Temporal earthquake forecasting

Veda Ong[1,6], Stefan Nielsen[2], Stefano Giani[3], Paul Johnson[4]

[1]Earth Sciences, University of Durham, Durham, UK
[6]Now at: Convedo, London, UK
[2]Earth Sciences, University of Durham, Durham, UK
[3]Engineering, University of Durham, Durham, UK
[4]Los Alamos National Laboratory, Los Alamos, New Mexico USA

Key Points:

- Background seismic signal undergoes subtle changes before an earthquake
- A neural network can be trained to identify the time intervals preceding earthquakes with a high success rate
- A slight increase in the precursor amplitude around frequencies of 0.2 Hz is found before the studied earthquakes

---

Corresponding author: Stefan Nielsen, `stefan.nielsen@durham.ac.uk`

Abstract

Convolutional Neural Networks (CNNs) can detect patterns that are otherwise difficult to identify and have been shown to excel in predicting fault characteristics in laboratory shear experiments and slow slip *in situ*. Here we show that during the precursory phase of some natural earthquakes, a subtle change in the seismic background signal occurs that can be identified by a suitably designed CNN, and used as a probabilistic forecasting tool. We use 31 earthquakes of $M_W \geq 6$ in Japan and vicinity, between March 2012 and February 2020, all recorded by station IU MAJO (Japan main island) except one recorded by station IU MA2 (Kamtchatka). The CNN is trained on 24 events, where a 16 mn time window preceding each earthquake is labelled as precursory (presumably containing a strong precursory signal), and another 16 mn time window far from the time of earthquake occurrence is labelled as noise (presumably containing weak or no precursory signal). The 7 remaining events were used for testing. The CNN achieves 98% training accuracy and a 96% testing accuracy in discriminating noise and precursor windows. Time windows in the $\sim$ 3 hours preceding the earthquakes are progressively interpreted by the model as precursors as earthquake time approaches. To characterize the signal detected by the CNN, we analyse spectra from noise and from precursory windows. Discriminative features appear most dominant over a frequency range of $\approx$ 0.1-0.9 Hz (in particular $\approx$0.16 and $\approx$0.21 Hz) coinciding with microseismic noise and recent observations of broadband slow earthquake signal (Masuda et al., 2020).

Plain Language Summary

Subtle signals may be emitted by faults in the hours preceding an earthquake. Here we test this hypothesis by training a convolutional neural network to identify time intervals preceding magnitude 6 earthquakes from broadband seismic signals.

1 Introduction

In natural earthquakes, precursors are thought to arise either when faults reach critical stress conditions preceding shear failure (Scholz, 2019), or when slow slip impacts an extended nucleation patch, triggering rupture of small asperities (Ruiz et al., 2017; Guérin-Marthe et al., 2019; Kato & Ben-Zion, 2021). Recently, systematic changes in seismic wave statistical characteristics (Rouet-Leduc et al., 2019, 2018, 2017; Lubbers et al., 2018; K. Wang et al., 2021; Shreedharan et al., 2020; Corbi et al., 2020; Scuderi et al., 2016) have been observed prior to lab fault failure. Laboratory experiments show systematic changes prior to stick-slip events on simulated faults, that may be regarded as a proxy for natural earthquake faults. Prior to fault failure, laboratory simulated earthquakes show an increase in small shear failures, each of which emit impulsive acoustic emissions (P. A. Johnson et al., 2013). Machine learning, a field used to analyse the statistical characteristics of large quantities of data, can also be used as a tool to investigate changes in the acoustic signal emitted prior to rupture. Machine learning on faults were initially tested on laboratory faults (Rouet-Leduc et al., 2017, 2018, 2019; Hulbert et al., 2019). This work demonstrated highly accurate prediction of lab earthquake instantaneous characteristics such as friction, as well as earthquake timing, by identifying statistical characteristics of the seismic signal emitted from the fault zone that were imprinted with information regarding fault slip.

A subsequent study (Rouet-Leduc et al., 2019) applied similar machine learning techniques to seismic data from the Cascadia subduction zone. By posing the problem as a regression between the statistical characteristics of the continuous seismic data and the surface GPS displacement rate, the study showed that the Cascadia megathrust continuously emits a tremor-like signal with statistical characteristics that reflect the displacement rate on the fault. Although this approach provides real-time access to the physical state of the slowly slipping portion of the megathrust, it has not successfully been applied to seismogenic earthquake prediction. Systematic precursors to seismogenic earthquakes are yet to be identified in the continuous signal applying machine learning (Mignan & Broccardo, 2020; C. W. Johnson & Johnson, 2021). Work on identifying impulsive precursors preceding fault failure using more classical means has been suggestive but not conclusive. For instance, Bouchon and colleagues (Bouchon et al., 2011, 2013) have shown potential for applying statistical approaches for some earthquakes but the observation is far from conclusive.

The difficulty in identifying natural precursors arises partly from the fact that without knowing the location of an impending earthquake, efforts cannot be focused towards detecting changes in the properties within and surrounding a specific fault zone prior to failure, especially where the signal to noise ratio is small (Scuderi et al., 2016; Rouet-Leduc et al., 2017). Additionally, precursors may often be masked by other earthquakes or earthquake swarms which are characterised by entirely different statistical properties (Ishibashi, 1988; C. W. Johnson & Johnson, 2021). There is hope that the significant increase in station density and sensitivity over the last 15 years will lead to advances in earthquake forecasting and precursor detection however (Rouet-Leduc et al., 2017), but an optimal location must be selected as a starting point. Meaning, a fault displacing measurably that is well instrumented.

Because CNNs can detect features of different scales (Zhao et al., 2017), we may expect that variations of the seismic signal over a wide interval of frequencies and amplitudes may be detected. CNNs have frequently been applied to earthquake detection, generating improved earthquake catalogues by efficiently analysing large quantities of seismic data (Van Quan et al., 2017; Perol et al., 2018; Mousavi et al., 2019; C. W. Johnson & Johnson, 2021). However, research into the potential of CNNs and complex neural network architectures to improve earthquake predictability is more limited. For instance, recent efforts applying an encoder-decoder model to analyze continous seismic data emanating from a seismogenic fault in Earth—the San Andreas Fault (SAF) at Parkfield—were unsuccessful in predicting fault instantaneous displacement and future earthquake timing (C. W. Johnson & Johnson, 2021). The study concluded that the seismic signals of interest, if they exist on this portion of the SAF, are too weak to identify within a noisy environment. Huang et al. (2018) utilised a simple CNN to investigate the seismic data prior to earthquakes in Taiwan. Taiwanese seismicity maps were transformed into 2D images by encoding earthquake magnitude as brightness. A classification-based approach was employed to detect differences within seismicity maps up to 30 days prior to large ($M_w \geq 6$) earthquakes, and seismicity maps up to 30 days prior to small ($M_w < 6$) earthquakes. Their algorithm yielded an R-score of 0.303 (where an R-score of 0 is the result of an entirely random prediction and an R-score of 1 is an entirely successful prediction). This suggests that the CNN captured some precursory seismic pattern, however, no further investigation was conducted into the patterns which led to this classification result. In addition, these results were not considered for probabilistic forecasting of earthquakes.

Although CNNs are commonly used on 2D images (Huang et al., 2018), here we investigate precursors based solely on features of the raw seismic signal. We apply neural networks to detect systematic, pattern-based changes in raw time series. We apply a statistical approach to test the potential of Deep Learning techniques in the short-term forecasting (minutes to hours) of an ensemble of earthquakes. Our goal in this investigation is to determine whether precursors can be detected without any substantial data pre-processing.

Rather than considering long-term changes such as decreases in seismic wave speed and increased foreshock activity, which do not systematically occur prior to large earthquakes, here we focus on short-term fluctuations and attempt to detect patterns within the seismic data that occur over a smaller time-frame (minutes to hours) prior to large earthquakes. Focusing on a smaller time-frame and training a complex CNN for classification as a first step, may more robustly enable the detection of previously undiscovered patterns in seismic signals. Evidence of novel pre-rupture patterns would indicate that some sort of mechanism is active during the earthquake nucleation phase, that may emit a subtle signal which is detectable with sophisticated methods. In short our goal is a proof-of-concept by applying deep learning to an ensemble of events, to determine if such an approach yields precursor information.

2 Description of the data

To increase the probability that predictive features are systematically present in the signal, seismic events that share a similar process and environment should be selected, possibly from the same region. This potentially reduces the generality of the training, but increases the chances of obtaining a positive proof of concept. However, confining the study to a limited geographical area reduces the total number of events available in the seismic catalogue for analysis. Therefore, a seismically active region is selected for our test.

We chose the Japan subduction region during a time interval of relatively high seismic activity in the years following the 2011, Tohoku M9 earthquake. Located at the junction of four tectonic plates, the zone experiences around 400 $M_w > 0$ earthquakes per day (McGuire et al., 2005). Additionally, earthquakes in Japan account for over 20% of all M6 or greater earthquakes worldwide (Mogi, 1981). The largest recorded earthquake was the 2011 M9 Tohoku Earthquake, which ruptured the central section of the Japan Trench to a depth of approximately 50 km (Ozawa et al., 2012). In addition to the dense seismic network and the high recurrence of relatively large earthquakes, aseismic slip with transient timescales of days to months has recently been observed in the Japan subduction zone using continuously monitored GPS arrays (McGuire et al., 2005). A continuously slipping subduction zone should increase the potential for precursors, however, it might also result in a significant number of foreshocks that could substantially mask precursors in the seismic signal (McGuire et al., 2005).

