



# A study on the effect of input data length on a deep-learning-based magnitude classifier

Megha Chakraborty<sup>1,4</sup>, Wei Li<sup>1</sup>, Johannes Faber<sup>1,2</sup>, Georg Rumpker<sup>1,4</sup>, Horst Stoecker<sup>1,2,3,5</sup>, and Nishtha Srivastava<sup>1,4</sup>

<sup>1</sup>Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany

<sup>2</sup>Institute for Theoretical Physics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany

<sup>3</sup>Xidian-FIAS International Joint Research Center, Giersch Science Center, 60438 Frankfurt am Main, Germany

<sup>4</sup>Institute of Geosciences, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany

<sup>5</sup>GSI Helmholtzzentrum für Schwerionenforschung GmbH, 64291 Darmstadt, Germany

**Correspondence:** Nishtha Srivastava (srivastava@fias.uni-frankfurt.de)

Received: 14 February 2022 – Discussion started: 30 May 2022

Revised: 6 September 2022 – Accepted: 28 September 2022 – Published: 10 November 2022

**Abstract.** The rapid characterisation of earthquake parameters such as its magnitude is at the heart of earthquake early warning (EEW). In traditional EEW methods, the robustness in the estimation of earthquake parameters has been observed to increase with the length of input data. Since time is a crucial factor in EEW applications, in this paper we propose a deep-learning-based magnitude classifier based on data from a single seismic station and further investigate the effect of using five different durations of seismic waveform data after first P-wave arrival: 1, 3, 10, 20 and 30 s. This is accomplished by testing the performance of the proposed model that combines convolution and bidirectional long short-term memory units to classify waveforms based on their magnitude into three classes: “noise”, “low-magnitude events” and “high-magnitude events”. Herein, any earthquake signal with magnitude equal to or above 5.0 is labelled as “high-magnitude”. We show that the variation in the results produced by changing the length of the data is no more than the inherent randomness in the trained models due to their initialisation. We further demonstrate that the model is able to successfully classify waveforms over wide ranges of both hypocentral distance and signal-to-noise ratio.

## 1 Introduction

The earthquake magnitude, defined as a logarithmic measure of the relative strength of an earthquake, is one of the most fundamental parameters in its characterisation (Mousavi and Beroza, 2020). The complex nature of the geophysical processes affecting earthquakes makes it very difficult to have a single reliable measure for its size (Kanamori and Stewart, 1978), and hence, magnitude values measured in different scales often differ by more than 1 unit. This is especially true for larger events due to saturation effects (Howell Jr, 1981; Kanamori, 1983). Owing to the above-mentioned reasons and the empirical nature of the majority of the magnitude scales, it is one of the most difficult parameters to estimate (Chung and Bernreuter, 1981; Ekström and Dziewon-ski, 1988). Some of the classical approaches to obtain first estimates of earthquake magnitude have used empirical relations for parameters such as predominant period  $\tau_{\text{max}}^{\text{p}}$  (Nakamura, 1988; Allen and Kanamori, 2003), effective average period  $\tau_{\text{c}}$  (Kanamori, 2005; Jin et al., 2013) in the frequency domain and parameters such as peak displacement (Pd) (Wu and Zhao, 2006; Jin et al., 2013) in the amplitude domain calculated from the initial 1–3 s of P waves. These relations form the basis of existing earthquake early warning (EEW) systems in Japan, California, Taiwan, etc. (Allen et al., 2009 and the references therein). The accuracy of such estimates has been shown to increase with the duration of data used to calculate them (Ziv, 2014).

The recent developments in the area of deep learning (LeCun et al., 2015), combined with the availability of affordable high-end computational power through graphics processing units (GPUs), have led to state-of-the-art results in image recognition (Krizhevsky et al., 2017; He et al., 2016), speech recognition (Mikolov et al., 2011; Hinton et al., 2012) and natural language processing (Peters et al., 2018; Collobert et al., 2011). In fields such as seismology, where the volume of available data has increased exponentially over the last decades (Kong et al., 2018), deep learning has achieved great success in tasks such as seismic phase picking (Zhu and Beroza, 2019; Liao et al., 2021; Li et al., 2021), event detection (Wang and Teng, 1995; Mousavi et al., 2020; Meier et al., 2019), magnitude estimation (Mousavi and Beroza, 2020), event location characterisation (Perol et al., 2018; Panakkat and Adeli, 2009; Kuyuk and Susumu, 2018) and first-motion polarity detection (Ross et al., 2018).

