Geophysical Research Letters

RESEARCH LETTER
10.1029/2018GL081119

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
- The three-component data are typically more accurate but not required for prediction

Supporting Information:

- Supporting Information S1

Correspondence to:
L. Linville,
linville@seis.utah.edu

Citation:
Linville, L., Pankow, K., & Draelos, T. (2019). Deep learning models augment analyst decisions for event discrimination. *Geophysical Research Letters*, *46*. https://doi.org/10.1029/2018GL081119

Received 6 NOV 2018
Accepted 20 FEB 2019
Accepted article online 25 FEB 2019

Deep Learning Models Augment Analyst Decisions for Event Discrimination

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

[1]University of Utah Seismograph Stations, University of Utah, Salt Lake City, UT, USA, [2]Geophysics Department, Sandia National Laboratories, Albuquerque, NM, USA

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 three-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 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.

1. 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 semiautomated 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 & Pulli, 1990; Baumgardt & Young, 1990; Mayeda, 1993; Rg: O'Rourke & Baker, 2016; Tibi et al., 2018). Although amplitude ratio methods perform well when tuned for specific data sets, their use can be limited when ratios rely on wave phases that are unavailable (e.g., at larger distances), 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 (Su et al., 1991). For example, Koper et al. (2016) 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.

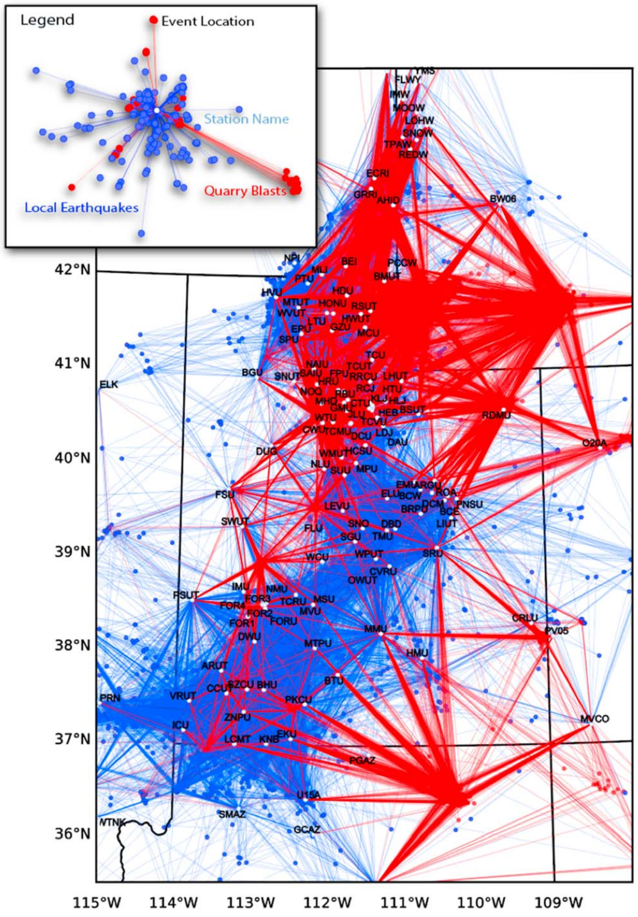


Figure 1. Map of events (circles) and source-receiver paths (lines) from University of Utah Seismograph Stations for quarry blasts (red) and local earthquakes (blue). Receivers (white circles) are labeled by station name. The University of Utah Seismograph Stations catalog includes events outside of the authoritative catalog boundaries (latitude: 37° to 42°N, longitude: 114° to 109°W). The data used in this study include 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.

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 subkilometer scale up to 400 km. Although deep NNs are clearly competent at event discrimination using Utah data, they do so at the cost of transparency; the complex data routing through the nonlinear modular hierarchy in model space obfuscate 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.

2. Data

This study uses finalized (manually reviewed) event solutions including phase arrival times, earthquake locations and magnitudes 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 s) for local earthquakes, and 2 or higher (±0.15 s) for quarry blasts. The number of examples from each class is balanced within 8% (46% quarry blasts and 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 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 one to three 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 1 Hz (Butterworth, four-corner). For three-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. Three-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.

