CREIME – A Convolutional Recurrent model for Earthquake Identification and Magnitude Estimation

Megha Chakraborty¹, Darius Fenner¹, Wei Li¹, Johannes Faber¹, Kai Zhou¹, Georg Rümpker², Horst Stöcker¹, and Nishtha Srivastava¹

¹Frankfurt Institute for Advanced Studies ²Goethe University Frankfurt

November 23, 2022

Abstract

The detection and rapid characterisation of earthquake parameters such as magnitude are important in real time seismological applications such as Earthquake Monitoring and Earthquake Early Warning (EEW). Traditional methods, aside from requiring extensive human involvement can be sensitive to signal-to-noise ratio leading to false/missed alarms depending on the threshold. We here propose a multi-tasking deep learning model – the Convolutional Recurrent model for Earthquake Identification and Magnitude Estimation (CREIME) that: (i) detects the earthquake signal from background seismic noise, (ii) determines the first P-wave arrival time and (iii) estimates the magnitude using the raw 3-component waveforms from a single station as model input. Considering, that speed is essential in EEW, we use up to two seconds of P-wave information which, to the best of our knowledge, is a significantly smaller data window compared to the previous studies. To examine the robustness of CREIME we test it on two independent datasets and find that it achieves an average accuracy of 98\% for event-vs-noise discrimination and can estimate first P-arrival time and local magnitude with average root mean squared errors of 0.13 seconds and 0.65 units, respectively. We compare CREIME with traditional methods such as short-term-average/ long-term-average (STA/LTA) and show that CREIME has superior performance, for example, the accuracy for signal and noise discrimination is higher by 4.5\% and 11.5\% respectively for the two datasets. We also compare the architecture of CREIME with the architectures of other baseline models, trained on the same data, and show that CREIME outperforms the baseline models.







100 150 200 250 300 Hypocentral distance (km)

A0000 Prequencies for the second sec

60 80 SNR of data (dB)

CREIME – A Convolutional Recurrent model for 1 Earthquake Identification and Magnitude Estimation

Megha Chakraborty ^{1,4}, Darius Fenner ^{1,6}, Wei Li ¹, Johannes Faber ^{1,2}, Kai Zhou ^{1,2,3}, Georg Rümpker ^{1,4}, Horst Stoecker ^{1,2,3,5}, Nishtha Srivastava ^{1,4} 3

¹Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany 5 ²Institute for Theoretical Physics, Goethe Universität, 60438 Frankfurt am Main, Germany ³Xidian-FIAS international Joint Research Center, Giersch Science Center, 60438 Frankfurt am Main, 6 7 Germany 8 ⁴Institute of Geosciences, Goethe-University Frankfurt, 60438 Frankfurt am Main, Germany 9 ⁵GSI Helmholtzzentrum für Schwerionenforschung GmbH, 64291 Darmstadt, Germany ⁶Johannes Gutenberg-Universität Mainz, 55122, Mainz,Germany 10 11

Key Points:

2

12

13	•	We use a novel sequence-to-sequence mapping to train a deep learning model to
14		detect an earthquake, pick the P-wave arrival and estimate its magnitude.
15	•	The proposed model can perform reasonably well with 5 second windows contain-
16		ing only up to 2s of P-wave data.
17	•	We show that our model can outperform traditional methods like STA/LTA and
18		the existing deep learning models.

Corresponding author: Nishtha Srivastava, srivastava@fias.uni-frankfurt.de

19 Abstract

The detection and rapid characterisation of earthquake parameters such as magnitude 20 are important in real time seismological applications such as Earthquake Monitoring and 21 Earthquake Early Warning (EEW). Traditional methods, aside from requiring extensive 22 human involvement can be sensitive to signal-to-noise ratio leading to false/missed alarms 23 depending on the threshold. We here propose a multi-tasking deep learning model – the 24 Convolutional Recurrent model for Earthquake Identification and Magnitude Estimation 25 (CREIME) that: (i) detects the earthquake signal from background seismic noise, (ii) 26 determines the first P-wave arrival time and (iii) estimates the magnitude using the raw 27 3-component waveforms from a single station as model input. Considering, that speed 28 is essential in EEW, we use up to two seconds of P-wave information which, to the best 29 of our knowledge, is a significantly smaller data window compared to the previous stud-30 ies. To examine the robustness of CREIME we test it on two independent datasets and 31 find that it achieves an average accuracy of 98% for event-vs-noise discrimination and 32 can estimate first P-arrival time and local magnitude with average root mean squared 33 errors of 0.13 seconds and 0.65 units, respectively. We compare CREIME with traditional 34 methods such as short-term-average / long-term-average (STA/LTA) and show that CREIME 35 has superior performance, for example, the accuracy for signal and noise discrimination 36 is higher by 4.5% and 11.5% respectively for the two datasets. We also compare the ar-37 chitecture of CREIME with the architectures of other baseline models, trained on the 38 same data, and show that CREIME outperforms the baseline models. 39

⁴⁰ Plain Language Summary

The detection of earthquakes and rapid determination of parameters such as mag-41 nitude is crucial in Earthquake Monitoring and Earthquake Early Warning (EEW). Ex-42 isting methods used to make such estimations are empirical and require expert analysts 43 to define involved parameters, which is quite challenging. They are also sensitive to noise, 44 which could lead to erroneous results. In this paper we propose a the Convolutional Recurrent 45 model for Earthquake Identification and Magnitude Estimation (CREIME) which is ca-46 pable to detect an earthquake within 2 seconds of the first P-wave arrival and provides 47 a first estimate for its magnitude. We test the model on two independent datasets to demon-48 strate its generalizability. CREIME successfully discriminates between seismic events and 49 noise with an average accuracy of 98% and can estimate first P-arrival time and local 50 magnitude with average root mean squared errors of 0.13 seconds and 0.65 units, respec-51 tively. We also show that CREIME can perform better than traditional methods like STA/LTA 52 and previously published deep learning architectures in the context of rapid character-53 isation. 54

55 1 Introduction

According to its original definition (Richter, 1935) the magnitude of an earthquake 56 is the logarithm of the maximum trace amplitude expressed in microns measured by a 57 standard short-period torsion seismometer at an epicentral distance of 100km. It is one 58 of "the most important and also the most difficult parameters" involved in real-time seis-59 mology (Jin et al., 2013) particularly since most magnitude scales such as local magni-60 tude (m_L) , body wave magnitude (m_B) , surface wave magnitude (m_S) are empirical and 61 saturate at different magnitude ranges (Chung & Bernreuter, 1981; Ekström & Dziewon-62 ski, 1988). This, coupled with the complexity of the nature of the geophysical processes 63 affecting earthquakes, makes it very difficult to have a single reliable measure for the size 64 of an earthquake (Kanamori & Stewart, 1978). Magnitude values measured in different 65 scales may thus differ by more than 1 unit, particularly for extremely large events due 66 to saturation effects (Howell Jr, 1981; Giardini, 1988; Geller, 1976; Kanamori, 1983). Even 67 for the same magnitude scale, values reported by different agencies may differ by up to 68

0.5 units (Mousavi & Beroza, 2020). Traditionally, frequency-domain parameters such 69 as predominant period τ_p^{max} (Nakamura, 1988; R. Allen & Kanamori, 2003), effective 70 average period τ_c (Kanamori, 2005; Kuyuk & Allen, 2013; Jin et al., 2013) and ampli-71 tude domain parameters such as peak displacement (P_d) (Wu & Zhao, 2006; Kuyuk & 72 Allen, 2013; Jin et al., 2013) calculated from the initial 1-3 seconds of P-waves have been 73 shown to provide reliable estimates of (body wave) magnitudes through empirical rela-74 tions. Such methods have been applied to Earthquake Early Warning (EEW) systems 75 in Japan, California, Taiwan etc. (R. Allen et al. (2009) and the references therein). It 76 has further been shown that the correlation of such parameters increases steadily upon 77 increasing the duration of data used (Ziv, 2014). Thus, there is an "inherent trade-off 78 between speed and reliability" (Meier et al., 2019). 79

