SAND2019-14938J

Benchmarking Current and Emerging Approaches to Infrasound Signal Classification

Sarah Albert[1, a)] **and Lisa Linville**[1]

Sandia National Laboratories, Albuquerque, NM 87123, USA

salber@sandia.gov,

llinvil@sandia.gov

(Dated: 5 November 2019)

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.

Abstract: Low frequency sound $\leq$ 20 Hz, known as infrasound, is generated by a variety of natural and anthropogenic sources. Following an event, infrasonic waves travel through a dynamic atmosphere that can change on the order of minutes. This makes infrasound event classification a difficult problem as waveforms from the same source type can look drastically different. Event classification usually requires ground truth information from seismic or other methods. This is time consuming, inefficient, and does not allow for classification if the event locates somewhere other than a known source, the location accuracy is poor, or ground truth from seismic data is lacking. Here we compare the performance of the state of the art for infrasound event classification, support vector machine (SVM), to the performance of a convolutional neural network (CNN), a method that has been proven in tangential fields such as seismology. For a 2-class catalog of only volcanic activity and earthquake events, the 4-fold average SVM classification accuracy is 75%, while it is 74% when using a CNN. Classification accuracies from the 4-class catalog consisting of the most common infrasound events detected at the global scale are 55% and 56% for the SVM and CNN architectures, respectively. These results demonstrate that using a CNN does not increase performance for infrasound event classification. This suggests that SVM should be the preferred classification method as it is a simpler and more trustworthy architecture and can be tied to the physical properties of the waveforms. The SVM and CNN algorithms described in this paper are not yet generalizable to other infrasound event catalogs. We anticipate this study to be a starting point for development of large and comprehensive, systematically labeled, infrasound event catalogs as such catalogs will be necessary to provide an increase in the value of deep learning on event classification.

Keywords: infrasound, support vector machine, convolutional neural network, classification, global monitoring

[a] Author to whom correspondence should be addressed.

1. Introduction

Infrasound ($\leq$ 20 Hz) is generated by a variety of natural and anthropogenic sources including ocean waves, volcanoes, mountains, chemical explosions, mining blasts, and rockets. Infrasonic waves travel through a dynamic atmosphere where temperature, wind speed, and wind direction can change drastically on the order of minutes. A change in atmospheric structure ultimately changes the shape of the recorded waveform, even when the source-receiver path remains the same. Changes in waveform morphology due to atmospheric structure have been observed from repeating sources over several days (Gibbons *et al.*, 2015), down to 20 minutes (Kulichkov, 2004). Signal durations and frequency characteristics are also dependent on the structure of the waveguide (Ceranna *et al.*, 2009; Green and Nippress, 2019). Lastly, amplitude attenuation occurs as signals propagate through the atmosphere. This attenuation is highly dependent on the structure of the upper atmosphere (Smets and Evers, 2014). The dependance of waveform morphology on atmospheric structure results in signals from the same source showing a variety of waveform characteristics. This makes infrasound signal classification a difficult problem. We are trying to classify a property (the source type) that is unaffected by atmospheric propagation. It is nearly impossible for an analyst to identify the source type based on the infrasound waveforms alone because of their dependence on atmospheric structure. Analysts may be able to use the event location, if known, to understand the source (i.e. if the location matches a known source such as a volcano). However, ground truth information from seismic data is often used to classify infrasound signals. This is time consuming, inefficient, and does not allow for a classification if the event locates somewhere other than a known source, the location accuracy is poor, or ground truth from seismic data is lacking. For the purposes of this paper we will refer to an infrasound waveform as a “signal”, while a collection of signals resulting from the same source will be referred to as an “event”.

Common methods from seismology, such as waveform cross correlation and template matching, work poorly for infrasound signal classification because infrasonic waveforms are so strongly affected by the atmospheric regime through which they travel. Bowman and Albert (2018) show that changes in the atmosphere (such as a developing storm) can affect both the amplitude and shape of infrasonic waveforms with the same source-receiver geometry, even when they are only separated in time by 90 minutes. Therefore, it is nearly impossible to distinguish between source types from time-series waveforms alone. However, the physical mechanism for each source type is fundamentally different, so their signals may contain unique frequency components that can be exploited using robust methods.

Infrasound signal classification remains of particular interest for national security and hazard mitigation. The International Monitoring System (IMS) was developed by the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) to monitor for nuclear explosions. It remains under development and will eventually consist of 60 global infrasound stations. At the time of this publication, the IMS infrasound network consists of 52 infrasound stations, though fewer were present in the past when most of the signals in our catalog were recorded. The International Data Centre (IDC), which is also part of the CTBTO, is responsible for processing all of the IMS data. Arrowsmith (2018) computed a “Genuine

False Alarm" (GFA) rate, using the same infrasound event processing method that the IDC uses, by attempting to associate simulated, unassociated, detections on two or more stations. Results suggest a GFA rate of ∼200 events per day, generating a high false alarm rate and placing a large workload on analysts. However, it is important to note that this is an estimate from Arrowsmith (2018). The IDC applies post-processing methods in an attempt to further reduce the false alarm rate. A reliable classification algorithm would allow analysts to focus on events of interest rather than false alarms. From a hazard mitigation standpoint, signal classification could provide eruption warnings in areas where seismic or other data is limited.

