

# Pervasive Label Errors in Seismological Machine Learning Datasets

Albert Leonardo Aguilar Suarez \*, Gregory C. Beroza <sup>1</sup>

<sup>1</sup>Department of Geophysics, Stanford University, Stanford, California

Author contributions: *Conceptualization*: Aguilar Suarez, Beroza. *Methodology*: Aguilar Suarez. *Formal Analysis*: Aguilar Suarez, Beroza. *Writing - Original draft*: Aguilar Suarez, Beroza.

**Abstract** The recent boom in artificial intelligence and machine learning has been powered by large datasets with accurate labels, combined with algorithmic advances and efficient computing. The quality of data can be a major factor in determining model performance. Here, we detail observations of commonly occurring errors in popular seismological machine learning datasets. We used an ensemble of available deep learning models PhaseNet and EQTransformer to evaluate the dataset labels and found four types of errors ranked from most prevalent to least prevalent: (1) unlabeled earthquakes; (2) noise samples that contain earthquakes; (3) inaccurately labeled arrival times, and (4) absent earthquake signals. We checked a total of 8.6 million examples from the following datasets: Iquique, ETHZ, PNW, TXED, STEAD, INSTANCE, AQ2009, and CEED. The average error rate across all datasets is 3.9 %, ranging from nearly zero to 8 % for individual datasets. These faulty data and labels are likely to degrade model training and performance. By flagging these errors, we aim to increase the quality of the data used to train machine learning models, especially for the measurement of arrival times, and thereby to improve the reliability of the models. We present a companion list of examples that contain problems, aiming to integrate them into training routines so that only the reliable data is used for training.

## 1 Introduction

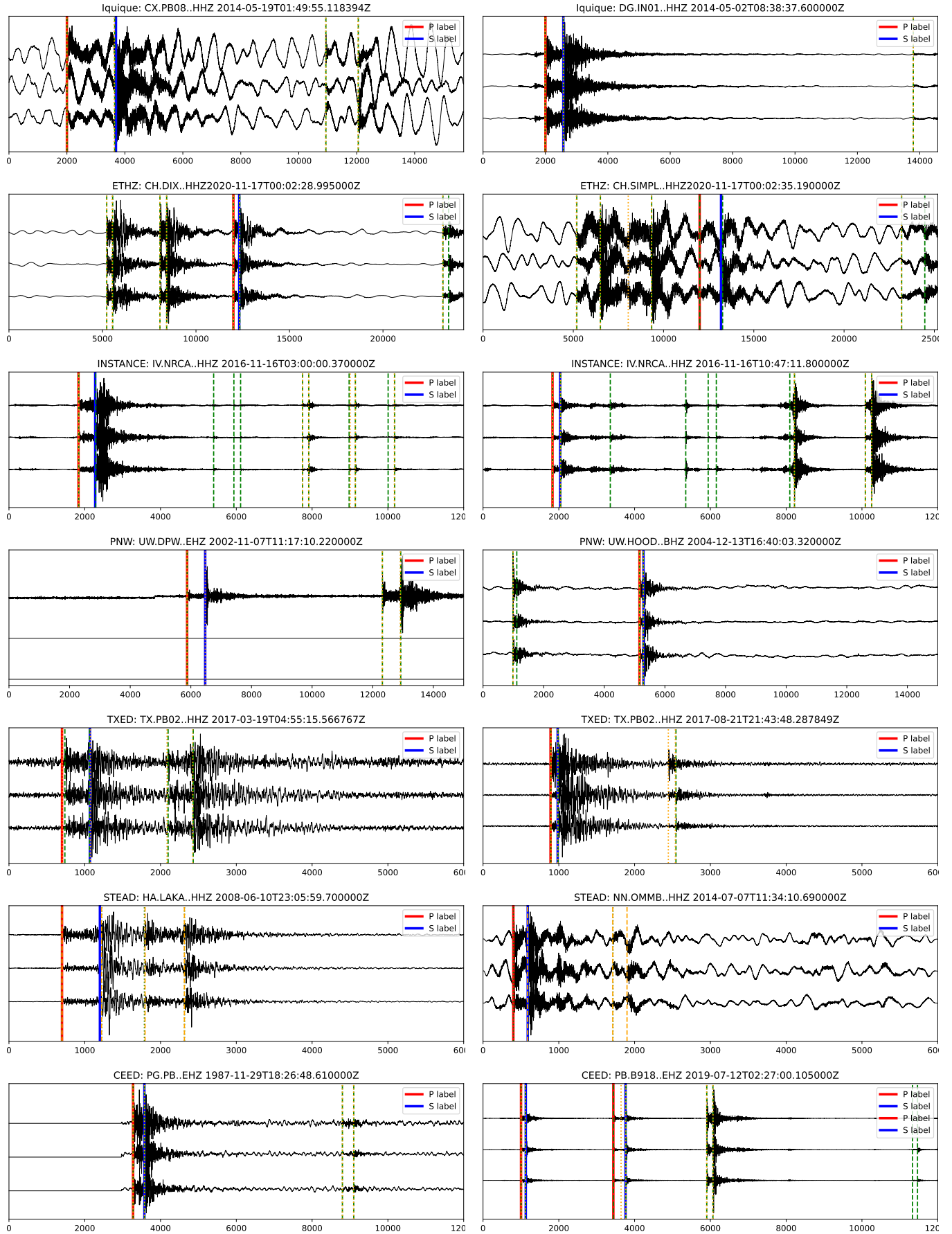
Data and algorithms are the pillars of the recent explosion in machine learning implementations for seismology (Arrowsmith et al., 2022; Mousavi and Beroza, 2022). The volume, variety, and veracity of data available to seismologists have fueled many deep learning models for the measurement of arrival times, which in turn have fueled a new generation of earthquake catalogs that contain many more earthquakes than previously known in routine catalogs (Tan et al., 2021; Schultz et al., 2023; Tan et al., 2025). The landscape of datasets for observational seismology is dominated by local earthquake recordings from densely instrumented regions (Mousavi et al., 2019; Michelini et al., 2021; Zhu et al., 2025). There is a growing number of datasets that aim to fill the gaps in other observational settings, including seismograms from: regional distances (Aguilar Suarez and Beroza, 2024; Kong et al., 2024), moderate magnitude earthquakes (Tang et al., 2024), ocean bottom seismometers (Bornstein et al., 2024), magmatic systems (Zhong and Tan, 2024), and induced seismicity (Chen et al., 2024). These datasets are the basis for many models (Zhu and Beroza, 2019; Mousavi et al., 2020; Niksejil and Zhang, 2024; Aguilar Suarez and Beroza, 2025) that are trained to identify and measure the arrival times of seismic waves. Arrival time measurements are the basis for the downstream tasks of phase association, location, and magnitude estima-