Considering that timeliness is of the essence in rapid earthquake characterisation, it becomes important to find an optimum duration for the input data that can provide a reliable and statistically significant estimate for various earthquake parameters while using a minimum amount of P-wave data. In this study, we present a deep learning model to perform time series multiclass classification (Fawaz et al., 2019; Aly, 2005) that classifies seismic waveforms as “noise”, “low-magnitude” or “high-magnitude”. Here a local magnitude of 5.0 is taken to be the boundary between the low-magnitude and high-magnitude classes. We further investigate the effect of using different lengths of data on the model performance. Please note that the boundary of 5.0 is arbitrarily chosen and can be modified depending on the purpose of the model and the local geology (which influences the correlation between earthquake magnitude and intensity). Magnitudes of 3 and 4 were also experimented with as decision boundaries, and accuracy, precision and recall values in either case were found to be similar to those for magnitude 5. Thus, the decision boundary in itself does not seem to influence the model performance. Unlike Saad et al. (2020), who use data from three seismic stations to characterise different earthquake parameters, the model discussed in this paper only uses three-component data from a single station.

## 2 Methodology

### 2.1 Generating training and testing datasets

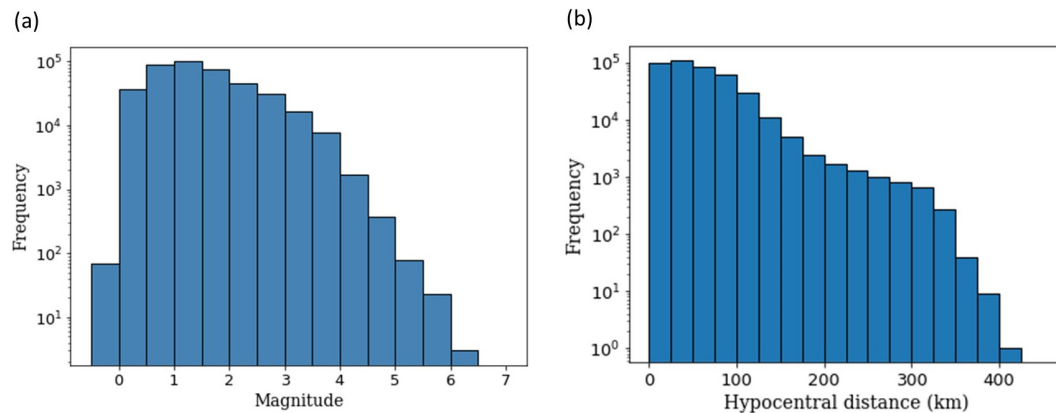
We use data from the STanford EArthquake Dataset (STEAD) (Mousavi et al., 2019) (see “Data availability”) to train and test our model. STEAD is a high-quality benchmarked dataset created for machine learning and deep learning applications and contains seismic event and noise waveforms of 1 min duration recorded by over 2500 seismic stations across the globe. The waveforms have been detrended and filtered with a bandpass filter between 1.0 and 40.0 Hz,

followed by a resampling at 100 Hz. Metadata consisting of 35 attributes for earthquake traces and 8 attributes for noise traces are provided by the authors.

To ensure consistency in magnitude we only use traces for which the magnitude is provided in “ml” (local magnitude) scale (as this is the case for most of the traces in the dataset). We also discard traces with signal-to-noise ratio less than 10 dB for quality control. We divide the noise and earthquake traces into training, validation and test sets in the ratio 60 : 10 : 30. Care is taken to make sure that the three aforementioned datasets are non-overlapping. This means that traces corresponding to a particular earthquake (represented by the “source\_id” attribute) but recorded at different stations are included in only one of the three sets. For noise traces, recordings from a particular seismic station are included in only one of the three sets. In this paper, we propose a classifier model for rapid earthquake characterisation. Furthermore, we investigate the effect of using different lengths of data after the first P arrival (1, 3, 10, 20 and 30 s) on the performance of this classifier model. In each case the P-wave data are preceded by 2.8–3.0 s of pre-signal noise, so the model can learn the noise characteristics of the station (Münchmeyer et al., 2020). The data labels 0, 1 and 2 are used to denote the classes noise, low-magnitude and high-magnitude, respectively.

