Reviewer#3:

Dear reviewer#3, We sincerely appreciate your detailed comments, which have been very helpful. Below, we address your suggestions one by one.

This paper deals with the automatic classification of seismic signals in volcanic environment. The authors suggest that weakly supervised machine learning approaches can be used to improve the detection and classification of signals, in comparison to direct transfer learning methods. Although the subject is very interesting, I agree with previous reviewers that the work must be improved before being considered for publication.

We appreciate the reviewer's comment as it allows us to clarify the general idea of our work, which we may not have conveyed effectively. Our work does not propose that weakly supervised machine learning approaches are superior to direct transfer learning methods for improving the detection and classification of signals. Instead, we argue that an automatic recognition system (whether trained from scratch or using a transfer learning approach) with a high recognition rate may still be biased. This means that it might have learned the information in the training catalog very well, without showing signs of overfitting, but could still be limited in scope.

To demonstrate this, we designed an experiment in the context of transfer learning (TF). We trained a system initially on Deception data, then retrained it using data from Popocatépetl (POPO). The recognition rate exceeded 85% at the frame level (applying a grammar would further improve the results). However, when we applied a weakly supervised algorithm, we observed that while the system accurately recognized cataloged events, it failed to detect many other events present in the seismic traces but not recorded in the catalog.

Therefore, our point is not that weakly supervised learning is better or worse than TF for detecting or classifying seismo-volcanic events. Rather, we suggest that a weakly supervised approach, applied to systems trained on less biased catalogs, can be a robust solution for addressing the issue of catalog completeness.

1) As stated in RC2, the manuscript lacks a description and a discussion of the phenomena to put the results in the perspective of volcanic monitoring and eruption forecasting. In its present form, you focus on signal classification, and not on the understanding of "volcano dynamics" as stated in the conclusion (l.455). A more detailed description of the data acquisition and catalog construction methodology is also missing.

First of all, we would like to emphasize that our work focuses on recognition (detection + classification), not just signal classification. This is crucial because classification inherently involves a process of isolating and segmenting potential events that will later be classified. This process is very expensive, which is why we operate under the premise that existing catalogs, although highly reliable, are incomplete.

Secondly, when we mention that a more complete catalog can improve the understanding of volcanic dynamics, we are referring to the idea that having a more detailed view of the frequency and occurrence of events during different volcanic phases (eruptive, pre-eruptive) allows observatories to enhance their knowledge of the volcanic dynamics associated with their volcano. As is well known, volcanological observatories count seismic events during risk assessments and compare behaviors with previous eruptions. By increasing the quality and completeness of seismic catalogs, this can help deepen the understanding of the volcano's behavior and apply this knowledge during future crises.

2) In the introduction, the authors mainly rely on their own publications for transfer learning approaches (e.g. "Based on our experience" l.86), but are they really the only team working on transfer learning methods for seismic signal classification? It is not clear either the extent to which this paper is novel compared to previous works on the same subject by the authors research team (citations l.86), or to other applications / studies in the litterature. The litterature review in the introduction is also, in my opinion, lacking key elements, in particular regarding existing fully unsupervised approaches. They have proven effecetive in volcanic context (e.g. Steinmann et al. (2024). Machine learning analysis of seismograms reveals a continuous plumbing system evolution beneath the Klyuchevskoy volcano in Kamchatka, Russia. JGR: SOlid Earth, 10.1029/2023JB027167). What are the limitations of fully unsupervised machine learning? How is your approach complementary? Similarly, you only refer to recent publications on early-warning systems based on seismic monitoring (Rey-Devesa et al. (2023) : you must be more explicit on the different approaches, and use more references.

In the field of automatic seismo-volcanic signal recognition, there are other colleagues making significant contributions. However, when we refer to "our previous experience," we are not suggesting that only our group works in this area. What we want to highlight is that our group has extensive experience in this field, and thus, the problem addressed is well-established and thoroughly tested.

Regarding the novelty of our work compared to others in the literature, we can structure the response into two main blocks. The first, less technical and more methodological, focuses on highlighting the issue that while AI systems can effectively learn from the information contained in catalogs, this does not necessarily mean that the system is correctly classifying all the information within the seismic trace. This issue is what we refer to as bias. The system simply adjusts to the information learned from the catalog. Thus, the more descriptive and complete the catalog is, the less bias the system will have. However, creating less biased and more complete catalogs is a very costly and complex task. Therefore, a methodology is needed to facilitate this process. This leads us to the second block, which is more technical and less methodological, where we propose using a system trained with a defined and minimally biased catalog to

obtain reliable preliminary catalogs, which, with minimal effort, help improve the construction of less biased systems.