Traditional machine-learning algorithms were "limited by their inability to process 80 data in its raw format" (LeCun et al., 2015) and the need for hand-crafted features. This 81 challenge has been overcome by the emergence of deep learning. Deep learning comprises 82 hierarchical feature learning methods (LeCun et al., 2015), whereby several simple non-83 linear mathematical functions are applied to the raw data, to extract an increasingly ab-84 stract representation of the data at each level. It is the job of the deep learning model 85 to learn the parameters of these functions. The advent of deep learning, coupled with 86 the availability of large volumes of data and affordable computational power in the form 87 of GPUs, have led to state-of-the-art results in image recognition (Krizhevsky et al., 2017; 88 He et al., 2016), speech recognition (Mikolov et al., 2011; Hinton et al., 2012), and nat-89 ural language processing (Peters et al., 2018; Collobert et al., 2011). In fields such as seis-90 mology, which have been data-intensive since their very origin and are witnessing an ex-91 ponential increase in the volume of data (Kong et al., 2018), deep learning has proven 92 successful in several tasks such as event detection (Perol et al., 2018; Z. Li et al., 2018; 93 Meier et al., 2019; W. Li et al., 2022; Fenner et al., 2022) and phase picking (W. Zhu & 94 Beroza, 2019; Mousavi et al., 2020; Liao et al., 2021; W. Li et al., 2021; Zhou et al., 2019), 95 event location characterisation (Perol et al., 2018; Panakkat & Adeli, 2009; Kuyuk & Susumu, 96 2018), first motion polarity detection (Ross et al., 2018; Hara et al., 2019), among oth-97 ers. 98

A deep learning based approach for magnitude estimation was presented by Mousavi 99 and Beroza (2020). The model presented in that paper focuses on estimating the mag-100 nitude for an earthquake waveform, using a window length of 30 seconds that includes 101 both the P- and S-wave information. The input to the model are earthquake traces, and 102 event-vs-noise discrimination and first P-arrival are not included in its goals. The use 103 of deep learning facilitates the learning of the most relevant features directly from the 104 waveform. This approach suffers from under-estimation at high magnitudes as these mag-105 nitudes are rare in nature and, hence, under-represented in the training data. In order 106 to overcome this drawback we propose a two-pronged approach – resampling the data 107 to get a more uniform magnitude distribution and penalising the underestimation of high 108 magnitudes during model training. As already mentioned the model presented in Mousavi 109 and Beroza (2020) uses S-wave information which makes it unsuitable for the purpose 110 of rapid characterisation and EEW, where the information from the faster P-waves is 111 leveraged to issue a warning before the slower, and more devastation S-waves hit the sur-112 face (Cremen & Galasso, 2020)). 113

In this paper we present a novel approach to achieve multi-tasking Convolutional 114 Recurrent model for Earthquake Identification and Magnitude Estimation (CREIME), 115 which can simultaneously perform earthquake identification, local magnitude estimation 116 and first P-wave arrival time regression solely based on 1-2 seconds P-wave recording. 117 Unlike J. Zhu et al. (2021) which uses a set of twelve features extracted from 3 seconds 118 of data to perform magnitude estimation, CREIME is end-to-end using a combination 119 of Convolutional and Recurrent neural network to extract features directly from the raw 120 waveform. The motivation for using such a small duration of P-wave data lies in its po-121

tential utility in applications such as rapid earthquake characterisation for EEW systems
(R. Allen et al. (2009); R. M. Allen and Melgar (2019) and references therein). While
multiple-station based approaches are generally more robust and reliable, single station
approaches are faster and therefore can be more useful in places where human settlements
may lie very close to the earthquake epicenter, such as Southern California.

The model presented here can be seen as a prototype that can be adapted into EEW 127 systems and has a potential to provide reliable first estimates. We demonstrate the ro-128 bustness of our model, by testing it on two datasets. It is ensured that these datasets 129 have no overlap in terms of the traces they contain to assert the generalizability of the 130 model. We also compare the effects of using different types of ground motion data as the 131 input to the model. As a final step we test the model on S-wave arrivals which are not 132 encountered by the model during training, to verify that S-wave arrivals from low mag-133 nitude events do not get wrongly identified as P-arrivals for high magnitude events. This 134 implies that the model can easily be adapted on real time data. 135

136 **2 Data**

137

2.1 STEAD

The data used to train and test CREIME has been obtained from the STanford 138 EArthquake Dataset (STEAD) (Mousavi et al., 2019). It is a high-quality benchmarked 139 global dataset of labelled seismograms which have been detrended, bandpass filtered be-140 tween 1.0-40.0 Hz and resampled to 100 Hz. There are a total of 7 different types of in-141 struments in which the data has been recorded, of these, 99.5% are either high-gain broad 142 band or extremely short period. Each seismogram is of duration 1 minute and is rep-143 resented in the form of NumPy arrays (Harris et al., 2020) of dimensions 6000×3 . All 144 earthquake waveforms are associated with local earthquakes with epicentral distance no 145 greater than 350km. The metadata includes 35 attributes for each earthquake waveform 146 and 8 attributes for each noise waveform. 147

For the sake of uniformity in magnitude, of the 23 different magnitude scales in which 148 earthquakes are reported, we only choose events for which the magnitudes are reported 149 in the 'ml' scale, i.e., local magnitude as these events constitute the majority (above 70%) 150 of the dataset. To ensure that extremely noisy data is left out from the training and test-151 ing process only waveforms with a signal-to-noise ratio (provided in the metadata) above 10 dB are used (similar to Mousavi and Beroza (2020) where 20 dB is the cutoff signal-153 to-noise ratio). The noise and earthquake traces are roughly divided in the ratio 60:10:30 154 for training, validation and test sets. A total of 32,356 traces are used for training. For 155 earthquake waveforms, it is made sure that all traces associated with one earthquake event 156 are present in only one of the aforementioned three sets with the help of the 'source_id' 157 attribute from the metadata. For noise waveforms, traces corresponding to a particu-158 lar station can be present in only one of the three sets. This ensures that the test dataset 159 is "truly unseen" to the model and hence, can give a reliable evaluation of the model's 160 performance. 161

In accordance with the Gutenberg-Richter power-law (Gutenberg & Richter, 1944), 162 high magnitude earthquakes are rare in nature. This power-law is reflected in the dataset 163 as well (with a magnitude of completeness around 1-1.5). The distribution of magnitudes 164 in the original dataset is similar to that of the testing data shown in Figure 1. This kind 165 of imbalance in the distribution of the target variable in a regression problem tends to 166 bias the model's performance towards lower magnitudes (<2.5) (Krawczyk, 2016) as ob-167 served in Mousavi and Beroza (2020). So, to make sure that the model can perform a 168 reliable estimation over all magnitude ranges, we perform random under-sampling up 169 to magnitudes of 4.0 and random over-sampling for magnitudes above 4.5. For this, dif-170 ferent rates (chosen by trial and error) of undersampling or oversampling (achieved by 171

Figures/Figure1.png

Figure 1. Distribution of magnitudes in training data (in slate blue) and chunk of STEAD(Mousavi et al., 2019) data used for testing (in orange). Note that the y-axis on the left corresponds to the training data distribution and that on the right corresponds to the test data distribution. While random undersampling and oversampling are applied to different magnitude ranges for training data in an attempt to get a uniform distribution, the original magnitude distribution of the test dataset is retained.

using windows with different starting time between 312-412 samples before P-arrival time)
are applied to different magnitude ranges on the training and validation sets. This results in a training set with a magnitude distribution as shown in Figure 1. No such augmentation is applied to the test set (Figure 1) to retain the real world distribution of earthquake magnitudes encountered by the system. Furthermore, for training and validation,
the number of noise traces chosen is exactly equal to the number of event traces.

