SAND2019-1319J

Deep Learning Models Augment Analyst Decisions for Event Discrimination

Lisa Linville[1], Kristine Pankow[1], Timothy Draelos[2]

1. University of Utah Seismograph Stations University of Utah 115 South 1460 East, Room 211 FASB Salt Lake City, Utah 84112-0102 U.S.A.

2. 2. Sandia National Laboratories. Geophysics Department MS 0758. P.O. Box 5800. Albuquerque, NM 87185-0758

KEY POINTS:

- Neural network models achieve above 99% classification accuracy between surface mining and tectonic events in Utah
- Model predictions are sensitive to P, S and coda wave energy
- 3-component data is typically more accurate but not required for prediction

ABSTRACT

Long-term seismic monitoring networks are well positioned to leverage advances in machine learning because of the abundance of labeled training data that curated event catalogs provide. We explore the use of Convolutional and Recurrent Neural Networks to accomplish discrimination of explosive and tectonic sources for local distances. Using a 5-year event catalog generated by the University of Utah Seismograph Stations we train models to produce automated event labels using 90 s event spectrograms from 3-component and single-channel sensors. Both network architectures are able to replicate analyst labels above 98%. Most commonly, model error is the result of label error (70% of cases). Accounting for mislabeled events (~1% of the catalog) model accuracy for both models increases to above 99%. Classification accuracy remains above 98% for shallow tectonic events, indicating that spectral characteristics controlled by event depth do not play a dominant role in event discrimination.

PLAIN LANGUAGE SUMMARY

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Seismic events observed using sensor networks are typically reviewed manually by seismic analysts to determine which source generated each event. In Utah, two of the most common event types are tectonic, or naturally occurring earthquakes, and quarry blasts from surface quarry operations. Since analysts in Utah have been reviewing events in Utah for more than 50 years, a large catalog of events labeled by source type exists. In this work we explore methods to leverage labeled event types from part of the catalog to automate event labelling for future events. Our approach includes two neural network variations (Recurrent and Convolutional) to identify events as either quarry blasts or earthquakes. Both methods achieve similar classification accuracies above 99%, which rivals the accuracy of human analysts on the same task.

INTRODUCTION

Discrimination between local tectonic events and mining or quarry blasts is a challenge common to seismic networks and researchers working to build seismicity catalogs in regions where both tectonic seismicity and anthropogenic sources exist (Astiz et al., 2014) because published seismic catalogs are typically intended to contain only events that pertain to the study of elastic strain partitioning and release in the crust. Within Utah, both earthquakes and quarry blasts occur in abundance, often in close spatial proximity, making event labelling a significant task for analysts.

Many of the strategies developed for automated or semi-automated approaches to source discrimination at local to regional scales exploit amplitude and spectral ratios of wave phases that are sensitive to source and/or path characteristics. For example, explosive sources preferentially excite P wave energy compared to S (Stump et al., 2002). At frequencies around 1-2 Hz, the amplitudes of P waves are typically much smaller than those of later arriving Lg phases

(further distances) or Rg phases (closer distances), and many strategies exploit these relationships (Lg: Dysart and Pulli, 1990; Baumgardt and Young, 1990; Mayeda, 1993; Rg: O'Rourk and Baker, 2016; Tibi et al., 2018). Although amplitude ratio methods perform well when tuned for specific datasets, their use can be limited when ratios rely on wave phases that are unavailable (at larger distances, for example), difficult to isolate (at very short distances), or occur in high-noise environments. Other recent approaches to event discrimination avoid issues related to phase isolation through the use of coda waves. For example, Koper et al. (2017) show that the duration of coda waves can help discriminate surface events from deeper tectonic earthquakes at local scales. However, methods that rely on waveform characteristics dominantly controlled by depth can have limited ability to resolve tectonic from anthropogenic sources in regions where tectonic events are shallow.

In this work we test two neural network (NN) architectures (Convolutional and Recurrent) on the task of binary event classification for tectonic earthquakes and quarry blasts at local scales. Both architectures achieve accuracies above 99% without waveform segmentation (by phase), feature engineering, or path corrections, avoiding manual feature selection and threshold tuning. The Utah models are also able to generalize over many source-path combinations (rather than a model per station approach) for many thousands of discrete sources and for distances between sub-km scale up to 400 km. Although deep neural networks are clearly competent at event discrimination using Utah data, they do so at the cost of transparency; the complex data routing through the non-linear modular hierarchy in model space obfuscates any direct or intuitive links to the underlying physics that lead to successful discrimination. In an effort to increase trust and transparency in trained models, we suggest a simple approach that leads to interpretable links between successful model predictions and underlying physics.

DATA

This study uses event solutions (phase arrival times, earthquake locations and magnitudes, etc.) from the University of Utah Seismograph Stations (UUSS) between October 2012 and July 2017 for events labeled as local earthquakes or quarry blasts by UUSS analysts (13,313 events; 103,944 phases). We use phase pick qualities of 1 or higher (expected pick errors small than +-0.06 sec) for local earthquakes, and 2 or higher (+- 0.15 sec) for quarry blasts. The number of examples from each class are balanced within 8% (46% quarry blasts, 54% local earthquakes), with more quarry blasts at intermediate distances and local earthquakes at near-source distances (Figure 1).

Waveforms are downloaded for each event from the Incorporated Research Institutions for Seismology (IRIS) using web-services for 10 s of data prior to the first P arrival and 2 min following the first arrival. For all examples (an event recorded on a single station with 1-3 channels), we resample the data to a uniform sample rate of 100 Hz. We then detrend, taper (Hann window, 1%), correct for instrument sensitivity, and apply a high-pass filter at 1Hz (Butterworth, 4-corner). For 3-component data, we rotate the recorded north-south and east-west horizontal channels to the radial (R) and tangential (T) orientations, respectively, using the azimuth from the receiver to the source. We then calculate a spectrogram using 2.56 s windows (256 samples, 12% overlap), keeping only frequencies between 1 and 20 Hz and times up to 90 s for all available channels.

