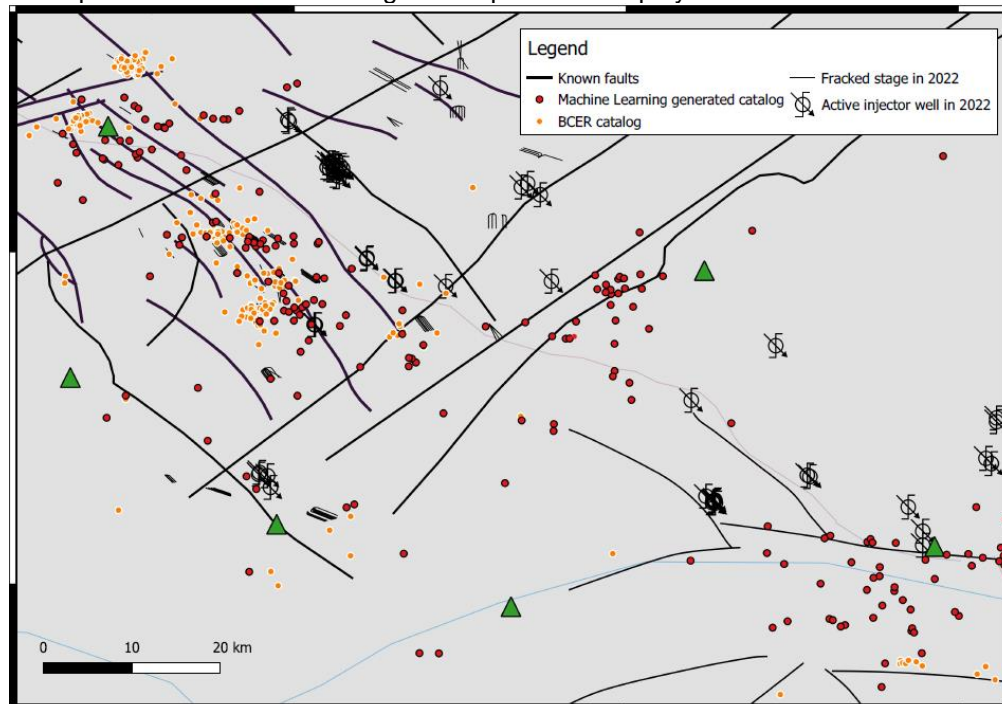


## Building a year-long seismic catalog using machine learning in British Columbia.

*Jesus Rojas Parra, David Eaton, and Rebecca Swinscoe  
University of Calgary*

### Summary

The creation of earthquake catalogs usually requires a lot of human effort. Mainly due to the lack of precision of mathematical based automatic pickers. Consequently, a human must refine the picks that were done by a machine automatically. This can be exhausting and boring for a person. However, a new generation of automatic pickers has been recently developed in the seismology community. These methods are based on machine learning techniques that try to simulate how a human would pick up phase arrivals. In this way, human level precision can be achieved with an automatic picking algorithm. The main goal of this presentation is to show how I tested this machine learning algorithms with public available continuous waveform. I delivery chose a zone with high unconventional fossil fuel development to see if there is any induced seismicity. The results show that seismicity aligns with regionally mapped paleo fault. This proves that machine learning algorithms can help to produce comprehensive and feasible seismic catalogs that can be used for passive seismic monitoring of in exploitation oil plays.