tion, for which there are also deep learning implementations.

Despite the many benefits of deep learning accelerated research, there are unaddressed issues with the high volume of machine learning enabled algorithms (Kapoor and Narayanan, 2023; Naranayan and Kapoor, 2024) in the broader research community. Data sets across domains, spanning text, image, and audio, contain errors that degrade model performance and evaluation (Northcutt et al., 2021a). Strategies have been devised to improve datasets and learn from them in the presence of faulty labels (Cordeiro and Carneiro, 2020; Northcutt et al., 2021b) in the so-called Data-Centric AI approach (Oala et al., 2024). Data quality has been highlighted as a major factor determining model performance (Gunasekar et al., 2023), as model performance gains from improved datasets outpace those from increasing model complexity. The goal of this paper is to improve the quality of seismological datasets for deep learning.

As established phase picking models are applied to challenging settings, several studies have documented some shortcomings (Ruppert et al., 2022; Park et al., 2023; Noel and West, 2025; Aguilar Suarez and Beroza, *in prep*) and solutions suggested (Park et al., 2025). Here, we analyze multiple data sets, including: the Stanford Earthquake Dataset (STEAD) (Mousavi et al., 2019), the Iquique aftershocks dataset (Woollam et al., 2019), the Curated Pacific Northwest AI-ready seismic dataset (PNW) (Ni et al., 2023), the Italian seismic dataset for

\*Corresponding author: aguilar@stanford.edu



**Figure 1** Samples that contain more earthquakes than those labeled. Dotted lines indicate modelensemble inferred arrivals.

machine learning (INSTANCE) (Michelini et al., 2021), the Texas Earthquake Dataset for AI (TXED) (Chen et al., 2024), the California Earthquake Event Dataset (CEED) (Zhu et al. (2025)), and the L'Aquila 2009 aftershocks dataset (AQ2009) (Bagagli et al., 2023). This work was facilitated by the common access point to these datasets provided by Seisbench (Woollam et al., 2022). We analyzed the training, development, and test partitions of these datasets.