UUSS maintains a variety of sensor types, some of which are not triaxial. 3-component data constitutes 50% of all examples. For single-channel (vertical) stations we fill the empty

horizontal channel spectrograms with zeros. Valid data channel spectrograms are log-scaled and normalized by the max spectral value.

METHOD

Neural networks are a well-established technique in seismic event processing, and some early implementations exhibit success on tasks that require minimal generalization despite predating some of the optimization and training strategies that make NN variations as successful as they are today (Del Pezzo et al., 2003; Dowla et al., 1990; Dysert and Pulli, 1990). We use two specific architectural adaptations to basic neural networks in the work presented here (Figure 2). Our first architecture is a recurrent neural network (RNN) variation called Long-Short-Term-Memory (LSTM; Hochrieter and Schmithuber, 1997). This adaptation of the RNN architecture solves issues introduced by back-propagating gradients used to train the network over many time-steps (Bengio, et al. 1994). We also use bidirectional layers to give both past and future context to current examples (Graves et al., 2005; Graves et al., 2013). The LSTM model routes input spectrograms (40*48 time-frequency indexed amplitude values) to output classes (local earthquake or quarry blasts encoded as 0 and 1, respectively) through 4 layers (node counts: 40, 40, 80, 80) in a many-to-one learning scenario (we take input from many time-steps to make one binary classification). Each LSTM node maintains a hidden state that is determined using weighted input and (typically) sigmoid activated gates with learned parameters that control information flow and output predictions (see Method Details the supplementary section of this document for additional details).

The second model architecture we test is a Convolutional Neural Network (CNN) (LeCun, et al. 1998). CNN methods have been widely adopted for seismic signal processing in recent years (Krizhevsky et al., 2012; Perol et al., 2017; Ross et al., 2018) and rely on the

assumption that characteristic features useful for classification over the entire dataset (in our case, pixel amplitude patterns from small image sections) can be learned and represented in hierarchical filters of finite size. Although previous work demonstrates that CNN's can utilize raw waveform data, for our data raw input CNN models underperformed compared to models that used spectrograms as input. Our final CNN models used 4 convolutional layers (filter count: 18, 36, 54, 54), one fully connected layer, Rectified Linear Units (ReLu; $f(x) = \max(0, x)$) and (2x2) max-pooling between each layer.

For the final layer in both architectures, we interpret output activations as a probability over all possible outputs using the softmax function (i.e. the probability of each class is computed by forcing the values of the output layer in the network to sum to 1 and be between 0 and 1). The final model performance is reported as the median of model accuracy and the standard deviation between all models using 10-fold cross-validation (80% training, 10% validation, 10% test).

Each event is associated with examples from as few as 1 station and up to 48 stations. In order to avoid class dominance over each training iteration, we randomize at the sample level using a batch size of 16, after partitioning at the event level. We cease training when accuracy on the validation set failed to increase over 8 iterations (epochs) through the training data and choose as a final model the earliest epoch with the highest validation accuracy.

RESULTS

Figure 3 presents cross-validation results for each model (LSTM, CNN), with accuracy reported at both the station-level and network-levels for data within the UUSS authoritative review boundary (Lat: 37° to 42° N, Lon: 114° to 109° W). Station-level accuracy is derived

using all examples, regardless of which event they associate to. Network-level accuracy combines all station-level examples from a single event, using the maximum of the summed class probabilities (explosion or earthquake) to determine one classification per event.

Trained models demonstrate notable similarities, all achieving ~96% on station-level classification, ~98% using more than 1 station, and additional gains on the order of fractions of a percent from additional stations. Restricting test data to events within Utah increases station-level accuracies by 0.8% and 0.9% for LSTM and CNN models, respectively. Differences between event level accuracy for earthquakes and quarry blasts when using equally weighted predictions for each station (instead of allowing highly confident examples to dominate predictions; see supplementary section S1 for a more detailed explanation) suggests that trained models are more adept at generalizing earthquake sources than quarry blasts in Utah.

We manually evaluated misclassified events from each model for data within Utah. Out of 167 misclassified events, 120 were misclassified by both CNN and LSTM models. Of the 120 misclassified events, 70% were mislabeled by analysts (based on manual review using waveforms, locations, multiple analysts and long-term catalog data for context). For the remaining 47 events misclassified by either CNN or LSTM, 9% (4/47) of the CNN misclassifications were label error and 19% (9/47) of the LSTM misclassifications were from label error. Removing mislabeled events or correcting event labels results in event-level classification accuracies of 99.1%.

There are 103,944 examples available for model training and testing in the Utah. 50,581 examples (49%) are from single component (vertical) stations, with null horizontal channels values. For our data, models trained on individual channels give nearly the same reported accuracies (vertical=93.8%, radial= 93.9%, transverse=93.3% accuracy), but when used together

accuracy increases by ~1.5% (on single-fold trials). While null channels do not appear to be a significant impediment to accurate predictions, vertical-only cases are disproportionately represented in misclassified examples (62% from single-channel vs 38% from triaxial sensors).

We validate our event duration selection (90 s) using test data that includes a 5-minute window following the event onset. Models trained on 5-minute duration and a 90 s duration yield equivalent accuracy. Investigating how short a window could be used, we found that using just the first 4 spectral windows following event onset (~9 s), a trained CNN (2-4 layers) accomplished station level accuracy of 84.7 ± 0.7% and event level accuracy of 94.4 ± 0.9% using 10-fold cross-validation.

We evaluate event distributions for misclassified events to identify deviations between correctly and incorrectly identified examples (Figure 3; bottom). Overall differences between correct and incorrect distributions are minor, and individually neither magnitude, depth or distance account for a majority of example misclassifications. To further validate model insensitivity to event depth, we test model performance on local earthquakes with epicentral depths of 2 km or less and find that model accuracy remains above 98% for shallow tectonic earthquakes.