Finally, we would like to mention completely unsupervised approaches. These approaches have been successful in various fields, and automatic volcanic signal recognition is no exception. However, it is important to highlight that completely unsupervised approaches would assist us in several ways: 1) Clustering data into clusters or categories based on similarities. This would help with the task of clustering different windows and, consequently, events. However, this clustering would inherently require subsequent expert evaluation of each cluster, which essentially equates to supervised learning. 2) Exploring data and discovering effective underlying features that aid in the detection and classification process. In our case, feature engineering based on filter banks, derived from expert knowledge and experience, helps us avoid developing a preliminary stage to extract features with unsupervised systems and then use them in a supervised training process with catalogs as labels. 3) The use of unsupervised techniques deprives us of the expert knowledge contained within the catalogs themselves, which is crucial in a complex problem where signals are influenced by various external factors. This knowledge is vital and we believe it cannot be discarded from the training process. In summary, while unsupervised techniques can be complementary, our work relies on techniques that incorporate human expert knowledge as a foundational element, which can be transferred across volcanoes.

3) The description of the catalog must be improved. As suggested by other reviewers, you need to explain more clearly what the different classes of seismic events correspond to, both in terms of physical processes and features used for classification. Their names must be homogenized (e.g. use only GAR or BGN for background noise), and you should show in a single figure / table how many events of each class there are in the two catalogs. As this is not done, it is sometimes difficult to understand your results (e.g. what are the 5 and 7 seismic categories used in Table 2)? In the same perspective, the features used for classification must be given, as well as the methodology used to compute them. It is also not clear to me if the catalogs associate labels to successive and constant time windows on the full signal, or on time windows of various lengths, defined manually, and corresponding to specific events.

We will aim to improve the description of the catalogs and the presentation of the results to enhance their clarity. In the preprint submitted, we included the event names (categories) and the number of events exactly as they appear in the original catalogs. Regarding the physical description of the different events, in this work, we have followed the description proposed by Ibañez et al. (2000), where events are categorized based on their waveform and spectral content, as the catalogs reference the events using this categorization. However, in Table 1, we included other approaches for comparison purposes and to aid in comprehension. As we previously mentioned, in the new version of the manuscript, we will try to include an image that shows both the waveform and the spectrogram of each signal. This was not done in the first version because the

papers describing each event were cited, and we considered this redundant information.

Regarding the meaning of the 5 or 7 seismic categories in Table 2, this information is related to the concept of Transfer Learning, which might not have been clearly explained in the manuscript. We will now try to clarify its meaning in simple terms. The aim of the article, as mentioned, is to provide an effective and straightforward solution to the problem of obtaining robust catalogs. To explain the challenges associated with the bias in the catalogs used to develop robust models, we rely on the concept of Transfer Learning. Essentially, this concept describes the idea of using knowledge acquired by a system in one domain and applying it to another, related or unrelated, domain. Therefore, the model trained on a minimally biased catalog (master) comprising 5 seismic categories serves as a starting point to train another model in a different volcanic environment with the same or different seismic categories. In this context, we have two alternatives for training the new model: 1) keep the number of seismic categories from the source domain in the target domain or 2) change the number of seismic categories in the target domain (in our case, with neural networks, this involves changing the number of outputs in the output layer).

Since the seismic categories of the POPO volcano correspond to those of Decepción (with the exception that some are divided into subcategories, such as the TRE event, which is divided into 3 subtypes) and some additional categories are considered, we designed two types of experiments: 1) Group all event subtypes into one and maintain the number of seismic categories from Decepción in the new model, ignoring the events present in POPO but not in Decepción (5 categories), and 2) Group all event subtypes into one and change the output layer of the new model, adding the events that are present in POPO but not in Decepción (7 categories). Therefore, the results shown in Table 2 reflect the recognition and adaptation percentages of the original POPO catalog when applying a Transfer Learning approach from the model trained with Decepción, considering 5 and 7 seismic categories.

Regarding the use of labels during training, we applied a windowing approach to facilitate capturing the evolution of signal information. For this, we simply used the information contained in the catalog and associated the label of each window with that information. Suppose that in the catalog, a signal is composed of 3 different event types, all of different durations: BGN-VTE-BGN. Our approach would window the entire signal into segments of a given duration and overlap. Since the catalog contains the start and end information for each event, we can associate the label of each window with the event type annotated in the catalog, as we know the start and end times for each window. Finally, each window is parameterized using a filter bank on a logarithmic scale. This allows the models to be trained with signals of varying durations that contain very different information.