We exclude datasets for which the waveforms are too short for available models, specifically the LenDB dataset of 27 seconds, for which padding or other signal modifications would be required to input the data into existing models, potentially harming model performance. We also avoided datasets with waveforms that are too long, like CREW (Aguilar Suarez and Beroza, 2024), for which only one model is available (SKYNET Aguilar Suarez and Beroza (2025)). Following the same logic we exclude the MLAAPDE (Cole et al., 2023) and NEIC (Yeck et al., 2020) datasets, that cover local, regional, and teleseismic distances, because their length does not reach the 5 minutes required to run SKYNET without modifications or padding. We exclude data from volcanic settings and ocean bottom seismometer recordings because there are not enough models to create a meaningful ensemble. We also exclude the SCEDC dataset due to its redundancy with the CEED dataset.

In a separate study (Aguilar Suarez & Beroza, in prep) we document a number of shortcomings of ML-based earthquake catalogs. In an effort to understand their origin, we compared the phase picks used to construct ML catalogs to that that are used to create the USGS Comprehensive Catalog (ComCat). We found that about 10 percent of arrivals are missed by deep learning models, and as a result that up to 10 percent of known earthquakes in routine earthquake catalogs are missed in ML catalogs. This low recall rate is compounded by the occurrence of false positive detections. Both of these shortcomings motivate efforts to train more reliable phase picking models based on more reliable training datasets that are used to construct them.

## 2 Errors

We take a Data-Centric approach, using established deep learning models PhaseNet (Zhu and Beroza, 2019) and EQTransformer Mousavi et al. (2020) to flag potential label errors in the datasets by ensembling their outputs and comparing to the dataset labels. We recognize that this is an approximate approach in the sense that the models used for the evaluation were themselves trained on data with label errors and are not perfectly reliable indications of the presence, absence, or timing of phase arrivals. Nevertheless, they have sufficient predictive value to identify large numbers of label errors. We recognized four main categories of errors: (1) unlabeled earthquakes, (2) earthquakes in noise samples, (3) false earthquakes, and (4) inaccurately timed arrivals.

### 2.1 Unlabeled Earthquakes

The first and most prevalent category of error is the presence of earthquake arrivals for which there are no labels. That is, uncataloged earthquakes in the waveform data in addition to the ones that are labeled. Figure 1 shows a variety of examples from the analyzed datasets that contain more earthquakes than the labeled ones. These labels are faulty because they are incomplete. The uncataloged earthquakes have both low and high amplitude and are located before and after the labeled signal, such that we observed no systematics in the position or relative amplitude of the labeled vs. unlabeled events. Intuitively, small uncataloged earthquakes would follow larger cataloged ones, but during intense aftershock sequences there are both small and larger earthquakes occurring closely in time. The number of unlabeled earthquakes ranges from one to ten. Not surprisingly, the datasets with longer waveforms, like INSTANCE and CEED show a greater range. Table 1 summarizes error prevalence in the eight datasets examined. The error rate ranges from nearly zero to 8 %. The dataset with the lowest prevalence of this type of faulty label is STEAD, followed by TXED, PNW, ETHZ, Iquique, CEED, AQ2009, and finally INSTANCE, with the highest prevalence of this error at 7.98 %.

### 2.2 Earthquakes in Noise Samples

The second type of error is the presence of earthquake signals in samples that are labeled as noise, which can be considered false negatives. For this type of error we are restricted to the STEAD, TXED, INSTANCE, and PNW datasets, which contain noise samples, whereas the other datasets do not. In this case if our ensemble found and agreed on at least one arrival, the example was targeted for further analysis. Figure 2 shows some of those examples that are in the datasets as noise labels, but in fact do contain earthquake waveforms, from one or more events. STEAD contains 235,426 noise samples, from which we found arrivals in 208. After manual inspection we removed 14 samples because they are glitches rather than earthquake arrivals. The prevalence of this error modality in the INSTANCE dataset was estimated to be 1.92 %, with 2549 examples like this one out of the 132,288 total noise examples. Out of the 51,462 noise samples of the PNW dataset, 446 were found to contain earthquake arrivals, making up 0.87 % of the noise partition. Finally, out of the 207,458 noise samples of the TXED dataset, 458 were flagged. In contrast to the other datasets, many of these flagged samples do not contain clear earthquake signals, and we had to manually check each one, resulting in only 16 samples (less than 0.01 %) with clear earthquake signals. Table 2 summarizes the number of samples analyzed, alongside the error prevalence in the noise datasets.

