this later). Without such steps, the claim that these new detections are not false positives remains speculative and limits the broader impact of the proposed method.

The discussion attributes the weak performance of the model on volcano-tectonic events (VTEs) to discrepancies in labeling criteria, subjective annotation boundaries, and prototype mismatches. While labeling inconsistency is a known challenge in volcano seismology, VTEs are typically among the most well-defined and reliably detectable seismic signals due to their impulsive, high-frequency nature. Numerous existing models (e.g., PhaseNet) have shown robust detection of such events across different volcanoes. The fact that the system recovers only 5% of annotated VTEs suggests that the problem may lie more in the modeling strategy or prototype selection than in catalog inconsistency alone. A more balanced discussion should consider whether the weak supervision framework fails to generalize to realistic variability within VTEs and whether model or prototype refinement could improve performance.

The comparison with PhaseNet in the third experiment raises concerns regarding methodology. The authors assess PhaseNet's performance by comparing the number of detected phases across different score thresholds, arguing that only detections above 0.8 correspond well with the labeled dataset. However, this approach overlooks the fact that many valid seismic picks—especially low-amplitude or emergent phases—often have lower phase scores (e.g., 0.3–0.6), yet still align with cataloged arrivals. Furthermore, raw pick counts do not constitute a meaningful evaluation metric unless aligned with ground truth picks using a timing tolerance. To make a valid comparison, the authors should report precision, recall, and pick timing accuracy against the labeled dataset across multiple thresholds. Without this, the argument that PhaseNet underperforms is not well supported and may misrepresent the model's actual capabilities.

- About Summary of Findings Section:

  - Figure 10: Clearly label differences between rows 3 and 4.
  - Ensure consistent PSD plotting style across Figures 10, 11, and 12 for clarity.

Other (Very) Minor Remarks:

- Abstract:
  - **Lines 2 & 4:** avoid unnecessary repetition of word "however" within the same paragraph; consider synonyms or rephrasing to improve readability.

- Introduction:
  - **Line 37:** a space is missing, "..crises.However"
  - **Line 51:** there is an extra space; "Canario et al., 2020 ;"
  - **Lines 57-63:** suggestion: another challenge is that upgrades and updates to seismic instrumentation over decades complicate the review of historical seismicity, as the digital signals may not share a consistent framework.
  - **Lines 56, 66, 71, 95, etc.:** There are inconsistencies in citation formatting throughout the manuscript. Please ensure that references within parentheses follow the standard format, e.g., "(Weiss et al., 2016)2, rather than "(Weiss et al. (2016))".
  - **Lines 168:** space missing at "MASTER-DEC(1-50HZ)"

- Methodology and experimental framework:
  - **Line 247:** repetitive vocabulary again (stream).

- Discussion.
  - **Lines 445-446:** review grammar of "According to such table, on average, only 5% of the analysis windows labeled as VTE in the original catalog were recognized by the retrained systems."
  - **Line 473:** please check the text "(Fig. 7)a."