*Figure 1 The study zone located in British Columbia, portraying our inferred epicenters (red circles) and the ones from the British Columbia Energy regulator (small orange dots). Locations of the regional seismic stations are depicted as green triangles. Locations of well pads that were completed by hydraulic fracturing are shown in black, and disposal wells are shown as black circles crossed with an arrow.*

## Introduction

Machine learning techniques have been increasingly used in the last two decades, for applications range from voice recognition to fully self-driving cars. Very few areas of science remain untouched by machine learning, and seismology is no exception. For example, Romeo et al. (1994) used a Perceptron to detect P-wave arrivals in a continuous waveform. His results demonstrated the potential of artificial neural networks (ANN) in the earthquake detection problem. At that time, the limitation was the computation time, which was higher than more conventional event detection methods such as Long-Term Average and Short Term Average (STA/LTA). However, these limitations have since been overcome, including using high-performance computing with graphical processing units (GPU) (Steinkraus et al., 2005) and other fit-for-purpose hardware architectures applied to ANN (Jeon et al., 2021).

There are now numerous examples of Machine Learning (ML) applied to the earthquake detection problem (e.g., Huang et al., 2018; Ross et al., 2019; Stork et al., 2020; Wu et al., 2018). These studies generally take advantage of GPU computing with real or synthetic training data sets. Synthetic data sets allow one to train a neural network without human-made labels, but at the expense of reduced accuracy (Cortez et al., 2020).

On the other hand, the use of time-consuming human-made labels for phase picks improves the accuracy and generalizes the trained models, rendering them capable of classifying inputs they have never seen before (Mousavi et al., 2020). In practice, investigating the reasons why a certain machine learning model works can be challenging, as it depends on the input training data set and the model architecture (Mousavi et al., 2020). SeisBench (Woollam et al., 2022) is a recently developed software package that combines a group of ANN and training data sets; this platform facilitates the testing and creation of ANN for application in seismology.

The data used for training this ANN have primarily been from natural earthquakes; hence, the application of this approach to induced seismicity must be tested. Knowledge of how these technologies perform in an unconventional hydrocarbon setting is also needed to evaluate their feasibility in commercial applications. The characteristics of a seismic catalog built mainly using machine learning techniques have not been addressed with depth for fossil fuel development. Accurate hypocentre locations for induced seismicity are important from a regulatory point of view, as their spatial correlations with industrial activities can activate seismic mitigation actions (Kao et al., 2018). This raises the question: can these methods generate a sufficiently reliable seismic catalog?

This work aims to validate the usage of a previously trained ANN, in the context of the development of unconventional fossil fuels. For this effort, the spatial similarity of the earthquakes of different catalogs will be the major factor. On this occasion, we will use a seismic catalog produced by the British Columbia Energy Regulator (BCER). We acknowledge the drawbacks that these types of comparison have; however, we believe that the comparison is necessary to learn what kind of result we obtain when using these modern techniques. In short, do the locations calculated using the machine learning technique have any spatial relationship with regional faults, wastewater disposal, or hydraulic stimulation? And if they do, how different are the results compared to those obtained by BCER?