### 2.3 False Earthquakes

The third type of error we identified is for samples that contain labeled arrivals, but the waveforms do not show earthquake signals because the signal is absent or incomplete, such that these can be considered as false

| Dataset  | Total Examples checked | Faulty examples | Prevalence |
|----------|------------------------|-----------------|------------|
| Iquique  | 13,400                 | 515             | 3.84 %     |
| ETHZ     | 36,743                 | 949             | 2.58 %     |
| PNW      | 183,909                | 4,970           | 2.15 %     |
| TXED     | 312,231                | 4,672           | 1.49 %     |
| STEAD    | 1,030,231              | 149             | 0.01 %     |
| INSTANCE | 1,159,249              | 92,561          | 7.98 %     |
| AQ2009   | 1,258,006              | 87,008          | 6.92 %     |
| CEED     | 4,140,530              | 253,309         | 6.12 %     |

**Table 1** Summary numbers of the multiplet error rate.

| Noise Dataset | Total Examples | Faulty examples | Prevalence |
|---------------|----------------|-----------------|------------|
| STEAD         | 235,426        | 194             | 0.09 %     |
| TXED          | 207,458        | 16              | 0.008 %    |
| INSTANCE      | 132,288        | 2,549           | 1.92 %     |
| PNW           | 51,462         | 446             | 0.87 %     |

**Table 2** Prevalence of errors in noise examples that contain unlabeled earthquakes.

positives. We initially recognized this failure mode by using a low threshold of 0.2 for phase picking. Any example for which PhaseNet and EQTransformer predicted no picks with the low threshold of 0.2 was flagged. These types of samples come mostly from accelerometer data, as shown in Figure 3. For the sample on the right panel of Figure 3, the arrivals are in a section of the seismogram that has constant amplitude, most likely due to an instrument outage, some form of padding during postprocessing, and/or an archival artifact. Due to our filtering criteria not translating well from one dataset to another, and a tendency to flag single channel data excessively, we did not quantify the prevalence of this type of error.

## 2.4 Inaccurate Timing

The fourth type of error is labeled arrivals with inaccurate timing, that is, P or S labels that are off by several seconds from their actual arrival times in the waveforms. We compared the labeled arrivals with the predictions from the deep learning ensemble, and marked those with a discrepancy larger than 2 seconds for further analysis. Figure 4(left) shows an example of a badly timed S arrival, which disagrees with the ensemble by several seconds, and is clearly labeled as arriving a couple of seconds before the onset of the S wave energy. We identified only a few such examples, but due to the process of automating the screening not easily translating from one dataset to another, we decided to leave this detailed analysis for future research.

## 3 From Big Data to Good Data

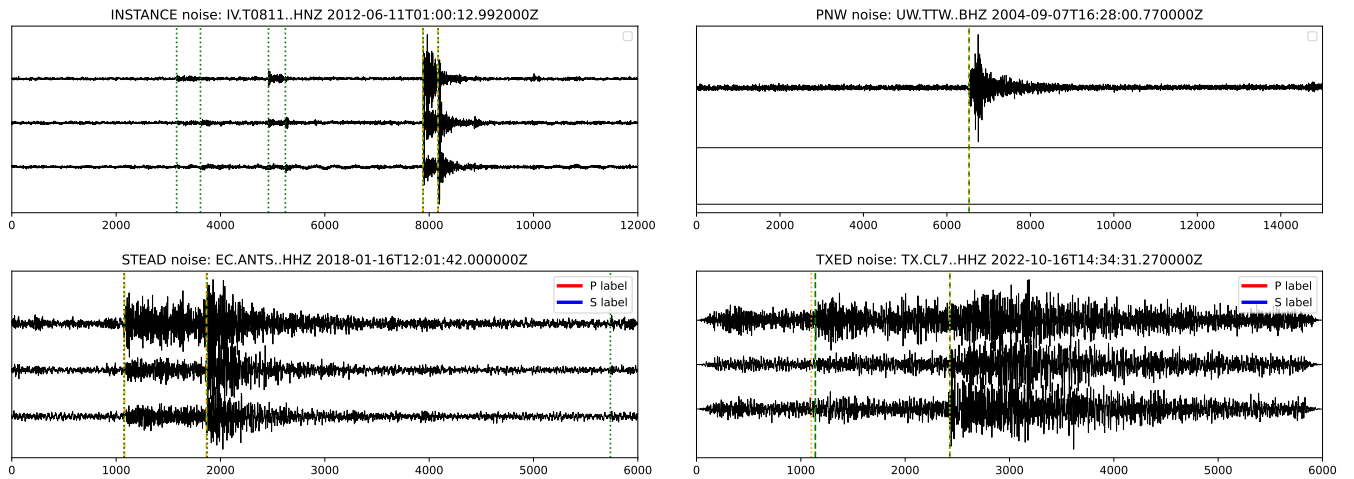
The findings outlined here are intended to lead to improvements in performance for deep learning models, which will have positive effects on downstream tasks, such as phase association and location, and will improve the reliability of automated earthquake catalogs. Northcutt et al. (2021a) estimated an average error rate of 3.3 % for the labels in the test sets of the 10 most

widely used datasets for image, text and audio for machine learning. Our study is most likely a lower bound of the error rate, which is on average 3.9 % across all the datasets analyzed. This number does not include all recognized types of errors, but only those for which we could create an automated, trustworthy, and translatable procedure.