3. Method

NNs 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; Dysart & Pulli, 1990). We use two specific architectural adaptations to basic NNs in the work presented here (Figure 2). Our first architecture is a recurrent neural network variation called Long-Short-Term-Memory (LSTM; Hochreiter & Schmidhuber, 1997). This adaptation of the recurrent neural network architecture solves issues introduced by back-propagating gradients used to train the network over many time

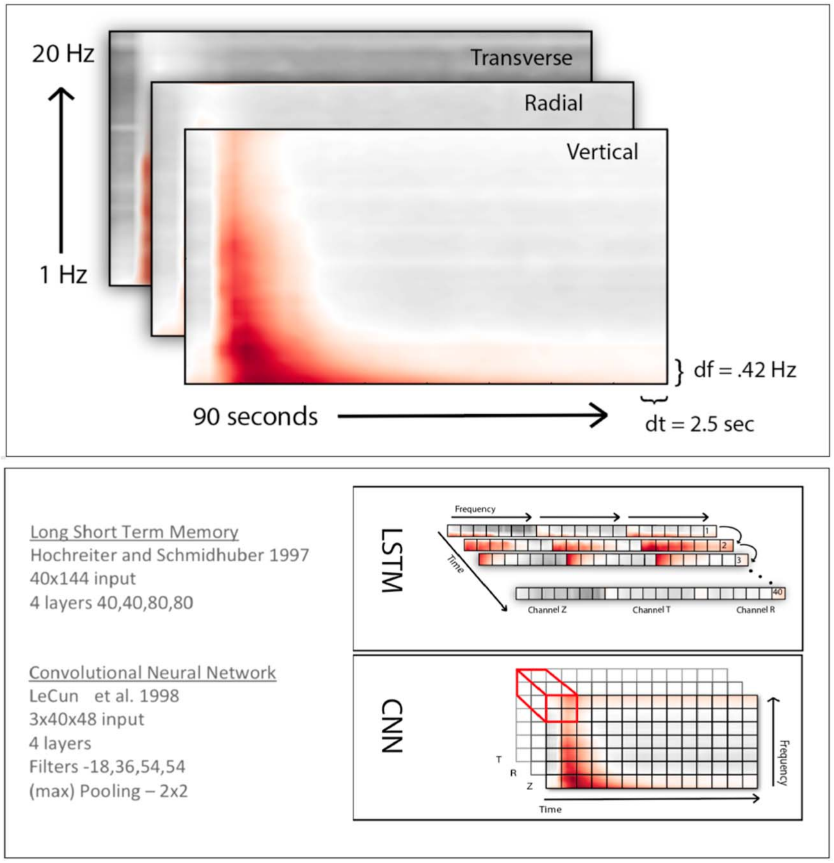


Figure 2. (top) Three-channel 90-s event spectrogram model input. (bottom) Permutations of three-channel spectrograms for model input. In the Long-Short-Term-Memory (LSTM) architecture the spectrogram enters the model as a 40 × 144 matrix (time step and frequency feature) where the 144 dimension is a three-channel frequency power spectral density vector for each time step. The convolutional neural network (CNN) architecture takes example spectrograms as 40 × 48 × 3 volumes (time, frequency, and channel). The red box indicates the starting computation using a 2 × 2 pixel filter that shares weights along the depth axis of the volume.

steps (Bengio et al., 1994). We also use bidirectional layers to give both past and future context to current examples (Graves et al., 2005, 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 four layers (node counts: 40, 40, 80, and 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 in the supporting information 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 data set (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, and 54), one fully connected layer, rectified linear units ($f(x) = \max(0, x)$) 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 tenfold cross validation (80% training, 10% validation, and 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 eight iterations (epochs) through the training data and choose as a final model the earliest epoch with the highest validation accuracy.