Machine learning has shown high accuracy in classifying infrasound signals at local ($<$ 15 km), regional (15 - 250 km), and global ($>$ 250 km) distances. For example, Ham and Park (2002) showed that their neural network (NN) classified infrasound signals as a volcano, mountain associated waves, impulsive, or no event, and were able to achieve accuracies greater than 90%. It is important to note that this study used only signals detected at a single station and the same source-station pairs (signals of a specific class are always detected at the same station) for the volcano and mountain associated-wave classes. Also, for the impulsive events class, many signals were recorded at local distances (100 - 1000 m). Therefore, atmospheric structure played a smaller role in waveform morphology and the impulsive signals likely looked very similar. Two previous studies implemented a method using support vector machine (SVM) to reach accuracies of 97.7% and 86.36%, depending on the feature extraction method. They classified infrasound signals as being from an earthquake, volcano, or tsunami (Li *et al.*, 2016; Liu *et al.*, 2014). Both studies use the same catalog, which is also limited to single-station signals, though source-station pairs differ slightly. They use 2, 4, and 3 source-station pairs for volcano, tsunami, and earthquake signals, respectively.

These previous studies are the state of the art for classifying infrasound signals, and work well on simple datasets that are limited in geographic area and/or source diversity. It is unclear how well these methods transfer to more complex datasets representative of those encountered in real-time monitoring operations. For example, catalogs consisting of signals detected at a variety of stations and/or multiple source-station pairs. In tangential fields such as seismology, deep neural network (DNN) based approaches to data driven problems have been highly successful (Linville *et al.*, 2018; Perol *et al.*, 2018; Ross *et al.*, 2018). However, typically the success of these methods is due to the large and comprehensive datasets available for model training. While the generation of large labeled event catalogs is often standardized in other fields, this has yet to happen within the infrasound community. At many data centers, infrasound data processing is often not done at the same level and only simple catalogs are built. Therefore, many infrasound catalogs are plagued by non-standard labels, automated arrival picks, and/or quick analyst review. There is also no guidance as to how large a catalog must be to be considered comprehensive for deep learning analysis.

Despite the inherent difficulties of infrasound signal processing, routine infrasound processing is moving towards realization. Many data centers, including the IDC, are hiring infrasound analysts and building event catalogs. Development of more general methods of source identification may be required for use in these larger scale monitoring environments. Achieving better source generalization through deep learning techniques has previously been inaccessible due to the quantity and quality of event catalogs needed for model training. For this study we obtain a comprehensive labeled event catalog produced by the IDC. With this new catalog we evaluate approaches to automate event classification with consideration for realistic challenges encountered in global scale infrasound event monitoring. We explore classification of the four most common infrasound source types using two data-driven approaches. As a community we currently have limited guidance on what is possible with the quality and size of catalogs and emerging methods for predictive modeling. Therefore, this study provides (1) analysis to benchmark infrasound signal classification methods and (2) cautions and suggestions as we move toward more complete and comprehensive catalogs and the need for automated data processing strategies in the future.

2. Infrasound Event Catalog

We make use of the Infrasound Reference Event Database (IRED) produced by the IDC on July 9, 2010. The catalog contains signals from infrasound events detected at IMS stations found worldwide (Fig. 1). Up to 42 stations were installed, depending on the date, since stations are added over time. It was designed to serve as a reference catalog for analysts and is therefore not comprehensive, but rather contains a subset of events that is representative of the those detected on IMS stations. All events are reviewed by an analyst and have been verified with other ground truth information from seismic data, satellite data, etc. The catalog contains a total of 786 signals from a variety of sources. Table 1 shows the number of signals for each event type. We note that the classes are imbalanced with respect to the number of signals in each one. The classes themselves can also vary considerably in regards to the physical mechanisms generating each event. For example, the anthropogenic activity class contains signals from fireworks *and* trains. Classifying signals as anthropogenic activity then becomes more difficult since the labels are somewhat arbitrary. Splitting the class into two does not fix the problem because there are so few example signals. Therefore, we chose to focus on a subset of the most abundant signals including those from mines and quarries, chemical/accidental explosions, earthquakes, and volcanic activity. This provided us with a total of 615 signals from 519 infrasound events. Of these events, over 10% (69 signals) were detected at more than one station providing 21, 26, 25, and 16 source-station pairs for mines and quarries, volcanic activity, earthquakes, and accidental/chemical explosions, respectively. The average distance from source to sensor is 1185 km, the median is 860 km, and the standard deviation is 1492 km. There is considerable variability in source-sensor distances, but most of the signals in the catalog are detected at regional and global distances (15 - 250+ km). Figure 1 shows a map of the source locations and straight-line paths connecting origins to various detecting stations. Note that some locations share paths from multiple events.