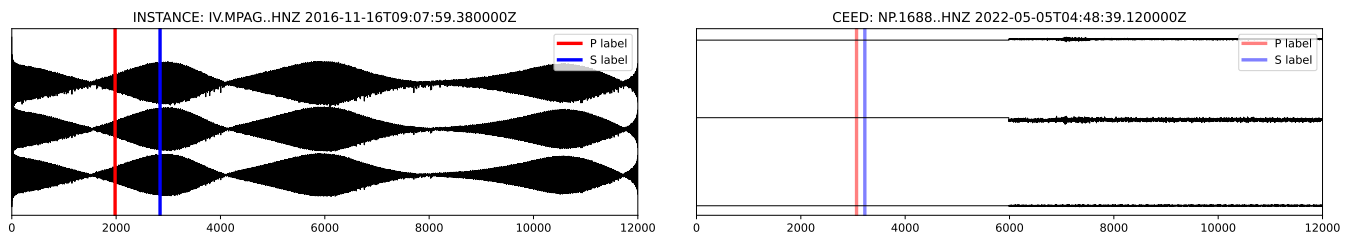
In our approach, we noticed that PhaseNet declares many more arrivals than EQTransformer. We used the conservative version of EQTransformer for our test. We observed several examples that do contain multiple earthquakes, but only PhaseNet and not EQTransformer identified them, so we do not treat them as errors of the multiplet type based on the ensemble. Thus, we believe the numbers presented here should be taken as a lower bound of the error prevalence.

The samples with multiple arrivals from multiple events highlight the complexity of earthquake sequences, with numerous earthquakes occurring closely in time. Specifically for the INSTANCE samples in Figure 1 in the 120 s long waveforms there can be up to 10 earthquakes in quick succession, for which accurate labeling is a challenge, even for state of the art machine learning pickers. We also note that the models we are using for evaluation were themselves trained with data that contains label inaccuracies. The AQ2009 and CEED datasets attempt to work around this shortcoming by allowing the dataset to contain labeled arrivals from more than one single earthquake, but they still fall short during intense seismicity sequences (Figure 1 bottom right). Models can be trained to deal with such dense earthquake signals with appropriate data augmentation (Zhu et al., 2020), as exemplified in Figure 2 of Aguilar Suarez and Beroza (2025) and Figure S3 of Tan et al. (2025). While we found no systematics in the relative amplitude or positions of the labeled and unlabeled earthquakes, we suspect we are missing samples for which the amplitude contrast is extreme, that is a very low amplitude signal next to a very high amplitude signal.

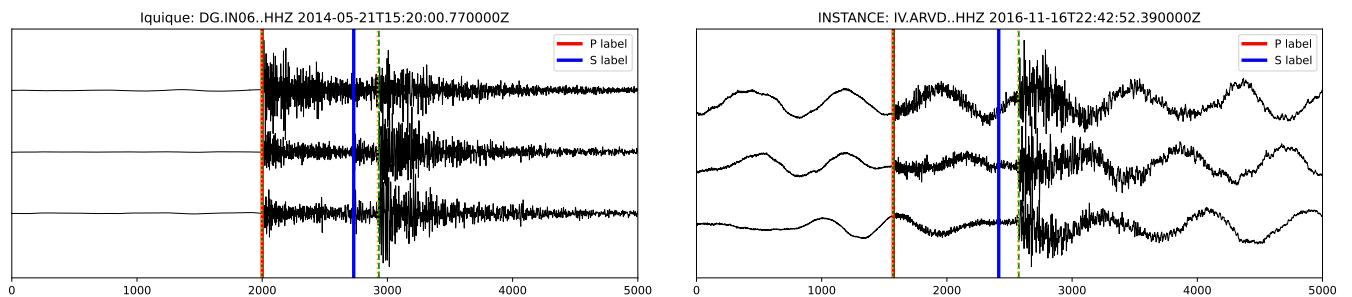
Errors due to unlabeled earthquakes are an order of



**Figure 2** Noise examples for which earthquake arrivals were found. PhaseNet and EQTransformer arrivals are shown with the dotted lines.