Next we choose the minimum magnitude of the target events whose precursory phase is investigated. Using small magnitude target events would limit the amplitude of the possible precursory signal, while using large magnitude limits the number of available target events. We settle for a threshold of $M_w \geq 6$, the highest possible value still allowing for a reasonable number of earthquakes available within the geographical area and time interval investigated.

The database includes earthquakes that occurred between March 2012 and February 2020. The upper time limit, February 2020, was driven by the date at which the data were downloaded, and followed by the configuration, training and testing of the network. The lower time limit, March 2012, was selected to reduce the influence of significant stress changes and afterslip from the March 2011 M9 earthquake on the features learnt by the neural network during training and to improve generality of the algorithm.

Changes in stress before, during and after the 2011 M9 Tohoku earthquake have been extensively investigated (see for example Becker et al. 2018 and references therein), confirming that the most significant modification to the stress field occurred at the time of the M9 Tohoku earthquake. As expected from stress changes occurring during a megathrust cycle (Herman & Govers, 2020), the region between northern Honshu and the Japan trench, previously under compressive horizontal stress, became extensional after the earthquake (Becker et al., 2018). Additionally, there was indication of a short-term transient increase of horizontal stress in the months following the Tohoku earthquake, until a plateau was reached after approximately one year. The frequency of aftershocks was similarly investigated and a sudden, short-term increase of the seismicity rate was observed immediately after the Mw 9 earthquake (Toda, 2019). This was followed by an approximately exponential decrease in the seismicity rate which is compatible with Omori's law of aftershock decay (Utsu et al., 1995). Roughly one year following the start of the Mw 9 earthquake, the rate had become stable and lower compared to the rate prior to the earthquake.

Data recorded by station IU MAJO (Fig. 1) preceding earthquakes that took place within 20^o of the station, were used for training and testing of the network. An additional earthquake outside of the 20^o radius was analyzed using recording from the IU MA2 station. The data for each event consists of ten hours of continuous seismic data preceding each magnitude 6 earthquake, recorded on the three channels (BH1, BH2, BHZ) of Streckeisen STS-2 High-gain instruments at 40 Hz.

If any precursory changes in the seismicity exist, they would likely be detectable by a station in the relative vicinity of the generating process, but attenuate with increasing distance.

Inspection of seismograms from $M_w \approx 6$ earthquakes at different distances from the station of interest, shows that the attenuation is significant (signal to noise ratio decreases significantly) by 2500 km from the station. The threshold of 20^o (approximately 2500 km) was selected assuming that the attenuation shown in the earthquakes' seismograms is a proxy, or possibly an upper limit, for the attenuation of unidentified, and weaker signals in the data that the model may identify.

Having defined the region, magnitude range and time interval, all corresponding events were inspected and some were excluded from the database based on the following criteria. Some of the events were found to contain impulsive earthquake signals arising from smaller ($M_w < 6$) earthquakes. The presence of highly impulsive earthquakes may alter the characteristics of the seismic data and affect the features learnt by the neural network during training (Ishibashi, 1988). This issue would be particularly significant when investigating very short-term precursors where the quantity of data input is very limited and therefore should be well representative of each class. Such events were discarded to encourage the network to analyse features of the background signal, removing the influence of earthquake waveforms (Rouet-Leduc et al., 2019). Events where the data recording was discontinuous or otherwise corrupted in the ten hours preceding the earthquake were also eliminated (Examples of discarded data in Fig. 3). Under such constraints, 31 events with $M_w \geq 6$ remained to develop and test the deep learning model (Fig. 1 and Table 1).

3 Description of the CNN algorithm

We applied a classification procedure to determine if and when precursors are present, and separate them from background noise. We tested different existing network architectures, notably: Residual Networks (He et al., 2016); Dilated Residual Networks (Yu et al., 2017); Long-Short-Term-Memory Fully Convolutional Networks (Karim et al., 2018). None of these networks performed well, and therefore we integrated features from several of these model-types into a single Convolutional Neural Network (CNN) (see Fig. 2). We gradually increased the complexity of a simple network. Typically, experimenting with different techniques proves to be the best method for generating a network that performs well. The modifications made to obtain the final network structure (Fig. 2) are detailed as follows.

Often, a convolution block in a CNN consists of one or two convolutional layers followed by a batch normalisation layer and a ReLU activation layer. Here, we added a max-pooling layer to each block after the convolution operation. Max-pooling enhances the strong activations from the convolution output (feature map) and discards the weak ones. All but one of the max pooling layers have a stride of 1, to avoid a change in the dimensions of the feature map (Fig. 2). This prevents a loss of information which occurs when the dimension of the output are reduced by using a stride > 1; however, we found that using a stride of 2 in a single convolutional block improved the performance on the test data.

The number of convolutional layers was increased to 7 resulting in an increased number of filters, up to a maximum of 256 filters in the final two convolutional blocks.

Dilation was added to all but the first 2 convolutional blocks, and was increased with depth in the network. Dilation produced only marginal improvement, possibly because the receptive field of the network was already large enough to contain the required information from the input. To avoid gridding artifacts (Yu et al., 2017) it was proposed (P. Wang et al., 2018) to use hybrid dilated convolution (where dilation rate increases and decreases in a

Time	Lat.(^{o})	Lon.(^{o})	z (km)	Catalog	M$_w$	Δ(^{o})
Training database:						
2019-06-18T13:22:19	38.6370	139.4804	12.0	NEIC PDE	6.4	2.32
2019-04-11T08:18:21	40.4096	143.2985	18.0	NEIC PDE	6.0	5.55
2019-01-08T12:39:31	30.5926	131.0371	35.0	NEIC PDE	6.3	8.43
2018-09-05T18:07:59	42.6861	141.9294	35.0	NEIC PDE	6.6	6.78
2018-01-24T10:51:19	41.1034	142.4323	31.0	NEIC PDE	6.3	5.62
2017-11-09T07:42:11	32.5208	141.4380	12.0	NEIC PDE	6.0	4.83
2017-10-06T07:59:32	37.5033	144.0201	9.0	NEIC PDE	6.2	4.74
2017-09-20T16:37:16	37.9814	144.6601	11.0	NEIC PDE	6.1	5.33
2017-09-07T17:26:49	27.7829	139.8041	451.0	NEIC PDE	6.1	8.87
2016-04-14T12:26:35	32.7880	130.7042	9.0	NEIC PDE	6.2	7.18
2016-01-14T03:25:33	41.9723	142.7810	46.0	NEIC PDE	6.7	6.48
2016-01-11T17:08:03	44.4761	141.0867	238.8	NEIC PDE	6.2	6.93
2015-05-12T21:12:58	38.9005	142.0217	39.3	ISC	6.8	3.83
2015-04-20T01:42:58	24.0574	122.4319	28.1	ISC	6.4	18.43
2015-02-20T04:25:23	39.8189	143.6157	13.3	ISC	6.2	5.37
2014-11-09T14:38:15	46.9300	140.6300	10.0	ISC	7.6	10.54
2014-08-10T03:43:18	41.1340	142.2790	50.6	ISC	6.1	5.60
2014-03-13T17:06:51	33.6222	131.8077	83.4	ISC	6.3	5.99
2014-03-02T20:11:22	27.4238	127.3279	118.9	ISC	6.5	12.96
2013-04-21T03:22:16	29.9644	138.9741	431.3	ISC	6.1	6.61
2013-04-05T13:00:02	42.7359	131.0640	571.3	ISC	6.3	8.27
2012-12-07T08:18:23	37.8201	144.1594	35.3	ISC	7.2	4.91
2012-07-08T11:33:05	45.4209	151.3906	37.7	ISC	6.0	13.31
2012-05-23T15:02:27	41.3569	142.1267	64.1	ISC	6.0	5.70
Test database:						
2018-11-14T21:21:50 (*)	55.6324	162.0008	50.2	NEIC PDE	6.1	7.18
2017-07-26T10:32:57	26.8975	130.1836	12.0	NEIC PDE	6.0	11.81
2016-11-11T21:42:59	38.4973	141.5658	42.4	NEIC PDE	6.1	3.30
2016-10-21T05:07:23	35.3676	133.8148	5.7	NEIC PDE	6.2	3.74
2016-09-20T16:21:16	30.5017	142.0478	9.0	NEIC PDE	6.1	6.84
2013-12-08T17:24:54	44.4691	149.1330	34.1	ISC	6.1	11.46
2013-10-25T17:10:17	37.1457	144.7540	14.7	ISC	7.1	5.27

Table 1. Events in the training and testing database. The event with an asterisk in the test database was recorded by station IU MA2, while all others were recorded by station IU MAIO.

sawtooth pattern); however, the latter performed slightly worse, so continuously increasing dilation was kept in the final model.

A dropout layer with a rate of 0.02 was added after the fully connected layer to regularise the network (Hatami et al., 2018). This slightly improved its generalisation to the test data.

The kernel initialiser was changed from the default to *random normal* which uses a normal distribution to initialise the weights.