The IRED catalog is small compared to the size of typical catalogs used to train predictive models in tangential domains. These catalogs benefit from a hundred thousand (Linville *et al.*, 2019) to a million signals (Ross *et al.*, 2018) and achieve impressive performance for limited geographic areas. The IRED catalog by comparison spans local and global distance sources, contains a limited number of examples for each source type, and is faced with additional complexity from dynamic travel paths.

Table 1. IRED Signals by Class

Source	Number of Signals
Mines and quarries	256
Chemical/accidental explosions	152
Earthquakes	103
Volcanic activity	104
Rocket launch/re-entry	57
Anthropogenic activity	46
Bolides and meteorites	26
Unknown	14
Aircraft	13
Cultural noise	8
Avalanches and landslides	7

3. Methods

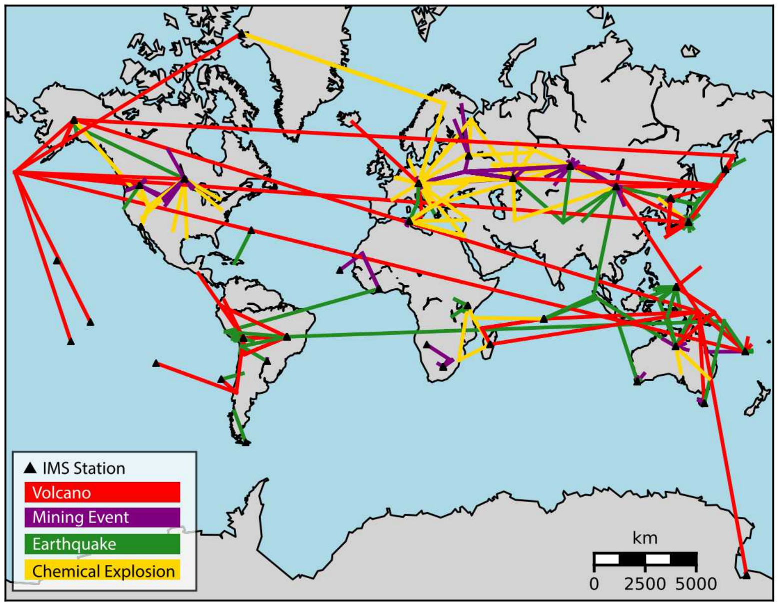

Fig. 1. A map of the infrasound events used in this study. Lines originate from the source location and connect to all stations (black triangles) that detected the event. Note that lines are straight paths and do not represent the actual path traveled by the infrasonic wave. Line colors represent event type. Most of the events are detected at global distances ($\geq$ 250 km). Note that some of the locations correspond to multiple events that may have been detected at a variety of sensor locations.

The event catalog used in this study is the largest labeled global infrasound event catalog currently available. The goal of this publication is to evaluate deep learning strategies from tangential fields in the context of the state of the art. Therefore, we use a convolutional neural network (CNN) and compare this to classification accuracies using SVM, which we consider to be the state of the art classification method for infrasound. We do this comparison for two types of catalogs. The first is a subset of the larger catalog consisting of only volcanic activity and earthquakes, designed to be similar to the catalog used by Li *et al.* (2016) and Liu *et al.* (2014). This will be referred to as the "2-class" catalog. The second is the "4-class" catalog of mines and quarries, volcanic activity, earthquakes, and accidental/chemical explosions, representing the largest currently available labeled global infrasound event catalog. We classify signals in each catalog using both SVM and CNN, and compare results 1) within each catalog and 2) from each method between the two catalogs.

3.1 Data Preparation

Each station consists of an infrasound array containing four or more infrasound microphones. The catalog entry for each event contains the following information: number of arrivals, duration, backazimuth, and trace velocity for each group of signals at each station that detected the event. These values were derived by the IDC using the Progressive Multi-Channel Cross Correlation (PMCC) detection and location algorithm (Cansi, 1995). We started by making use of the time-series waveforms and detection information from each signal. First, the waveforms were selected to begin 5 seconds prior to the earliest arrival time for each sensor. The waveforms from each station were then time-aligned based on the detection backazimuth and trace velocity from the reference sensor (given in the catalog metadata), which we will refer to as the "delay-and-sum beam". The delay-and-sum beam was used to represent the signal from that station. Therefore, events detected at multiple stations were represented by multiple waveforms. Signal durations ranged from 9 seconds to 3.5 hours depending on source variability and the number of arrivals detected by the IDC. For SVM analysis, we used the full duration of each signal and used the delay-and-sum beam as the input for feature extraction. For CNN analysis we used a fixed signal duration of 475 seconds (2 std. of median signal length). Some of the signals included multiple arrivals. This was common in the volcanic activity class, where many signals included multiple arrivals corresponding to increased volcanic activity over a period of time. We chose a signal duration of 475 seconds to minimize the risk of including these multiple arrivals in what was considered to be a single example of a signal from the source type. For signal durations less than 475 seconds long, we zero-padded the signal out to that time. We detrended the signals and applied a 1% taper for a gradual transition to zero at the edges. The 1% taper on our 475 s window only effects out to about 2 s at each edge. This avoids interference with the signal onset since signals were collected 5 seconds prior to the given arrival time. Then we computed a normalized spectrogram from the delay-and-sum beam. While signal length is variable by source and not fixed by class, some classes exhibit characteristic lengths that require caution when used as input for DNNs. We suggest that future studies may benefit from either alternative methods capable of variable signal length learning such as recurrent neural networks, or that secondary arrivals, signals, and noise are included in model input.