**Figure 3** Labeled arrivals for which there are no earthquake waveforms.



**Figure 4** Samples with inaccurately timed arrivals. Dotted lines show the machine learning ensemble. The name of the dataset is indicated on the title of each panel.

magnitude more prevalent than errors due to faulty noise samples (3.8 % vs 0.7 %), as detailed in Tables 1 and 2. The multiplet error rate is lower for STEAD and TXED, which are the datasets with the shortest waveforms, at 60 seconds each. This contrasts with the other datasets, which contain 120 or 150 second long waveforms, which can inherently contain more unlabeled earthquakes. This might also be behind the fact that both STEAD and TXED have the lowest prevalence of earthquake signals in samples labeled as noise. We believe, however, that the low prevalence of error in STEAD is a consequence of the careful quality control performed while assembling that dataset as reported by Mousavi et al. (2019), who estimated the error rate to be less than 1 %. Our findings suggest that the error rate might be dramatically lower, closer to 0.02 %.

The errors found in this study likely include the biases of the data used to train both PhaseNet and EQ-Transformer, which will limit their ability to detect all errors comprehensively. We identified a number of inconsistencies with more regional earthquake waveforms, which are not well represented in STEAD or the Northern California data used to train PhaseNet. We manually checked random samples of flagged data and adjusted our selection criteria in order to avoid valid data being declared as faulty. There are many examples for which the arrivals are visible but neither PhaseNet nor EQ-Transformer declare a pick for all labeled phases, such that there is only partial quality control performed on such samples.

One of the reasons for the prevalence of errors is that the process of assembling these datasets begins with gathering metadata from available catalogs. The corollary is that catalogs assembled on a routine basis have a variety of shortcomings, including but not limited to, bad measurements of arrival times, and mis-association of false picks. On an event-by-event basis, not all visible waveforms are picked, especially when analysts are dealing with intense seismicity sequences.

## 4 Future Directions

We documented the pervasive nature of label errors across seismological machine learning datasets, and quantified it for the two most prevalent types. We find an average error rate of 3.9 %. While the automated nature of our work allowed us to process several million waveforms, selected data was visually inspected, which suggests that we are not yet ready for full automation. Our approach has some limitations due to the nature of the ensemble, which can be biased to agree with data it was trained with. Nevertheless, the results shown here testify to the utility of our approach. The data that has been declared as faulty can be ignored, or potentially fixed, which might entail a combination of semisupervised learning and careful manual supervision. The results outlined here aim to improve the quality of seismological machine learning datasets, which will in turn improve the performance of models trained on those datasets.

We have focused on benchmarks for seismic phase picking, but downstream tasks would benefit from the

creation of additional benchmarks for such tasks as catalog building (Linville et al., 2019), phase association (Pennington et al., 2025; Puente Huerta et al., 2025), and event location (Yu et al., 2025).

## Acknowledgements

The authors thank two anonymous reviewers and the editors for helping improve this manuscript. This work was supported by the Department of Energy (Basic Energy Sciences; Award Number DE-SC0026319)

## Data and code availability

Files containing the names of the examples in each dataset, alongside the reasons for their selection are available here: <https://github.com/albertleonardo/labelerrors/> We are in the process of integrating a new metadata field to Seisbench such that the examples that should be avoided can be removed from training or testing routines.

## Competing interests

The authors declare no competing interests. This is an LLM free manuscript.

## References

- Aguilar Suarez, A. L. and Beroza, G. Curated Regional Earthquake Waveforms (CREW) Dataset. *Seismica*, 3(1), May 2024. doi: 10.26443/seismica.v3i1.1049.
- Aguilar Suarez, A. L. and Beroza, G. Picking Regional Seismic Phase Arrival Times with Deep Learning. *Seismica*, 4(1), Apr. 2025. doi: 10.26443/seismica.v4i1.1431.
- Aguilar Suarez, A. L. and Beroza, G. The problem with machine learning earthquake catalogs and the path forward. Unpublished, Manuscript in preparation, in prep.
- Arrowsmith, S. J., Trugman, D. T., MacCarthy, J., Bergen, K. J., Lumley, D., and Magnani, M. B. Big Data Seismology. *Reviews of Geophysics*, 60(2):e2021RG000769, 2022. doi: 10.1029/2021RG000769.
- Bagagli, M., Valoroso, L., Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V. AQ2009 – The 2009 Aquila Mw 6.1 earthquake aftershocks seismic dataset for machine learning application., 2023. doi: 10.13127/AI/AQUILA2009.
- Bornstein, T., Lange, D., Münchmeyer, J., Woollam, J., Rietbrock, A., Barcheck, G., Grevenmeyer, I., and Tilmann, F. Pick-Blue: Seismic Phase Picking for Ocean Bottom Seismometers With Deep Learning. *Earth and Space Science*, 11(1), 2024. doi: 10.1029/2023EA003332.
- Chen, Y., Savvaidis, A., Saad, O. M., Dino Huang, G., Siervo, D., O’Sullivan, V., McCabe, C., Uku, B., Fleck, P., Burke, G., Alvarez, N. L., Domino, J., and Grigoratos, I. TXED: The Texas Earthquake Dataset for AI. *Seismological Research Letters*, 95(3):2013–2022, 01 2024. doi: 10.1785/0220230327.
- Cole, H. M., Yeck, W. L., and Benz, H. M. MLAAPDE: A Machine Learning Dataset for Determining Global Earthquake Source Parameters. *Seismological Research Letters*, 94(5):2489–2499, 06 2023. doi: 10.1785/0220230021.

