# Detection of earthquake precursors using neural networks

**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 some large natural earthquakes
- A neural network can be trained to identify the time intervals preceding earthquakes
- A slight increase in 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 a suitably designed CNN can be trained to identify some precursory change in the seismic signal preceding some large natural earthquakes by up to a few hours, with a variable success rate. We use 65 $M_w \geq 6$ events in the NE pacific in and around Japan from 2012 to 2022. By repeating the training/testing cycle with variable random initial weights, we obtained up to 98% in training accuracy and 96% in testing accuracy in discriminating noise and precursor windows. In the $\sim$ 3 hours preceding the earthquakes, the network identifies precursors progressively more frequently as earthquake time approaches. A final subset of more recent seismic events was used for a further verification, with mixed results. While the network appears to differentiate noise and precursor with a statistically positive incidence, the results are highly variable depending on the events that are analysed, with poor potential for generalisation. This may indicate that not all earthquakes in the catalog contain precursor signals, or at least no signal similar to the training subset. Discriminative features between precursor and noise windows appear most dominant over a frequency range of $\approx$ 0.1-0.9 Hz (in particular $\approx$0.16 and $\approx$0.21 Hz) broadly coinciding with observations made elsewhere of microseismic noise and 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, one of the most 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 63 earthquakes in total; 31 earthquakes that occurred between March 2012 and February 2020 (set A, see table 1) used for a first training and testing of the network; and additional 32 events between February 2020 and May 2022 (set B, see table 2) used for additional verification. 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. 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 sawtooth pattern); however, the latter performed slightly worse, so continuously increasing dilation was kept in the final model.

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

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

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 (although it will increase the 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

| Time | Lat.($^o$) | Lon.($^o$) | z (km) | $M_w$ | $\Delta$($^o$) |
|---|---|---|---|---|---|
| Additional verification events: | | | | | |
| 2022-05-22T15:17:31 | 33.228 | 141.412 | 12.0 | 6.0 | 4.61 |
| 2022-05-09T06:23:03 | 24.022 | 122.501 | 27.02 | 6.3 | 20.09 |
| 2022-03-22T17:41:39 | 23.387 | 121.607 | 24.0 | 6.7 | 21.18 |
| 2022-03-16T14:36:33 | 37.73 | 141.595 | 59.85 | 7.3 | 3.59 |
| 2022-03-16T14:34:27 | 37.647 | 141.673 | 57.18 | 6.0 | 3.64 |
| 2022-01-21T16:08:37 | 32.744 | 132.043 | 39.0 | 6.3 | 7.24 |
| 2022-01-03T09:46:36 | 23.994 | 122.26 | 19.0 | 6.2 | 20.29 |
| 2021-12-09T02:05:08 | 29.413 | 129.385 | 7.0 | 6.0 | 11.34 |
| 2021-11-29T12:40:49 | 31.1 | 142.8 | 10.0 | 6.6 | 7.13 |
| 2021-11-29T12:40:46 | 31.186 | 142.478 | 22.07 | 6.3 | 6.86 |
| 2021-11-10T15:45:17 | 23.7 | 126.4 | 10.0 | 6.6 | 17.45 |
| 2021-11-10T15:45:14 | 23.593 | 126.448 | 12.0 | 6.6 | 17.49 |
| 2021-10-24T05:11:34 | 24.511 | 121.831 | 69.0 | 6.2 | 20.32 |
| 2021-09-29T08:37:06 | 38.894 | 135.444 | 364.0 | 6.1 | 3.62 |
| 2021-09-29T08:37:05 | 45.8 | 153.5 | 401.2 | 6.5 | 17.88 |
| 2021-09-20T20:25:27 | 46.397 | 152.483 | 35.0 | 6.1 | 17.35 |
| 2021-08-24T05:37:52 | 48.862 | 154.91 | 37.36 | 6.0 | 20.76 |
| 2021-05-13T23:58:14 | 37.708 | 141.778 | 32.0 | 6.0 | 3.76 |
| 2021-05-01T01:27:28 | 38.23 | 141.665 | 47.3 | 6.8 | 3.85 |
| 2021-03-20T09:09:44 | 38.475 | 141.633 | 43.0 | 7.0 | 3.93 |
| 2021-02-13T14:07:51 | 37.76 | 141.72 | 51.89 | 7.1 | 3.72 |
| 2020-12-20T17:23:23 | 40.867 | 142.581 | 35.0 | 6.3 | 6.15 |
| 2020-12-10T13:19:59 | 24.763 | 122.01 | 73.17 | 6.1 | 20.03 |
| 2020-11-30T22:54:35 | 48.252 | 140.797 | 589.0 | 6.4 | 11.99 |
| 2020-09-12T02:44:11 | 38.751 | 142.25 | 34.0 | 6.1 | 4.61 |
| 2020-06-13T15:51:24 | 28.859 | 128.271 | 165.0 | 6.6 | 12.56 |
| 2020-04-19T20:39:06 | 38.896 | 142.032 | 38.0 | 6.3 | 4.49 |
| 2020-04-18T08:25:37 | 27.127 | 140.131 | 453.0 | 6.6 | 9.61 |
| 2020-02-13T10:33:44 | 45.631 | 148.929 | 144.0 | 7.0 | 14.06 |
| 2019-11-20T08:26:08 | 53.163 | 153.685 | 486.81 | 6.3 | 22.71 |
| 2019-08-04T10:23:04 | 37.76 | 141.609 | 38.0 | 6.3 | 3.61 |
| 2019-07-27T18:31:08 | 33.146 | 137.325 | 367.0 | 6.3 | 3.51 |

