Reviewer#8:

Dear reviewer#8, We are very thankful for your thoughtful suggestions. Below, we present how we have addressed them.

General Comments:

The manuscript presents a highly relevant approach that combines machine learning with weakly supervised methods for seismic-volcanic event detection. The application of these techniques to geophysical event detection is an exciting and promising field of study, and I commend the authors for their effort in tackling such a complex problem. The subject matter is particularly valuable given the growing interest in leveraging machine learning models for natural hazard monitoring, and the use of weak supervision opens new possibilities for working with limited labeled data, a common challenge in seismology and volcanology.

However, while the approach is interesting, the manuscript, in its current form, requires substantial rewriting to improve clarity, structure, and the strength of its arguments. There are several critical issues that need to be addressed before the manuscript can be considered for publication:

1. Methodology Section Reconstruction: The methodology section lacks sufficient clarity and structure. Key concepts such as UMAP, the Leave-One-Out cross-validation method, and the iterative processes involved in the pseudo-labeling task are either insufficiently explained or poorly integrated into the overall narrative. The methodology needs to be rewritten to clearly define these elements and their role in the overall framework, ensuring that readers can follow the steps taken in the model development and evaluation process.

   We will aim to improve the wording of this section to streamline the reading and comprehension of the proposed ideas. To achieve this, we will include a schematic representation of the algorithm itself and define some of the mentioned concepts to make the work more self-contained.

2. Justification for Using a Single Dataset in Transfer Learning: The authors attempt to justify the use of a single dataset in their transfer learning approach, but the arguments presented are not convincing. As the authors themselves note, ‘it could change when using a different test dataset,’ suggesting that model performance may not generalize well to other geological settings. The authors need to make a stronger case for why the use of a single dataset is valid for this weakly supervised learning approach. Ideally, the manuscript should explore the potential limitations of this approach or, alternatively, incorporate multiple datasets from different volcanic settings to demonstrate broader applicability.

   We completely agree with the reviewer's suggestion. The only reason our work focuses on a single volcano in TF is that we do not have access to other available seismic catalogs and data from which to build a comparative base. As we have extensively mentioned to the majority of

reviewers, we are fully open to testing our system with different databases and volcanoes of varying nature. We invite the reviewers to join this initiative and collaborate on a project that could result in a universally applicable work. Otherwise, we kindly ask for guidance on where we can obtain reliable seismic data and catalogs for further experimentation.

3. Overall Structure and Writing Quality: The manuscript, though scientifically significant, suffers from poor structure and unclear writing, which detracts from its scientific contributions; this has resulted in several instances where key ideas are poorly expressed or ambiguously presented. A thorough revision of the manuscript is needed to ensure that the concepts and findings are communicated effectively. I suggest the authors consider restructuring the entire manuscript to enhance readability, focusing particularly on tightening the introduction, improving transitions between sections, and making the arguments in the discussion more robust.

   We will work on improving the structure and writing of the article to meet the expectations of the several reviewers who have suggested this.

In conclusion, while the study introduces an interesting and timely approach to seismic-volcanic event detection using machine learning, the manuscript requires significant rewriting to better articulate its methodology and address critical gaps in the explanation of its approach. I recommend a major revision to enhance clarity, strengthen the justification for key methodological choices, and improve the overall presentation of the research.

Specific comments & Technical corrections:

We will address the key issues raised by the reviewer. The remaining suggestions will be implemented without further discussion in order to expedite the review process, as most of them are technical and grammatical corrections.

- 1. Introduction.

  line 50:"Bayesian" misspelled
  line 99: ¿references for master dataset?
  lines 99-100: "*has already been successfully applied in different DL architectures*"; ¿references?
  line 102: references for the Popo dataset?, and ¿why it is of high quality?
  line 105: It would be very useful to provide more information about the volcanic dynamics observed in the proposed datasets, especially as machine learning developments and methodologies are evolving to incorporate physics-based input.

- 2. Seismic data and catalogues.

  line 125: "..on the applicationof HMM models, etc."; ¿references?
  line 130: “*While it is true that not all types of signals are present in this 'Master database’, especially those associated with ongoing eruptive processes.*”, so,

perhaps it would be important to have a master dataset that includes this information as well. It is crucial to incorporate datasets representing different stages of volcanic unrest and to clarify which specific stages the machine learning models are most useful for.

line 145: A more detailed description of Popocatépetl's volcanic activity is needed, including its cyclical behavior of effusive activity, dome formation followed by explosive events, tremor signals, and other relevant features.

line 148: Are there any references available for this group of geophysicists or their work?