Many modern deep NNs are known to be poorly calibrated (Guo et al., 2017) and can offer untrustworthy assessments of prediction certainty especially for out-of-distribution samples (Lee et al., 2018). Most (92%) station-level predictions from Utah models are made with 90% certainty (softmax probability) or greater. Network-level predictions for events that agree and events that disagree with analyst labels report high levels of certainty (96% are above a median class difference of .8, where 1 is completely sure and at 0.5 the prediction is equally split between classes) indicating that thresholds on certainty estimates are not likely to be a viable

way to limit events that require manual review following model classification. Despite high certainties for misclassifications, which in many cases is due to event mislabeling, uncertain event locations generally follow the locations of incorrectly classified events when label error is accounted for. Utah models are more likely to be uncertain about local earthquakes (278/454 unsure events are local earthquakes), and most of the unsure local earthquake events occur in mining areas where tectonic seismicity and mining induced seismicity occur together. The Bingham mine dominates spatial event densities where the model incorrectly classifies quarry blast examples at the station level (Figure S2 available in the electronic supplement to this article).

Although class activation mapping or saliency methods are a common approach to identifying aspects of the input that are influential for prediction (Adebayo et al., 2018; Shrikumar, et al., 2017; Simoyan et al., 2013), we use a simple occlusion method that generates spatial heatmaps of accuracy deflections and inflations after we iteratively mask sections of the input signal. We remove signal (zero fill) 2x5 non-overlapping pixel blocks from randomly selected input examples and stack the resulting accuracy over 1000 iterations for each class. Figure 4 outlines deflections in accuracy when the model is denied access to information from regions of the input over many examples form each class. It is apparent that removing non-event related signal tends to increase prediction certainty, verifying that features more important for prediction are not directly related to processing artifacts or ambient noise conditions. The absence of 5-10 Hz S-wave energy on the transverse component causes the largest decrease in prediction certainty for local earthquakes at the 50-75 km source-receiver distances, while for quarry blasts earlier arriving phases at lower frequencies on vertical channels are more important. Although Figure 4 uses low frequency-time resolution, pixel-scale resolution provides

211 a way to link fluctuations in model certainty to specific wave phases (Figure S3 available in the
212 supplementary section of this article).
213
214 DISCUSSION
215 There is a clear temporal progression expressed in the records of seismic sources captured
216 by distant receivers. This is the result of a complex source coupled with wave propagation
217 through a complex earth medium. In our data, examples sample discrete portions of the crust
218 over various distances with at least 4 distinct velocity profiles (Roller, 1965; Keller et al., 1975;
219 Pechmann et al. 1984; Loeb, 1986). Previous work has navigated some of the complexity in
220 waveforms through feature engineering, and data segmentation (e.g., using only information
221 from specific wave phases that are pre-cut from the data). The diversity of training examples
222 available in even modest duration catalogs appears to be sufficient to generalize well for
223 earthquakes sources in Utah. Quarry blast event accuracy depends more highly on representative
224 samples with higher classification certainty, indicating that model retraining may be required
225 over time for new mine sites, mine products or extraction techniques.
226 Although LSTM models are a more intuitive choice for time-dependent signal modeling,
227 CNN models appear to be equally as adept at modeling the variation within Utah data given the
228 data representation. Our results suggest that LSTM models may have the capacity to model time
229 dependence that far exceeds the capacity required to model the 40 time-steps of our input.
230 Likewise, we achieve no additional gains in accuracy by increasing CNN model depth beyond 4
231 layers (the 11-layer CNN model of Simoyan and Zisserman, 2014 for example gives equivalent
232 accuracies). It is generally observed that feature complexity increases with CNN model depth
233 (Figure S3 available in the supplementary section to this article). The ability to learn and

discriminate meaningful patterns at scales that exceed the fixed resolution of our event spectrograms seems neither possible, nor helpful.

When mislabeled events exist in the test data, even sparsely, they can obscure true estimates of model performance because each event has on average 8, and up to 48 associated examples, and because events remain associated within each data split. We conservatively estimate a ~1% label error rate (LER) for all data (based on the difference between accuracy for all data vs Utah-only data for models trained on all data). Quarterly reviews by UUSS analysts offer secondary quality checks on examples within the UUSS authoritative boundaries. This additional review may account for part of why events outside of Utah have a higher label error rate. Other viable reasons for decreased label accuracy outside Utah include lack of comprehensive mine databases and reduced communication with mine operators compared to sites within Utah.

There is remarkable consistency between station-level classifications for events even when the model disagrees with the analyst label. Inside Utah, station-level agreement that disagrees with network-level analyst labels often identifies events that were mislabeled by analysts. The ability to identify mislabeled events with reasonable accuracy (70% using ensemble predictions from both CNN and LSTM models) makes this method useful for the identification of analyst error going back (or with caution, forward in time) for cataloged events.

Identification and exclusion of mislabeled events precludes an examination of meaningful patterns in misclassified examples. Manual review of a subset of model misclassifications in Utah (where both CNN and LSTM models agreed) was possible because we use a modest sized training catalog (5 years, 13k events) and because trained models maintained low error rates. We mention this to highlight difficulties that arise in assessing model performance as it approaches

or exceeds human error rates and data size increases. In our data, excluding mislabeled events and evaluating misclassifications suggests that misclassifications are not easily explainable systematically, but require scrutiny at the individual example level.

The main objective of this work was to build a model capable of reproducing analyst classifications on new incoming data from the UUSS network in near real-time. We chose to partition training data randomly but acknowledge that future studies may benefit from a temporal partitioning that may more directly assess the performance of models under realistic monitoring conditions. With a classification latency of ~80 seconds after the first arrival on a station we achieve network level accuracies above 99%. It is possible to reduce signal length to a few seconds following the first arrival, however lack of P and S energy for most samples reduces accuracy to 94.4% at the event level (CNN, using events with more than 2 associated station examples).