4. Results

Figure 3 presents cross-validation results for each model (LSTM and CNN), with accuracy reported at both the station level and network levels for data within the UUSS authoritative review boundary (latitude 37° to 42°N, longitude: 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 one 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 supporting information 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%.

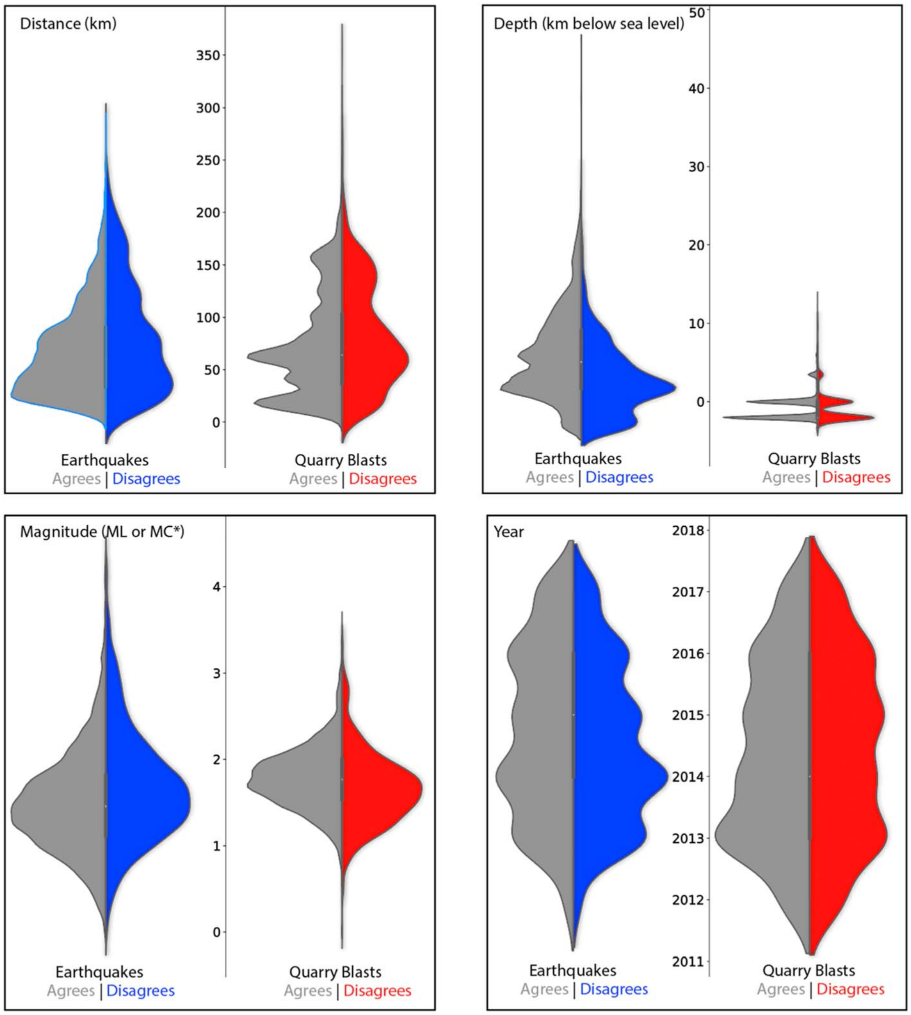
There are 103,944 examples available for model training and testing in the Utah. The 50,581 examples (49%) are from single-component (vertical) stations, with null horizontal channels values. For our data, models trained on individual channels give nearly the same reported accuracies (vertical = 93.8%, radial = 93.9%, and 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-min window following the event onset. Models trained on 5-min duration and a 90-s duration yield equivalent accuracy. Investigating how short a window could be used, we found that using just the first four spectral windows following event onset (~9 s), a trained CNN (two to four layers) accomplished station-level accuracy of 84.7 ± 0.7% and event-level accuracy of 94.4 ± 0.9% using tenfold 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, magnitude, depth, or distance does not 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., 2017). Most (92%)

	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



*local magnitude (ML) or coda magnitude (MC)

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 normalized distributions for misclassified (disagrees; 3401 examples) and the correctly classified (agrees; 81,626 examples) station examples in Utah as a function of distance, depth, magnitude, and time. We find that misclassified local earthquakes on average have source-receiver distances 18 km longer than correct classifications, 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. In spite of these subtle differences, no single characteristic or simple combination accounts for a majority of misclassifications. CNN = convolutional neural network; LSTM = Long-Short-Term-Memory.