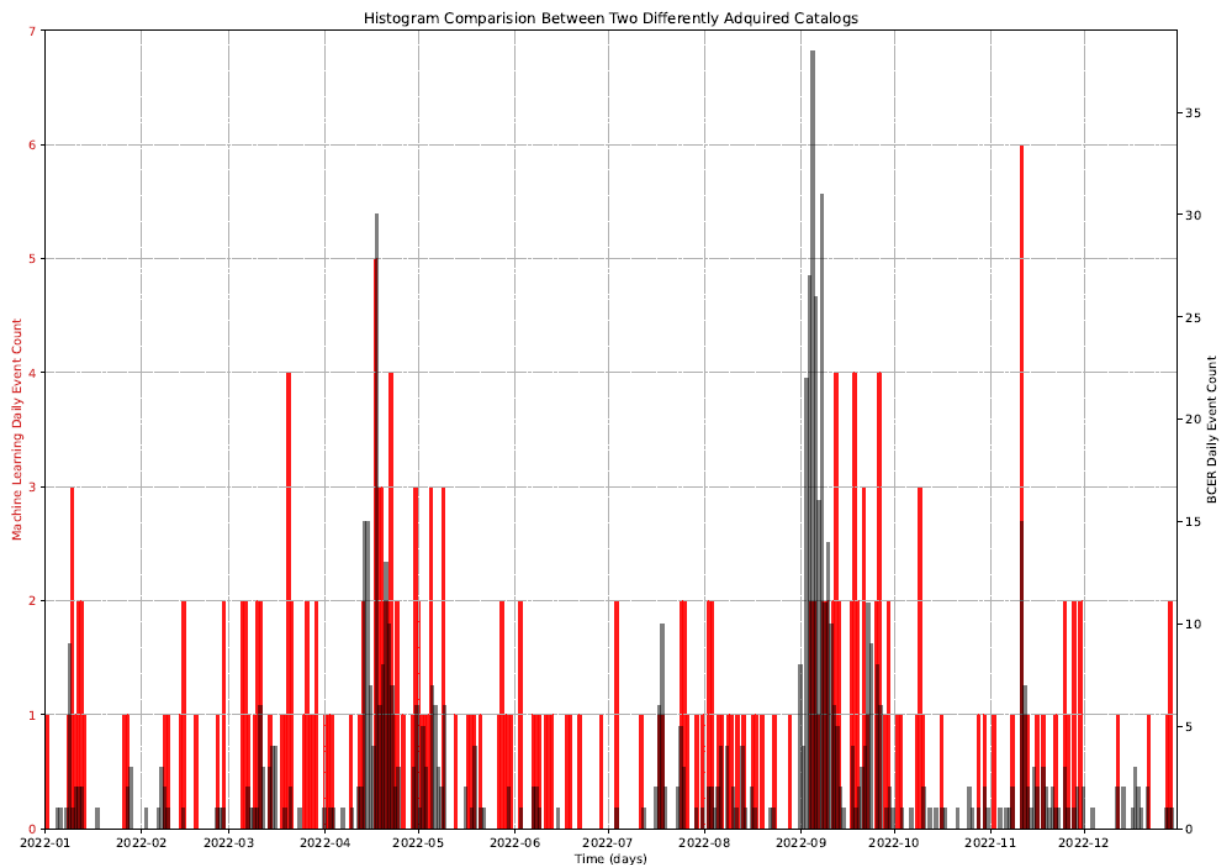


Figure 2 Occurrence times for events depicted in Figure 1, comparing our catalog (red bars) with the one provided (using more stations) by the British Columbia Energy Regulator (black bars). The numerical range of our catalog is shown on the left with a red colored font; likewise, the one for reference is shown on the right with black font. We detected a total of 216 events for 2022, in contrast to the 677 detected by the regulator.

To respond to these questions, this chapter is organized as follows. The Methods section introduces the machine learning technologies used to process the continuous waveform of six regional broadband seismic stations for the year 2022. This is followed by the Results section, which compares event epicenters of the two catalogs. Next, possible reasons for differences between the two catalogs are discussed. Finally, the conclusion section lists the main lessons obtained from this study.

## Method

We downloaded continuous waveform data from the Incorporated Research Institutions for Seismology (IRIS) for six permanent broadband seismic stations. All available data from January 1, 2022, to December 31, 2022, were downloaded. After downloading, we obtain P and S phase arrivals using PhaseNet (Zhu & Beroza, 2019). We used an ANN that was trained using the STanford Earthquake Dataset (STEAD) (Mousavi et al., 2019). We decided to use this ANN, as it is built from a global set of seismic events. Thus, it should have a better generalization or”

knowledge” of a variety of events. In addition, this ANN is available on SeisBench (Woollam et al., 2022), together with a collection of different architectures and training data sets.