**Table 2.** Additional events used for verification (set B). All recorded at station IU MAIO.

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 proxy for probability. (Note that here we 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 of set A (table 1) 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 time intervals of 1000 s (16 minutes and 40 s, or 40000 time samples). For each time interval, 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 interval 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 intervals 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 interval 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 window length and overlap as above, resulting in 259 ($37 \times 7$) noise and 259 precursor windows, for a total of 518 windows.

To avoid ambiguity in the following text we will refer to the 36 intervals of 40k samples (1000 s) as "time intervals" and to the smaller overlapping subdivisions of 16384 samples each as "windows".

## 5 Results from training and testing of set A

### 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 in the TensorFlow library (Martín Abadi, 2015) used here for deep learning, the maximum accuracy achieved can vary slightly depending on the version used (here 1.13.1) due to the availability of specific functions in specific versions. The accuracy on the individual test events are reported in Table (3). 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, a possible indication that the event in question belongs to a distinct category emitting different or no precursor signal. However, we did keep the event in the test database, as a strategy to test the robustness and generality of the model. More limitations of the generality will be discussed in the section on further tests. We also stress that the accuracy results of Table (3) regard a single training/testing cycle among several realisations with different initial random weights, the one shown is the one where optimal results were achieved. This may indicate that the minimisation process converges in local minima of the misfit function between model and ground-truth; using many different starting models increases the chances of avoiding such shortcoming.

**Table 3.** 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

The training described above was performed using only the first and the last time intervals in the ten hours. However, 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, we perform two different tests. In the first one, both training and testing are conducted on each individual time interval. The aim is to understand which time intervals contain features that best discriminate them from the noise data in interval 1 (how different from the initial background noise?). In the second one, we train the network on intervals 1 and 36, but test it on all time intervals. The aim is to understand which time intervals contain

features that best associate them to the precursor data of interval 36 (how similar to the final precursory interval?).

#### *5.2.1 Training and testing on each individual time interval*

For this test we trained the network on the entire ten hours. In the training, interval 1 of all training events was labelled as noise, while intervals (2-36) as precursors. After training, all time intervals (including 2-35) of the test events was evaluated by the network. The whole training/testing process was repeated five times independently, where the system was re-trained using the first and last time intervals, but with randomly initialised weights every time; this allows the computation of a mean and a standard deviation.

The fraction of windows classed as precursor within each time interval is represented as a function of time, in the ten hours preceding the earthquake (Fig. 5). The average of all test events in all five train-test cycles is shown for each time interval, as an estimate of how effectively the network discriminates signal from the background noise of interval 1 as the time passes. We find that the precursory character increases as each earthquake rupture time approaches, with a trend that outbounds the standard deviation at about 3.3 hours before the earthquake. In a similar test, the network was trained using noise from time intervals unrelated to the earthquake, instead of the first time interval of the ten hours. In this case the increase starts 2.5 hours before the earthquake (not shown).

This increase in precursory character can be interpreted as the marker of an increasing different signal from the background noise of interval 1 as the time passes. 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.

#### *5.2.2 Training on 1 and 36, testing each individual time interval*

Here, similar to section 5.1, we use the network trained on intervals 1 (noise) and 36 (precursor) alone. However, we perform the test on all intervals (including 2-32). In this case, a gradual increase from 0.5 to 0.8 in the fraction of windows classified as precursor within each time interval (sup mat. ??).

## 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. However, we note that occluding specific time windows may in fact constitute an adversarial attack on an unstable network. The degraded performance therefore does note necessarily indicate that the specific time window contains a discriminating signal.

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 $8^{th}$ 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, Short Time Fourier Transform (STFT) spectrogram was generated 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.

We further comment that an increase in small foreshocks (where corner frequencies are generally high) should in principle generate an increased amplitude in the frequency band above 1 Hz, if such were to happen in the precursory phase. This is not quite compatible with the results of frequency occlusion discussed above.

Likewise, Low Frequency Earthquakes (LFEs) or tremor generally observed in Japan are in the band 1-10 Hz, also above the differences identified at a fraction of an Herz (016-0.21 Hz). However, LFEs overlap with the frequency band where the effective accuracy is observed to increase (0-17 Hz, see Figure 7).

Finally, Very Low Frequency Eartquakes (VLFEs) radiate frequencies that do overalp with 016-0.21 Hz. The duration of VLFEs is typically one minute or more, with the highest amplitude peaks (or strong signal envelope) lasting about half as long. Within the time resolution of the occlusion tests of Figure (9), the amplitude of the precursory signal appears stronger over a duration of 10 to 20 s, close, but slightly low with respect to the expected typical VLFEs.