A random layer (Fig. 2) was applied directly to the input (Lee et al., 2020) and this improved the performance of the network. A random signal was obtained by convolution with a kernel which was randomized before each epoch, then added to the input signal. The root mean square of the random signal was about 20%-30% that of the original signal. The random layer produced slightly different versions of the inputs with each epoch. The addition of noise is a proven regularization technique to reduce its generalization error but not its training error. This is achieved by presenting slightly different data every epoch forcing the model to learn the more general features or those which remain consistent epoch after epoch. In addition, it aids in generalising the network; by increasing the number of different inputs, the model learns the more general features or those which remain consistent in the randomly augmented inputs. Also, randomisation prevents overfitting.

The values in the two output neurons (bottom layer, Fig. 2) represent a score, with values between 0 and 1, for the two classes, noise and precursor. The scores are obtained by applying the Softmax activation function to the two values from the dropout layer, after computing their dot product with their weights. This process is repeated for each time window (16348 samples). The Softmax function is defined in such a way that the sum of the outputs is always 1 for each time window, so that they can be interpreted as probabilities. (Note that here we chose to show scores as [0-100%] rather than [0-1]).

During training, the class with highest score is elected as the class to which the sample belongs. The success or failure to classify the windows correctly is used to improve the network during the training. In addition, the fraction of windows correctly predicted allows estimation of the accuracy of the network performance both in the final training run and in the test, as described in section (5.1).

4 Data formatting for training and testing

The 31 earthquakes selected (section 2) were split into two groups: 24 events were chosen randomly for the neural network training, while the remaining 7 were used for testing.

Each event was split into 36 macro-windows of 1000 s (16 minutes and 40 s, or 40000 time samples). For each window, we implemented mean removal (standardization) and normalisation of all three components jointly. As a result, the relative static offset of the three components was preserved. (Normalizing and standardizinq by individual components was also tested, but resulted in a lesser accuracy of the network). For the scope of the neural network training, the data in window no. 36 of each event (1000 s immediately preceding the earthquake) was labelled as precursor. This decision assumes that precursor energy progressively increases as failure is approached, as has been observed in laboratory studies (P. A. Johnson et al., 2013) and field studies (Bouchon et al., 2013). Note that the time interval is arbitrary and other time intervals could have been selected. In addition, windows classified as noise may also contain the same precursory signature signal, only with a lesser amplitude than the time windows immediately preceding the earthquake. In short we are assuming the exponential increase in precursor activity observed in laboratory, Earth and simulation studies will be sufficiently pronounced for the classification procedure to work. Data labelled as noise was taken from either (A) the macro-windows no. 1 of each event (10 hours to 9 hours 43’20” before the earthquake) or (B) a random 1000 s in a time interval unrelated to any of the earthquakes (at least 48 hours before or after any of the earthquakes). Two types of training were conducted, using noise (A) or (B), but the same precursor in both.

To increase the number of samples in the training, a data augmentation technique was implemented. Each of the 1000 s (40000 samples) intervals classed as noise or precursor was split into 37 windows of 16384 samples with an overlap of 15374 samples. As a result, a total of 888 time windows (with three channels and 16384 time samples each) classed as precursor was obtained from the 24 earthquakes to train the neural network during the semantic segmentation training. Equally, 888 noise windows were obtained, resulting in a combined number of 1776 windows of both noise and precursors. The test data set –comprised of seven seismic events– was split according to the same sub-window length and overlap as above, resulting in 259 (37 × 7) noise and 259 precursor windows, for a total of 518 windows.

5 Results

5.1 Evaluating the performance of the network

To illustrate the performance of the network, we compute f as the percentage of correctly classified widows, defined as:

$$f = 100 \times (T_P + T_N)/N_{tot}$$

to measure the accuracy of the model across the entire dataset. T_P is the number of correctly identified precursor windows (they fall within the 1000 s before the earthquake), T_N is the number of correctly identified noise windows (either ten hours before the earthquake for test A, or in time intervals unrelated to earthquakes for test B); N_{tot} is the total number windows (precursory or noise). A random classification would result in a score $f \approx 50\%$, while a perfectly accurate classification would result in $f = 100\%$.

When using noise windows taken from signal ten hours before the earthquake (test A), the final network achieved an average training accuracy of 97% (1725/1776 correctly classified) an average test accuracy of 90% (467/518 correctly classified). The performance can also be visualised applying a confusion matrix as shown in Fig. (4). When using noise windows taken from signal unrelated to the earthquake (test B), the performance is not significantly different. In such a case, the final network achieved an average training accuracy of 98% and a test accuracy of 96%.

Note that the maximum accuracy achieved, although generally high, can vary slightly depending on the TensorFlow (Martín Abadi, 2015) version used (here 1.13.1). The accuracy on the individual test events are reported in Table (2). All events except event 39 are from station IU MAJO, event 39 is from the IU MA2 station. The IU MA2 station is not been used to construct the train dataset, therefore, the network has never seen data from that station before. An accuracy of 97% suggests that what the network learned from the IU MAJO station is also relevant for other stations like in this case the IU MA2 station.

One of the events from IU MAJO shows a low accuracy of 0.5 that we attribute to a classification error done by the operator during the pre-processing. However, we did keep the event in the test database, as such error would be a likely occurrence in any real-life application of the model (faulty data, operator error, etc). Therefore, including this can be a considered as a strategy to test the robustness of the model.

Table 2. Accuracy of the network on the events in the test dataset.

Event	Accuracy
12	1.00
13	0.89
14	0.50
15	0.95
29	1.00
30	1.00
39	**0.97**

5.2 Results obtained over the entire ten hour time span

Although the training was performed using only the first and the last time windows in the time series of ten hours, as previously mentioned, we presume that precursors identified in the final minutes of signal may be present and detectable in earlier time windows. Thus we further investigate the potential of the trained network to discriminate noise from precursory signal on all the time intervals (1-36) in the ten hours preceding the earthquake. To this end, the network was tested on the entire time interval (including intervals 2-35), producing the fraction of windows classified as precursors in each interval. This process was repeated five times independently (re-training the system on wfirst and last time windows with randomly initialised weights every time), to allow computation of a mean and a standard deviation. The fraction of precursory windows is represented as a function of time in the ten hours preceding the earthquake in Fig. 5, as an indicative measure of the precursory character of each time interval.

We find that the precursory character increases persistently as the earthquake rupture time approaches, with a trend that clearly bounds the standard deviation at about 3.3 hours before the earthquake (in the case where the network was trained using interval 1 of 36 as noise, Fig. 5) or 2.5 hours (in the case where the network was trained using noise from time intervals unrelated to the earthquake, not shown).

This increase in precursory character can be interpreted as the marker of an increasing intensity of precursory signal preceding the earthquake. The result can also be viewed as a proxy for the increasing probability that the network may detect a precursor as the time of the earthquake approaches.

6 What pattern does the network detect?

Ideally, the detection and the identification of earthquake premonitory signals should inform our understanding of the earthquake source mechanics, in particular of the nucleation phase and how it integrates in the seismic cycle. However, the end-to-end learning strategy of CNNs make their representations a black box, meaning that it is difficult to understand the logic of their predictions. As a consequence, it is not always straightforward to identify or to isolate the features, or the combination of features or the general pattern that triggers the CNN.

CNN representations can be investigated with a number of techniques which fall under visualisation. These help to reveal what specific patterns and which segments of the data allow neural networks to detect features and classify samples. These techniques typically include feature map visualisation, feature map inversion, saliency maps, filter visualisation and occlusion. In our case most of these techniques were ineffective in shedding light on the network workings, with the exception of occlusion. In addition, we were able to find some characteristics of premonitory time windows by using more classical Fourier transform techniques and spectral analysis.

6.1 Occlusion of time intervals

Occlusion sensitivity is a simple technique for understanding what features in the input are most important for classification. In our case, different portions of the time series are excluded from the time window that is analysed, with the aim of quantifying the relative importance of different portions of the input in the classification result.

For the occlusion exercise, we investigate a series of precursor windows in the test dataset, with a high gradient in the prediction score (a significant change in the prediction score from one window to the next). As the analysis time window is shifted ahead, the prediction score increases rapidly (Fig. 6), indicating that the newly incorporated time interval contains features specific to the precursor class (precursor-related features) or that the removed interval contained features associated with the noise class (noise-related features).

We investigate the window with the greatest increase in certainty relative to the previous window (the window with a certainty of 81.1% in Fig. 6). We apply an occlusion mask, a short length of zeros that is moved along the input at a fixed stride, and determined the prediction score for each position of the mask along the input (Fig. 7). All 3 channels contributed to the prediction scores. It was evident that when the data points between 15600 and 16000 were removed, the network predicted the input as noise rather than precursor. This exercise demonstrated that the addition of information to the end of the window as opposed to removal of information from the start increased the prediction score of the window to the precursor class.

The occlusion output did not drop below 0.5 unless all 3 channels were occluded. This indicates that the network used patterns between channels such as similarities or differences as well as channel specific patterns. However, when comparing the occlusion outputs for each individual channel, it became clear that channel 0 (horizontal, North component of the station) had a greater contribution to the network's decision, enough to reduce the certainty to the precursor class from 81.1% to 56.2% certainty, while the other 2 channels did not reduce the prediction score as significantly. Therefore, at least in the case of the event investigated with occlusion, channel 0 appears to provide more precursor-related information than channels 1 and 2.