3.2 SVM and CNN Architectures

First, we classified signals using the state of the art, SVM. We extracted features using the spectral entropy method developed by Li *et al.* (2016), as well as other, more easily interpretable features such as distance from source to sensor, waveform duration, etc. We began with many features and selected eight using the random forest method to determine feature importance (Brieman, 2001). It is important to note that our analysis show the two most important features are distance from source and waveform duration. This suggests there may be a range dependency on signal classification. Of the original features we selected the top eight: distance from source, waveform duration, wavelet singular spectrum entropy, spectral spread, wavelet energy spectrum entropy, number of zero crossings, energy, and wavelet power spectrum entropy (Fig. 2). Each of the SVM features that include entropy aim to quantitively capture the uncertainty of the signal energy distribution in various domain. The values calculated from wavelet singular spectrum entropy reflect the uncertainty of the signal energy distribution in the time-frequency domain. The wavelet power spectrum entropy is calculated using the power spectral density of the signal and therefore reflects the uncertainty of the signal energy distribution in the frequency domain. The wavelet energy spectrum entropy uses the energy spectrogram, reflecting the uncertainty of the signal energy distribution in the time-frequency domains. These features describe physical properties of the waveforms, providing a tangible link to the source and propagation physics, as opposed to the pattern discovery approach as used by a CNN. SVM is a powerful tool for classification because it identifies the hyperplane in a high-dimensional space that maximizes the distance between points within the given classes. When the data is characterized by nonlinear relationships, SVM requires a kernel function to transform the data prior to classification. Once the features were extracted from each waveform, we input them into the SVM algorithm for training and classification using a radial basis function (rbf) kernel (Vapnik, 1995). We chose the rbf kernel because it is widely generalizable and has been proven to provide accurate classifications in other domains.

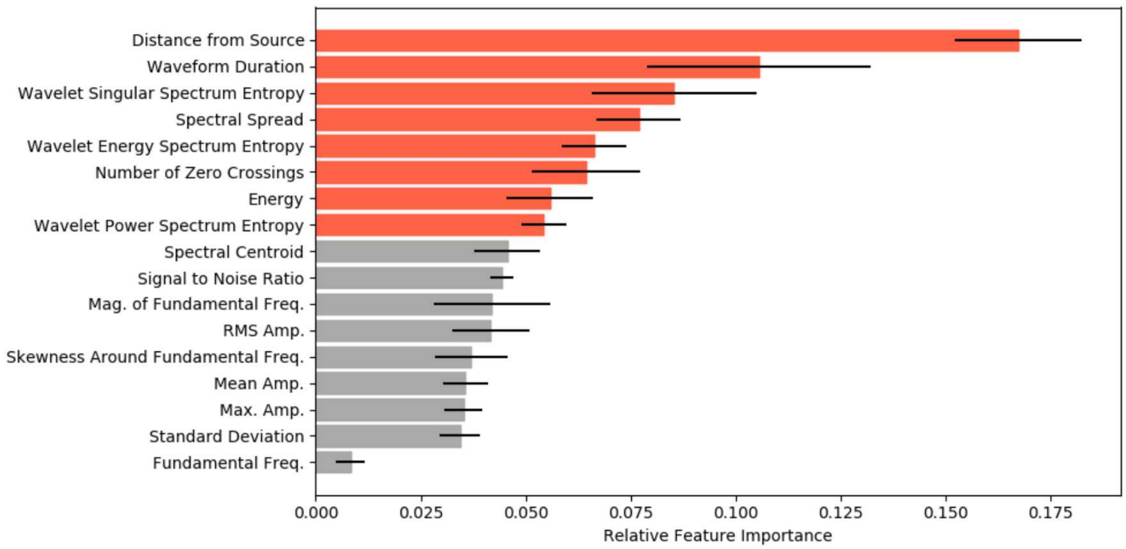

Fig. 2. Feature importance calculated using the random forest method. Features selected in this study are highlighted in red. Note that the two most important features are distance from source and waveform duration, suggesting a range dependency on classification.