178 2.2 INSTANCE

We further test our model on the INSTANCE dataset (Michelini, Cianetti, Gaviano,
 Giunchi, Jozinovic, & Lauciani, 2021), which is a recently published dataset comprising

1.2 million three-component waveform traces and 130,000 noise traces, each with a du-181 ration of 2 minutes, recorded primarily by the Italian National Seismic Network (net-182 work code IV). Corresponding to each trace 100 metadata, including magnitude and P-183 wave arrival sample, are provided. To make sure that there is no overlap with the train-184 ing data, we exclude data from stations that are part of the STEAD dataset. We choose 185 only traces for which magnitudes are provided in the 'ML' scale. For a fair evaluation 186 of our model, we use only those traces with a single event and with distance and depth 187 each within the corresponding maximum value present in the training data. Once again, 188 traces with signal-to-noise ratio lower than 10 dB are not used. This leaves us with 135,347 189 traces corresponding to events between April 2005 to January 2020 and having a mag-190 nitude distribution as shown in Figure B1 in the appendix. The preprocessing steps for 191 this data are very similar to those of the STEAD data except the bandpass filtering, so 192 we apply a bandpass filter between 1.0 to 40.0 Hz using the bandpass function from ob-193 spy.signal.filter (Beyreuther et al., 2010). 194

¹⁹⁵ 3 Methodology

We use supervised learning (Chollet, 2017, Chapter 4) in this work to achieve earthquake identification and magnitude estimation, together with P-arrival time regression, based upon short records of P-wave data. The local magnitude (which is provided in the metadata for both the STEAD and INSTANCE datasets) or Richter scale magnitude (Richter, 1935) has the form:

219

$$M_L = \log A - \log A_0 + S \tag{1}$$

where, A is the peak horizontal amplitude measured on a Wood-Anderson seismograph, 202 and A_0 and S are empirically determined distance and station correction terms derived 203 from amplitude-distance relations representing attenuation and site functions respectively. 204 While the peak amplitude can be directly obtained from the input data as we do not ap-205 ply normalisation, it is expected that the model will learn the distance parameters, which 206 are not provided explicitly, from the frequency content of the data itself. All three com-207 ponents are provided, to facilitate the learning of site effects (a similar approach has been 208 followed by Mousavi and Beroza (2020)). We provide the data in units of 'counts' and 209 do not perform instrument corrections, which gives the advantage that the analysis can 210 be done in real-time. 211

A sequence-to-sequence approach is developed – the input to our model being 512 samples (5.12s) from 3 channels and the output is an array of the same length (512 samples). The data window for earthquake waveforms is chosen in such a way as to include 1 to 2 seconds of P-wave data, preceded by pre-signal noise (for noise waveforms the window has 512 samples of noise). This type of windowing allows the model to learn the noise characteristics (Münchmeyer et al., 2020). The Y-label for each X is a 512 × 1 array. These values are defined as follows:

$$y_i = \begin{cases} M & \text{if } i \ge i_p \\ -4 & \text{otherwise} \end{cases}$$
(2)

where M is the magnitude of the event and i_p denotes the P-arrival sample (in case 220 of earthquake waveforms). The value -4 representing noise is arbitrary and chosen em-221 pirically by testing model performance on the validation data. The use of an arbitrary 222 negative number to represent noise was explored by Yanwei et al. (2021). An example 223 of this labelling for event and noise data is shown in Figure 2a. We have also tried mod-224 ifying the final layer of the model to output two numbers corresponding P-arrival sam-225 ple and magnitude instead of a sequence, similar to the approach of Yanwei et al. (2021) 226 (not shown in the paper). However, our observation was that the sequence-to-sequence 227 mapping approach leads to smaller errors. 228

Figures/Figure_2.jpg

Figure 2. (a) Example of labelling for an event trace (left) and a noise trace (right); the label value is set to -4 for all samples before the P-arrival and the event magnitude for the P-arrival sample onward; for the noise trace it is set at -4 for all samples. (b) A schematic showing the architecture of the CREIME model; each convolution layer has a kernel size 16 and the number of filters are 32, 16 and 8; each Maxpooling layer reduces the dimension of the data by a factor of 4 and the Bi-LSTM layers have dimensions of 128 and 256 respectively.



Figure 3. The variation of the training and validation loss as training progresses. The validation and training losses remain close to each other, which shows that the training is quite robust and there is no discernible overfitting.

The architecture of the CREIME model consists of three sets of 1D Convolution 229 (Kiranyaz et al., 2015) and Maxpooling (Nagi et al., 2011) layer followed by two bidi-230 rectional Long-Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) layers 231 of dimensions 128 and 256, which is followed by the output layer of dimensionality 512 232 (Figure 2b). The convolutional and maxpooling layers are used to extract and retain the 233 relevant features while downsampling the data volume. Bidirectional LSTMs are used 234 because of their ability to detect temporal dependencies for sequential data such as earth-235 quake waveforms. Each convolution has a kernel size 16, a stride of 1, and padding type 236 "same"; the number of filters is 32, 16 and 8, respectively. Each maxpooling layer reduces 237 the size of the data by a factor of 4. Unlike the approach in Lomax et al. (2019) we find 238 the model performance to be better when we use the original data without any normal-239 isation. The model has a total of 1,454,992 trainable parameters and is trained using RMS 240 Propagation optimiser (Tieleman & Hinton, 2012), with a batch size of 512. The model 241 is implemented using Keras (Charles, 2013). On an NVIDIA A100GPU the training pro-242 cess takes less than 1 second per epoch. Each hyperparameter, including the number of 243 layers in the model was chosen through meticulous experimentation by running several 244 iterations of training and subsequent testing on the validation data. 245

We use early stopping (Prechelt, 2012) during the training to prevent overfitting. The validation loss is monitored and the training stops if it does not reduce for 15 consecutive epochs. We have an initial learning rate of 10^{-3} and reduce it by a factor of 10 until it reaches 10^{-6} if the validation loss does not go down for 10 consecutive epochs. The model with the lowest validation loss is saved. With these conditions the model trains for 71 epochs. The training history (i.e. learning curve) is shown in Figure 3.

For the cost function, we customized a combination of three losses, as different loss functions proved to be working better for different tasks and for different ranges of magnitude. The weights were determined by a trial and error method.