## 7 Results including verification set B

To further test the network performance, in particular to evaluate the potential for generalisation, we augmented the catalog of set A (May 2012 to June 2019, Table 1) with more recent earthquakes of set B (July 2019 to May 2020, table 2) from the same region. We integrated set B events to implement two different evaluations.

### 7.1 Training the network on set A and testing on set B

First, we used the network trained on set A (section 5.1) to test the events in set B. We do recall that the tests performed both here (on set B) and in section 5.1 (on set A) are conducted on events that were not used in the training; therefore, one may expect that the accuracy from both sets is not significantly different.

However, although the majority (60%) of set B events do show a test accuracy $> 0.5$, the average accuracy (0.56) is degraded with respect to the results of the test on the events of set A (0.97). This disparity in the results may highlight the difficulty of a network trained and tested on one limited catalog, to generalise the result to a wider set of events. The most evident difference is the time span between the two sets. Events tested in A fall within the same years as the events used to train the network, whereas events tested in B are all in a subsequent time period.

Further scrutiny of the test of set B suggests that the distribution of accuracy is bimodal. A cluster of 19 values $> 0.5$ with mean 0.77 and standard deviation 0.13, is completed by tail toward very low values of accuracy, including four zero accuracy results. This may indicate that some of the earthquakes belong to a similar type as the ones trained in set A, with an overall fair response to the network test, and that other fall in a different category.

It is likely that the preparatory phase of earthquakes will not develop according to a single, universal dynamic, but that the change in seismic signal preceding an earthquake is either not always the same, or not always present, or not always detectable. Furthermore, the dynamics of the seismic cycle are expected to gradually evolve in time, because the regional tectonic stress changes in the years after a very large earthquake. As a consequence any premonitory change in the seismic noise preceding an earthquake may also evolve in such a way that the accuracy of a network trained in data from years before will decline.

To further explore what possible differences among set B events may result in such disparate network performance, we classed events by date, epicentral distance from the seismic station and depth of the event. No clear correlation is apparent between accuracy and any of the three variables above. However, it is notable that several events in set B belong to a different fault system than the Honshu subduction zone, while a majority of events in set A did belong to such zone (probably an increased zone activity in the aftermath of Tohoku). Indeed focal mechanism of a number of set B events are not thrust faulting, or indicate an almost perpendicular strike with respect to the events of the Honshu subduction zone. It may be argued that these events are atypical or different from the core events trained in set A. Most events that show an accuracy $\leq 0.5$ indeed fall within an atypical category. However, two events with 0 accuracy do belong to the thrust events of the Honshu zone, and three of the atypical events do show a relatively high accuracy.

### 7.2 Training and testing the complete catalog (sets A and B jointly)

In an attempt to improve the generality of the network performance, we conducted 100 cycles of training, testing and verification on the whole dataset (set A and B), each cycle with 100 epochs, in order to explore more systematically the minimization of the misfit function which, as mentioned for the set A simulations, appears to possess local minima, and to retain the best performing combination of weights for the trained network. Across the 100 cycles,

the events were separated in test, train and verification groups using ten different random combinations.

Relatively high accuracy was achieved in some of the cycles for either training (max = 1), testing (max = 0.751) or verification (max = 0.718). Inspection of the ensemble of results, however, reveals that the optimum accuracies are lower than in set A alone. In addition, a high accuracy in all three groups combined (test, train and verification) is difficult to achieve. The maximum combined average of test, train and verification accuracy obtained in a given cycle is 0.769, where individual accuracies for test, train and verification are 0.997, 0.591, and 0.718, respectively.

Returning to the remarks of section (7.1) on the difficulty to generalise the results of the trained network, we may surmise that an improvement in generality can be obtained by offering a larger number and a larger variability of events to the network. However, the improvement in generalisation comes at the cost of a loss in accuracy. This may be the sign that this type of network struggles to train on different typologies of events simultaneously, at least with a database of only a few tens of events.

## 8 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 63 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 36 intervals of 1000 s each (40k time samples). Each interval was further split into 37 windows of 16384 samples each, with an overlap of 15374. For the training, windows were labelled as either noise or precursor, depending to their proximity to hypocentral time, under the hypothesis that the signal change (or precursor) was more significant in the proximity of the earthquake.

After testing several network prototypes we settled on a fully connected network comprising seven convolutional layers.

We first analysed events in the time interval October 2013 – June 2019 (set A, 31 events), using 24 events for the training of the network, and keeping 7 more events separate to be used in the 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 as precursor in the final time interval. Furthermore, we trained the network to distinguish the windows of the noise interval from any other window in the 10 hours. This resulted in a significant (above error bar) accuracy starting about 3 hours before the earthquakes; 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

554 portability of the network to other contexts, without re-training. However, this result for a
555 single distant event is not a statistically significant test of generality.

556 To further test the results we performed tests including the additional 31 events of the
557 catalog (set B). These occurred in the time interval July-2019 to May 2022, postdating the
558 events of set A.