The training input for our CNN algorithm consists of normalized spectrograms from each signal, shown in Figure 3. We utilize a 4-layer relu-activated CNN. More information on this method can be found in Linville *et al.* (2019). We chose a CNN because of demonstrated success in parallel domains such as discrimination, detection, and location of seismic signals (Linville *et al.*, 2018; Perol *et al.*, 2018; Ross *et al.*, 2018). It also has a higher capacity to model the data as opposed to other algorithms such as SVM. We experimented with adding gaussian and random noise to balance the classes, increase the dataset size, and augment the data, but these methods did not result in significant improvement in accuracy. Our dataset is small compared to others used for CNN classification so it is easy to overfit when augmenting the data. Strong regularization is likely an important requirement in limited data domains, but simple gaussian noise did not make up for the fundamental difficulty we face for infrasound signal classification: atmospheric effects on waveform morphology. There are alternate regularization approaches we did not try, such as aggressive dropout but we did not pursue these based on the observation that ensembles of trained models did little to increase our prediction accuracy.

For both methods, we used 4-fold cross validation to determine the accuracy of the data. The total catalog was partitioned into 4 data subsets consisting of 64 examples of mining signals, 38 examples of chemical/accidental explosions, 26 examples of volcanic eruptions, and 25 examples of earthquakes. We train 4 models for each method using 75% (or 3 of the 4 partitions, minus a randomly drawn validation set of 1%) of the data for training. We stop training models once the accuracy on the validation data ceases to increase over 10 cycles through the training data (epochs). We used the remaining 25% of signals to test the algorithm. We computed test accuracy by dividing the number of correctly classified signals by the total number of signals in the test set for each fold. We computed the 4-fold average accuracy by taking the mean of the accuracies for all of the folds. We then compared test accuracies for each catalog and results from the 2- and 4-class catalogs.

4. Results

First, we compared classification accuracies on the 2-class catalog consisting of only earthquakes and volcanic activity. When using SVM, we achieved an average 4-fold classification accuracy of 75%. CNN provides a similar result as SVM, giving a 4-fold average accuracy of

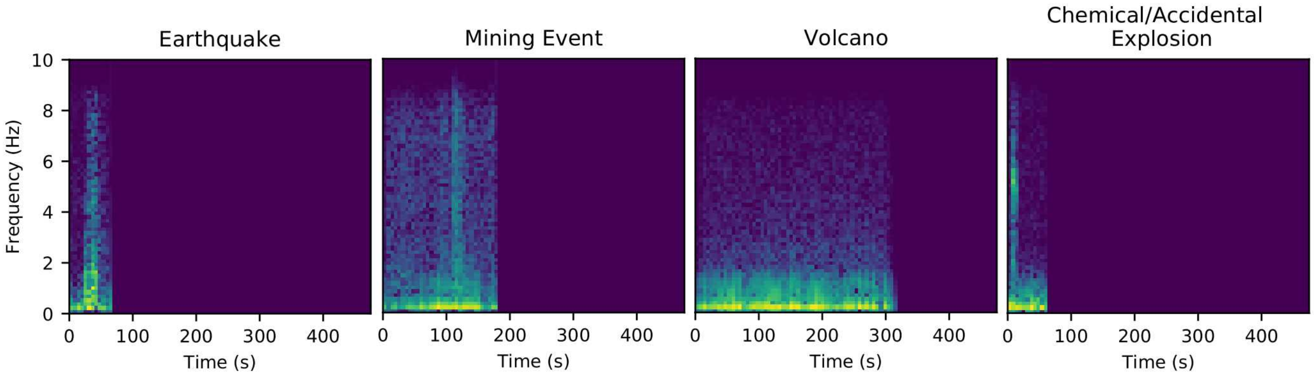

Fig. 3. Examples of spectrograms computed for each class and used to train the CNN algorithm. These signals are all shorter than 475 s so they include zero padding.

74%. However, as we will see with the 4-class catalog, results from the 2-class catalog show that we are able to achieve higher accuracies on a smaller catalog. Model performance can be described by a confusion matrix, averaged over all of the data partitions (Table 2). The confusion matrix shows the mean fraction of the data that is either correctly classified or misclassified. For example, Table 2 shows two confusion matrices corresponding to classification accuracies using the SVM and CNN architectures on the 2-class catalog. If we focus on the SVM confusion matrix, it shows that 68% of the earthquake signals were correctly classified, while 32% were misclassified as volcanic activity.