6.2 Occlusion of frequency

To investigate a type of frequency-related occlusion, we eliminated specific frequency bands in the signal, rather than specific time intervals. An 8$^{\text{th}}$ order (roll off = -48 dB /octave) low pass filter was applied. Starting at a cutoff frequency of 20 Hz (the maximum frequency in the input data), the cutoff frequency was reduced in intervals of 0.1 Hz until only frequencies below 0.1 Hz remained in the test data. Each time the cutoff frequency was reduced, the best weights obtained in the training were validated on the whole filtered test dataset, and the accuracy f was computed.

The change in accuracy as a function of cutoff is shown in Fig. 8. The significant frequencies in discriminating noise from precursors appear to be mostly below 3.5 Hz. Indeed little deterioration of the prediction is induced by cutting higher frequencies. In addition, the accuracy seems to increase in particular within the two intervals [0.1–0.8] and [1.8–2.7].

6.3 Spectral analysis

To determine the importance of the frequency anomalies in distinguishing noise from precursors prior to all of the investigated earthquakes, we analysed the frequency-amplitude spectra in three different ways.

First, a spectrogram was produced to verify if any relevant time change was detectable in the time window prior to one of the detected M6 earthquakes. The Fourier transform was computed within a sliding window of 1000 samples and a stride of 500. The result shown in Fig. (9) for the final 6.8 minutes before the earthquake, (corresponding to the time window where certainty increases to 97.3% in Fig. 6).

Second, the Fourier amplitude spectrum for noise and precursor windows were obtained separately for each event in the train and test datasets (all 3 channels). Then the cumulative sum of the frequency responses for all events and their 3 channels were calculated for either noise-labelled and precursor-labelled data, and plotted on the same figure for comparison (Fig. 10). No obvious differences were evident when comparing the cumulative frequency responses for noise and precursor data in the training and test datasets. Although some small differences between noise and precursor data occurred in the very low frequencies of the training dataset ($\approx$ 0.02 Hz - 0.04 Hz), these did not occur in the test data. Any non-systematic differences (differences not evident in both datasets) would unlikely have contributed to the classification result.

Third, to magnify any possible difference between noise and precursor spectra, the relative percentage difference between the cumulative noise and precursor frequency responses were calculated for all 3 channels in the train and test datasets. The relative percentage difference was obtained by computing the difference between the cumulative precursor and noise spectra and normalising by the cumulative noise spectrum. The results for both the train and test datasets are shown in Figure 11 where the dots are the results and the curves are smoothed versions of the results. The smoothed versions were obtained using Savitzgy-Golay smoothing with a width of 0.062 Hz. Differences between the cumulative precursor and noise frequency responses become clearer when represented as the relative difference, for both the test and train datasets. Significant and systematic differences occur at approximately 0.16 Hz and 0.21 Hz, as indicated by the two vertical dashed lines.

The results obtained in Fig. 11 indicate two low frequencies that provided information for discriminating precursor windows from noise windows in the train and test data. From these investigations, it can be concluded that frequencies of $\approx$ 0.16 Hz and 0.21 Hz were significant during classification. The huge spike in amplitude at $\approx$ 12 Hz in the smoothed test plot (Fig. 11) is irrelevant to the classification result, as can be concluded from Fig. (8) which demonstrates that frequencies above 3.5 Hz did not significantly affect the prediction score on the test dataset.

7 Discussion and conclusions

We tested a classification Convolutional Neural Network to detect precursor activity in the hours preceding earthquakes. We used seismic data in the ten hours preceding 31 earthquakes of magnitude 6 and above. The earthquakes took place in the Western Pacific area immediately surrounding Japan, and were recorded at a single three-channel broadband station UI-MAYO (Japan) of the Global Seismic Network. An additional earthquake recorded at station IU MA2 (Kamchatka) was processed to test the performance of the network on a different setting. We also used background signal recorded at time intervals distant from any moderate of large earthquake occurrence.

The ten hours data preceding each earthquake was split into windows of 1000 s. The network was trained to classify each time window as either precursor or noise from direct input of data windows with all three components. The last window before each earthquake was labelled as precursor. Windows taken either ten hours before the earthquake, or from random time intervals unrelated to any moderate of large earthquake, were labelled as noise.

After several network prototypes were tested without success, we designed a fully connected network comprising seven convolutional layers. 27 events were used for the training of the network, and seven were kept separate to be used in testing.

The performance of the network was evaluated by counting the number of windows correctly classified, producing an accuracy percentage as the fraction of windows correctly classed. A significant (above error bar) accuracy is obtained about 3 hours before the earthquakes, and the accuracy increases with approaching earthquake time up to about 85% in the test batch (an accuracy of 50% represents a null result).

Another measure of the system performance is a score consisting in the difference of the fraction of correctly predicted precursory windows to total precursory windows, minus the fraction of falsely predicted precursors to total noise windows. Such a score can be used to evaluate the confidence of the network; in the very last window before the earthquake the score is typically 97% (a score of 50% represents a null result). A positive outcome was also obtained by testing the network on an earthquake outside of the training area (Japan) and using another seismic station (Kamchatka); this is encouraging indication of possible portability of the network to other contexts, without re-training. However, all earthquakes analysed belong to a similar context (subduction zone tectonics within the same area of the Western Pacific rim).

These results show –for the database analysed– that (1) there must be a change in the signal in the hours preceding the earthquakes and (2) the change can be detected by the convolutional network. Such a change is very subtle and elusive to more traditional signal analysis. However, a convolutional network can be qualitatively compared to a series of frequency filters and cross-correlations. Although the interpretation of the inner workings of the neural network is not trivial, future work may focus on a detailed analysis of the convolutions constructed by the network during the training, with the aim of revealing features of the precursory signal in terms of specific waveforms, patterns (sequences of small impulsive sources) and frequency shifts (starting or stopping of tremor in specific frequency bands). So far, it has not been possible to identify from the deep network layers the precise pattern (or combination of patterns) associated with the positive forecasting.

Here, instead, we directly investigated the signal previous to the earthquake by conducting a number of spectral analyses: by selectively filtering different frequency bands, we show that the significant band is below 3.5 Hz. Analysis of spectral amplitude also reveals tiny relative changes in amplitude at low frequency, in particular in the range 0.16–0.21 Hz.

We can offer tentative interpretations in terms of the physics of earthquakes process and their nucleation:

Slow slip episodes preceding earthquakes have been documented from timescales of seconds (Tape et al., 2018) to weeks or months (Ruiz et al., 2014, 2017; Socquet et al., 2017; McGuire et al., 2005; Bouchon et al., 2011, 2013; Hasegawa & Yoshida, 2015) in natural earthquakes, revealing the progressive growth of instability on a fault patch, and can be simulated in laboratory experiments (Nielsen et al., 2010; Latour et al., 2013; Guérin-Marthe et al., 2019), or numerical models (Ampuero & Rubin, 2008). Fault instability has been analysed previously in the framework of rate-and-state friction, where a critical length h_{RR} for the nucleation patch can be theoretically derived (Dieterich, 1992; Ruina, 1983; Rice & Ruina, 1983; Uenishi & Rice, 2003; Rubin & Ampuero, 2005; Ampuero & Rubin, 2008). Close to the instability, oscillations in the slip can take place with increasing amplitude. These may radiate a low-amplitude, low-frequency tremor that increases the relative amplitude of a given frequency range in the background noise.

Slow slip will also trigger tiny foreshocks within the sliding area (Ruiz et al., 2017), or at the front of the expanding slip patch (Kato et al., 2012). These tiny foreshocks correspond to stick-slip episodes on small sticky patches of the fault (e.g., areas where the friction is locally velocity-hardening). Although their amplitude may be below the noise and these tiny events are not detectable individually, they will contribute to a background tremor and alter the quality of the background noise. The background chatter of these tine events can be amplified through constructive interference or resonance due to guided waves within a slab of lower seismic impedance around the fault zone.

Several interesting follow-up paths arise from the present study:

The generalisation and portability of the method to other regions should be tested; here, an initial portability test was conducted using the Kamchatka earthquake and station, with promising results.

Further scrutiny of the network's convolutional filters may help to identify the nature of the precursory signal, rather than using the network as a black box. This task may be facilitated by seeking for a simplified version of the current neural network, as all parts of its complex architecture may not be necessary. In addition, the design of an effective CNN is in essence a random search by trial and error. Is it possible to optimise the network design process by understanding why a subset of architectures are more successful for this problem?

Finally, modelling of nucleation and preslip with a fault zone may help to elucidate the process. Models can be used to produce synthetic seismic data, which is then analysed by the CNN after addition of background noise. Selecting the models that produce similar results to natural earthquakes when analysed with CNN may allow one to constrain regarding what type earthquake nucleation physics is most realistic. Henceforth, by looking at the signal of the model without the added noise, interesting features of the precursory pattern may be revealed.