As mentioned earlier, we take a local magnitude 5.0 to be the decision boundary between high-magnitude and low-magnitude events. However, the training dataset originally has a magnitude distribution as shown in Fig. 1; this would lead to a high imbalance between the low-magnitude and high-magnitude classes (a ratio of nearly 3300 : 1). It is widely agreed by the machine learning community that most classifiers assume an equal distribution between the different classes (Batista et al., 2004). Although examples from some domains where models perform reasonably well, even in highly imbalanced datasets, show that there are other factors at play, imbalanced datasets not only are a major hindrance in the development of good classifiers but can also lead to misleading evaluations of the accuracy of the model (Batista et al., 2004). To tackle this imbalance problem, we apply resampling of the data (Krawczyk, 2016) as follows.

- Events with magnitude equal to or above 5.0 are represented 20 times in the dataset by using a shifting window starting from 300 samples to 280 samples before the first P-arrival sample, the window being shifted by 2 samples for each representation. Each of these traces are also flipped; i.e. their polarity is reversed, since it does not affect the magnitude information of the data. Such data augmentation techniques used for images have also been found to be useful for time series data (Batista et al., 2004; Wen et al., 2021).



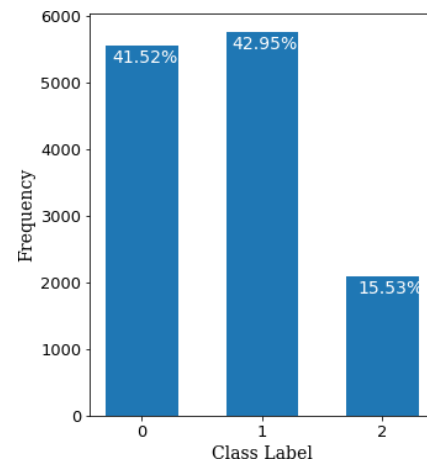
**Figure 1.** Original distribution (prior to data augmentation) of (a) local magnitudes and (b) hypocentral distances in the chunk of STEAD (Mousavi et al., 2019) data used for training.

- For low-magnitude events the following strategy of random undersampling is adopted.
  1. All events with magnitude between 4.5 and 5.0 are used.
  2. A total of 1/3 of events with magnitude between 4.0 and 4.5 are used.
  3. A total of 1/50 of events with magnitude between 2.0 and 4.5 are used.
  4. A total of 1/100 of events with magnitude less than 2.0 are used.
- A total of 1/25 of the available noise traces are used.

Note that special care is taken to include more events close to the decision boundary (magnitude 5.0) so that the model can learn to differentiate between events of magnitude, say, 4.0 to 5.0, which is more difficult compared to differentiating between events of magnitude, say, 2.0 and 5.0. The corresponding distribution of the different classes is shown in Fig. 2. The validation and test datasets follow a similar distribution. As one can see, in spite of the resampling techniques employed, the high-magnitude class is still under-represented in the dataset, as compared to the other two classes. So, we apply a class weight (Krawczyk, 2016) of 1 : 1 : 10, chosen experimentally, for classes 0, 1 and 2 while training the model. The data are used without instrument response removal. Unlike Lomax et al. (2019) we do not normalise the data. Only the waveform information is provided to the model. Since the dataset includes waveforms from different types of instruments, choosing only one type of instrument would significantly reduce the amount of training data, thereby limiting the learning; therefore we use data from different instruments to train the model.

## 2.2 Model architecture and model training

The model architecture (Chakraborty et al., 2021) consists of two sets of 1D convolution (Kiranyaz et al., 2021), dropout



**Figure 2.** The distribution of classes in the training dataset obtained by undersampling “noise” (represented by class 0) and “low-magnitude” (represented by class 1) data and applying data augmentation to “high-magnitude” (represented by class 2) events. A similar distribution of classes is seen in the validation and test datasets as well.

(Srivastava et al., 2014) and max-pooling (Nagi et al., 2011) layers, followed by three bi-directional long short-term memory (LSTM) layers (Hochreiter and Schmidhuber, 1997). Convolutional neural networks have often been found to be useful for seismological data analysis as they are capable of extracting temporally independent patterns in the data (features). When combined with LSTMs the temporal relations between these features can be obtained. In applications such as magnitude-based classification of earthquakes, this aids in the effective analysis of signal features as compared to the pre-signal background noise. The dropout layers are used to prevent the model from overfitting, and the max-pooling layer is a method to reduce the data dimensionality so that only relevant features can be retained. The final layer is a softmax layer (Goodfellow et al., 2016), which gives a three-

element array of the form  $[P_0, P_1, P_2]$ , where  $P_i$  is the probability of the waveform belonging to the class  $i$  (Fig. 3). A detailed description of the model architecture is provided in the caption for Fig. 3.