Next we compared the classification accuracies of the two architectures on the 4-class catalog consisting of signals from mines and quarries, chemical/accidental explosions, earthquakes, and volcanic activity. When using the feature extraction method previously described with the SVM classification scheme, we achieve a 4-fold average accuracy of 55%. CNN classification provides only a one percent increase in accuracy, giving a 4-fold average

Table 2. Mean Confusion Matrices for the 2-Class Example

SVM

		Predicted	
		Earthquake	Volcanic Activity
True	Earthquake	0.68	0.32
	Volcanic Activity	0.19	0.81

CNN

		Predicted	
		Earthquake	Volcanic Activity
True	Earthquake	0.80	0.20
	Volcanic Activity	0.32	0.69

of 56%. Again, we compare confusion matrices from the two architectures in Table 3. In both cases the models struggle with classifying earthquakes and volcanoes, but do well with classifying signals from mines and quarries. When comparing the classification of earthquakes using the two methods, the CNN algorithm gives a 12% higher accuracy. This suggests that there are shared feature characteristics between the earthquake class and the explosions and mines and quarries classes. As can be seen in the 2-class examples, the binary classification of earthquakes and volcanoes can be done with an accuracy of about 75%, but their classification becomes difficult when a more complex catalog is used.

5. Discussion

When comparing the state of the art in infrasound classification (SVM) to CNN, a deep learning method proven in tangential fields, we see that CNN does not outperform SVM. When using the SVM method on the 2-class catalog, we achieved an average 4-fold classification accuracy of 75%. CNN provides a similar result as SVM, giving a 4-fold average accuracy of 74%. For the 4-class catalog, our accuracies are 55% and 56% for SVM and CNN, respectively. Both the SVM and CNN algorithms perform worse than the architecture by Li *et al.* (2016) where they achieve 86% accuracy (though it is unclear how many iterations generated this number). It is important to note, however, that we have limited insight regarding the event catalog of Li *et al.* (2016). Their publication does not describe the source locations or source-receiver distances. Therefore, we suggest that the performance gap between our results and theirs is due to variation in geographic area and source type diversity. This is somewhat expected since their method was not designed to be generalized. The CNN algorithm outperforms SVM at classifying earthquakes, suggesting the that our features used in SVM do not fully capture the earthquake signal characteristics. Both the SVM and CNN models in our study struggle with classifying earthquakes and volcanoes, but do well in classifying mining activity. This is likely due to shorter propagation distances. Mining activity is recorded at distances of 581 km on average, making the signals less affected by propagation (Figure 1). In contrast, chemical/accidental explosion, earthquake,

Table 3. Mean Confusion Matrices for the 4-Class Example

SVM

		Predicted			
		Earthquake	Explosion	Mines and Quarries	Volcanic Activity
True	Earthquake	0.25	0.37	0.33	0.05
	Explosion	0.05	0.54	0.35	0.07
	Mines and Quarries	0.04	0.17	0.78	0.00
	Volcanic Activity	0.07	0.37	0.34	0.23

CNN

		Predicted			
		Earthquake	Explosion	Mines and Quarries	Volcanic Activity
True	Earthquake	0.47	0.15	0.20	0.18
	Explosion	0.09	0.59	0.25	0.08
	Mines and Quarries	0.02	0.14	0.79	0.04
	Volcanic Activity	0.22	0.16	0.25	0.37

and volcanic activity signals are recorded at distances of 1571, 1199, and 2084 km, respectively. This suggests that if more IMS stations were installed to decrease propagation paths, our algorithms may have performed better. Unfortunately, we could not analyze this idea as we had a limited number of locally recorded signals. However, our feature importance results suggest that distance from source to sensor is the most important feature of the data, supporting the idea that more IMS stations would produce better classifications. Individual mines also set off the same explosive source for every event at that mine, making the source process stable from one event to another.

We attempted to identify a distance at which signals are increasingly misclassified. A separation in distance between correctly classified and misclassified signals exists only for the mining events class, where regional signals show the lowest misclassification rate. The other classes do not have such a clear separation. In fact, for the earthquake and volcano classes, signals from all distances are misclassified more frequently than they are correctly classified. This may be due to variance in atmospheric conditions for station-signal pairs, smaller class size, or fundamental source characteristics that are not being adequately captured by our models. More signal examples in each class would likely solve the misclassification problem for these classes.

In the 4-class example, CNN generates an increase in accuracy of only 1%. Though both algorithms provide lower accuracies than previous research, we argue that unlike previous studies, we are aiming to solve a more complex problem. For one, we use waveforms from multiple stations that detected the same event. Also, a variety of stations recorded each signal type. Previous studies have used only single-station signals, a small number of source-station pairs, and some have recorded the same source at the same station for all class examples. This reduces the complexity of the problem, making classification an easier task for the model. It also limits the reach of the model, making it less generalizable to other event catalogs.

On that same note, it is expected that the 2-class example outperforms the 4-class example. Classifying only two types of signals is a much simpler task than classifying four. The confusion matrices for the 4-class SVM example show that earthquakes and volcanic activity are often classified as explosions and mines and quarries. This is also true for the CNN algorithm, although misclassified events are more equally spread between the other classes. This suggests that the earthquakes and volcanic activity classes share similar features with the explosions and mines and quarries classes (both physical from SVM and discovered using CNN). Therefore, removing those classes provides an accuracy increase.