An additional expectation was that models handle data from both single-component and 3-channel stations. While sufficient information for classification exists in each channel individually (R, T, Z demonstrate nearly equal performance), together they perform better (by ~1.5%) and station level accuracy is 2% higher for triaxial examples (97% and 95%, for triaxial and vertical, respectively). Allowing models to learn from all three channels but still make predictions on examples where only vertical data are available extends the usage to events that occur outside of areas with broadband coverage (or fractionally about half of the network stations).

We expect significant variation at the source level within each class. The mining industry employs a variety of blasting methods (Langefors, 1978; Persson et al., 1994; Dowding, 1985). For local earthquakes, variation can come from fault and local stress conditions. We do not attempt to resolve dependence on these variations for any of the misclassified examples, but we note that model uncertainty for local earthquakes (less than 90% confident) is most highly

concentrated in areas where mining-induced seismicity and tectonic seismicity are both known to occur. For quarry blasts, the model uncertainty is concentrated around the Bingham copper mine.

CONCLUSIONS

The models presented here were developed to demonstrate that deep learning can be a highly accurate way to classify seismic events for monitoring networks with stable station characteristics, a diversity of tectonic event locations and static mine site locations. We show that several model architectures can successfully accomplish binary event classification at accuracies above 99%. We allow our models access to both source and path-controlled information, in addition to ambient background noise and null data (zero fill for vertical only stations), requiring each model to learn, through training, which aspects of the input domain are most important for prediction. We verify that learned features are directly related to event signal, and that while model predictions exploit multiple aspects of each signal including those related to source and path, the resulting models performs event discrimination that goes beyond event depth.

In addition, we demonstrate that events that may be ambiguous in the analyst domain, are often clear outliers within the NN model domain. Each model is able to identify human errors within the event catalog. We also highlight that events outside of Utah have a higher label error rate (~1%) than events within Utah. Our best model trained on Utah-only data performs above 99% for network-level classification and above 96% on station-level classification when mislabeled events are excluded. The identification of mislabeling as a significant source of model misclassification highlights ongoing difficulties researchers face in assessing model performance as it approaches or exceed human error rates. Although confidence assessments for individual examples should be interpreted with caution, locations within Utah that produce event predictions with the highest model confidence on average occur in regions where both mining induced seismicity and tectonic earthquakes or quarry blasts exist. Variation in

prediction confidence for specific source types or geographic areas illuminates a level of complexity in the target domain that is not represented in the binary event labelling schema.

ACKNOWLEDGEMENTS AND RESOURCES

We rely on the Center for High Performance Computing at the University of Utah for computational resources. We use Tensorflow and Theano for early implementations and iterate model parameters using Keras. Training data and final models are available on git (https://github.com/quapity/Utah). Our figures were generated using Matplotlib, Basemap and Seaborn. UUSS data were accessed from IRIS using webservices and Obspy. Data processing relied on python libraries scipy and numpy. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

FIGURES AND TABLE CAPTIONS

Figure 1. Map of events (circles) and source-receiver paths (lines) from UUSS for quarry blasts (red) and local earthquakes (blue). Receivers (white circles) are labeled by station name. The UUSS catalog includes events outside of the authoritative catalog boundaries (Lat: 37° to 42° N, Lon: 114° to 109° W). The data used in this study includes events from outside the authoritative review boundary, but model accuracy is reported using data inside Utah only because of the difficulty in assessing true event labels for events outside the review boundary. See text for further discussion on event mislabeling.

Figure 2. Top: 3-channel 90 second event spectrogram model input. Bottom: Permutations of 3-channel spectrograms for model input. In the LSTM architecture the spectrogram enters the model as a 40x144 matrix (time-step, frequency feature) where the 144 dimension is a 3-channel frequency power spectral density vector for each time time-step. The CNN architecture takes example spectrograms as 40x48x3 volumes (time, frequency, channel). The red box indicates the starting computation using a 2x2 pixel filter that shares weights along the depth axis of the volume.

Figure 3. Top: model accuracy (by class) for station examples and network level predictions, where each station example associated with an event votes for a final event classification. Results are restricted to data within the authoritative review boundary for Utah. Bottom:

violin plots showing the distribution of misclassified events (colored by class label and normalized to equal area owing to the sparsity of misclassified examples) and the correctly classified events as a function of distance, depth, magnitude and time for all data. For the distance view (top left), earthquakes are slightly more likely to be misclassified if the receiver is further away from the source than the average for correctly predicted examples. For the depth view (top right), there is a clear skew toward shallow events for the earthquake class. These events predominantly occur outside of Utah and are misclassified by analysts based on time-of-day and location. Misclassified examples in Magnitude space (bottom left) more closely resemble the opposite class for both local earthquakes and quarry blasts. In time space (bottom right), month scale resolution indicates that there is no clear temporal bias to misclassified events, although for the earthquake class, misclassified events disproportionately occur slightly earlier in the event catalog. We find that misclassified local earthquakes on average have 18 km longer source-receiver distances, magnitudes that are 0.46 magnitude units larger, and occur at depths less than 2.5 km on average.

Misclassified quarry blasts have average distances that are 3.8 km greater, 0.2 magnitude units smaller, and are 1.22 km shallower.

Figure 4. Regions of the spectrogram most important for model predictions. To identify the regions of input spectrograms that have the largest influence on average prediction we generate heatmaps for specific source-receiver distances by removing signal from 2x5 patches and evaluating inflections in model accuracy over 1000 iterations.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (pp. 9525-9536).

Alasonati, P., Wassermann, J., & Ohrnberger, M. (2006). Signal classification by wavelet-based hidden Markov models: application to seismic signals of volcanic origin. *Statistics in Volcanology*, (1), 161-174.

Astiz, L., Eakins, J. A., Martynov, V. G., Cox, T. A., Tytell, J., Reyes, J. C., ... & Davis, G. A. (2014). The Array Network Facility seismic bulletin: Products and an unbiased view of United States seismicity. *Seismological Research Letters*, *85*(3), 576-593.