1. Mean Squared Error (MSE) with a weight of 40%: This is the average of squared 255 values of errors corresponding to each data point in a minibatch. For k output val-256 ues and a batch size n it has the form: 257

$$\mathcal{L}_{MSE} = \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{k} \sum_{i=0}^{k-1} (y_{true}^{i,j} - y_{pred}^{i,j})^2$$
(3)

Here $y_{true}^{i,j}$ and $y_{pred}^{i,j}$ represents the true and predicted y values of the i-th sam-259 ple for the j-th example in the minibatch, respectively. 260 2. Mean Absolute Error (MAE) with a weight of 40%: This is the average of abso-261

lute errors corresponding to each data point in a minibatch. For k output values and a batch size n it has the form:

$$\mathcal{L}_{MAE} = \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{k} \sum_{i=0}^{k-1} |y_{true}^{i,j} - y_{pred}^{i,j}|$$
(4)

3. Magnitude Estimation Loss with a weight of weight: 20%: As mentioned in the 265 Introduction, we penalise the underestimation of magnitude, for high magnitude 266 events (and overestimation for noise traces). To achieve this we define a third loss 267 function. For k output values and a batch size n it has the form: 268

$$\mathcal{L}_{ME} = \frac{1}{n} \sum_{j=0}^{n-1} \alpha^j \frac{1}{k} \sum_{i=0}^{k-1} (y_{true}^{i,j} - y_{pred}^{i,j})$$
(5)

where,

258

262

263

264

269

276

277

278

279

280

281

$$\alpha^{j} = \begin{cases} \text{Event Magnitude,} & \text{for events} \\ -4, & \text{for noise} \end{cases}$$

As already mentioned, we utilise the output from our model to perform three tasks: 270 discrimination between seismic event and noise, magnitude estimation, P-arrival sam-271 ple detection. Based on a manual investigation of the output data and a subsequent test-272 ing on the validation dataset we used the following analysis to extract the desired pa-273 rameters from the 512 sample sequence output by the model: 274

1. Predicted magnitude, 275

$$M_{pred} = \frac{1}{10} \sum_{i=(k-9)}^{k} y_{pred}^{i}$$
(6)

where k is the number of samples, in our case, 512

2. Considering the first sample point in the data window as zeroth sample, we define Ρ

-arrival sample =
$$i_{pred}^p$$
 such that $y_{pred}^i > -0.5$ for all $i \ge i_{pred}^p$ (7)

3. If M_{pred} calculated by equation (6) is less than -0.5 then it is classified as noise.

The value -0.5 is chosen empirically based on the magnitude range of the data. For 282 a detailed description of the metrics please refer to the Appendix A. 283

4 Results 284

The model was tested on a chunk of the STEAD dataset (Figure 1). Figure 4a shows 285 the confusion matrix for noise/event classification; Figure 4b shows that the predicted 286 magnitudes for noise data wrongly classified as event tend to be low (mostly ≤ 1) indi-287 cating that the possibility of false alarms caused by noise is low and events which are 288 wrongly classified as noise are usually of low magnitude (≤ 2) indicating a low risk of 289

Figures/Figure_4.jpg

Figure 4. Analysis of CREIME model performance as a classifier on STEAD Data. It achieves an accuracy of 98.58%. The true magnitude of events misclassified as noise and predicted magnitude of events misclassified as noise tends to be low, which reduces the chance of missed or false alarms. (a) Confusion matrix for classifier performance. (b) Distribution of predicted magnitudes of noise misclassified as event and true magnitude of events misclassified as noise. Figures/Figures_5.jpg

Figure 5. (a) Example of correct classification of an event trace (left) and a noise trace (right); one can see that the predicted magnitude for the event trace is very close to the true magnitude. (b) Example of incorrect classification of an event trace as noise trace(left) and a noise trace as an event trace (right); the event is a low magnitude one, and quite difficult to identify in this frequency range; the noise level in case of the noise trace is quite high gets classified as a low magnitude earthquake.

Figures/Figures_6.jpg

Figure 6. Analysis of model performance as a regressor on STEAD Data. The density plot shows that the highest density of points lies close to the zero error line; in spite of our penalization of under-estimation of high magnitudes, some under-estimation is observed above a magnitude of 5.5. In over 90% of the cases, the error in predicted magnitudes is less than 1 unit. (a) Relationship between true and predicted magnitude values. (b) Distribution of errors in predicted magnitudes. (c) Distribution of errors in P-arrival estimation.

false alarms, which is reassuring. Figure 5 shows examples of input and corresponding outputs for correctly and incorrectly classified traces.

The scatter plot for predicted versus true magnitudes is shown in Figure 6a. It is 292 worth to note that for majority of the events (shown with higher relative density in the 293 plot) the prediction reproduce well for the true magnitudes up to 5.5. For higher mag-294 nitudes events, some degree of underestimation is observed in spite of the penalty incor-295 porated in the loss function. The result here, however, is an improvement over Mousavi 296 and Beroza (2020), where magnitude underestimation starts to occur from a magnitude 297 of 4. It should be taken into account that the data used to train the MagNet Mousavi and Beroza (2020) has a signal-to-noise ratio above 20 dB whereas, we use a lower thresh-200 old of 10 dB in our analysis. The histogram for errors in magnitude (Figure 6b) has a 300 mean of -0.06 units, and a slight left skew, reflecting our penalisation of underestima-301 tion of magnitudes. The histogram for errors in predicted P-arrival (Figure 6c), is also 302 unimodal, with a higher negative skewness indicating, that the P-arrival is more often 303 predicted to be at a *later* time than it really is. The kurtosis for errors in P-arrival pre-304 diction is also much higher than that for magnitude prediction, indicating that errors 305 in P-arrival predictions have a much narrower peak compared to errors in magnitude pre-306 diction. Similar results are observed for the INSTANCE dataset. We refer interested read-307 ers to Appendix B for the corresponding figures. 308

309 Comparison with other models

We compare our model with ones published in the papers listed below. It is important to note here, that the input data for the models in these studies differs from our data in terms of length, pre-processing etc. Therefore, for an unbiased comparison, all models have been retrained on the same training data that we use for our model. This is essentially a comparison between different architectures and not between the methodology presented by the respective authors.

- 1. MagNet (Mousavi & Beroza, 2020): This paper presents a deep learning model 316 to perform only magnitude estimation using 30 seconds of data including both P 317 and S phases. While both MagNet and CREIME use a combination of CNNs and 318 bidirectional LSTMs, they differ significantly in the number of layers (MagNet uses 319 2 Convolutional layers and 1 bi-LSTM whereas CREIME uses 3 Convolutional lay-320 ers and 2 bi-LSTMs), the model output (MagNet outputs the estimated magnitude 321 and the aleatoric uncertainty whereas CREIME outputs a 512 dimensional array) 322 and the choice of hyperparameters (such as number of filters in the Convolutional 323 layers and dimension of LSTM). Unlike MagNet, CREIME does not use dropout 324 layers. The only modification we make to the original architecture of MagNet, while 325 re-training it, is to change the input shape from (3000,3) to (512,3). We then com-326 pare this model with CREIME in terms of estimation of event magnitudes. 327
- 2. CNN model for signal noise discrimination (Meier et al., 2019): The model presented in this paper originally takes 4s of data, starting 0.5 to 1.5 seconds before
 the P-arrival to discriminate between earthquake signals and noise. We train it
 on our data while keeping the architecture intact except a change in the input dimensions. Unlike the original paper, however, we do not impose a lower limit on
 the magnitudes of the events.
- 3. ConvNetQuake_INGV (Lomax et al., 2019): This model is inspired by the Con-334 vNetQuake (Perol et al., 2018), and uses 10 seconds of data to perform multiclass 335 classification to identify seismic events and characterise earthquake parameters 336 such as magnitude, distance, depth and azimuth. While the original architecture 337 uses 9 CNN layers, each downsampling the data by a factor of 2, we use only 8 338 (similar to Perol et al. (2018)) since the length of data in our case is almost half 339 of that in the original paper. Further, in the last layer we use 31 classes for mag-340 nitude instead of 20 in the original paper giving a total of 32 nodes (one for sig-341

		Metric							
		A	Precision (%)		Recall (%)		F1-score (%)		
Dataset	Model Architecture	Accuracy (76)	Event	Noise	Event	Noise	Event	Noise	
	CREIME Model	98.58	99.64	96.25	98.31	99.18	98.97	97.70	
	CNN Model	89.72	99.18	75.37	85.93	98.37	92.08	85.35	
SIEAD	$ConvNetQuake_INGV$	96.56	99.12	91.30	95.91	98.05	97.49	94.55	
	STA/LTA Algorithm	94.08	96.03	89.70	95.43	91.00	95.73	90.34	
	CREIME Model	97.59	98.66	95.75	97.53	97.68	98.10	96.71	
INCOM	CNN Model	91.71	96.77	84.33	90.00	94.71	93.23	89.22	
INSTANCE	$ConvNetQuake_INGV$	86.48	96.00	74.16	82.79	93.47	88.90	82.70	
	STA/LTA Algorithm	86.03	90.87	78.49	86.81	84.66	88.79	81.46	

Table 1. Comparison between the performance of CREIME model and other baseline modelsas a classifier for events and noise. CREIME model outperforms the other models.