While the current study focused on data from a single station, a processing based on multiple stations would be a natural enhancement of the method. We also note that the current network output is limited to a binary answer (precursor / no precursor). No forecast for magnitude value has been attempted (and the data was from a narrow range of magnitudes anyways) nor for distance from the source or location of the events (the latter would require joint analysis of several stations). In addition, large time intervals between earthquakes have not been analysed, therefore it is difficult to evaluate how similar signals could occur without the consequence of an earthquake. This prevents an accurate estimate of the likelihood of obtaining a true detection of the signal combined with a false positive earthquake forecast. Therefore, while there is potential to develop CNN for probabilistic forecasting with risk mitigation and early warning techniques, this prototype network will need to be tested on more data and enhanced to allow a more robust output.

Acknowledgments

PAJ acknowledges support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division under grant 89233218CNA000001. All seismic data used here are from the the IU network (GSN; Albuquerque, 1988) and were downloaded through the IRIS Web Services (https://service.iris.edu/). For convenience we replicate the data before and after formatting on the online repository 4TU.ResearchData https://doi.org/10.4121/20101901.

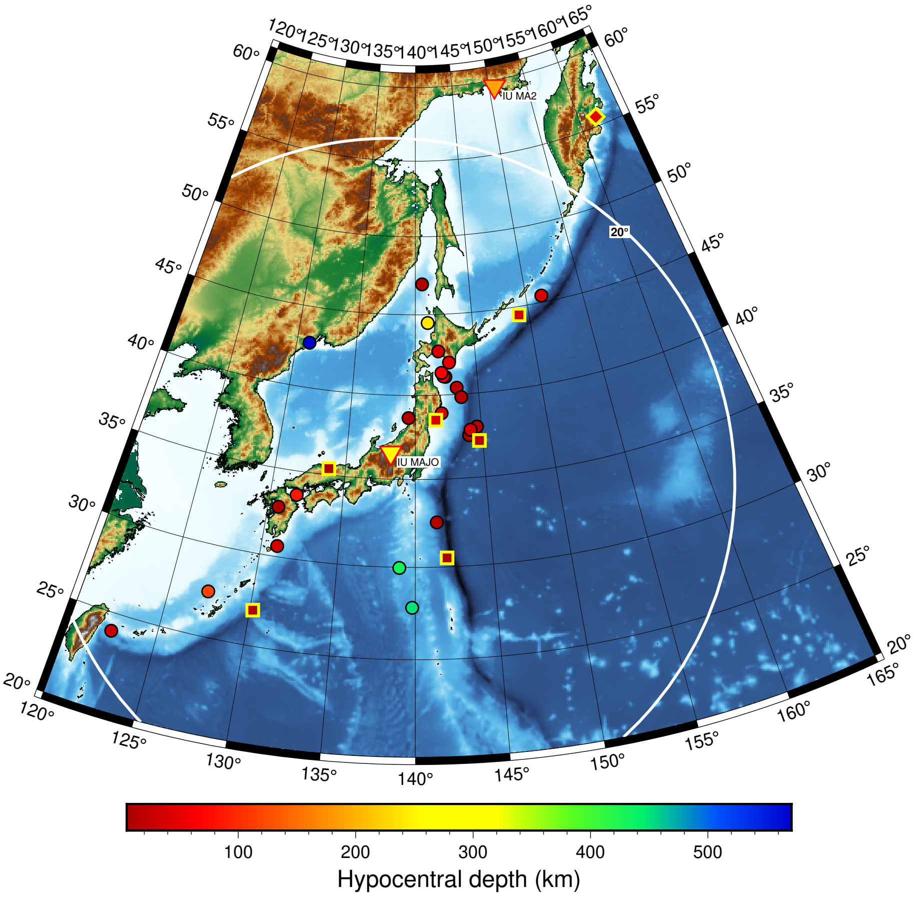

Figure 1. The 31 $M_w \geq 6$ earthquakes in the northwest Pacific region surrounding Japan used for training and testing of the CNN. Filled circles indicate the epicenter of earthquakes in the training dataset. Filled squares indicate the epicenter of earthquakes in the unseen (test) dataset. The earthquakes were selected within 20^o of the station of interest, IU MAJO, indicated by a yellow triangle (the 20^o radius circle around IU MAJO is represented in white). An additional test earthquake outside of the 20^o radius was tested (filled diamond), using recording from the IU MA2 station (orange triangle). The color of the earthquake symbols corresponds to hypocentral depth as indicated by the colorscale.

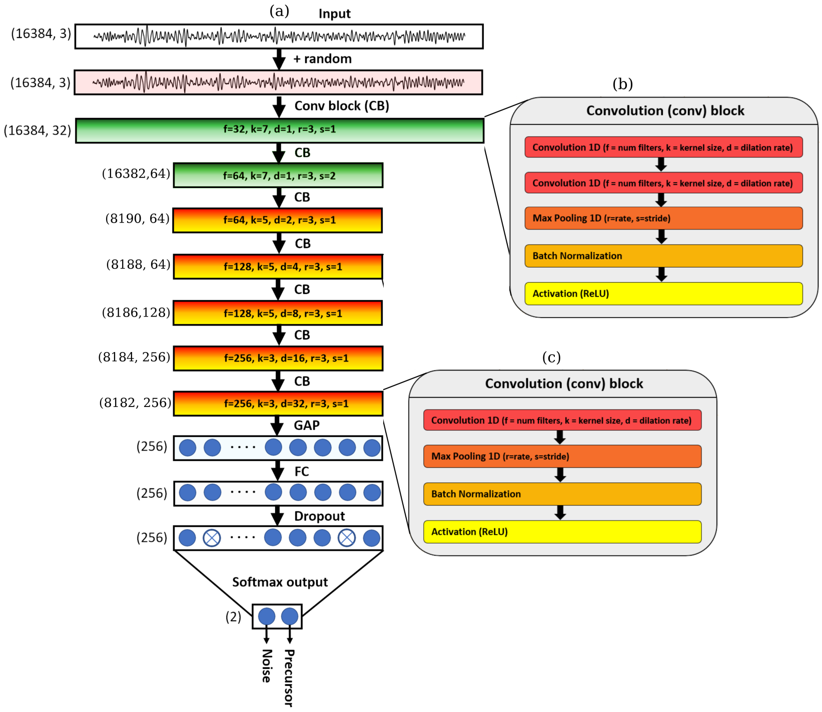

Figure 2. CNN used to detect earthquake precursor patterns. (a) Representation of all layers in the network. f is the number of filters, k is the length of the filter kernel and d is the dilation rate in the convolution operation. r and s are the length of the window and the stride of the max-pooling operation. (Stride is set to 1 in all convolution and max-pooling operations, except in convolution bloc 2 where it is set to 2 only for max-pooling). The first number in parenthesis represents the size of the input vector. Note that there are 3 components in the input vectors, they correspond to the three motion components of the seismic station (vertical, horizontal 1 and horizontal 2). The second number is the same as f, or number of filters. (b-c) Detail of structures of the different types of convolutional block in the network.

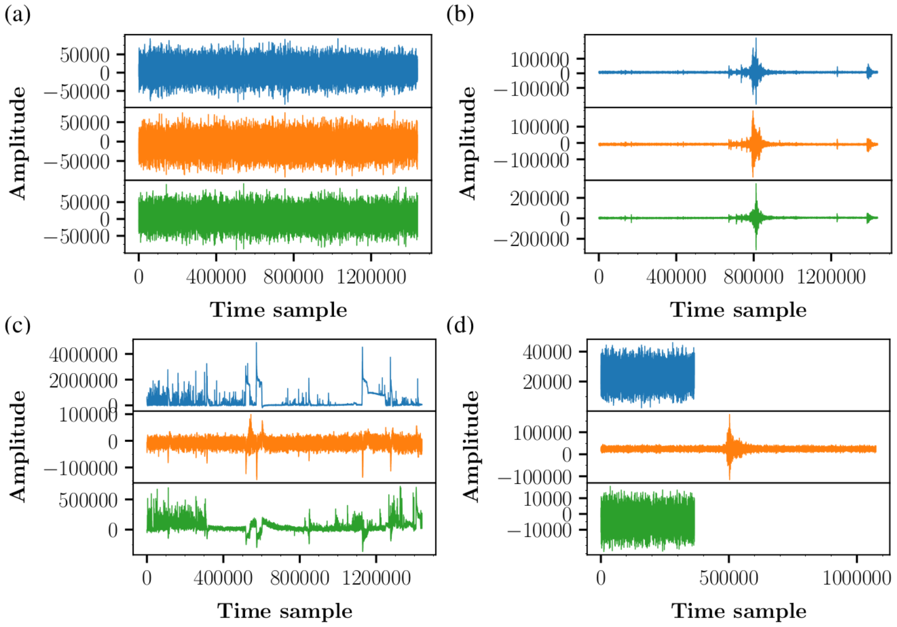

Figure 3. Examples of the 3 channels of seismic data (vertical: blue; North-South: orange; East-West: green curves) over the 10-hour period prior to the selected $M_w \geq 6$ earthquakes. (a) 10-hour period with no impulsive earthquake signal above the noise level. Events such as this were included in the investigation. The following are examples of excluded events: (b) Event containing impulsive signal above the background noise level. (c) Spikes unrelated to earthquakes. (d) Files with channels of varying lengths.