Baumgardt, D. R., & Young, G. B. (1990). Regional seismic waveform discriminants and case-based event identification using regional arrays. *Bulletin of the Seismological Society of America*, *80*(6B), 1874-1892.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157-166.

Beyreuther, M., C. Hammer, J. Wassermann, M. Ohrnberger, and T. Megies (2012), Constructing a Hidden Markov Model based earthquake detector: application to induced 946 seismicity, Geophys. J. Int., 189, 602–610, doi:10.1111/j.1365-246X.2012.05361.x

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*(Aug), 2493-2537.

Dammeier, F., Moore, J. R., Hammer, C., Haslinger, F., & Loew, S. (2016). Automatic detection of alpine rockslides in continuous seismic data using hidden Markov models. Journal of Geophysical Research: Earth Surface, 121(2), 351-371.

Dowla, F. U., Taylor, S. R., & Anderson, R. W. (1990). Seismic discrimination with artificial neural networks: preliminary results with regional spectral data. *Bulletin of the Seismological Society of America*, *80*(5), 1346-1373.

Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M., & Scarpetta, S. (2003). Discrimination of earthquakes and underwater explosions using neural networks. *Bulletin of the Seismological Society of America*, *93*(1), 215-223.

Dong, L., Li, X., & Xie, G. (2014, February). Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive Bayes classification. In *Abstract and Applied Analysis* (Vol. 2014). Hindawi Publishing Corporation.

Dowding, C. H. (1985). *Blast vibration monitoring and control* (Vol. 297). Englewood Cliffs: Prentice-Hall.

Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on* (pp. 273-278). IEEE.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, 753-753.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Guo, J. (2013). Backpropagation through time. *Unpubl. ms., Harbin Institute of Technology*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hammer, C., Ohrnberger, M. and Schlindwein, V. (2015). Pattern of cryospheric seismic events observed at Ekström ice shelf, Antarctica , *Geophysical Research Letters*, 42 . doi: 10.1002/2015GL064029

Hammer, C., M. Ohrnberger, and D. Faeh (2013), Classifying seismic waveforms from scratch: a case study in the alpine environment, *Geophysical Journal International*, 192(1), doi:10.1093/gji/ggs036

Ibáñez, J. M., Benítez, C., Gutiérrez, L. A., Cortés, G., García-Yeguas, A., & Alguacil, G. (2009). The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes. *Journal of Volcanology and Geothermal Research*, *187*(3), 218-226.

Keller, G. R., R. B. Smith, and L. R. Braile (1975). Crustal structure along the Great Basin Colorado Plateau transition from seismic refraction studies, J. Geophys. Res. 80, 1093-1098.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2018). Machine learning in seismology: turning data into insights. *Seismological Research Letters*.

Knapmeyer-Endrun, B., and C. Hammer (2015), Identification of new events in Apollo

16 lunar seismic data by Hidden Markov Model-based event detection and classification, J. Geophys. Res. Planets, doi:10.1002/2015JE004862.

Koper, K. D., Pechmann, J. C., Burlacu, R., Pankow, K. L., Stein, J., Hale, J. M., ... & McCarter, M. K. (2016). Magnitude-based discrimination of man-made seismic events from naturally occurring earthquakes in Utah, USA. *Geophysical Research Letters*, *43*(20).

Kortström, J., Uski, M., & Tiira, T. (2016). Automatic classification of seismic events within a regional seismograph network. *Computers & Geosciences*, *87*, 22-30.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Langer, H., Falsaperla, S., Masotti, M., Campanini, R., Spampinato, S., & Messina, A. (2009). Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy. *Geophysical Journal International*, *178*(2), 1132-1144.

Langefors, U. (1978). Kihlstrom. *The Modern Technique of Rock Blasting, Sweden*.

Lee, K., Lee, H., Lee, K., & Shin, J. (2017). Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Li, C. J., Li, L., Qian, J., & Liu, J. G. (2017). Batch Size Matters: A Diffusion Approximation Framework on Nonconvex Stochastic Gradient Descent. *arXiv preprint arXiv:1705.07562*.

Loeb, D. T. (1986). The P-wave velocity structure of the crust-mantle boundary beneath Utah, M. S. Thesis, University of Utah, Salt Lake City, Utah, 126 pp.

Loffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (pp. 448-456).

Mayeda, K. (1993). mb (LgCoda): A stable single station estimator of magnitude. *Bulletin of the Seismological Society of America*, *83*(3), 851-861.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529.

O'Rourke, C. T., & Baker, G. E. (2016). A Spectrogram-Based Method of Rg Detection for Explosion Monitoring. *Bulletin of the Seismological Society of America*.

Ohrnberger, M., (2001) Continous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia. Ph.D. thesis, Institut fu¨r Geowissenschaften, Universita¨t Postdam.

Pechmann, J.C., W. D. Richins, and R. B. Smith (1984). Evidence for a “double moho” beneath the Wasatch front, Utah, EOS, Trans. Am. Geophys. Union 65, 988.

Perol, T., Gharbi, M., & Denolle, M. (2017). Convolutional Neural Network for Earthquake Detection and Location.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93-104.

Roller, J.C. (1965). Crustal structure in the eastern Colorado Plateau province from seismic-refraction measurements, Bull. Seism. Soc. Am. 55, 107-119.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4580-4584). IEEE.

Scarpetta, S., Giudicepietro, F., Ezin, E. C., Petrosino, S., Del Pezzo, E., Martini, M., & Marinaro, M. (2005). Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. *Bulletin of the Seismological Society of America*, *95*(1), 185-196.