Table 2. Comparison between magnitude estimation by CREIME model and other baselinemodels. The smallest errors are shown by CREIME model.

		Metric				
Dataset	Model Architecture	Mean Error	St. dev. of Error	RMSE	MAE	
	CREIME Model	-0.06	0.60	0.61	0.46	
STEAD	MagNet ConvNetQuake_INGV	-0.29 0.41	$\begin{array}{c} 0.65 \\ 1.05 \end{array}$	$0.72 \\ 1.13$	0.53 0.94	
	CREIME Model	-0.02	0.69	0.69	0.54	
INSTANCE	MagNet	-0.33	0.80	0.86	0.68	
	$ConvNetQuake_INGV$	0.78	0.98	1.25	1.04	

nal vs noise discrimination). To compare the magnitude regression performance with CREIME we take the predicted magnitude to be the arithmetic mean of the boundaries for the predicted class.

In addition to these deep learning models, we also compare our model with the Short-Term Average/Long-Term Average method (STA/LTA)(R. V. Allen, 1978), to evaluate the performance of our model in terms of classification and P-arrival time prediction. This is done by using the *classic_sta_lta* from *Obspy* (Beyreuther et al., 2010). The best set of parameters, determined on the basis of a grid search on the training data are: shortterm window length = 20 samples (0.2s), long-term window length = 200 samples (2s), and trigger threshold = 4.0.

The performance metrics for the CREIME classifier in comparison with other classification models and the STA/LTA algorithm are summed up in Table 1. CREIME outperforms all the other architectures trained on the same data, and the conventional STA/LTA algorithm. The performance of the model in estimating magnitude and P-arrival time is summarised in tables 2 and 3 respectively. CREIME model outperforms MagNet and ConvNetQuake_INGV in terms of magnitude estimation. It also gives lower values for both RMSE and MAE compared to STA/LTA algorithm.

359 5 Discussion

342

343

344

We investigated the different factors that influence the results of our model. Figure 7a shows the variation of errors with the signal-to-noise ratio in the data. It is observed that the errors in magnitude and P-arrival time show highest density within ± 1 units and ± 0.1 seconds, respectively, and tends to be lower for higher signal-to-noise ratios.

			Metric		
Dataset	Model Architecture	Mean Error (s)	St. dev. of Error (s)	$\begin{array}{c} \text{RMSE} \\ \text{(s)} \end{array}$	MAE (s)
STEAD	CREIME Model	-0.05	0.10	0.12	0.08
	STA/LTA	0.01	0.37	0.36	0.18
INSTANCE	CREIME Model	-0.04	0.13	0.14	0.09
	STA/LTA	0.01	0.52	0.52	0.29

Table 3. Comparison between CREIME model and STA/LTA method in terms of P-arrivalpicking. CREIME model outperforms STA/LTA

Figures/Figure_7.jpg

Figure 7. Factors affecting the error in estimation of magnitude and P-arrival times; errors in both magnitude and P-arrival are lower for higher signal-to-noise ratios; the magnitude of events seems to be under-estimated for higher hypocentral distances owing to their under-representation in the data. (a) Variation of errors with signal to noise ratio. (b)Variation of errors with hypocentral distance.

Figure 7b shows the variation of errors with hypocentral distance. We see that the errors tend to be close to zero over a wide range of hypocentral distances (up to 200km). There is a tendency for the model to underestimate the magnitude for higher hypocentral distances, which are under-represented in the training data. Both these figures are generated using STEAD data, and the corresponding figures for INSTANCE data can be found in the Appendix B.

We further looked into the effect of using different types of ground motion data as input (by removing instrument response), a summary of which can be found in Appendix C.

To make sure, that S-arrivals for low magnitude earthquakes do not get detected 374 as high-magnitude events, we test the model on S-arrival data. We do not notice any sys-375 tematic overestimation, only in 9% of the cases in the overestimation more than 1 unit. 376 377 This means that our model can be applied to the incoming seismogram in real time for rapid characterisation. Comparing the performance of the CREIME model with our ob-378 servations in Chakraborty et al. (2021), we find that providing data labels in the form 379 of a series and including the first P-arrival information is beneficial for the model, in es-380 timating the earthquake magnitude. 381

382 6 Conclusion

We present a novel deep learning model, CREIME, which successfully unifies the 383 tasks of event and noise discrimination, P-arrival time estimation and magnitude esti-384 mation using a smaller window (up to 2 seconds) of P-wave data as compared to pre-385 viously published models. The model in its current form, however, is restricted by the 386 fact that was trained specifically on data windows where the P-wave arrival is between 387 3.12 and 4.12 seconds of the starting sample. This restriction can be overcome in a fu-388 ture version of the model by modifying the training dataset to include a wider range of 389 arrival times. Nevertheless, this model can be seen as an important first step to a fully 390 automated earthquake characterisation approach in real time. We show that it performs 391 better than baseline models re-trained on the same duration of data. It also outperforms 392 traditional event discrimination algorithms such as STA/LTA. We demonstrate the ro-393 bustness of our model by testing it on two independent datasets, and show that it can 394 provide reliable estimates over a wide range of hypocentral distances and signal-to-noise 395 396 ratios. The model is designed to handle seismological waveform data in its raw format, which makes it very efficient in handling big data. Such models can also find their util-307 ity in smartphone applications to issue timely warnings to the public, as smartphone sen-398 sors have been shown to be capable of detecting seismic events (Kong et al., 2016). 300

400 Appendix A Metrics used for model evaluation

A1 Classification Metrics

We use different kinds of metrics to evaluate the classification and regression tasks. The performance of a classifier is often visualised with the help of a confusion matrix (Ting, 2017). The metrics we use to evaluate our model performance are described below. The abbreviations used are: TP: True positives

406 TN: True negatives

401

- 407 FP: False positives
- 408 FN: False negatives

• Accuracy: The accuracy of a classifier is the ratio of the number of correct predictions to the total number of predictions made by the model.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
(A1)

• **Precision**: The precision of a classifier is the ratio of the number of correct predictions for a particular class to the total number of times that class is predicted.

$$Precision = \frac{TP}{TP + FP}$$
(A2)

• **Recall**: The recall of a classifier is the proportion of the number of instances of a class in the data set that are correctly predicted.

$$Precision = \frac{TP}{TP + FN}$$
(A3)

• **F1 Score**: By definition, there is an inherent trade-off between the precision and the recall of a classifier. Therefore, it is often worthwhile to look at the harmonic mean of the two. This metric is called the F1-score of the classifier.