The model is trained using an Adam optimiser (Kingma and Ba, 2015), categorical cross-entropy loss (Murphy, 2012) and a batch size of 256. Early stopping (Prechelt, 2012) is used to prevent overfitting, whereby the validation loss is monitored, and the training stops when there is no reduction in it for 20 consecutive epochs. We start with a learning rate of  $10^{-3}$  and reduce it by a factor of 10 if the validation loss does not reduce for 15 consecutive epochs until it reaches  $10^{-6}$ . The model for the epoch corresponding to the lowest validation loss is retained.

### 3 Results

To analyse the effect of different lengths of data on the performance of the classifier model, we use the metrics listed below to evaluate the model performance. The metrics are calculated in terms of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) samples.

- *Accuracy*. The accuracy of a classifier is the proportion of testing samples that are correctly classified. Mathematically, it can be defined as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}. \quad (1)$$

- *Precision*. This is the ratio of the number of times the model *correctly* predicts a class to the total number of times it predicts that class. Mathematically it is defined as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

- *Recall*. This is the ratio of the number of times the model correctly predicts a class to the total number occurrences of that class in the dataset. Mathematically it is defined as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

Figure 4 shows three waveforms (one from each class) that have been correctly classified. The softmax probabilities, as described in Sect. 2.2, are also shown. In each case the highest probability is predicted for the corresponding class. Figure 5 shows the softmax probabilities, predicted by the model for different lengths of the same waveform. Although the waveform is correctly classified in each case, the predicted probabilities are different and show no dependence on the length of input data.

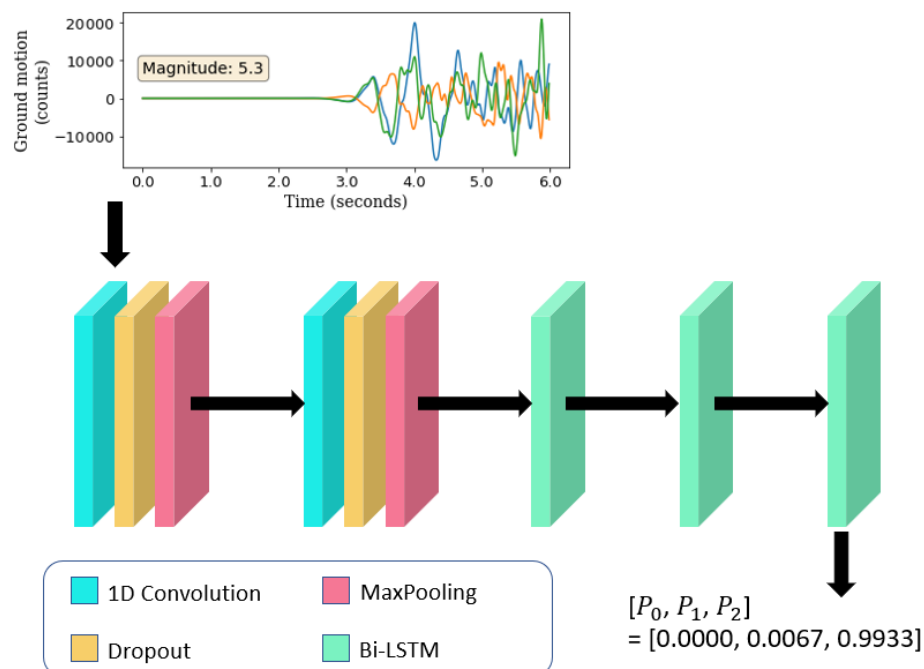
### 4 Discussions

We investigated the possible factors that might be influencing the model performance. Figure 6a shows the variation in the model performance with respect to the duration of P-wave data used as an input. As we do not tune a random seed during model training (Bengio, 2012; Madhyastha and Jain, 2019), we also looked at the randomness in the performance when the model is trained on the same data five times (Fig. 6b). Thus, we can see that the variation in the results caused by changing the length of data is comparable to the randomness in the results due to random initialisation upon retraining the model on the same data.