Shore, J., & Johnson, R. (1981). Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, *27*(4), 472-482.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

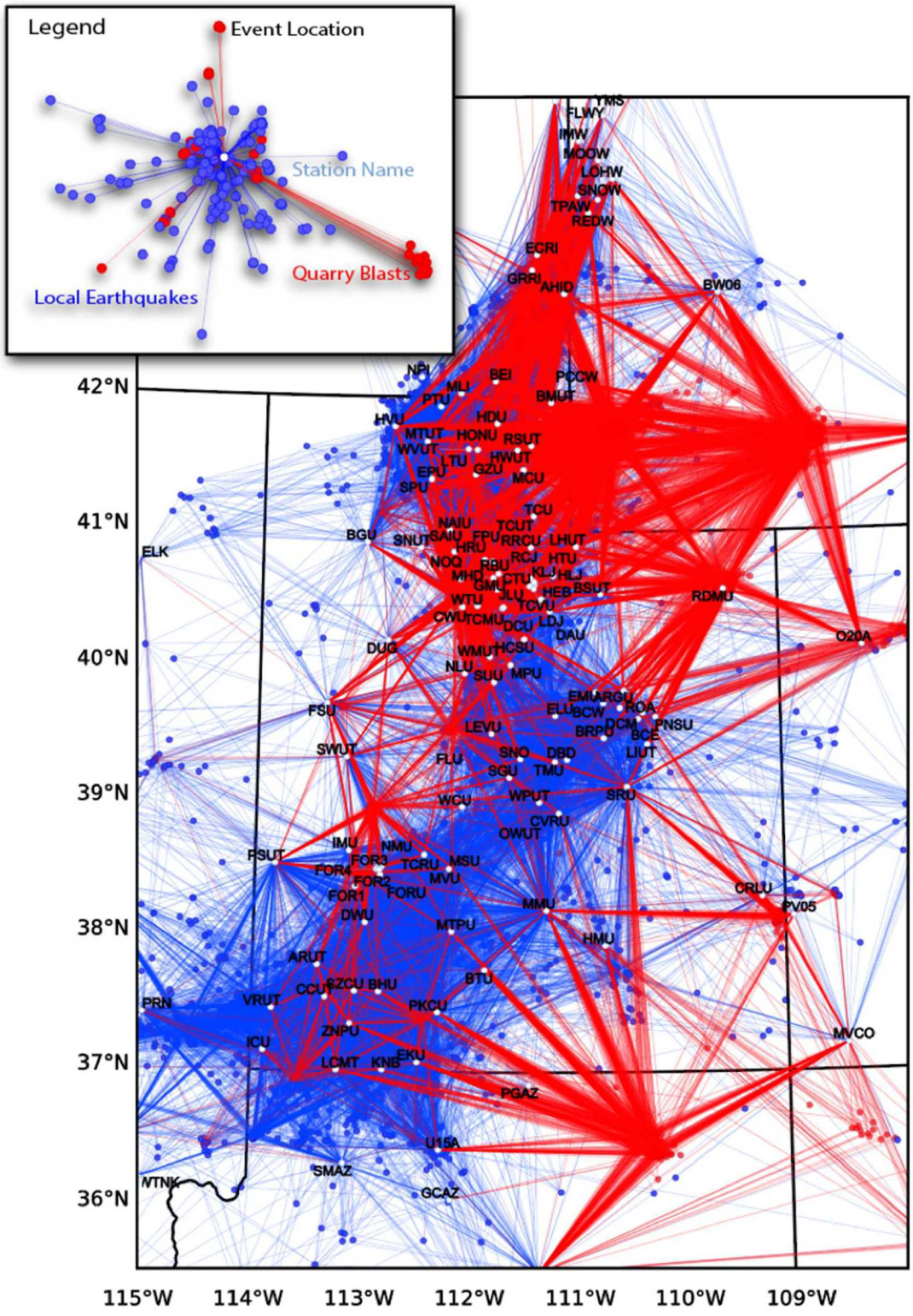
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

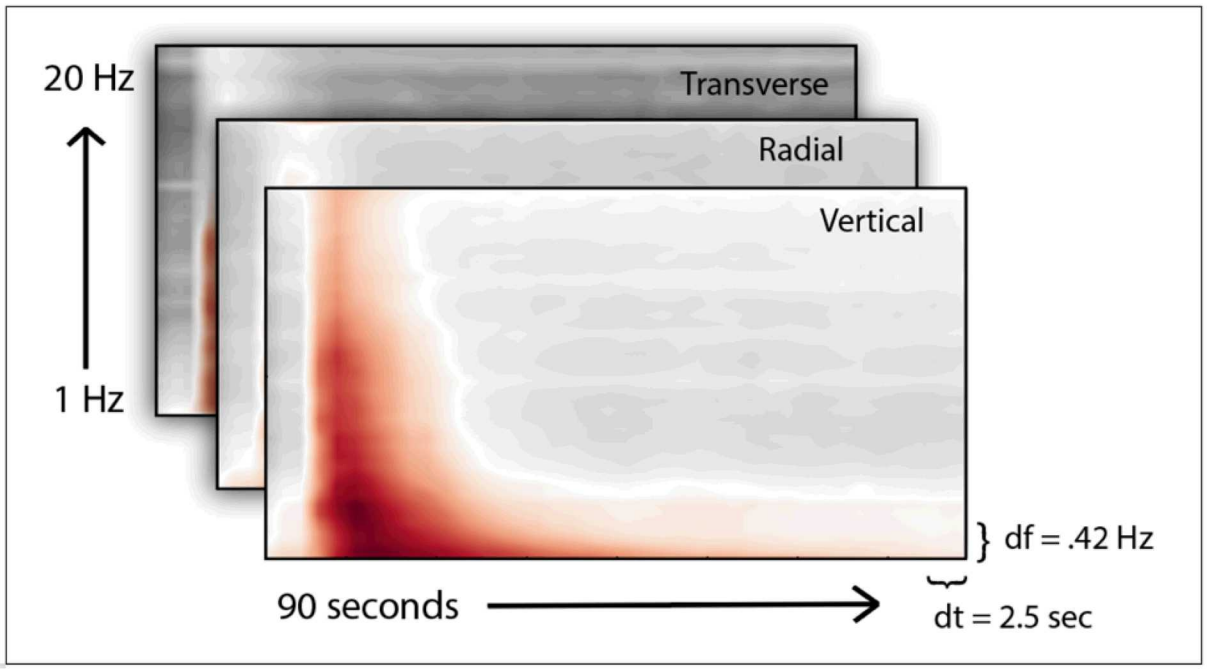
Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, *15*(1), 1929-1958.