- Cordeiro, F. R. and Carneiro, G. A Survey on Deep Learning with Noisy Labels: How to train your model when you cannot trust on the annotations? *Proceedings - 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2020*, pages 9–16, 2020. doi: 10.1109/SIBGRAPI51738.2020.00010.
- Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks Are All You Need, 2023. <https://arxiv.org/abs/2306.11644>.
- Kapoor, S. and Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.
- Kong, H., Xiao, Z., Lü, Y., and Li, J. Seis-PnSn: A Global Million-Scale Benchmark Data Set of Pn and Sn Seismic Phases for Deep Learning. *Seismological Research Letters*, 07 2024. doi: 10.1785/0220230379.
- Linville, L., Brogan, R. C., Young, C., and Aur, K. A. Global- and Local-Scale High-Resolution Event Catalogs for Algorithm Testing. *Seismological Research Letters*, 90(5):1987–1993, 07 2019. doi: 10.1785/0220180345.
- Michellini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V. INSTANCE-the Italian seismic dataset for machine learning. *Earth System Science Data*, 13(12):5509–5544, 2021. doi: 10.5194/essd-13-5509-2021.
- Mousavi, S. M. and Beroza, G. C. Deep-learning seismology. *Science*, 377(6607):eabm4470, 2022. doi: 10.1126/science.abm4470.
- Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. STanford Earthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. *IEEE Access*, 7:179464–179476, 2019. doi: 10.1109/ACCESS.2019.2947848.
- Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1):1–12, 2020. doi: 10.1038/s41467-020-17591-w.
- Narayanan, A. and Kapoor, S. *AI snake oil*. Princeton University Press, 2024.
- Ni, Y., Hutko, A. R., Skene, F., Denolle, M. A., Malone, S. D., Bodin, P., Hartog, J. R., and Wright, A. K. Curated Pacific Northwest Already Seismic Dataset. *Seismica*, 2, 2023. doi: 10.26443/seismica.v2i1.368.
- Niksejel, A. and Zhang, M. OBSTransformer: a deep-learning seismic phase picker for OBS data using automated labelling and transfer learning. *Geophysical Journal International*, 237(1): 485–505, feb 2024. doi: 10.1093/gji/ggae049.
- Noel, S. K. and West, M. E. Challenges and Opportunities of Machine Learning Earthquake Detection for Regional Monitoring. *Bulletin of the Seismological Society of America*, 06 2025. doi: 10.1785/0120250021.
- Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv preprint arXiv:2103.14749*, pages 1–16, 2021a. <http://arxiv.org/abs/2103.14749>.
- Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021b. doi: 10.1613/JAIR.1.12125.
- Oala, L., Maskey, M., Bat-Leah, L., Parrish, A., Gürel, N. M., Kuo, T.-S., Liu, Y., Dror, R., Brajovic, D., Yao, X., Bartolo, M., Rojas, W. A. G., Hileman, R., Aliment, R., Mahoney, M. W., Risdal, M., Lease, M., Samek, W., Dutta, D., Northcutt, C. G., Coleman, C., Hancock, B., Koch, B., Tadesse, G. A., Karlaš, B., Alaa, A., Dieng, A. B., Noy, N., Reddi, V. J., Zou, J., Paritosh, P., van der Schaar, M., and Bollacker, K. DMLR: Data-centric Machine Learning Research – Past, Present and Future, 2024. <https://arxiv.org/abs/2311.13028>.
- Park, Y., Beroza, G. C., and Ellsworth, W. L. A Mitigation Strategy for the Prediction Inconsistency of Neural Phase Pickers. *Seismological Research Letters*, 94(3):1603–1612, 03 2023. doi: 10.1785/0220230003.
- Park, Y., Armstrong, A. D., Yeck, W. L., Shelly, D. R., and Beroza, G. C. Divide and conquer: separating the two probabilities in seismic phase picking. *Geophysical Journal International*, 243 (2):ggaf333, 08 2025. doi: 10.1093/gji/ggaf333.