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$
(A4)

A2 Regression Metrics

409

410

411

412

For the regression task, the following metrics will be used to measure the CREIME performance:

• Mean Error: This is the mean value of errors corresponding to each example in the data set.

Mean Error,
$$\bar{\mathcal{E}} = \frac{1}{N} \sum_{i=0}^{N-1} \mathcal{E}_i = \frac{1}{N} \sum_{i=0}^{N-1} y_{true}^i - y_{pred}^i$$
 (A5)

where N is the total number of examples in the dataset.

• **Standard Deviation of Error**: This is the standard deviation of the errors in the predictions.

Standard Deviation of Error,
$$\sigma_{\mathcal{E}} = \sqrt{\frac{\sum_{i=0}^{N-1} (\mathcal{E}_i - \bar{\mathcal{E}})^2}{N}}$$
 (A6)

• Root Mean Squared Error (RMSE): As the name says, this is the square root of the mean of squares of errors in prediction.

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N-1} \mathcal{E}_i^2}{N}}$$
(A7)

• Mean Absolute Error: This is the mean of the absolute values of the errors in prediction.

$$MAE = \frac{\sum_{i=0}^{N-1} |\mathcal{E}_i|}{N}$$
(A8)

413 Appendix B Model performance on INSTANCE Dataset

Figures/Figure_B1.jpg

Figure B1. Distribution of magnitudes in chunk of INSTANCE data used for testing. Once again, no resampling is applied to the dataset based on magnitude.

Figures/Figure_B2.jpg

Figure B2. Analysis of model performance as a regressor on INSTANCE Data, here the events misclassified as noise, reflect the imbalanced distribution of magnitudes in the dataset itself, whereas the predicted magnitude of noise waveforms follows a similar trend as in case of STEAD data. (a) Confusion matrix for classifier performance on Instance Data. (b) Distribution of predicted magnitudes of noise misclassified as event and true magnitude of events misclassified as noise.

Figures/Figure_B3.jpg

Figure B3. Analysis of model performance as a regressor on INSTANCE Data. (a) Relationship between true and predicted magnitude values. (b) Distribution of errors in predicted magnitudes. (c) Distribution of errors in P-arrival estimation.

Figures/Figure_B4.jpg

Figure B4. (a) Variation of errors with signal to noise ratio. (b) Variation of errors with hypocentral distance.

Appendix C Effect of using different types of ground motion data as input

We compared the performance of the model when trained on different kinds of ground 416 motion data viz. acceleration (in $\mu m s^{-2}$), velocity (in nm s⁻¹) and displacement (in nm) 417 to investigate the effects of instrument response removal. This part of the analysis was 418 done only on the STEAD data. We lose roughly one fourth of the data due to unavail-419 ability of the instrument response. In each case, the models were trained on roughly the 420 same number of traces, alongside which we also compare it with the model whose results 421 are discussed in the Results (trained on more traces). For a fair comparison, we also train 422 a model on raw data, using roughly the same number of traces as in case of ground mo-423 tion data, accounting for the loss of data due to unavailability of instrument response 424 (this model is referred to as raw data (smaller) in the tables). The reason behind doing 425 this is to highlight one of the advantages of using counts data without instrument response 426 removal, which is the availability of more traces for training and testing. All five mod-427 els have been tested on the same traces. 428

Tables C1-C3 show the comparison between different types of input data. Even though certain ground motion parameters perform better in some metrics, using the raw data gives us an advantage that the data can be used in real time, and it is much readily available.

Table C1. Summary of classification performance for different types of ground motion data; removing instrument response does not seem to provide a significant advantage over using raw data

			Metr	ic			
	A	Precision (%)		Recall (%)		F1-score $(\%)$	
Type of input data	Accuracy (76)	Event	Noise	Event	Noise	Event	Noise
Raw Data	98.33	99.90	85.90	98.25	99.09	99.06	92.03
Raw Data (smaller)	97.85	99.91	82.36	97.71	99.17	98.79	89.99
Acceleration	98.15	99.79	93.62	97.75	99.37	98.76	96.41
Velocity	97.81	99.65	92.74	97.42	98.98	98.53	95.76
Displacement	96.52	99.61	88.54	95.74	98.86	97.64	93.41

Table C2. Summary of magnitude estimation for different types of ground motion data

	Metric					
	Mean Error	St. dev. of Error	RMSE	MAE		
Ground Motion						
Raw Data	-0.19	0.63	0.65	0.50		
Raw Data (smaller)	0.01	0.64	0.64	0.49		
Acceleration	-0.11	0.56	0.57	0.44		
Velocity	-0.09	0.62	0.63	0.47		
Displacement	-0.32	0.65	0.72	0.54		

		Metric		
	Mean Error (s)	St. dev. of Error (s)	RMSE (s)	MAE (s)
Ground Motion				
Raw Data	-0.04	0.11	0.12	0.08
Raw Data (smaller)	-0.06	0.11	0.13	0.09
Acceleration	-0.07	0.12	0.14	0.10
Velocity	-0.06	0.14	0.15	0.11
Displacement	-0.06	0.18	0.19	0.13

Table C3. Summary of P-arrival estimation for different types of ground motion data

433 Acknowledgments

This research is supported by the "KI-Nachwuchswissenschaftlerinnen" - grant SAI 01IS20059 434 by the Bundesministerium für Bildung und Forschung - BMBF. Calculations were per-435 formed at the Frankfurt Institute for Advanced Studies' new GPU cluster, funded by BMBF 436 for the project Seismologie und Artifizielle Intelligenz (SAI). The authors are also thank-437 ful to Dr. KiranKumar Thingbaijam, Dr. Jan Steinheimer and Jonas Köhler for their 438 kind suggestions. H.St. gratefully acknowledges the Judah M. Eisenberg Laureatus - Pro-439 fessur at Fachbereich Physik, Goethe Universität Frankfurt, funded by the Walter Greiner 440 Gesellschaft zur Förderung der physikalischen Grundlagenforschung e.V. 441

442 Open Research

The seismic waveforms used in our research are a part of two datasets – STanford
 EArthquake Dataset (STEAD) (Mousavi et al., 2019) which was downloaded https://
 github.com/smousavi05/STEAD (last accessed January 2022) and INSTANCE (Michelini,
 Cianetti, Gaviano, Giunchi, Jozinović, & Lauciani, 2021) which was downloaded from

https://doi.org/10.13127/instance (last accessed January 2022).

448 References

449	Allen, R., Gasparini, P., Kamigaichi, O., & Böse, M. (2009). The status o	f earth-
450	quake early warning around the world: An introductory overview.	Seismol.
451	Res. Lett., 80, 682–693. doi: 10.1785/gssrl.80.5.682	

- Allen, R., & Kanamori, H. (2003, 06). The potential for earthquake early warning
 in southern california. Science (New York, N.Y.), 300, 786-9. doi: 10.1126/
 science.1080912
- Allen, R. M., & Melgar, D. (2019). Earthquake early warning: Advances, scientific challenges, and societal needs. Annu Rev Earth Planet Sci, 47(1), 361-388. Retrieved from https://doi.org/10.1146/annurev-earth-053018-060457 doi: 10.1146/annurev-earth-053018-060457
- Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces. Bull. Seismol. Soc. Am., 68(5), 1521–1532. Retrieved from
 https://doi.org/10.1785/BSSA0680051521
- Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J.
 (2010, 06). Obspy: A python toolbox for seismology. *Seismol. Res. Lett.*, 81,
 530-533. doi: 10.1785/gssrl.81.3.530
- ⁴⁶⁵ Chakraborty, M., Li, W., Faber, J., Rümpker, G., Stöcker, H., & Srivastava, N.
- 466 (2021). A study on the effect of input data length on deep learning based mag-467 nitude classifier.
- Charles, P. (2013). Project title. GitHub. Retrieved from https://github.com/
 charlespwd/project-title
- ⁴⁷⁰ Chollet, F. (2017). *Deep learning with python*. Manning.
- 471 Chung, D. H., & Bernreuter, D. L. (1981). Regional relationships among earthquake