Our algorithms produced lower than expected accuracies, which could possibly be increased by pre-processing the input data in different ways. For example, zero padding of the spectrograms is likely not the best solution because it can elongate a short signal and truncate a long signal. Future studies should consider either including the addition of ambient noise for a fixed signal length or use learning strategies capable of processing variable length signals. It is also important to note that Green and Nippress (2019) show that signal duration generally increases with propagation distance. We use the signal onset time provided in the IRED catalog, though it may be more appropriate to use a 475 s window surrounding the maximum amplitude since we classify signals that have propagated far distances. No pre-processing was performed on the data prior to classification, apart from calculating the delay-and-sum beam. Therefore, it is likely that long-duration microbarom signals, which serve as noise, are contained within the training input. We experimented with filtering waveforms prior to classification though this did not increase accuracy. Using alternate spectral processing methods, such as the Hilbert-Huang Transform (Huang and Zhaohua, 2008), for input into the CNN algorithm may have allowed for the algorithm to better discriminate between long period noise and signal. These avenues exist to improve the performance of deep learning methods, though in the absence of an adequate catalog these approaches likely will not increase performance. The catalog used in this study was designed to serve as a reference for analysts so it contains useful examples, but they may not fully capture the variances within each signal class. We also have a limited number of examples in our training data. CNNs usually require orders of magnitude more examples in order to reach high accuracies. Therefore, a much larger labeled catalog is necessary in order to design a generalized method for global infrasound signal classification.

We recommend future catalogs use a standardized method that aims to fully capture the complexities of a signal as well as its propagation path. The IRED catalog uses CSS3.0 standard static and dynamic tables that hold information about both the station and the signal such as: source label, event identification number, event location, velocity, backazimuth, station location, station, and instrument. We recommend including this information in a standardized catalog. Infrasound signals detected at global distances often have gradual onsets that occur some time following the signal's first arrival at the station. Therefore, capturing both the timing of the first arrival and the timing of the maximum amplitude ensures the characteristics of both impulsive and gradual-onset signals are captured. We assume that all detections in an infrasound catalog are infrasound. However, infrasound signals are sometimes detected on seismic stations. This should be logged in the catalog using a station flag (examples include 'I' or 'S'). Lastly, we recommend the addition of the nomenclature developed by Brown *et al.* (2002) and expanded by Hedlin *et al.* (2018) to describe any path an infrasonic wave has traveled through the atmosphere. Including this in a standardized catalog would allow users to better understand how the morphology of an individual signal was affected by atmospheric structure.

6. Conclusions

The CNN algorithm gives virtually the same classification accuracy as the SVM algorithm. Therefore, SVM should be the preferred method on datasets such as the ones described in this study because it is the simpler method. SVM is a trustworthy architecture that is physically interpretable - the features relate to physical properties of the waveforms. It also requires less computation time to produce classifications (though the feature extraction process can take a considerable amount of time if complex).

The SVM and CNN algorithms discussed in this paper are not yet generalizable for global infrasound event catalogs. Turning these ideas into an operational classification algorithm requires more infrasound data to be collected in systematic ways. An example of how this could be done comes from seismic data centers where large and comprehensive event catalogs are compiled by analysts. Consistent, systematically labeled infrasound data should contain standardized labels, analyst reviewed phase arrival picks, and in-depth data quality review. Generating a global infrasound catalog with these qualities would likely increase the value of deep learning and data driven strategies when used for classification. Until then, deep learning strategies may do little more than overfit small datasets and perform poorly in real use cases. However, the strength of DNNs for infrasound classification lies in their ability to determine the waveform attributes most meaningful for prediction. This is not something we know as infrasound analysts, and may prove to be more beneficial to infrasound classification than using physically derived features. As catalog sizes increase, we expect the performance of SVM and DNN to diverge in favor of DNN. This, however, can only be proven once larger, more refined catalogs are compiled. We encourage readers to refer to this study when motivation is needed for the development of large and comprehensive, systematically labeled, infrasound event catalogs for classification purposes.

7. Data and Resources

This catalog can be accessed by member states (states that have signed the Comprehensive Nuclear Test Ban Treaty) by requesting it from the IDC through their principal point of contact. More information can be found at the following link: *https://www.ctbto.org/verification-regime/the-international-data-centre/distribution-of-data-and-data-bulletins-to-member-states/.*

Acknowledgments

The authors would like to thank the International Data Centre for providing the infrasound event catalog. We would also like to thank Tim Draelos for his helpful and insightful advice. 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. This research was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, 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-NA-0003525.

References and links

Arrowsmith, S. (**2018**). "False alarms and the ims infrasound network: understanding the factors influencing the creation of false events," Geophys. J. Int. **215**(2), 1322–1337, doi: 10.1093/gji/ggy350.

Bowman, D., and Albert, S. (**2018**). "Acoustic event location and background noise characterization on a free flying infrasound sensor network in the stratosphere," Geophys. J. Int. **213**(3), 1524–1535, doi: 10.1093/gji/ggy069.