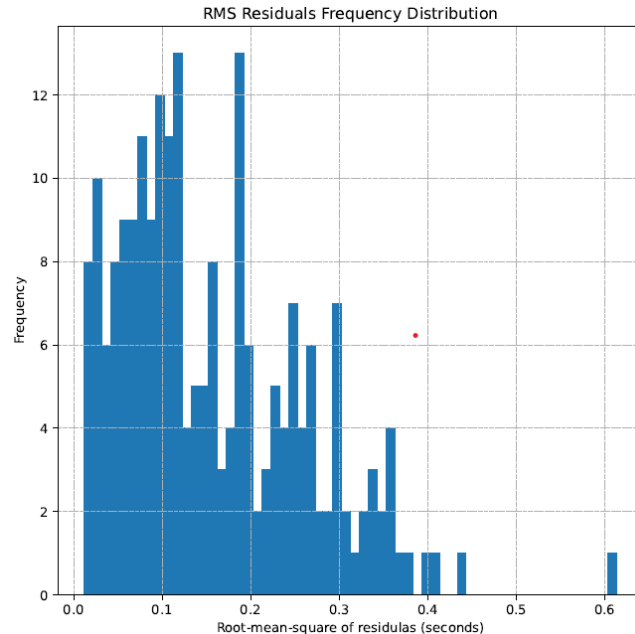


Figure 3 Root-Mean-Square (RMS) time residuals for our catalog obtained from the  $L_2$  inversion scheme used by NonLinLoc to locate the events. The width of each bin is 0.01 seconds. A decrease in frequency follows as the residual RMS increases. One event shows RMS residual of 0.6 seconds, which can be interpreted as an outlier.

For one year of continuous records for six stations, the number of arrivals is on the order of hundreds of thousands. Thus, we used GaMMA, an unsupervised machine learning technique Zhu et al., 2022 to associate arrivals with events. This algorithm requires an average P-velocity as well as an S-velocity that was calculated based on a locally calibrated layered velocity model. We configured GaMMA such that a minimum of three P-arrivals and S-arrivals are required to classify a cluster of arrivals into a seismic event. At the end of this step, we have an initial seismic catalog with crude origin times and approximate locations.

In the last processing step, the events were exported to the appropriate format to be located using NonLinLoc (Lomax et al., 2000) to produce the final catalog used for analysis of the seismicity in the study zone. Additionally, to see if there is a relationship between seismicity and hydraulic stimulation or wastewater disposal, we downloaded a database of active well pads and disposal wells during 2022. This database was downloaded with the use of geoSCOUT (geoLOGIC systems ltd, 2023). Finally, we used the seismicity catalog of the BC energy regulator for 2022 to make a comparison.

Additionally, to help analyze the results, considering mainly the spatial relationship, a buffer was calculated using QGIS (QGIS Development Team, 2023). The distance from the well pads, in the case of the fracking stages, was established at 2.5 km. For wastewater disposal wells, we used a distance of five km from the toe and heel of each well. We consider this to be a reasonable proxy for identifying potential spatial correlations between these activities and seismicity. Furthermore, all map views in this chapter include a vector layer of faults identified in the zone by Hayes et al.,2021.