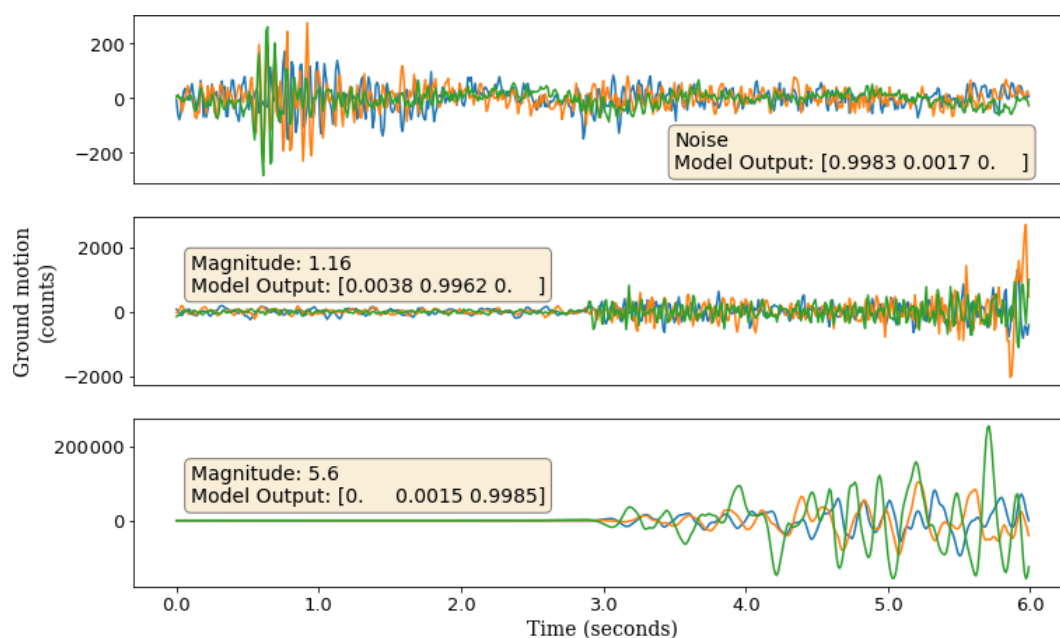
Figure 7 shows the classification statistics for one of the iterations of the model trained on the 3 s data. The events classified as noise tend to be of low magnitude, while the misclassification of low-magnitude events as high-magnitude and vice versa is most pronounced at the decision boundary of 5.0. Another important observation is that the degree of misclassification of low-magnitude events is much higher than the reverse case; approximately 65 % of events with magnitude between 4.5 and 5.0 and 35 % of events with magnitude between 4.0 and 4.5 are classified as high-magnitude, while fewer than 10 % of events with magnitude between 5.0 and 5.5 are classified as low-magnitude; this is intentional as a missed alarm is considered more dangerous than a false alarm in this context (Allen and Melgar, 2019) and is achieved by giving the high-magnitude class more weight during model training.

Figure 8 visualises the classification of events across different hypocentral distance (Fig. 8a) and signal-to-noise ratios (Fig. 8b). We observe that there are instances of correct classification across a wide-range of hypocentral distances and signal-to-noise ratios (SNRs), which means that the model is capable of learning the frequency characteristics of waveforms to some extent and does not directly correlate the amplitude or SNR with magnitude. We do observe some clustering of low-magnitude events classified as noise for SNRs below 20 dB. But for the demarcation between low-magnitude and high-magnitude events the misclassification seems to be close to the decision boundary and spread across a wide range of hypocentral distances and signal-to-noise ratios.