559 Set B was used first as an independent verification group (analysed by the network that
560 was trained on set A). The results were mixed, showing that the network struggled to
561 generalise the results obtained on the events from the first $\approx 5$ years of the catalog, to those
562 belonging to the last $\approx 4$ years. One possible interpretation is that the regional stress regime
563 and the state of the fault may have changed in the years after Tohoku, evolving from a
564 particular post-earthquake stage to the inter-seismic phase in the seismic cycle. This regime
565 shift would events of a different nature than those used for the training batch of set A would
566 be generated.

567 Set B was then used jointly with set A in a complete 63 event catalog that was split in
568 train, test and verification subgroups to perform a fully new evaluation of the network. A
569 hundred cycles were conducted to fully explore the minimisation space. The maximum
570 combined average of test, train and verification accuracy obtained in a given cycle was 0.769,
571 where individual accuracies for test, train and verification are 0.997, 0.591, and 0.718,
572 respectively. The maximum in all cycles for the test accuracy was 0.751. These accuracies
573 are lower than in the case of set A alone, suggesting that the network can be trained on a
574 wider dataset with gain in generality but with loss in accuracy.

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

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

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

592 Slow slip episodes preceding earthquakes have been documented from timescales of
593 seconds (Tape et al., 2018) to weeks or months (Ruiz et al., 2014, 2017; Socquet et al., 2017;
594 McGuire et al., 2005; Bouchon et al., 2011, 2013; Hasegawa & Yoshida, 2015) in natural
595 earthquakes, revealing the progressive growth of instability on a fault patch, and can be
596 simulated in laboratory experiments (Nielsen et al., 2010; Latour et al., 2013; Guérin-Marthe
597 et al., 2019), or numerical models (Ampuero & Rubin, 2008). Fault instability has been anal-
598 ysed previously in the framework of rate-and-state friction, where a critical length $h_{RR}$ for
599 the nucleation patch can be theoretically derived (Dieterich, 1992; Ruina, 1983; Rice & Ru-
600 ina, 1983; Uenishi & Rice, 2003; Rubin & Ampuero, 2005; Ampuero & Rubin, 2008). Close
601 to the instability, oscillations in the slip can take place with increasing amplitude. These may
602 radiate a low-amplitude, low-frequency tremor that increases the relative amplitude of a given
603 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-weakening). 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 further. Because the network results are very uneven, and do not perform as well on all seismic events, it would be interesting to take forward an analysis to separate the earthquakes in different categories. Here we have sought whether accuracy varies with either distance, date, or depth of the earthquakes, without finding any apparent systematic correlation. We note however that most of the events that perform poorly do belong to different focal mechanisms or regions that are not in the same section of the subduction zone.

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. In particular, the lack of generality in the results may prevent extrapolation in time, region or tectonic environment without specific re-training of a neural network, which may be difficult in face of the limited number of large earthquakes in a bounded interval.

## 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.

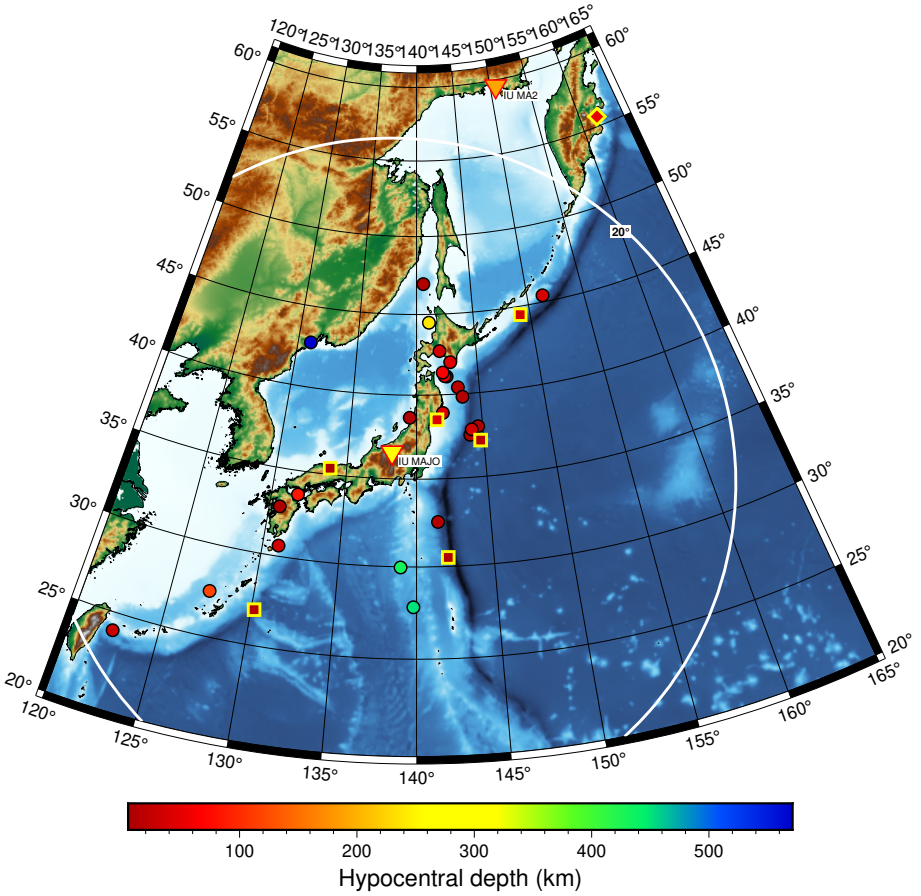


**Figure 1.** Northwest Pacific region surrounding Japan used for training and testing of the CNN. We show 31 $M_w \geq 6$ earthquakes belonging to set A (set B is omitted to avoid cluttering). Filled circles indicate the epicenter of earthquakes in the training dataset. Filled squares indicate the epicenter 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.