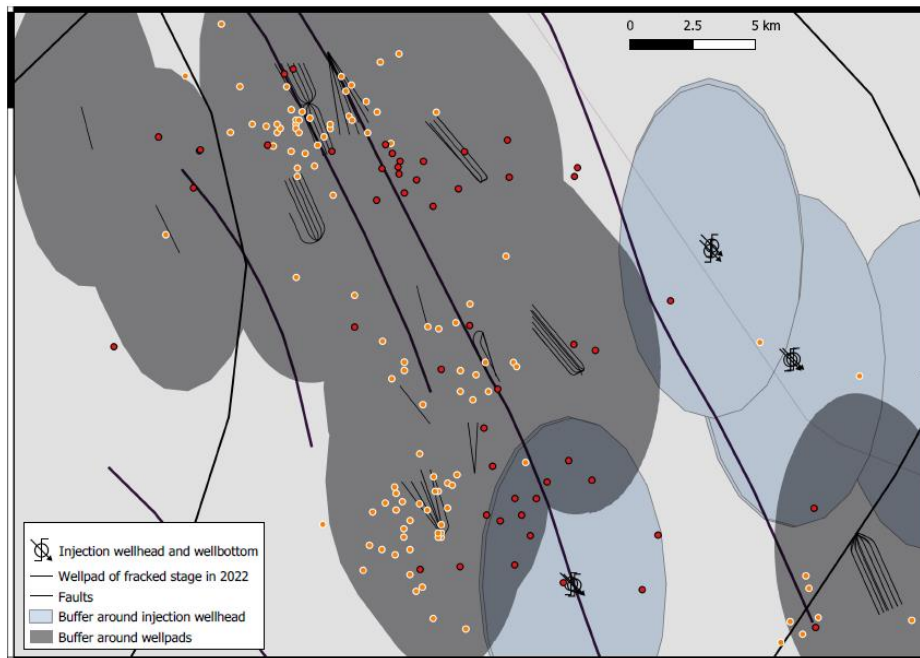


Figure 4 A zoom in of Figure 1 to one of the seismicity clusters. Depicted are our inferred epicenters (in red), the ones for reference (in orange), the well pads (black thin lines), the disposal wells (black circles crossed with an arrow), and the geological faults (bold black lines) compiled by Hayes et al., 2021. Additionally, this Figure shows the spatial buffers calculated 2.5 kilometers away from the hydraulically fractured well pads, and five kilometers from the wellheads and well bottoms of the disposal wells.

## Results

Figure 1 shows the epicenters of the events found by the machine learning approach. For comparison, Figure 1 also shows the epicenters provided by the BCER. A preliminary observation of the faulting system in the study zone is shown as bold black lines. The well pads that were completed by hydraulic fracturing in 2022 are shown as thinner black lines. Wastewater disposal wells are represented by black circles crossed by arrows.

Similarly, Figure 4 shows a zoom-in to one of the events groups in Figure 1. To aid in the initial observations, Figure 4 presents buffers around the well pads (transparent black) and around the wellheads and bottoms of the disposal wells (transparent blue).

A more qualitative perspective is presented in Figure where daily counting is presented in the form of histograms for both catalogs. In red, the catalog produced for this chapter, and in black, the one downloaded from the BCER website (British Columbia Energy Regulator, 2023). Furthermore, consider that Figure 2 presents different y axes, one on the left (in red) and the other on the right (in black).

Finally, Figure 3 presents the root mean square of the residuals from the NonLinLoc inversion scheme (Lomax et al., 2000). In this figure the resolution of the bins reaches 0.01 seconds. Please note that this calculation is particularly useful as all events have different amounts of phase arrivals for their locations, and this calculation factors this consideration.

## Discussion

These results need to be interpreted with care, because the choice of velocity model, geometry, and processing strongly affects the epicenter's locations (Eaton, 2018). However, we cropped the data sets into space to find a common area of interest. In other words, the multiphysics phenomena for the year 2022 in that area are unique in time and space. Thus, despite the difference in noise introduced, the tendencies in seismicity should be the same. In that direction, the maximum number of events in Figure 2 for BCER is four times higher than we determined using machine learning. However, there is an overall consistency between the catalogs, portraying similar increases between April and May 2022 and between September and October 2022. There are other tranches of times where that relation could be still true; however, is harder to argue it robustly.

In Figure 4 the similarity of the spatial distribution can be inferred, as events from both the BCER catalog and ours lie within the spatial buffer range of the wells. These observations alone can build on the strong argument that both catalogs perceived the same phenomena, at least in space. Furthermore, in Figure 1, we observe events that are not detected by BCER. The main driver for this difference is likely due to differences in methodology, velocity model, and the number of stations used since the BCER catalog used the full network, whereas our approach was limited to 6 stations. Given the requirement of 3 P- and S-wave picks used for event association, as well as the use of moveout characteristics of a uniform average velocity, further refinement of our approach may be possible.