4) The methodology of the weakly supervised learning must be more clearly explained, at least in an Appendix. It is not clear how the assumptions stated l. 236 to 241are important, and how the results can be interpreted if they are not verified (l.242-243 -> are the marginal distributions indeed the same? l.247-249: I don't understand the logical link suggested by "therefore", between the assumptions and the possibility to use weakly suppervised learning). Figure 1 must also be improved. In particular, the iterative refinement process is not displayed. Following remark 3), it is also not clear in the Figure what the signal in B) corresponds to : a portion of the signal identified manually in the catalog, or continuous data? For the same reason, it is not clear to me what the "dataset" mentionned in the text and in D) corresponds to. You should also explain how you define the threshold used for the drift adaptation method (l.266), and how you choose to stop the iterative refinement (what is the "desired result" l.270?). More generally, there are many terms that are technical and could be clarified for non experts readers (e.g. "self-consistency" (equivalent to accuracy?), "softmax", "argmax softmax", "confusion", accuracy" ...).

We will attempt to more clearly and thoroughly explain the proposed methodology, either within the manuscript itself or in an appendix. Regarding Figure 1, we will improve its description. However, we would like to clarify the reviewer's concerns by briefly explaining the figure: We start with a model trained in a given volcanic environment, Deception, and aim to obtain a pseudo-catalog of events for a volcano in a different environment, POPO. Therefore, the signals in the POPO dataset are analyzed using the model trained on Deception. The events recognized in POPO by the model trained on Deception, with a user-defined membership probability, will be included in the new database with the label suggested by the Deception-trained model. In this way, after analyzing the POPO dataset, we will have a new dataset with pseudo-labels assigned by the Deception model. This database and the new labels will be used to retrain the model. Once trained, the process is repeated from the beginning (an iterative process), so that in each iteration, events are recognized with greater probability, as the model has been trained with data from this new environment.

Regarding the technical terms, we will aim to provide clear and formal definitions so that any reader can understand their meaning. In the initial version, we did not do so because we believed these terms were well-known within the community and that it was not strictly necessary.

5) The presentation of the Results can also be improved. As mentionned above, as the event classes are not clearly defined, it is not always easy to understand the results. Regarding to the cross-validation : you use it for the direct TL approach, but not the weakly supervised TL, why? Besides, isn't it interesting to look at the variations of the accuracy to see if the learning is stable or not, in addition to the mean accuracy? Table 5 must be presented and discussed in more details : it is referred to only in the discussion and after a reference to Table 6.

We will aim to improve the readability of the results in the revised version of the manuscript. Regarding the use of cross-validation, in the direct application of Transfer Learning (TL), it is used to assess how well the systems fit the original POPO2002 catalog. When applying weakly supervised learning, we could also use cross-validation, but this would involve training 16 models (4 models for each of the 4 models that make up the cross-validation for TL). Since this work is primarily focused on highlighting the issue of bias in models trained on catalog information and how that bias can potentially be reduced, we believe that, as the models derived from the weakly supervised approach show very similar performance, applying cross-validation in this context would yield very similar results. By observing the difference between the events detected using the weakly supervised approach and those originally annotated, we conclude that the use of any of the models obtained through cross-validation would result in similar outcomes.

Regarding Table 5, we included it in the manuscript to give the reader an idea of how the event labels differ between those assigned by the weakly supervised approach and the original POPO2002 catalog annotations. Essentially, what is observed is that all tested systems detect many events in traces that were labeled as BGN in the original catalog. Many of these events, if they meet the user-defined threshold criteria, will be included in the new training dataset, and the model will thus be able to find many of these events in the test partition. This is why Table 6 shows a significantly different number of initially annotated events compared to those recognized by the weakly supervised approach.

6)A major argument of your work is that catalogs can be biased, and that the accuracy of ML learning techniques should thus not be the only criterion of a classifyer efficiency. Although this is worth saying it, you say it repeatedly. E.g. l.360 to 375 is only about this point and is only a repetition of what is already said in the introduction : I don't see what is new in this paragraph. Another unclear point is that you present the Popo2002 catalog as a high-quality catalog, but then suggest that some VTE, LPE and TRE are missclasified (l.396-708). Then, you refer to difference between the catalog and your calssification as "an 'error' that was not really an error" (l.404), but as I understand it is is based only on the judgment of "a geophysical expert" (l. 406). Why is this jugemnt more reliable than the classification obtained thanks to the "quality of the human team" mentionned l.103?

We appreciate the reviewer’s comment, as it provides us with the opportunity to clarify the philosophy of our work and the conclusions we have drawn from it. It is true that throughout the manuscript we emphasize that the classification results, rather than highlighting the system's robustness, emphasize how well the system has learned the information contained in the catalog. This explains why systems trained with the POPO2002 dataset achieve results around and above 85% (Table 2).