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 0.8, 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

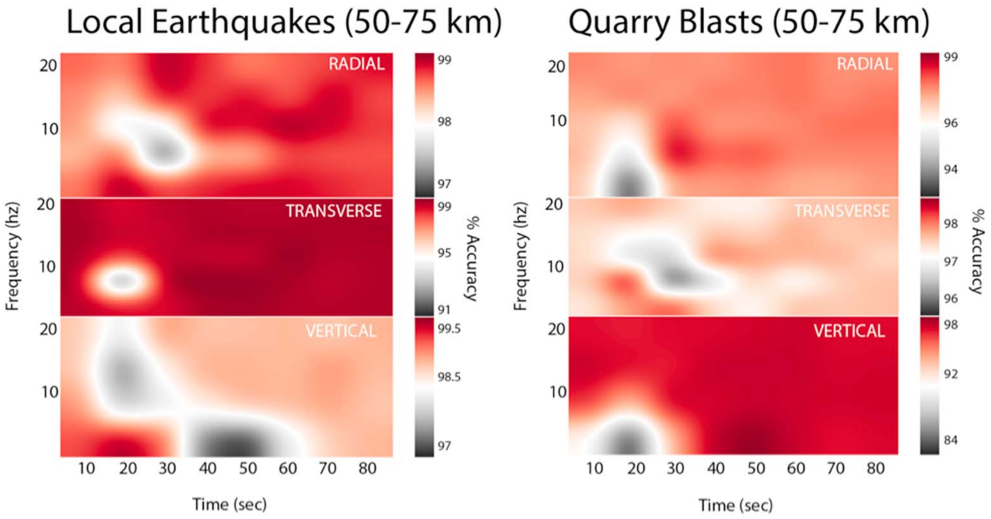


Figure 4. Regions of the spectrogram most important for model predictions. To identify the regions of event spectrograms that have the largest influence on average prediction we feed 1,000 randomly selected spectrograms from a fixed distance bin (in this example 50–75 km) for each class and use a randomly selected trained convolutional neural network model to predict the class accuracy when sections (2 × 5 patches) of the input include no signal (are zero). Gray sections of the heatmap for each class indicate regions where the largest decreases in accuracy occur in the absence of signal from corresponding locations of event spectrograms.

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 supporting information of 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; Simonyan 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) 2 × 5 nonoverlapping pixel blocks from randomly selected input examples and stack the resulting accuracy over 1,000 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 nonevent 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- to 10-Hz *S* wave energy on the transverse component causes the largest decrease in prediction certainty for local earthquakes at the 50- to 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 a way to link fluctuations in model certainty to specific wave phases (Figure S3 available in the supporting information of this article).

5. Discussion

There is a clear temporal progression expressed in the records of seismic sources captured by distant receivers. This is the result of a complex source coupled with wave propagation through a complex earth medium. In our data, examples sample discrete portions of the crust over various distances with at least four distinct velocity profiles (Keller et al., 1975; Loeb, 1986; Pechmann et al., 1984; Roller, 1965). Previous work has navigated some of the complexity in waveforms through feature engineering and data segmentation (e.g., using only information from specific wave phases that are precut from the data). The diversity of training examples available in even modest duration catalogs appears to be sufficient to generalize well for earthquakes sources in Utah. Quarry blast event accuracy depends more highly on representative samples with higher classification certainty, indicating that model retraining may be required over time for new mine sites, mine products or extraction techniques.