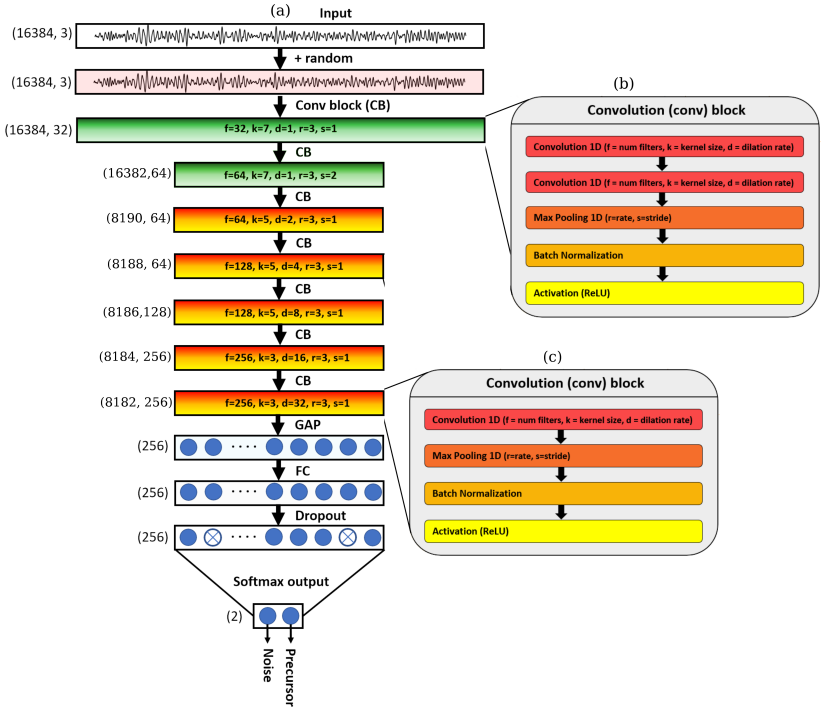


**Figure 2.** CNN used to detect earthquake precursor patterns. (a) Representation of all layers in the network. f is the number of filters, k is the length of the filter kernel and d is the dilation rate in the convolution operation. r and s are the length of the window and the stride of the max-pooling 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. (GAP stands for Global Average Pooling and FC for Fully Connected Layer).

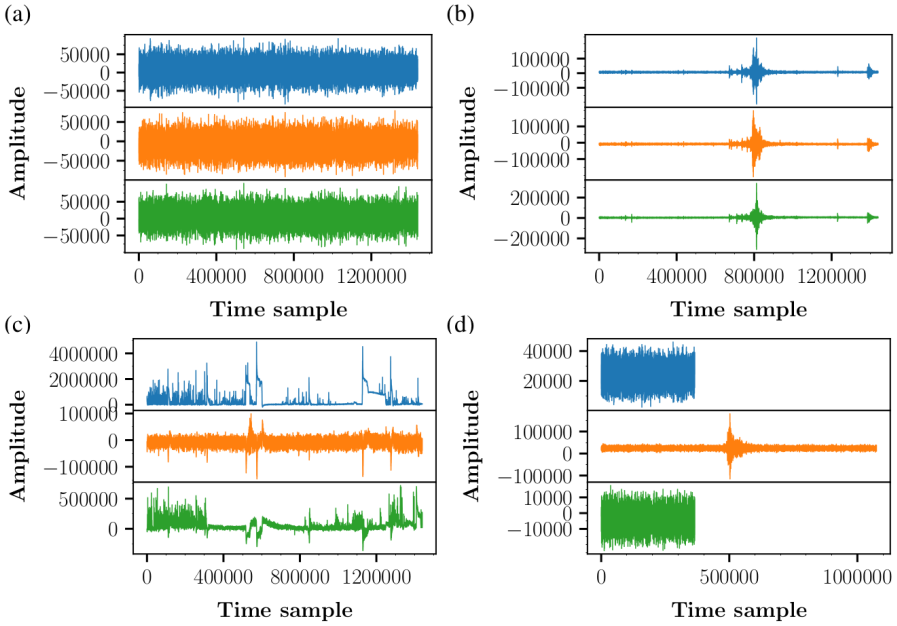


**Figure 3.** Examples of the 3 channels of seismic data (vertical: blue; North-South: orange; East-West: green curves) over the 10-hour period prior to the selected $M_w \geq 6$ earthquakes. (a) 10-hour period with no impulsive earthquake signal above the noise level. Events such as this were included 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.