It is worth noting, however, that most of our epicenters appear to coincide with regionally mapped faults Hayes et al., 2021. Moreover, most of our events occur at locations where geological faults and industrial activity coincide in space. This spatial correlation suggests that industrial activity as a probable cause for fault activation in this zone during 2022.

To investigate the precision of the velocity model, we used a plot of the Root Mean Square (RMS) of the residuals (see Figure 3). This histogram helps us to understand whether the chosen model can properly resolve the ML-picked phase arrivals. For reference, there are studies that call precise located events with a RMS of the residuals of less than one second (Castro et al., 2011).

All the RMS of the residuals in Figure 3 show values below 0.6 seconds. The bin resolution in Figure 3 is 0.01 seconds. With this resolution, may not be intuitive, but the frequency of values decreases as the RMS of the residuals increases. This behavior is expected when using an L2 inversion scheme as Nonlinloc does. There is an outlier at 0.6 seconds; this represents 1 of 216 located events. Whether this outlier constitutes a reason to change the L2 inversion scheme to an L1 scheme remains a subject of research. It may be that the layered velocity model is too simple to explain the structural geology complexity in the study zone, or it could be that the picked phase arrivals are not accurate enough.

In that sense, the fact that we are using an ANN that was trained on seismicity located elsewhere introduces an observation error or noise. A way around this is to retrain the ANN using local seismicity to better resolve for this case (Niksejel & Zhang, 2023). However, this could lead to human effort comparable to following more traditional methods. A preliminary solution would be to train on a 2-year continuous waveform to obtain a very accurate ANN to be used in years to come. For now, the locations are embedded with error sources from both the layered earth model and the ANN PhaseNet. Identifying the amount of error introduced by these sources of error requires further investigation.

Nevertheless, these results are encouraging as the results show relatively small residuals even when using an ANN trained elsewhere. This demonstrates that these methods have the potential to contribute to automating passive seismic monitoring. One fact to note is the relatively low processing time required to build this six-station, year-long catalog. Adding up the downloading time, the ML-phase arrival picking, and the Nonlinloc location, the processing time was approximately eleven hours. This time was achieved mainly due to the use of GPU processing that is optimized to run ML techniques.

## Conclusions

After using PhaseNet (Ross et al., 2019) for the selection of phase arrivals, GaMMA (Zhu et al., 2022) for phase association, and Nonlinloc (Lomax et al., 2000) for location, we can conclude that ML provides good potential to contribute to automating passive seismic monitoring in the context of unconventional fossil reservoir development. The distribution of the time residuals  $\leq 1$ s from NonLinLoc provides support for this conclusion.

In terms of comparison with the BCER catalog, a spatial correlation was observed. Association in time of events (in seconds difference) between the two catalogs was unsuccessfully attempted. Nevertheless, a comparison in days of the two catalogs shows apparent correlation. The BCER catalog reports four times more events for the same area of interest, likely because of the use of the full seismic network. This comparison allows us to corroborate that our observed epicenters lie near industry activity, geological faults, and the BCER detected events.

## Acknowledgements

Sponsors of the C-DaPS consortium are thanked for their ongoing support. Authors of open-source codes are sincerely thanked for making these codes available. GeoLOGIC systems ltd. is thanked for their contribution of data and software used in this study. All geoLOGIC systems ltd. data and software are 2023.