Although LSTM models are a more intuitive choice for time-dependent signal modeling, CNN models appear to be equally as adept at modeling the variation within Utah data given the data representation. Our results suggest that LSTM models may have the capacity to model time dependence that far exceeds the capacity required to model the 40 time steps of our input. Likewise, we achieve no additional gains in accuracy by increasing CNN model depth beyond four layers (the 11-layer CNN model of Simonyan & Zisserman, 2014, for example gives equivalent accuracies). It is generally observed that feature complexity increases with CNN model depth (Figure S3 available in the supporting information of 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 LER. 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 s 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 two associated station examples).

An additional expectation was that models handle data from both single-component and three-channel stations. While sufficient information for classification exists in each channel individually (R, T, and 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 (Dowding, 1985; Langefors, 1978). 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.

6. 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 LER (~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.

Acknowledgments
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 web services and Obspy. Data processing relied on python libraries scipy and numpy. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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 U.S. Government.

References

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

Astiz, L., Eakins, J. A., Martynov, V. G., Cox, T. A., Tytell, J., Reyes, J. C., et al. (2014). The Array network facility seismic bulletin: Products and an unbiased view of United States seismicity. *Seismological Research Letters*, *85*(3), 576–593. https://doi.org/10.1785/0220130141

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. https://doi.org/10.1109/72.279181

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. https://doi.org/10.1785/0120020005

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

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.

Dysart, P. S., & Pulli, J. J. (1990). Regional seismic event classification at the NORESS array: seismological measurements and the use of trained neural networks. *Bulletin of the Seismological Society of America*, *80*(6B), 1910–1933.

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 (pp. 753–753).

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649). IEEE.

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Keller, G. R., Smith, R. B., & Braile, L. R. (1975). Crustal structure along the Great Basin Colorado plateau transition from seismic refraction studies. *Journal of Geophysical Research*, *80*(8), 1093–1098. https://doi.org/10.1029/JB080i008p01093

Koper, K. D., Pechmann, J. C., Burlacu, R., Pankow, K. L., Stein, J., Hale, J. M., et al. (2016). Magnitude-based discrimination of man-made seismic events from naturally occurring earthquakes in Utah, USA. *Geophysical Research Letters*, *43*, 10,638–10,645. https://doi.org/10.1002/2016GL070742

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Langefors, U. (1978). *The modern technique of rock blasting*. Sweden: Kihlstrom.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

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

Loeb, D. T. (1986). The *P*-wave velocity structure of the crust-mantle boundary beneath Utah, (MS thesis, 126 pp.). University of Utah, Salt Lake City, UT.

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

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*.

Pechmann, J. C., Richins, W. D., & Smith, R. B. (1984). Evidence for a "double moho" beneath the Wasatch front, Utah, EOS. *Transactions of the American Geophysical Union*, *65*, 988.

Perol, T., Gharbi, M., & Denolle, M. (2017). Convolutional neural network for earthquake detection and location. *Science Advances*, *4*(2), e1700578.

Roller, J. C. (1965). Crustal structure in the eastern Colorado Plateau province from seismic-refraction measurements. *Bulletin of the Seismological Society of America*, *55*, 107–119.

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5A), 2894–2901.

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

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

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

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), 1011. https://doi.org/10.1029/1998RG000048

Su, F., Aki, K., & Biswas, N. N. (1991). Discriminating quarry blasts from earthquakes using coda waves. *Bulletin of the Seismological Society of America*, *81*(1), 162–178.

Tibi, R., Koper, K. D., Pankow, K. L., & Young, C. J. (2018). Depth discrimination using Rg-to-Sg spectral amplitude ratios for seismic events in Utah recorded at local distances. *Bulletin of the Seismological Society of America*, *108*(3A), 1355–1368. https://doi.org/10.1785/0120170257

References From the Supporting Information

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.

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.

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

Shore, J., & Johnson, R. (1981). Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, *27*(4), 472–482. https://doi.org/10.1109/TIT.1981.1056373

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.

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, *78*(10), 1550–1560. https://doi.org/10.1109/5.58337