472	magnitude scales. Rev. Geophys., 19(4), 649-663. Retrieved from https://
473	agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/RG019i004p00649
474	doi: https://doi.org/10.1029/RG019i004p00649
475	Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P.
476	(2011). Natural language processing (almost) from scratch. J. Mach. Learn.
477	<i>Res.</i> , 12(ARTICLE), 2493–2537. doi: 10.5555/1953048.2078186
478	Cremen, G., & Galasso, C. (2020). Earthquake early warning: Recent advances
479	and perspectives. Earth Sci Rev, 205, 103184. doi: https://doi.org/10.1016/j
480	.earscirev.2020.103184
481	Ekström, G., & Dziewonski, A. (1988). Evidence of bias in estimations of earth-
482	quake size. Nature, 332, 319-323. Retrieved from https://doi.org/10.1038/
483	332319a0
484	Fenner, D., Rümpker, G., Li, W., Chakraborty, M., Faber, J., Köhler, J.,
485	Srivastava, N. (2022). Automated seismo-volcanic event detection ap-
486	plied to stromboli (italy). Frontiers in Earth Science, 10. Retrieved from
487	https://www.frontiersin.org/article/10.3389/feart.2022.809037 doi:
488	10.3389/feart.2022.809037
489	Geller, R. (1976). Scaling relations for earthquake source parameters and magni-
490	tudes. Bull. Seismol. Soc. Am., bb(5), 1501-1523. Retrieved from https://doi
491	.org/10.1785/BSSA0660051501
492	Giardini, D. (1988). Frequency distribution and quantification of deep earth-
493	quakes. J. Geophys. Res. Solid Earth, $93(B3)$, 2095-2105. doi: 10.1029/ ID002:D02 $_{2}$ 02005.
494	JD095ID05p02095
495	Saturnal Soc. Am. 21 185 188
496	Setsmol. Soc. Am., 54, 105-100.
497	of waveform data in western Japan using deep learning Farth Planate Space
498	of waveform data in western Japan using deep rearring. Earth 1 tanets space, 71(127) doi: https://doi.org/10.1186/s/0623-010-1111-y
499	Harris C B Millman K I van der Walt S Commers B Virtanen P Cour-
500	napeau D Oliphant T E (2020 September) Array programming with
502	NumPy. Nature, 585(7825), 357-362. Retrieved from https://doi.org/
503	10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2
504	He, K., Ren, S., Sun, J., & Zhang, X. (2016). Deep residual learning for image
505	recognition. IEEE Conf Comput Vis Pattern Recognit(CVPR), 770-778. doi:
506	10.1109/CVPR.2016.90
507	Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Kingsbury,
508	B. (2012). Deep neural networks for acoustic modeling in speech recognition:
509	The shared views of four research groups. IEEE Signal Process. Mag., 29(6),
510	82-97. doi: 10.1109/MSP.2012.2205597
511	Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Com-
512	put., $9(8)$, 1735–1780. doi: 10.1162/neco.1997.9.8.1735
513	Howell Jr, B. (1981). On the saturation of earthquake magnitudes. Bull. Seis-
514	mol. Soc. Am., 71(5), 1401-1422. Retrieved from https://doi.org/10.1785/
515	BSSA0710051401
516	Jin, X., Zhang, H., Li, J., Wei, Y., & Ma, Q. (2013). Earthquake magnitude estima-
517	tion using the τ_c and p_d method for earthquake early warning systems. Earthq.
518	Sci., 26 (es-26-1-23), 23-31. doi: 10.1007/s11589-013-0005-4
519	Kanamori, H. (1983). Magnitude scale and quantification of earthquakes. <i>Tectono-</i>
520	<i>physics</i> , 93(3-4), 185-199. Retrieved from https://doi.org/10.1016/
521	0040-1951(83)90273-1
522	Kanamori, H. (2005). Real-time seismology and earthquake damage mit-
523	igation. Annu Rev Earth Planet Sci, $33(1)$, 195-214. doi: 10.1146/
524	annurev.earth. $33.092203.122020$
525	Kanamori, H., & Stewart, G. S. (1978). Seismological aspects of the guatemala
526	eartinguake of february 4, 1976. J. Geophys. Res. Solid Earth, 83(B7), 3427-

527	3434. doi: https://doi.org/10.1029/JB083iB07p03427
529	Kiranyaz S. Ince T. Hamila B. & Gabboui M. (2015). Convolutional neural net-
520	works for patient-specific ecg classification 2608-2611 doi: 10.1109/EMBC
529	2000 2011. doi: 10.1100/EMDC
530	Kong O Allen P. M. & Schwier I. (2016) Muchalice Initial observations from a
531	roles and a science of the second sec
532	giodal smartphone seismic network. Geophys. Res. Lett., 43, 9588–9594. doi:
533	10.1002/2010GL070955.
534	Kong, Q., Irugman, D. I., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoff,
535	P. (2018). Machine learning in seismology: Turning data into insights. Seis- L = L = L = 0.0(1) 2.14 $E = L = 0.0(1)$ 2.14 $E = 0.000$
536	<i>mol. Res. Lett.</i> , $90(1)$, 3-14. Retrieved from https://doi.org/10.1/85/
537	
538	Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future di-
539	rections. Prog. Artif. Intell., 5, 221–232. Retrieved from https://doi.org/10
540	.1007/s13748-016-0094-0
541	Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with
542	deep convolutional neural networks. Commun. ACM , $60(6)$, 84–90. Retrieved
543	from https://doi.org/10.1145/3065386
544	Kuyuk, H. S., & Allen, R. M. (2013). A global approach to provide magnitude es-
545	timates for earthquake early warning alerts. Geophys. Res. Lett., $40(24)$, 6329-
546	6333. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/
547	10.1002/2013GL058580 doi: https://doi.org/10.1002/2013GL058580
548	Kuyuk, H. S., & Susumu, O. (2018). Real-time classification of earthquake using
549	deep learning. Proc. Comput. Sci, 140, 298–305. Retrieved from https://doi
550	.org/10.1016/j.procs.2018.10.316
551	LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.
552	Retrieved from https://doi.org/10.1038/nature14539
553	Li, W., Chakraborty, M., Fenner, D., Faber, J., Zhou, K., Rümpker, G., Sri-
554	vastava, N. (2021). Epick: Multi-class attention-based u-shaped neural net-
555	work for earthquake detection and seismic phase picking. Retrieved from
556	https://arxiv.org/abs/2109.02567
557	Li, W., Sha, Y., Zhou, K., Faber, J., Rümpker, G., Stöcker, H., & Srivastava, N.
558	(2022). Deep learning-based small magnitude earthquake detection and seismic
559	phase classification. Retrieved from https://arxiv.org/abs/2204.02870
560	Li, Z., Meier, M. A., Hauksson, E., Zhan, Z., & Andrews, J. (2018). Ma-
561	chine learning seismic wave discrimination: Application to earthquake
562	early warning. Geophys. Res. Lett., 45(10), 4773-4779. Retrieved from
563	https://doi.org/10.1029/2018GL077870
564	Liao, W., Lee, E., Mu, D., Chen, P., & Rau, R. (2021, 03). ARRU Phase
565	Picker: Attention Recurrent-Residual U-Net for Picking Seismic P- and
566	S-Phase Arrivals. Seismol. Res. Lett., 92(4), 2410-2428. Retrieved from
567	https://doi.org/10.1785/0220200382_doi: 10.1785/0220200382
569	Lomax A Michelini A & Jozinović D (2019 02) An investigation of rapid
500	earthquake characterization using single-station waveforms and a convolutional
570	neural network Seismal Res Lett 90 517-529 doi: 10.1785/0220180311
570	Majer M. Boss Z. E. Bamachandran A. Balakrishna A. Nair S. Kundzicz
5/1	P Vue V (2010) Reliable real-time seismic signal/noise discrimina-
572	tion with machine learning I Geophyse Res Solid Earth 19/(1) 788 800
5/3	Retrieved from https://squpubs.onlinelibrary.uilov.com/doi/shc/
5/4	10 1029/2018 IB016661 doi: https://doi.org/10.1020/2018 IB016661
5/5	Michelini A Cianetti S Caviano S Ciunchi C Loginovic D & Laucieri V
576	(2021) Instance the italian gaigmin dataget for machine learning. Farth Curt
577	(2021). Instance – the manan seismic dataset for machine learning. Earth Syst.
578	Sci. Data Discuss., 2021 , 1–47. dol: 10.0194/essd-2021-104 Michelini A. Cionetti S. Conigna G. Cionetti O. La Sci V. D. & La Sci V.
579	Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinovic, D., & Lauciani, V.
580	(2021). Instance the Italian seismic dataset for machine learning, seismic wave-
581	forms and associated metadata. Istituto Nazionale di Geofisica e Vulcanologia