Table 1: nice.

Data & sensors: It would be ideal to provide a clearer explanation of the types of instruments being used, including whether all components are available, sampling, etc., as well as details on the sensors. For example, are all instruments capable of measuring all types of events in both datasets? Nowadays, seismic networks are densified with a combination of broadband and short-period sensors, which may influence data quality, coverage, and distance to volcanic sources. The proximity of sensors to the volcanic source is critical, as it directly affects the resolution and accuracy of the recorded data.

In the new version of the manuscript, all these suggestions will be addressed to improve the readability and understanding of the work

- 3. Methodology.

lines 234 - 246: about marginal and conditional distributions: a need for clarity:

The authors' explanation regarding the assumptions of marginal and conditional distributions in the pseudo-labeling task could benefit from greater clarity. Specifically, they state that the marginal distributions of the source and target domains are assumed to be the same (P_s (X_s) = P_t (X_t)), maybe implying that the input features (seismic windows) in both domains are similarly distributed? However, they also assume that the conditional distributions of the source and target domains are the same (Q_s (Y_s | X_s) = Q_t (Y_t | X_t)), suggesting that the relationship between input features and event types is identical across both datasets.

Key Challenge and, Potential Problem?:

The text acknowledges that while the marginal distributions of the input features may be the same, the conditional distributions might differ between the source and target domains. This introduces a key challenge: even though seismic signals may "look similar" across different datasets (i.e., the marginal distributions are similar), the relationship between these signals and the seismic events they represent (i.e., the conditional distribution) may vary.

This discrepancy can create a potential problem when using pseudo-labeling and transfer learning techniques. If the model is trained assuming that the conditional distributions are the same, it may misclassify events in the target domain, especially if the seismic signatures there correspond to different types of events than in the source domain. This issue could result in reduced accuracy and reliability of event detection in the target domain, undermining the effectiveness of the model's generalization.

Conclusion: The Need for Diverse Datasets

This challenge is crucial because it highlights a potential flaw in the transfer learning approach: the assumption that conditional distributions are the same across different volcanic settings may not always hold. To address this, it may be necessary to collect and incorporate datasets from a wider range of volcanic regions, where the relationships between seismic features and event types can vary. Doing so would enable the development of more robust models that can better generalize across domains, improving the accuracy and reliability of event detection in different geological contexts. This would strengthen the use of transfer learning techniques and ensure that models are more adaptable to varying volcanic behaviors.

We fully agree with the reviewer's comment, which is why we once again encourage the reviewers to join this initiative so that we can conduct a study that includes a wide range of volcanoes and catalogs. This would help reduce the gap between domain distributions and bring us closer to a more universal model.

Figure 1: bad quality figure in the PDF file. Do steps A, B, etc., correspond to the actual process in your proposed methodology? line 276: reference missing.

In the new version of the manuscript, this suggestion will be addressed to improve the quality of the image.

- 4. Results.

  line 291: review grammar ("..using as training..")
  line 302: The text on self-consistency should be explained and included in the methodology section ('*We apply the Leave-One-Out cross-validation method*').
  line 329: Should Section 4.3 be renumbered as Section 4.2?
  line 343: These iterations need to be clearly specified in the methodology section, as you mention the goal of "*until a reliable catalog is achieved*".
  Line 344: The authors mention that "*however, it could change when using a different test dataset*" which highlights an important point regarding model generalization. While their approach is based on a single dataset, this raises questions about its robustness across varying geological settings. To truly validate the effectiveness of the model, it would be crucial to demonstrate its performance using multiple datasets from different volcanic environments. By doing so, they could provide stronger evidence that the model can generalize across diverse conditions, rather than being tailored to a specific dataset. The authors need to convincingly argue why relying on a single dataset is sufficient, or alternatively, why incorporating multiple datasets might be necessary for ensuring broader applicability.

In the new version of the manuscript, all these suggestions will be addressed to improve the readability and understanding of the work

- 5. Discussion.

    line 357: It would be helpful to clarify the phrase “when effectively use” throughout the text to strengthen the main arguments. Perhaps the grammar could be reviewed in that sentence.
    line 365: Should Fig. 1 be renumbered as Fig. 2?
    2: UMAP should be introduced in the methodology section and connected to the general objectives.

In the new version of the manuscript, all these suggestions will be addressed to improve the readability and understanding of the work.