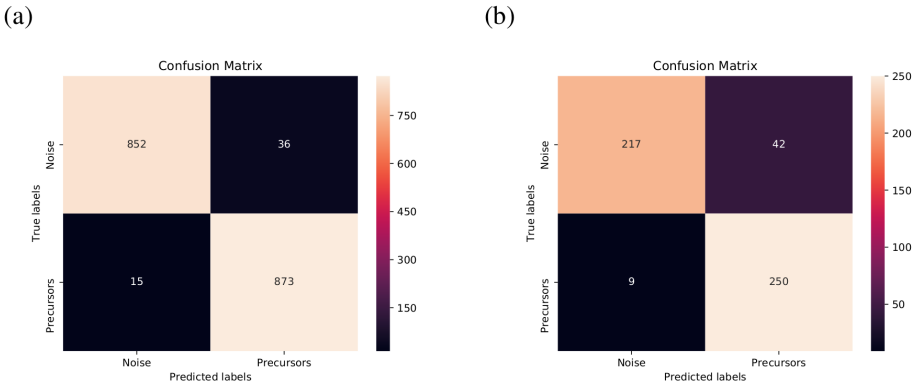


**Figure 4.** Confusion matrices indicating the performance of the classification model by summarising the prediction results on the train and test datasets — (a) Confusion matrix obtained with the best weights (89% test accuracy) on (a) the training dataset and (b) the test dataset. The confusion matrix 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).

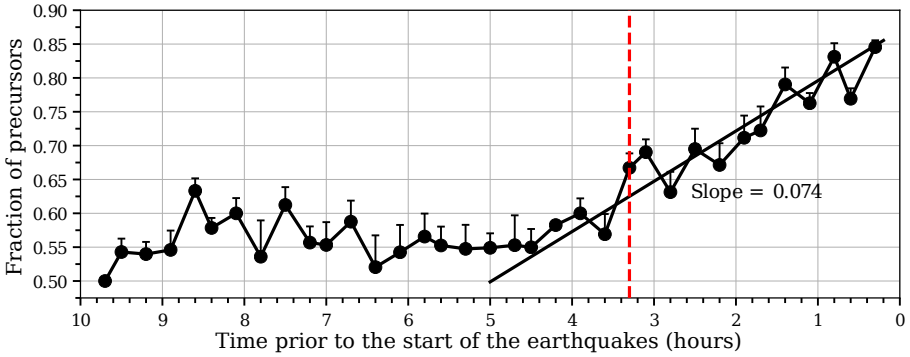


**Figure 5.** Changes in the fraction of windows classified as precursor in the ten hours preceding the earthquakes (in 36 time intervals). The average of five independent network iterations is represented, with the standard deviation shown on one side of the data points to improve clarity. The red,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

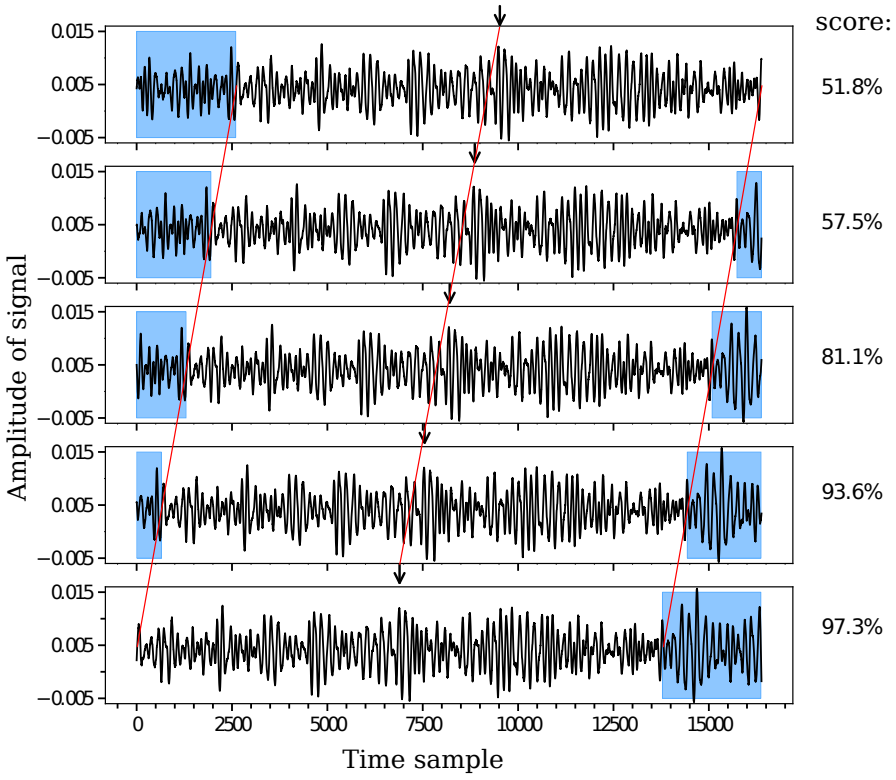


**Figure 6.** Example of sequential precursor windows (channel 0 only for simplicity) and increase in prediction score (probability) as a specific new time interval is incorporated. The windows stride is 650 time steps; this can be visualised by noticing that the region not shaded in blue is the same 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.