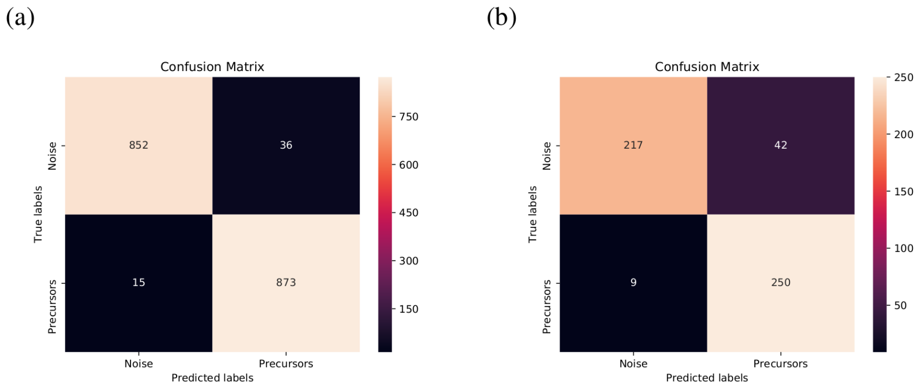

Figure 4. Confusion matrices indicating the performance of the classification model by summarising the prediction results on the train and test datasets — (a) Confusion matrix obtained with the best weights (89% test accuracy) on (a) the training dataset and (b) the test dataset. The confusion matrix indicates the relative accuracy of the network in terms of four possible scenarios: 1. accurate prediction of an event (bottom right), 2. failure to predict an event (bottom left), 3. false prediction of an event (top right), 4. accurate prediction of no event (top left). The numbers in each box indicate the number of windows classified in each scenario. (Case where noise windows are extracted 10 hours before each M6 earthquake).

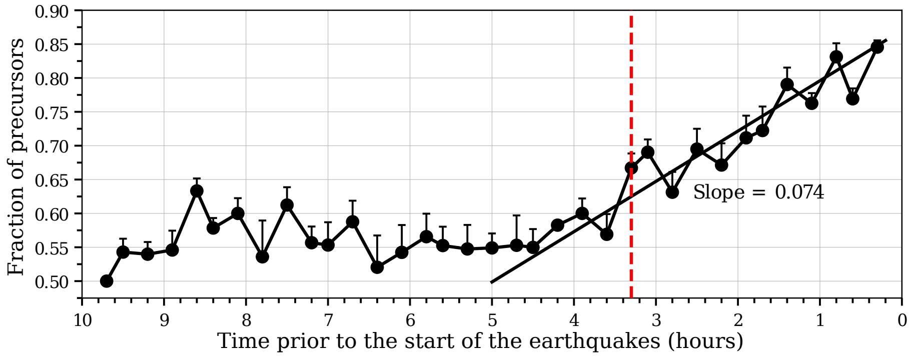

Figure 5. Changes in the fraction of windows classified as precursor in the ten hours preceding the earthquakes (in 36 time intervals). The average of five independent network iterations is represented, with the standard deviation shown on one side of the data points to improve clarity. The red, dashed line indicates the time when the increase in precursory character is significant (exceeding the standard deviation).

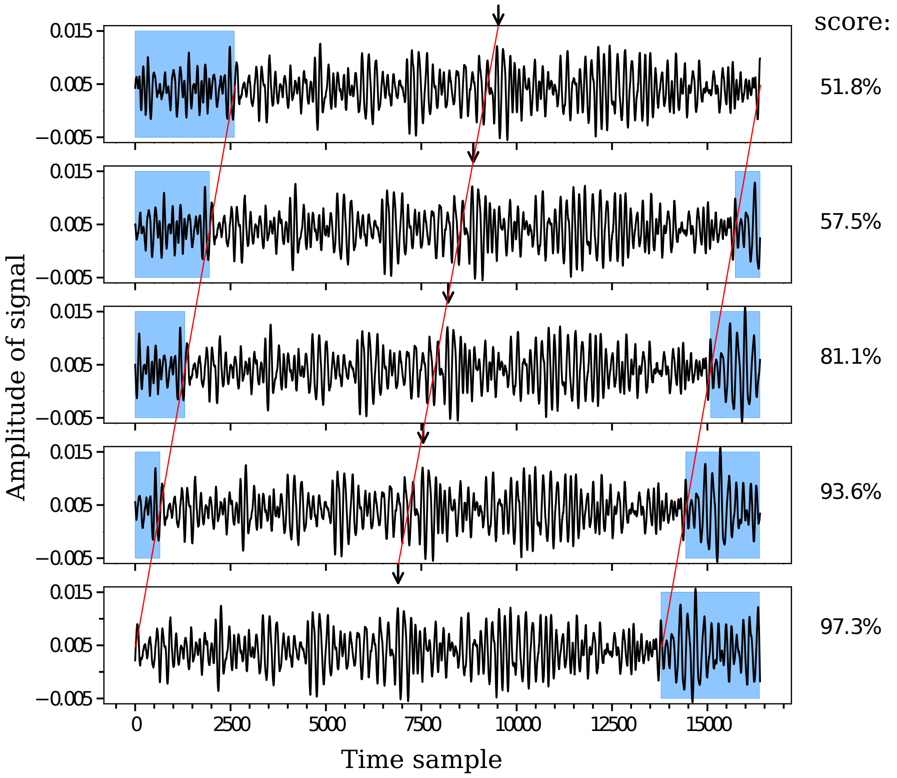

Figure 6. Example of sequential precursor windows (channel 0 only for simplicity) and increase in prediction score (probability) as a specific new time interval is incorporated. The windows stride is 650 time steps; this can be visualised by noticing that the region not shaded in blue is the same in each plot. The arrows indicate the same time on each window. The certainty or prediction score of the network when classifying each window (all 3 channels) as a precursor is indicated.

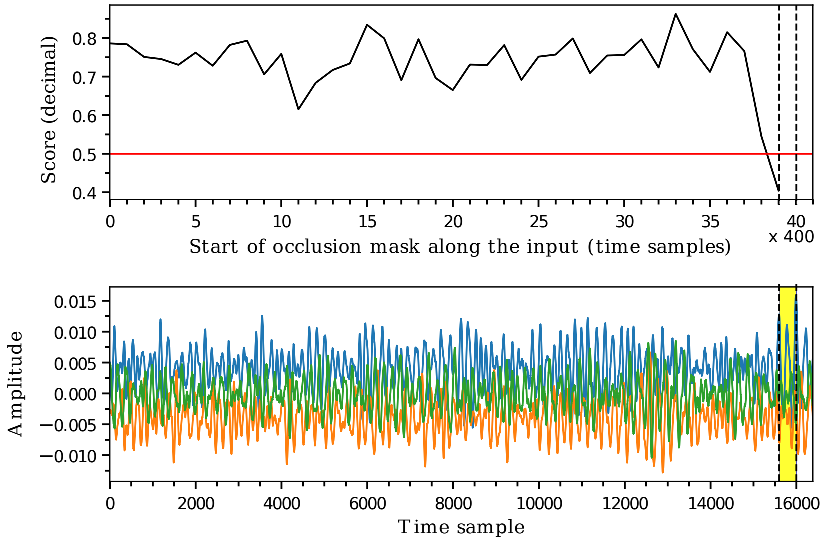

Figure 7. Score sensitivity (top) for different positions of the occlusion mask on the precursor window investigated (bottom). A mask length of 400 and a stride of 400 were used. Prediction scores below 0.5 (red line) line indicate regions of the input containing significant, precursor-related information. All 3 channels of the input (blue, orange, green) were used. The input with high importance is highlighted in yellow/dashed line. Window length 16384, a mask length and stride 400.

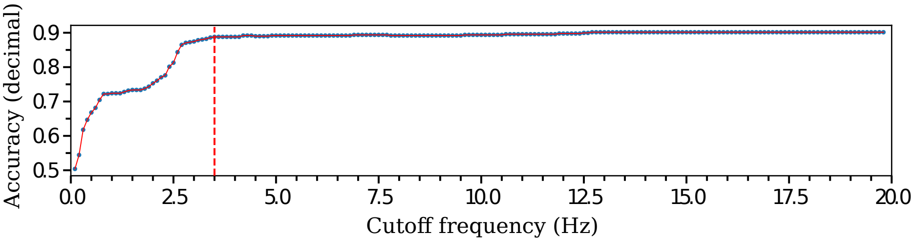

Figure 8. Changes in the test accuracy and test loss when applying the low pass filter to the test dataset with a variable cutoff frequency. The red, dashed, vertical line indicates the cutoff frequency at which the test accuracy started to decrease significantly.