582	(INGV). Retrieved from https://doi.org/10.13127/instance
583	Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černocký, J. (2011). Strategies for
584	training large scale neural network language models. In 2011 ieee workshop on
585	automatic speech recognition understanding (p. 196-201). doi: 10.1109/ASRU
586	.2011.6163930
587	Mousavi, S. M., & Beroza, G. C. (2020). A machine-learning approach for earth-
588	quake magnitude estimation. Geophys. Res. Lett., 47, e2019GL085976 Re-
589	trieved from https://doi.org/10.1029/2019GL085976
590	Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L., & Beroza, G. (2020).
591	Earthquake transformer—an attentive deep-learning model for simultane-
592	ous earthquake detection and phase picking. Nat. Commun., 11(3952). doi:
593	https://doi.org/10.1038/s41467-020-17591-w
594	Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). Stanford earth-
595	quake dataset (stead): A global data set of seismic signals for ai. <i>IEEE</i>
596	Access, 7, 179464-179476. Retrieved from https://doi.org/10.1109/
597	ACCESS.2019.2947848
598	Münchmeyer, J., Bindi, D., Leser, U., & Tilmann, F. (2020). The transformer earth-
599	quake alerting model: a new versatile approach to earthquake early warning.
600	Geophys. J. Int., 225(1), 646-656. Retrieved from https://doi.org/10.1093/
601	gji/ggaa609 doi: 10.1093/gji/ggaa609
602	Nagi, J., Ducatelle, F., Di Caro, G. A., Ciresan, D., Meier, U., Giusti, A., Gam-
603	bardella, L. M. (2011). Max-pooling convolutional neural networks for
604	vision-based hand gesture recognition. In 2011 ieee international confer-
605	ence on signal and image processing applications (icsipa) (p. 342-347).
606	10.1109/ICSIPA.2011.6144164
607	Nakamura, Y. (1988). On the urgent earthquake detection and alarm system (ure-
608	das). 9th world conference on earthquake engineering, VII(B7), 673-678.
609	Panakkat, A., & Adeli, H. (2009). Recurrent neural network for approximate
610	earthquake time and location prediction using multiple seismicity indica-
611	tors. ComputAided Civ. Infrastruct. Eng., 24(4), 280-292. Retrieved from
612	https://doi.org/10.1111/j.1467-8667.2009.00595.x
613	Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for
614	earthquake detection and location. Sci. Adv., 4(2), e1700578. doi: 10.1126/
615	sciadv.1700578
616	Peters, M. E., Neumann, M., Ivver, M., Gardner, M., Clark, C., Lee, K., & Zettle-
617	mover, L. (2018, June). Deep contextualized word representations.
618	2227-2237. Retrieved from https://aclanthology.org/N18-1202 doi:
619	10.18653/v1/N18-1202
620	Prechelt, L. (2012). Early stopping — but when? In Neural networks: Tricks of
621	the trade: Second edition (pp. 53–67). Springer Berlin Heidelberg. Retrieved
622	from https://doi.org/10.1007/978-3-642-35289-8_5 doi: 10.1007/978-3
623	-642-35289-8_5
624	Richter, C. F. (1935). An instrumental earthquake magnitude scale. Bull.
625	Seismol. Soc. Am., 25, 1-32. Retrieved from https://doi.org/10.1785/
626	BSSA0250010001
627	Ross, Z. E., Meier, M., & Hauksson, E. (2018). P wave arrival picking and first-
628	motion polarity determination with deep learning. J. Geophys. Res. Solid
629	Earth, 123(6), 5120-5129. Retrieved from https://doi.org/10.1029/
630	2017JB015251
631	Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop. coursera: Neural networks
632	for machine learning. University of Toronto, Technical Report.
633	Ting, K. M. (2017). Confusion matrix. In Encuclopedia of machine learning
634	and data mining (pp. 260–260). Boston, MA: Springer US. Retrieved from
635	https://doi.org/10.1007/978-1-4899-7687-1_50 doi: 10.1007/978-1-4899
636	-7687-1_50

- Wu, Y., & Zhao, L. (2006). Magnitude estimation using the first three seconds p wave amplitude in earthquake early warning. *Geophys. Res. Lett.*, 33(16). Re trieved from https://doi.org/10.1029/2006GL026871
- Yanwei, W., Xiaojun, L., Zifa, W., Jianping, S., & Enhe, B. (2021). Deep learning for P-wave arrival picking in earthquake early warning. *Earthq. Eng. Eng. Vib.*, 20, 391–402. doi: 10.1007/s11803-021-2027-6
- 643Zhou, Y., Yue, H., Kong, Q., & Zhou, S.(2019, 04).Hybrid Event Detection and644Phase-Picking Algorithm Using Convolutional and Recurrent Neural Net-645works.Seismological Research Letters, 90(3), 1079-1087.Retrieved from
- https://doi.org/10.1785/0220180319 doi: 10.1785/0220180319
 Zhu, J., Li, S., Song, J., & Wang, Y. (2021). Magnitude estimation for earthquake
 early warning using a deep convolutional neural network. Frontiers in Earth
 Science, 9, 341. doi: 10.3389/feart.2021.653226
- Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic
 arrival-time picking method. *Geophys. J. Int.*, 216(1), 261-273. Retrieved from
 https://doi.org/10.1093/gji/ggy423
- Ziv, A. (2014). New frequency-based real-time magnitude proxy for earthquake early
 warning. *Geophys. Res. Lett.*, 41(16), 7035–7040. Retrieved from https://
 doi.org/10.1002/2014GL061564

Figure1.



Figure2.



(b)



Figure3.



Figure4.

(a)

(b)



Figure5.



Figure6.



Error in P-arrival (seconds)

Figure7.



(a)

(b)

FigureB1.



FigureB2.

(a)

(b)



FigureB3.

(a)



FigureB4.



(a)

(b)