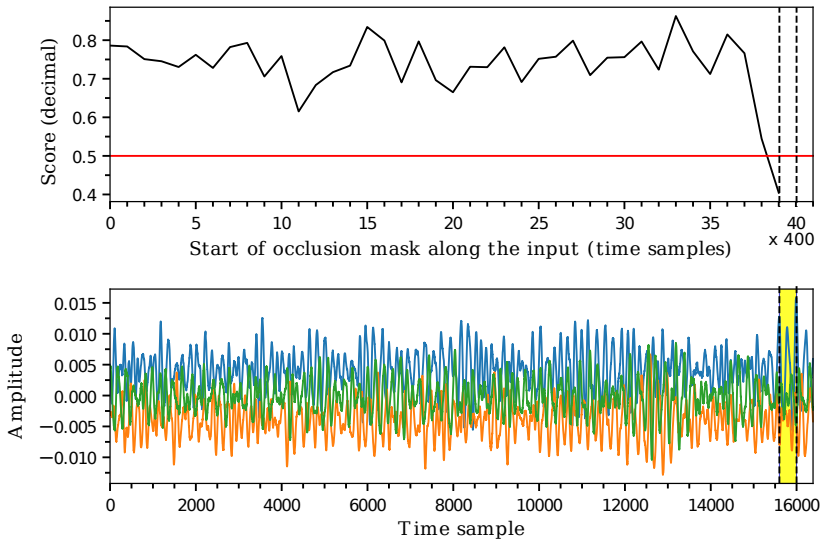


**Figure 7.** Score sensitivity (top) for different positions of the occlusion mask on the precursor window investigated (bottom). A mask length of 400 and a stride of 400 were used. Prediction scores below 0.5 (red line) line indicate regions of the input containing significant, precursor-related information. 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.

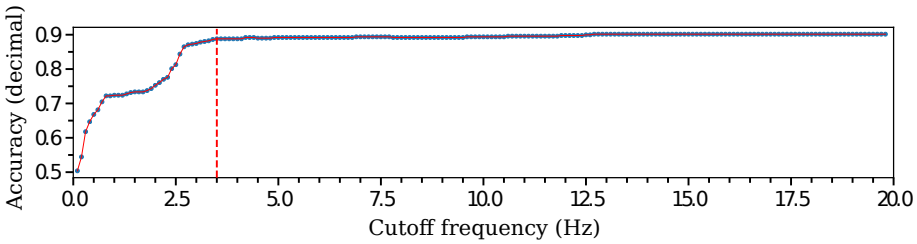


**Figure 8.** Changes in the test accuracy and test loss when applying the low pass filter to the test dataset with a variable cutoff frequency. The red, dashed, vertical line indicates the cutoff frequency at which the test accuracy started to decrease significantly.