Brieman, L. (**2001**). “Random forests,” Machine Learning **45**(1), 5–32.

Brown, D., Katz, C., Bras, R. L., Flanagan, M., Wang, J., and Gault, A. (**2002**). *Infrasonic signal detection and source location at the Prototype International Data Centre* (Burkhäuser, Basel).

Cansi, Y. (**1995**). “An automatic seismic event processing for detection and location: The pmcc method.,” Geophys. Res. Lett. **22**(9), 1021–1024, doi: 10.1029/95GL00468.

Ceranna, L., Pichon, A. L., Green, D., and Mialle, P. (**2009**). “The buncefield explosion: a benchmark for infrasound analysis across central europe,” Geophys. J. Int. **177**(2), 491–508, doi: 10.1111/j.1365-246X.2008.03998.x.

Gibbons, S. J., Ashming, V., Eliasson, L., Federoc, A., Fyen, J., Kero, J., Kozlovskaya, E., Kvaerna, T., Liszka, L., Nasholm, S. P., Raita, T., Roth, M., Tiira, T., and Vinorgadov, Y. (**2015**). “The european arctic: a laboratory for seismoacoustic studies.,” Seismol. Res. Lett. **83**(3), 917–928, doi: 10.1785/0220140230.

Green, D., and Nippress, A. (**2019**). “Infrasound signal duration: the effects of propagation distance and waveguide structure,” Geophys. J. Int. **216**(3), 1974–1988, doi: 10.1093/gji/ggy530.

Ham, F. M., and Park, S. (**2002**). “A robust neural network classifier for infrasound events using multiple array data,” in *Neural Networks, 2002*, IEEE, Proceedings of the 2002 International Joint Conference on, Vol. 3, pp. 2615–2619.

Hedlin, M., Drob, D., Walker, K., and deGroot Hedlin, C. (**2018**). “A study of acoustic propagation from a large bolide in the atmosphere with a dense seismic network,” J. Geophys.

Res. **115**(B11), doi: 10.1029/2010JB007669.

Huang, N., and Zhaohua, W. (**2008**). "A review on hilbert huang transform: Method and its applications to geophysical studies," Rev. Geophys. **46**(2), doi: 10.1029/2007RG000228.

Kulichkov, S. (**2004**). "Long-range propagation and scattering of low-frequency sound pulses in the middle atmosphere," Met. Atmos. Phys. **85**(1-3), 47–60, doi: 10.1007/s00703-003-0033-z.

Li, M., Liu, X., and Liu, X. (**2016**). "Infrasound signal classification based on spectral entropy and support vector machine," Applied Acoustics **113**(2016), 116–120, doi: 10.1016/j.apacoust.2016.06.019.

Linville, L., Pankow, K., and Draelos, T. (**2019**). "Deep learning models augment analyst decisions for event discrimination," Geophys. Res. Lett. **46**(7), 3643–3651, doi: 10.1029/2018GL081119.

Linville, L. M., Pankow, K. L., and Kilb, D. L. (**2018**). "Contour-based frequency-domain event detection for seismic arrays," Seismol. Res. Lett. **89**(4), 1514–1523, doi: 10.1785/0220170242.

Liu, X., Li, M., Tang, W., Wang, S., and Wu, X. (**2014**). "A new classification method of infrasound events using hilbert-huang transform and support vector machine," Mathematical Problems in Engineering **2014**, doi: 10.1155/2014/456818.

Perol, T., Gharbi, M., and Denolle, M. (**2018**). "Convolutional neural network for earthquake detection and location," Sci. Adv. **4**(2), e1700578, doi: 10.1126/sciadv.1700578.

Ross, Z. E., Meier, M.-A., and Hauksson, E. (**2018**). "P-wave arrival picking and first-motion polarity determination with deep learning," J. Geophys. Res. **123**(6), 5120–5129, doi: 10.1029/2017JB015251.

Smets, P., and Evers, L. (**2014**). "The life cycle of a sudden stratospheric warming from infrasonic ambient noise observations," J. Geophys. Res. **119**(21), 12–084, doi: 10.1002/2014JD021905.

Vapnik, V. (**1995**). *The nature of statistical learning theory* (Springer-Verlag, New York).

8. List of Figure Captions

Figure 1. A map of the infrasound events used in this study. Lines originate from the source location and connect to all stations (black triangles) that detected the event. Note that lines are straight paths and do not represent the actual path traveled by the infrasonic wave. Line colors represent event type. Most of the events are detected at global distances ($\geq$ 250 km). Note that some of the locations correspond to multiple events that may have been detected at a variety of sensor locations.

Figure 2. Feature importance calculated using the random forest method. Features selected in this study are highlighted in red. Note that the two most important features are distance from source and waveform duration, suggesting a range dependency on classification.

Figure 3. Examples of spectrograms computed for each class and used to train the CNN algorithm. These signals are all shorter than 475 s so they include zero padding.