Stump, B. W., Hedlin, M. A., Pearson, D. C., & Hsu, V. (2002). Characterization of mining explosions at regional distances: Implications with the International Monitoring System. *Reviews of Geophysics*, *40*(4).

552 Su, F., Aki, K., & Biswas, N. N. (1991). Discriminating quarry blasts from earthquakes using
553 coda waves. *Bulletin of the Seismological Society of America*, *81*(1), 162-178.
554
555 Tibi, R., Koper, K. D., Pankow, K. L., & Young, C. J. (2018). Depth Discrimination Using Rg-to-
556 Sg Spectral Amplitude Ratios for Seismic Events in Utah Recorded at Local
557 Distances. *Bulletin of the Seismological Society of America*, *108*(3A), 1355-1368.
558
559 Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings*
560 *of the IEEE*, *78*(10), 1550-1560.
561

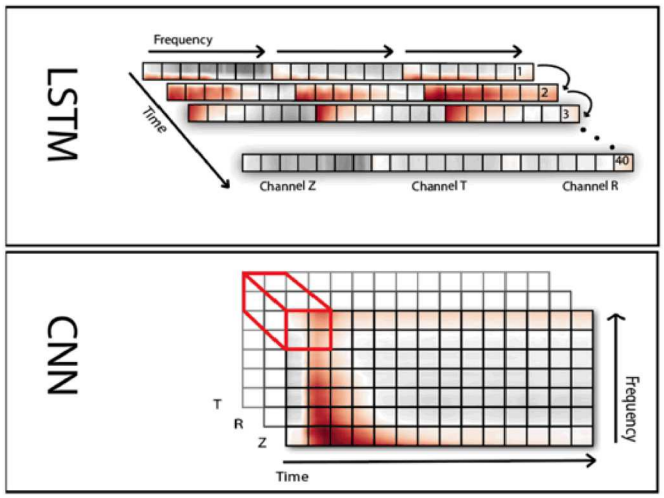

Legend
Event Location
Station Name
Quarry Blasts
Local Earthquakes
42°N
41°N
40°N
39°N
38°N
37°N
36°N
115°W
114°W
113°W
112°W
111°W
110°W
109°W
YMS
FLWY
IMW
MOOW
LOHW
SNOW
TPAW
REDW
ECRI
GRRI
AHID
BW06
NPI
BEI
PCCW
MLI
PTU
BMUT
HVU
HDU
MTUT
HONU
RSUT
WVUT
HWUT
LTU
GZU
EPU
MCU
SPU
TCU
NAIU
TCUT
BGU
SNUT
SAIU
FPU
RRCU
LHUT
HRU
ELK
NOQ
RBU
RCJ
HTU
MHD
CTU
KLJ
HLJ
GMU
JLU
HEB
BSUT
RDMU
WTU
TCVU
CWU
TCMU
DCU
LDJ
DAU
O20A
DUG
WMU
HCSU
NLU
SUU
MPU
FSU
EMU
RQU
ELU
BCW
ROA
LEVU
BRPU
DCM
PNSU
BCE
SWUT
FLU
SNO
DBD
LIUT
SGU
TMU
WCU
WPUT
SRU
CVRU
OWUT
PSUT
IMU
NMU
FOR3
TCRU
MSU
FOR4
FOR2
MVU
FOR1
FORU
CRLU
PV05
DWU
MMU
MTPU
HMU
ARUT
BTU
CCU
SZCU
BHU
PRN
VRUT
PKCU
ZNPU
MVCO
ICU
LCMT
KNB
EKU
PGAZ
U15A
SMAZ
MTNK
GCAZ

562



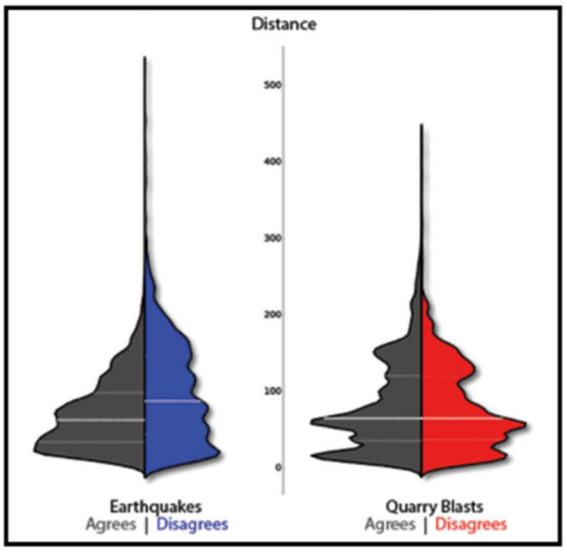
Long Short Term Memory
Hochreiter and Schmidhuber 1997
40x144 input
4 layers 40,40,80,80

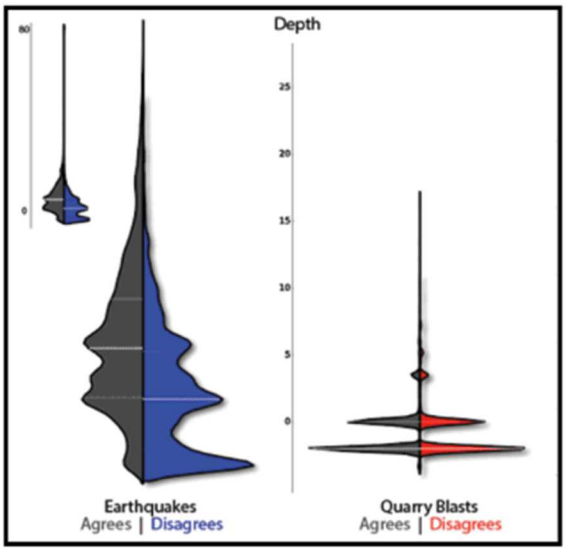
Convolutional Neural Network
LeCun et al. 1998
3x40x48 input
4 layers
Filters -18,36,54,54
(max) Pooling – 2x2