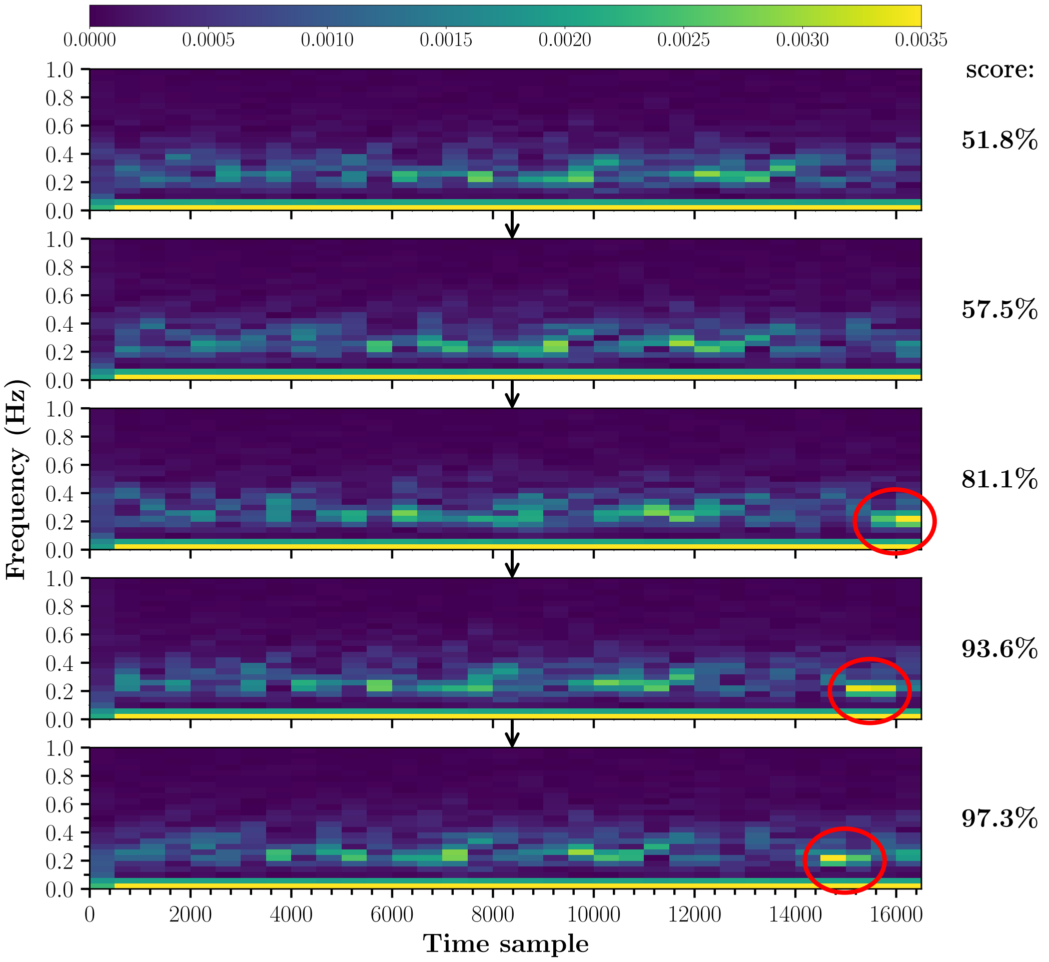

Figure 9. Spectrogram of the sequential precursor windows (channel 0 only). The Fourier amplitude was calculated within a sliding window of length 1000, stride 500 and plotted in colour. The prediction score of the network when classifying each window (all 3 channels) as a precursor is indicated. The red circles highlight a localised region of increased amplitude of frequencies 0.16Hz and 0.2Hz.

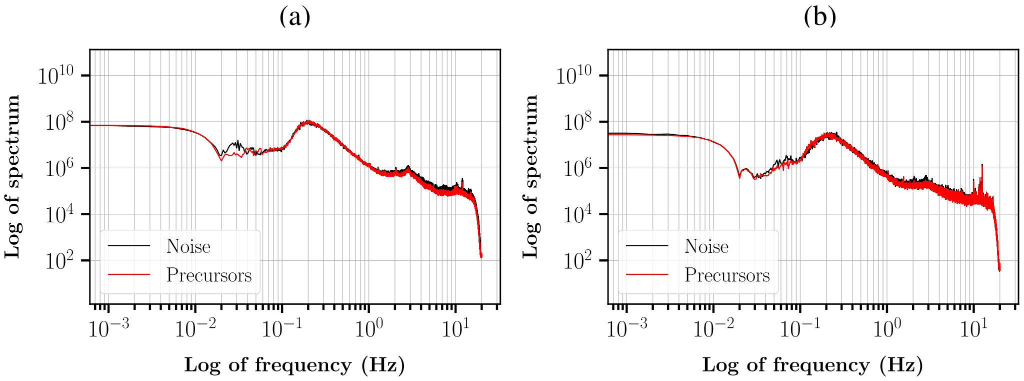

Figure 10. Amplitude spectrum of noise windows (black) and precursor windows (red). (a) The cumulative sum of the frequency responses for all events and their 3 channels were calculated separately for noise-labelled windows (black) and precursor-labelled windows (red) in the training dataset. (b) Same as (a) but for the test data set.

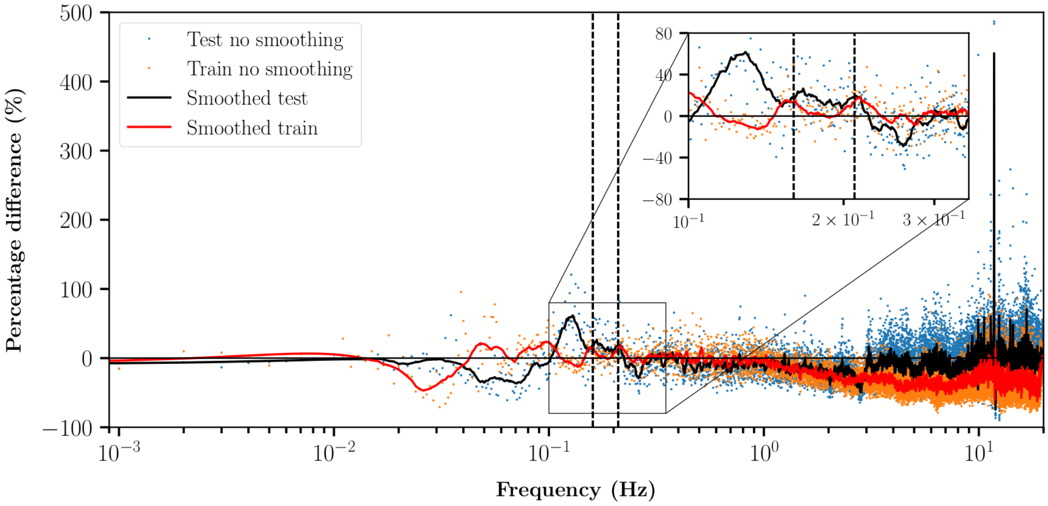

Figure 11. Relative percentage difference between the cumulative frequency spectra in Figure (10) for precursor-labelled and noise-labelled data. The test and train results are plotted on the same graph for ease of comparison. The discrete frequencies are shown as blue and orange dots, while a smoothed spectrum is shown as a red and a blue curve. The dashed, vertical lines are plotted at frequencies 0.16 Hz and 0.21 Hz coinciding with significant amplitude differences between precursor and noise data in both the train and test datasets (peaks in the smoothed plots). A horizontal, black line is plotted at a 0% difference.

References

Ampuero, J.-P., & Rubin, A. M. (2008). Earthquake nucleation on rate and state faults – Aging and slip laws. *J Geophys Res: Solid Earth*, *113*(B1). doi: 10.1029/2007JB005082

Becker, T. W., Hashima, A., Freed, A. M., & Sato, H. (2018). Stress change before and after the 2011 M9 Tohoku-oki earthquake. *Earth Planet Sc Lett*, *504*, 174–184. doi: 10.1016/j.epsl.2018.09.035

Bouchon, M., Durand, V., Marsan, D., Karabulut, H., & Schmittbuhl, J. (2013). The long precursory phase of most large interplate earthquakes. *Nature Geosci*, *6*(4), 299–302. doi: 10.1038/ngeo1770

Bouchon, M., Karabulut, H., Aktar, M., Özalaybey, S., Schmittbuhl, J., & Bouin, M.-P. (2011). Extended Nucleation of the 1999 Mw 7.6 Izmit Earthquake. *Science*, *331*, 877–880. doi: 10.1126/science.1197341

Corbi, F., Bedford, J., Sandri, L., Funiciello, F., Gualandi, A., & Rosenau, M. (2020). Predicting imminence of analog megathrust earthquakes with machine learning: Implications for monitoring subduction zones. *Geophysical Research Letters*, *47*(7), e2019GL086615. doi: 10.1029/2019GL086615

Dieterich, J. H. (1992). Earthquake nucleation on faults with rate-and state-dependent strength. *Tectonophysics*, *211*(1), 115–134. doi: 10.1016/0040-1951(92)90055-B

Guérin-Marthe, S., Nielsen, S., Bird, R., Giani, S., & Di Toro, G. (2019). Earthquake nucleation size: Evidence of loading rate dependence in laboratory faults. *Journal of Geophysical Research: Solid Earth*, *124*(1), 689-708. doi: 10.1029/2018JB016803

Hasegawa, A., & Yoshida, K. (2015). Preceding seismic activity and slow slip events in the source area of the 2011 Mw 9.0 Tohoku-Oki earthquake: a review. *Geoscience Letters*, *2*(1), 6. doi: 10.1186/s40562-015-0025-0

Hatami, N., Gavet, Y., & Debayle, J. (2018). Classification of time-series images using deep convolutional neural networks. In *Tenth International Conference on Machine Vision (ICMV 2017)* (Vol. 10696, pp. 242–249). SPIE. doi: 10.1117/12.2309486

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 770–778).