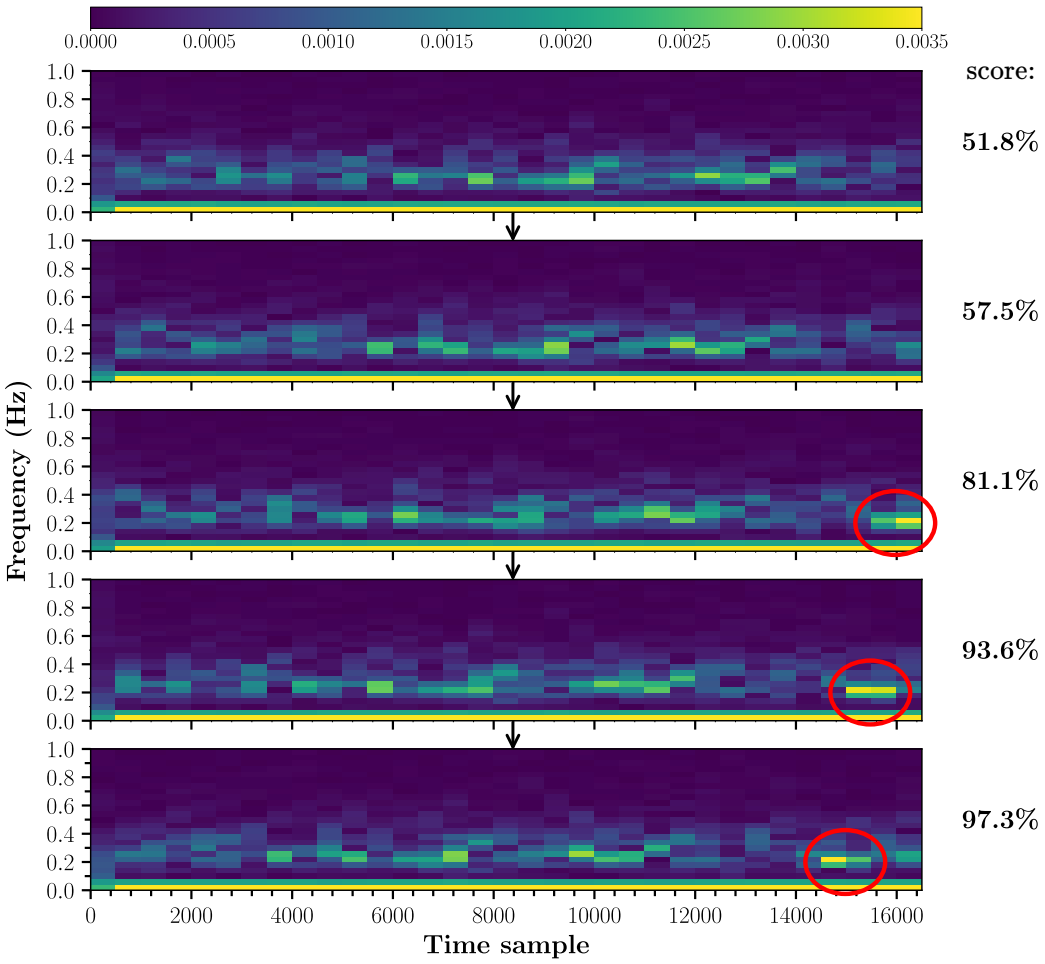


**Figure 9.** Spectrogram of the sequential precursor windows (channel 0 only). The Fourier amplitude was calculated within a sliding window of length 1000, stride 500 and plotted in colour. The prediction score of the network when classifying each window (all 3 channels) as a precursor is indicated.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

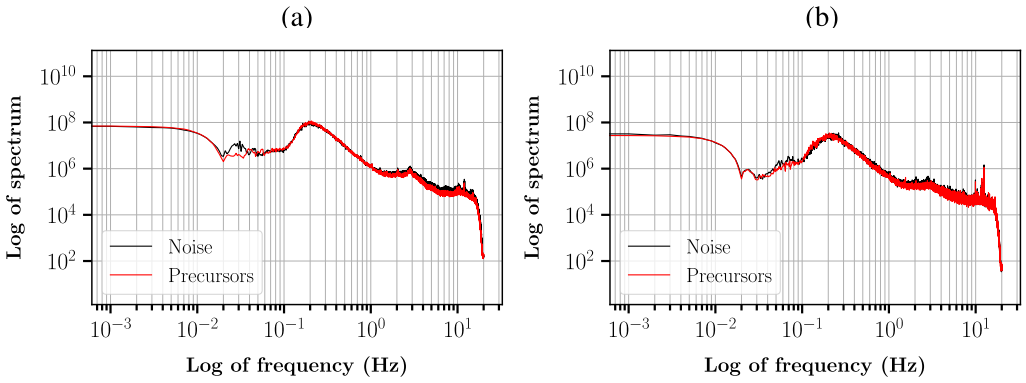


**Figure 10.** Amplitude spectrum of noise windows (black) and precursor windows (red). (a) The cumulative sum of the frequency responses for all events and their 3 channels were calculated separately for noise-labelled windows (black) and precursor-labelled windows (red) in the training dataset. (b)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

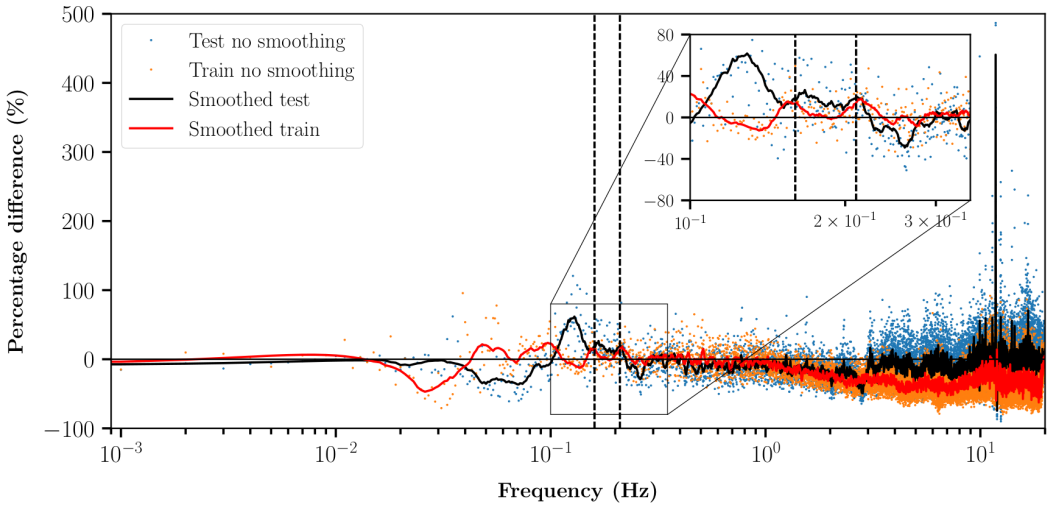


**Figure 11.** Relative percentage difference between the cumulative frequency spectra in Figure (10) for precursor-labelled and noise-labelled data. The test and train results are plotted on the same graph for ease of comparison. The discrete frequencies are shown as blue and orange dots, while a smoothed 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).