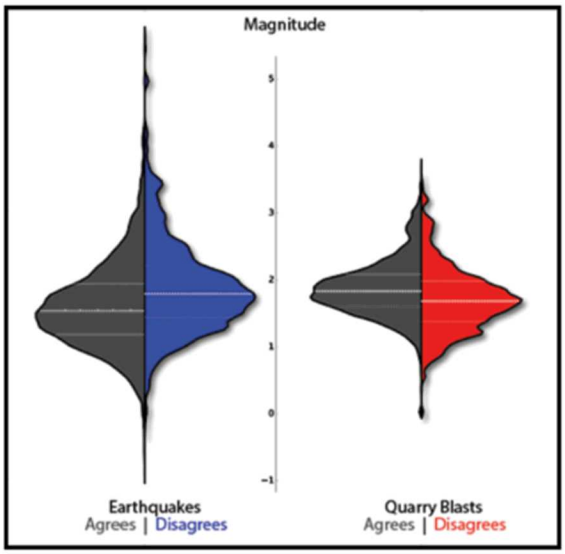


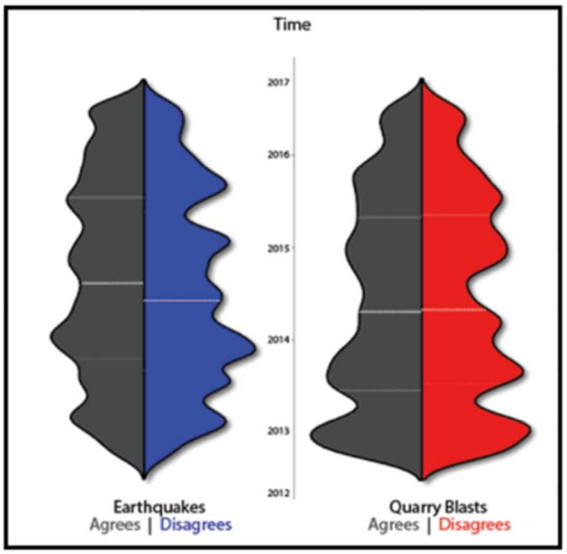
563

	CNN ACCURACY		LSTM ACCURACY	
	Station	Network	Station	Network
LE	95.8 ± .9	99.1 ± .5	96.0 ± .9	99.0 ± .5
QB	95.8 ± .8	99.3 ± .3	95.9 ± .8	99.5 ± .4



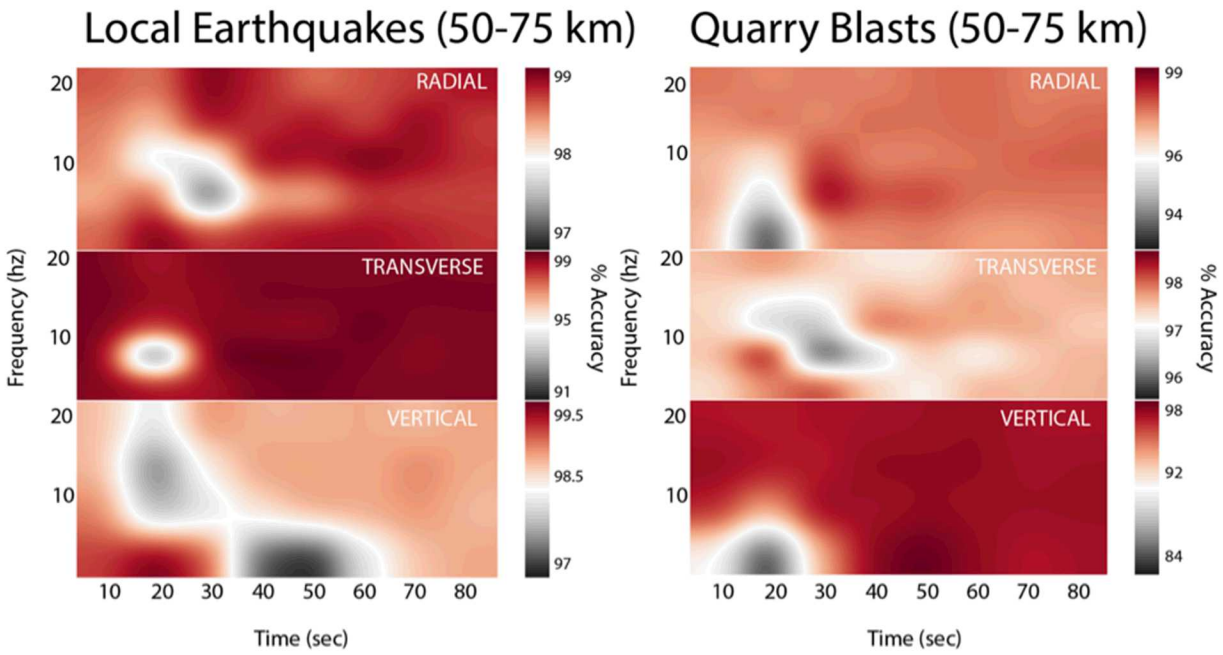






564

565


Local Earthquakes (50-75 km)
Quarry Blasts (50-75 km)
RADIAL
TRANSVERSE
VERTICAL
Frequency (hz)
Time (sec)
% Accuracy
20
10
10 20 30 40 50 60 70 80
99
98
97
95
91
99.5
98.5
96
94
92
84

566