Herman, M. W., & Govers, R. (2020). Stress evolution during the megathrust earthquake cycle and its role in triggering extensional deformation in subduction zones. *Earth Planet Sc Lett*, *544*, 116379. doi: 10.1016/j.epsl.2020.116379

Huang, J., Wang, X., Zhao, Y., Xin, C., & Xiang, H. (2018). Large earthquake magnitude prediction in taiwan based on deep learning newral network. *Neural Netw World*, *28*(2), 149–160. doi: 10.14311/NNW.2018.28.009

Hulbert, C., Rouet-Leduc, B., Johnson, P. A., Ren, C. X., Rivière, J., Bolton, D. C., & Marone, C. (2019). Similarity of fast and slow earthquakes illuminated by machine learning. *Nat Geosci*, *12*(1), 69–74.

Ishibashi, K. (1988). Two categories of earthquake precursors, physical and tectonic, and their roles in intermediate-term earthquake prediction. *Pure Appl Geophys*, *126*(2), 687–700. doi: 10.1007/BF00879015

Johnson, C. W., & Johnson, P. A. (2021). Learning the low frequency earthquake activity on the central san andreas fault. *Geophysical Research Letters*, *48*(13), e2021GL092951. doi: https://doi.org/10.1029/2021GL092951

Johnson, P. A., Ferdowsi, B., Kaproth, B. M., Scuderi, M., Griffa, M., Carmeliet, J., … Marone, C. (2013). Acoustic emission and microslip precursors to stick-slip failure in sheared granular material. *Geophysical Research Letters*, *40*(21), 5627-5631. doi: 10.1002/2013GL057848

Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2018). LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access*, *6*, 1662–1669. doi: 10.1109/ACCESS.2017.2779939

Kato, A., & Ben-Zion, Y. (2021). The generation of large earthquakes. *Nature Reviews Earth & Environment*, *2*(1), 26–39. doi: 10.1038/s43017-020-00108-w

Kato, A., Obara, K., Igarashi, T., Tsuruoka, H., Nakagawa, S., & Hirata, N. (2012). Propagation of Slow Slip Leading Up to the 2011 Mw 9.0 Tohoku-Oki Earthquake. *Science*, *335*(6069), 705–708. doi: 10.1126/science.1215141

Latour, S., Schubnel, A., Nielsen, S., Madariaga, R., & Vinciguerra, S. (2013). Characterization of nucleation during laboratory earthquakes. *Geophysical Research Letters*, *40*(19), 5064–5069. doi: 10.1002/grl.50974

Lee, K., Lee, K., Shin, J., & Lee, H. (2020). Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning. *arXiv:1910.05396 [cs.LG]*.

Lubbers, N., Bolton, D. C., Mohd-Yusof, J., Marone, C., Barros, K., & Johnson, P. A. (2018). Earthquake catalog-based machine learning identification of laboratory fault states and the effects of magnitude of completeness. *Geophysical Research Letters*, *45*(24), 13,269-13,276. doi: 10.1029/2018GL079712

Martín Abadi, e. a. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Masuda, K., Ide, S., Ohta, K., & Matsuzawa, T. (2020). Bridging the gap between low-frequency and very-low-frequency earthquakes. *Earth Planets Space*, *72*(1), 47. doi: 10.1186/s40623-020-01172-8

McGuire, J. J., Boettcher, M. S., & Jordan, T. H. (2005). Foreshock sequences and short-term earthquake predictability on East Pacific Rise transform faults. *Nature*, *434*(7032), 457–461. doi: 10.1038/nature03377

Mignan, A., & Broccardo, M. (2020). Neural Network Applications in Earthquake Prediction (1994–2019): Meta-Analytic and Statistical Insights on Their Limitations. *Seismol Res Lett*, *91*(4), 2330–2342. doi: 10.1785/0220200021

Mogi, K. (1981). Seismicity in Western Japan and Long-Term Earthquake Forecasting. In *Earthquake Prediction* (pp. 43–51). American Geophysical Union (AGU). doi: 10.1029/ME004p0043

Mousavi, S. M., Zhu, W., Sheng, Y., & Beroza, G. C. (2019). CRED: A Deep Residual Network of Convolutional and Recurrent Units for Earthquake Signal Detection. *Sci. Rep.*, *9*(1), 10267. doi: 10.1038/s41598-019-45748-1

Nielsen, S., Taddeucci, J., & Vinciguerra, S. (2010). Experimental observation of stick-slip instability fronts. *Geophys. J. Int.*, *180*, 697-702. doi: 10.1111/j.1365-246X.2009.04444.x

Ozawa, S., Nishimura, T., Munekane, H., Suito, H., Kobayashi, T., Tobita, M., & Imakiire, T. (2012). Preceding, coseismic, and postseismic slips of the 2011 Tohoku earthquake, Japan. *J Geophys Res: Solid Earth*, *117*(B7). doi: 10.1029/2011JB009120

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

Rice, J. R., & Ruina, A. L. (1983). Stability of steady frictional slipping. *J. Appl. Mech.*, *50*, 343–349. doi: 10.1115/1.3167042

Rouet-Leduc, B., Hulbert, C., Bolton, D. C., Ren, C. X., Riviere, J., Marone, C., … Johnson, P. A. (2018). Estimating Fault Friction From Seismic Signals in the Laboratory. *Geophys Res Lett*, *45*(3), 1321–1329. doi: 10.1002/2017GL076708

Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2019). Continuous chatter of the Cascadia subduction zone revealed by machine learning. *Nat Geosci*, *12*(1), 75–79.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine Learning Predicts Laboratory Earthquakes. *Geophys Res Lett*, *44*(18), 9276–9282. doi: 10.1002/2017GL074677

Rubin, A. M., & Ampuero, J.-P. (2005). Earthquake nucleation on (aging) rate and state faults. *J of Geophys Res: Solid Earth*, *110*(B11). doi: 10.1029/2005JB003686

Ruina, A. (1983). Slip instability and state variable friction laws. *J Geophys Res: Solid Earth*, *88*(B12), 10359–10370. doi: 10.1029/JB088iB12p10359

Ruiz, S., Aden-Antoniow, F., Baez, J. C., Otarola, C., Potin, B., del Campo, F., … Bernard, P. (2017). Nucleation Phase and Dynamic Inversion of the Mw 6.9 Valparaiso 2017 Earthquake in Central Chile. *Geophys Res Lett*, *44*(20), 10,290–10,297. doi:

10.1002/2017GL075675

Ruiz, S., Metois, M., Fuenzalida, A., Ruiz, J., Leyton, F., Grandin, R., … Campos, J. (2014). Intense foreshocks and a slow slip event preceded the 2014 Iquique Mw 8.1 earthquake. *Science*, *345*(6201), 1165–1169. doi: 10.1126/science.1256074

Scholz, C. (2019). *The mechanics of earthquakes and faulting*. Cambridge University Press.

Scuderi, M. M., Marone, C., Tinti, E., Di Stefano, G., & Collettini, C. (2016). Precursory changes in seismic velocity for the spectrum of earthquake failure modes. *Nat Geosci*, *9*(9), 695–700. doi: 10.1038/ngeo2775

Shreedharan, S., Bolton, D. C., Rivière, J., & Marone, C. (2020). Preseismic Fault Creep and Elastic Wave Amplitude Precursors Scale With Lab Earthquake Magnitude for the Continuum of Tectonic Failure Modes. *Geophys Res Lett*, *47*(8), e2020GL086986. doi: 10.1029/2020GL086986

Socquet, A., Pina Valdes, J., Jara, J., Cotton, F., Walpersdorf, A., Cotte, N., … Norabuena, E. (2017). An 8-month slow slip event triggers progressive nucleation of the 2014 Chile megathrust. *Geophys Res Lett*. doi: 10.1002/2017gl073023

Tape, C., Holtkamp, S., Silwal, V., Hawthorne, J., Kaneko, Y., Ampuero, J. P., … West, M. E. (2018). Earthquake nucleation and fault slip complexity in the lower crust of central Alaska. *Nat Geosci*, *11*(7), 536–541. doi: 10.1038/s41561-018-0144-2

Toda, S. (2019). Damaging aftershock hits japan after 55 years. *Temblor*. doi: http://doi.org/10.32858/temblor.030

Uenishi, K., & Rice, J. R. (2003). Universal nucleation length for slip-weakening rupture instability under nonuniform fault loading. *J Geophys Res: Solid Earth*, *108*(B1). doi: 10.1029/2001JB001681

Utsu, T., Ogata, Y., S, R., & Matsu'ura. (1995). The Centenary of the Omori Formula for a Decay Law of Aftershock Activity. *Journal of Physics of the Earth*, *43*(1), 1–33. doi: 10.4294/jpe1952.43.1

Van Quan, N., Yang, H.-J., Kim, K., & Oh, A.-R. (2017). Real-Time Earthquake Detection Using Convolutional Neural Network and Social Data. In *IEEE Third International Conference on Multimedia Big Data (BigMM)* (pp. 154–157). doi: 10.1109/BigMM.2017.58

Wang, K., Johnson, C., Bennett, K., & Johnson, P. (2021, 12). Predicting fault slip via transfer learning. *Nature Communications*, *12*. doi: 10.1038/s41467-021-27553-5

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., & Cottrell, G. (2018). Understanding Convolution for Semantic Segmentation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1451–1460). doi: 10.1109/WACV.2018.00163

Yu, F., Koltun, V., & Funkhouser, T. (2017). Dilated Residual Networks. In (pp. 472–480).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. In (pp. 2881–2890).