## References

- British Columbia Energy Regulator. (2023). *Northeast BC Seismicity App*. Retrieved October 22, 2022, from Northeast BC Seismicity App: <https://geoweb-ags.bc-er.ca/portal/apps/webappviewer/index.html?id=a1ecce14d6ae4c92a5295c62a3ee618b>
- Castro, R. R., Valdés-González, C., Shearer, P., Wong, V., Astiz, L., Vernon, F., . . . Mendoza, A. (2011). The 3 August 2009 M w 6.9 Canal de Ballenas region, Gulf of California, earthquake and its aftershocks. *Bulletin of the Seismological Society of America*, *101*, 929–939.
- Côrte, G., Dramsch, J., Amini, H., & MacBeth, C. (2020). Deep neural network application for 4D seismic inversion to changes in pressure and saturation: Optimizing the use of synthetic training datasets. *Geophysical Prospecting*, *68*, 2164–2185.
- Eaton, D. W. (2018). *Passive seismic monitoring of induced seismicity: Fundamental principles and application to energy technologies*. Cambridge University Press.
- geoLOGIC systems ltd. (2023, October 1). geoSCOUT. *geoSCOUT*. Retrieved from <https://www.geologic.com/geoscout/>
- Hayes, B. J., Anderson, J. H., Cooper, M., McLellan, P. J., Rostron, B., & Clarke, J. (2021). Wastewater disposal in the maturing Montney play fairway, northeastern British Columbia (NTS 093P, 094A, B, G, H). *Geoscience BC Summary of Activities 2020: Energy and Water, Geoscience BC, Rept. 2021-02*, 91–102.
- Huang, L., Li, J., Hao, H., & Li, X. (2018). Micro-seismic event detection and location in underground mines by using Convolutional Neural Networks (CNN) and deep learning. *Tunnelling and Underground Space Technology*, *81*, 265–276.
- Jeon, W., Ko, G., Lee, J., Lee, H., Ha, D., & Ro, W. W. (2021). Deep learning with GPUs. In *Advances in Computers* (Vol. 122, pp. 167–215). Elsevier.
- Kao, H., Visser, R., Smith, B., & Venables, S. (2018). Performance assessment of the induced seismicity traffic light protocol for northeastern British Columbia and western Alberta. *The Leading Edge*, *37*, 117–126.
- Lomax, A., Virieux, J., Volant, P., & Berge-Thierry, C. (2000). Probabilistic earthquake location in 3D and layered models: Introduction of a Metropolis-Gibbs method and comparison with linear locations. *Advances in seismic event location*, 101–134.
- Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, *11*, 3952.
- Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. *IEEE Access*, *7*, 179464–179476. doi:10.1109/ACCESS.2019.2947848
- Niksejel, A., & Zhang, M. (2023). OBSTransformer: A Deep-Learning Seismic Phase Picker for OBS Data Using Automated Labelling and Transfer Learning. *arXiv preprint arXiv:2306.04753*.
- QGIS Development Team. (2023). *QGIS Geographic Information System*. Retrieved from <http://qgis.org>
- Romeo, G., & others. (1994). Seismic signals detection and classification using artificial neural networks.

- Ross, Z. E., Yue, Y., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2019). PhaseLink: A deep learning approach to seismic phase association. *Journal of Geophysical Research: Solid Earth*, 124, 856–869.
- Steinkraus, D., Buck, I., & Simard, P. Y. (2005). Using GPUs for machine learning algorithms. *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, (pp. 1115–1120).
- Stork, A. L., Baird, A. F., Horne, S. A., Naldrett, G., Lapins, S., Kendall, J.-M., . . . Williams, A. (2020). Application of machine learning to microseismic event detection in distributed acoustic sensing data. *Geophysics*, 85, KS149–KS160.
- Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., . . . others. (2022). SeisBench—A toolbox for machine learning in seismology. *Seismological Research Letters*, 93, 1695–1709.
- Wu, Y., Lin, Y., Zhou, Z., Bolton, D. C., Liu, J., & Johnson, P. (2018). DeepDetect: A cascaded region-based densely connected network for seismic event detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57, 62–75.
- Zhu, W., & Beroza, G. C. (2019). PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216, 261–273.
- Zhu, W., McBrearty, I. W., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2022). Earthquake phase association using a Bayesian Gaussian mixture model. *Journal of Geophysical Research: Solid Earth*, 127, e2021JB023249.