- Pennington, C. N., McBrearty, I. W., Kong, Q., and Walter, W. R. A Comparison of Machine Learning Methods of Association Tested on Dense Nodal Arrays. *Seismological Research Letters*, 96(4): 2201–2212, 02 2025. doi: 10.1785/0220240290.
- Puente Huerta, J., Sippl, C., Münchmeyer, J., and McBrearty, I. W. Benchmarking seismic phase associators: Insights from synthetic scenarios. *Seismica*, 4(2), Sep. 2025. doi: 10.26443/seismica.v4i2.1559.
- Ruppert, N. A., Barcheck, G., and Abers, G. A. Enhanced Regional Earthquake Catalog with Alaska Amphibious Community Seismic Experiment Data. *Seismological Research Letters*, 94(1): 522–530, 11 2022. doi: 10.1785/0220220226.
- Schultz, R., Park, Y., Aguilar Suarez, A. L., Ellsworth, W. L., and Beroza, G. C. En echelon faults reactivated by wastewater disposal near Musreau Lake, Alberta. *Geophysical Journal International*, 235(1):417–429, 2023. doi: 10.1093/gji/ggad226.
- Tan, X., Tramelli, A., Gammaldi, S., Beroza, G. C., Ellsworth, W. L., and Marzocchi, W. A clearer view of the current phase of unrest at Campi Flegrei caldera. *Science*, 0(0):eadw9038, 2025. doi: 10.1126/science.adw9038.
- Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., Chiaraluce, L., Beroza, G. C., and Segou, M. Machine-Learning-Based High-Resolution Earthquake Catalog Reveals How Complex Fault Structures Were Activated during the 2016–2017 Central Italy Sequence. *The Seismic Record*, 1(1): 11–19, 2021. doi: 10.1785/0320210001.
- Tang, K., Chen, K., Chen, D., Chin, T., and Hsu, T. The CWA Benchmark: A Seismic Dataset from Taiwan for Seismic Research. *Seismological Research Letters*, dec 2024. doi: 10.1785/0220230393.
- Woollam, J., Rietbrock, A., Bueno, A., and De Angelis, S. Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network. *Seismological Research Letters*, 90(2 A):491–502, 2019. doi: 10.1785/0220180312.
- Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., Michellini, A., Saul, J., and Soto, H. SeisBench-A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*, 93(3):1695–1709, 2022. doi: 10.1785/0220210324.
- Yeck, W. L., Patton, J. M., Ross, Z. E., Hayes, G. P., Guy, M. R., Ambruz, N. B., Shelly, D. R., Benz, H. M., and Earle, P. S. Leveraging Deep Learning in Global 24/7 Real-Time Earthquake Monitoring at the National Earthquake Information Center. *Seismological Research Letters*, 92(1):469–480, 09 2020. doi: 10.1785/0220200178.
- Yu, Y., Ellsworth, W. L., and Beroza, G. C. Accuracy and precision of earthquake location programs: Insights from a synthetic controlled experiment. *Seismological Research Letters*, 96(3):1860–1874, 2025.
- Zhong, Y. and Tan, Y. J. Deep-Learning-Based Phase Picking for Volcano-Tectonic and Long-Period Earthquakes. *Geophysical Research Letters*, 51(12), 2024. doi: 10.1029/2024GL108438.

- Zhu, W. and Beroza, G. C. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 2019. doi: 10.1093/gji/ggy423.
- Zhu, W., Mousavi, S. M., and Beroza, G. C. Chapter Four - Seismic signal augmentation to improve generalization of deep neural networks. In Moseley, B. and Krischer, L., editors, *Machine Learning in Geosciences*, volume 61 of *Advances in Geophysics*, pages 151–177. Elsevier, 2020. doi: 10.1016/bs.agph.2020.07.003.
- Zhu, W., Wang, H., Rong, B., Yu, E., Zuzlewski, S., Tepp, G., Taira, T., Marty, J., Husker, A., and Allen, R. M. California Earthquake Dataset for Machine Learning and Cloud Computing, 2025. <https://arxiv.org/abs/2502.11500>.