However, we do not believe this creates a contradiction when we state that the POPO2002 catalog is of high quality, but there are events that were not included in its construction. As mentioned in the introduction, catalog creation is a very time-consuming task, and during periods of high activity, it is humanly impossible to analyze each registered event. In many cases, these events are grouped into more general traces. This does not conflict with the idea of a high-quality catalog, as we have shown that the models can learn the catalog's contents with high performance, demonstrating their self-consistency.

However, when we apply our algorithm and detect an event not annotated in the original catalog, from a statistical standpoint, this event is considered an insertion, which counts as an error in the confusion matrix. This is evident in Table 5. Upon expert analysis, though, this inserted event is not considered an error since it is correctly recognized.

In conclusion, while the original catalog is of high quality, given the complexity of the problem, it is humanly impossible to meticulously analyze all the information in the seismic trace. As a result, reliable but incomplete catalogs are typically produced.

7)You state l.355 that you have "verified" that weakly supervised approaches could "significantly enhance the detection and identification capabilities". However it is not clear what the enhancement refers to. In comparison to what? How do you quantify the enhancement? Thus, I don't think the Results section illustrates correctly this sentence. As a metter of fact, Table 4 shows that a weakly supervised approach improves the accuracy in comparison to a direct application of the MASTER-DC classifier to the Popo database, but then you state that the accuracy is not necessairly a good indicator of a classifier efficiency. On the contrary, if you consider the accuracy as a robust indicator, then the classic TL approach yields better results than the weakly supervised approach (compare Tables 2 and 4).

Regarding the previous point, we would like to emphasize that we were not able to clearly convey our idea in the text. When we conclude that weakly supervised approaches can improve detection and classification capabilities, we base this on the results shown in Table 6. As can be seen, the number of originally annotated events and the number of recognized events differ significantly. We believe that our conclusion is founded on a comparison with the original catalog, and the concept is clear.

Table 4 does not show that a weakly supervised approach improves accuracy compared to the direct application of the MASTER-DC classifier to the Popo database. Instead, Table 4 shows that if we compare the performance of a weakly supervised approach with the information annotated in the original catalog, the performance is quite poor, not exceeding 65% (while a TL approach can reach up to 85% or more). However, when we closely analyze the results, we observe that this performance decline is due to the insertion or detection of many events that, for various reasons, were not initially annotated in the catalog.

Therefore, what our work proposes, and what we want to emphasize, is that while accuracy is a measure of how robust a system is, in our specific and complex problem, accuracy might instead reflect how well the system aligns with the knowledge contained in the catalog. These two concepts are not incompatible, as a model achieving 85% or more average performance could be statistically robust, but a model with 65% accuracy could also be robust if the insertions it makes (if correct) lower the accuracy without compromising robustness. We recognize that explaining this concept is quite challenging, which is likely why our work has not been fully understood by the reviewers.

8) Another argument you put forward is that the weakly supervised approach allows to detect more events than in the catalog. However this is expected, as you apply your classifier to more time windows than the Popo2022 catalog (2139 labelled events in the Popo catalogue, more than 20,000 times windows labelled with your classifier). Besides the number of labelled events is different depending on the classifier (compare sum of columns in Table 6), why is that so? The real question, that I think you don't answer fully in your paper, would be : do weakly supervised classifiers allow to detect more events, and in a more robust way, than classical ML methods and direct TL approaches?

Once again, we would like to thank the reviewer for this comment, which provides us with an opportunity to further elaborate on our results. As the reviewer has pointed out, it is clear that we have not successfully conveyed our idea. Our detection results are not higher because we analyze more windows than events, as the reviewer suggests.

Our approach analyzes the signals by windows. For each window, a label is assigned, but each window is not considered a detected event. Instead, consecutive windows with the same label, meeting the minimum average duration required for each event type in the Deception catalog, are grouped into a single event. In other words, if our approach detects several consecutive windows for the same event, but their combined duration does not meet the average duration imposed by the master catalog, that event is considered unknown and is not even annotated in the table.

Thus, the table reflects recognized events, not windows. This explains why the total number of events in the columns of Table 6 differs for each system. Each system detects events differently. If we were considering windows, the three columns would add up to the same number, as the reviewer suggests. However, that is not the case, nor does such an approach make sense in our context. This methodology is thoroughly described in the references included in the text.

Finally, we would like to emphasize that our work does not argue that weakly supervised classifiers allow the detection of more events, or in a more robust way, than classical ML methods and direct TL approaches. Our argument is that by applying a weakly supervised approach and leveraging TF, we can build less biased catalogs and help develop rapid and robust automatic monitoring systems for seismo-volcanic signals.

For these reasons, I suggest the authors to review thouroughly their work before considering it agin for submission. Their work is of great interest and importance. However, its implications both in terms of pure classification problems, and in terms of volcano monitoring, are not sufficiently investigated.