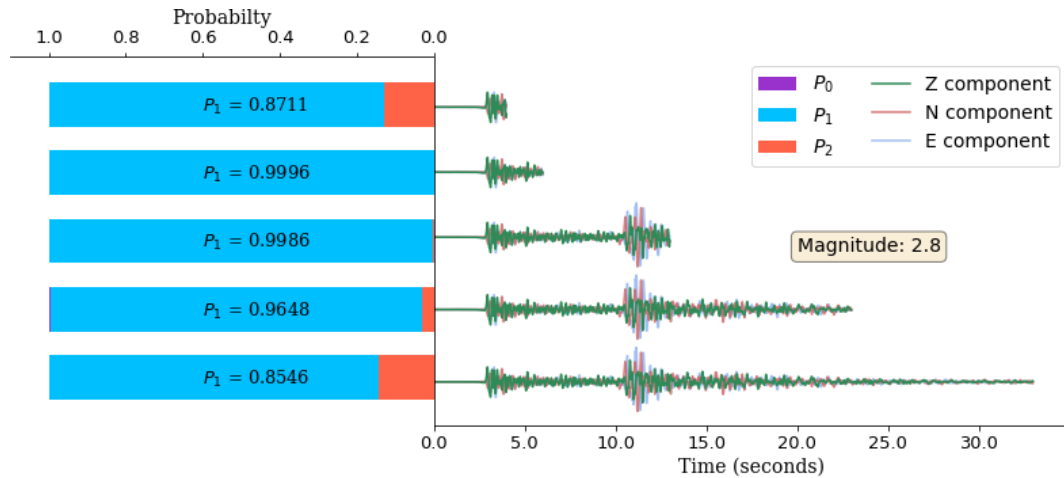
Despite maximising the amount of data on either side of the decision boundary between low and high magnitude, we find some incorrect classifications, most of which lie within a range of  $5.0 \pm 0.5$ , as can be seen in Fig. 8. However, considering that sometimes even magnitudes of the same scale reported by different agencies can vary by as much as 0.5 magnitude units (Mousavi and Beroza, 2020), it can be expected that the model would have difficulty in classifying traces close to the decision boundary. In a future version of the model, it might be helpful to treat this as a regression problem instead of classification, thereby providing the model more information about the exact value of the magni-



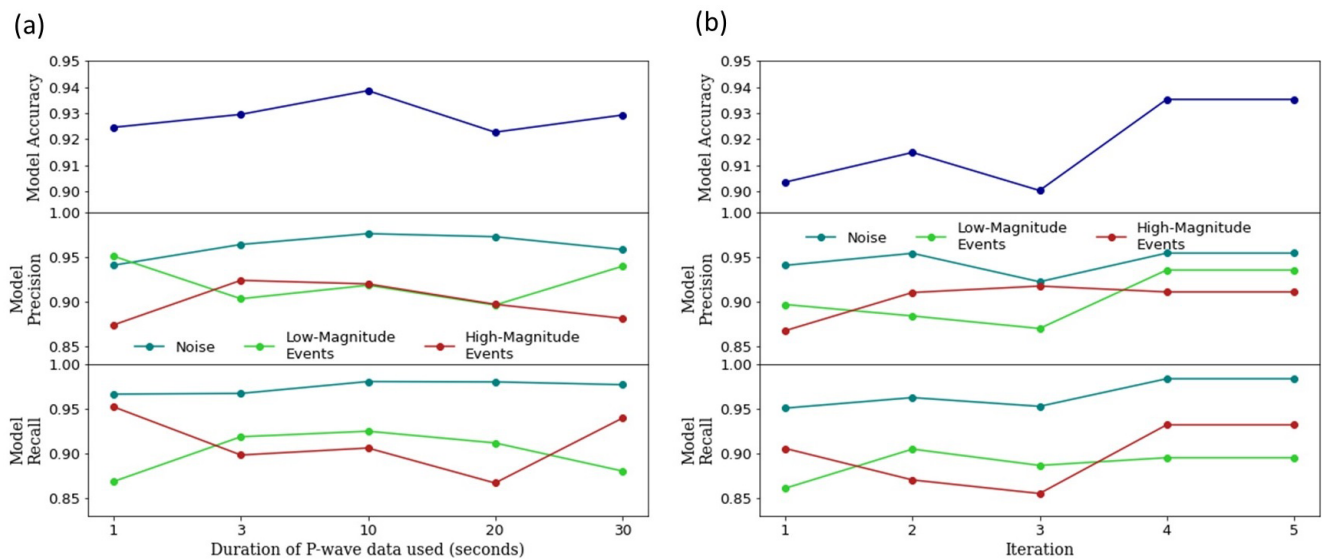
**Figure 3.** The architecture of the model used to perform the three-class classification. The input to the model is three-component seismic waveform data from a single station. The example shown here corresponds to the case where 3 s of P-wave data is used (the total length of data is, thus, 6 s). The 1D convolution layers have a kernel size of four and eight filters each; the drop rate for each dropout layer is 0.2, and each max-pooling layer reduces the size of the data by a factor of 4; the bi-LSTM layers have dimensions of 256, 256 and 128, respectively. The final layer is a softmax layer that outputs the probability of the trace belonging to classes 0 (noise), 1 (low-magnitude) and 2 (high-magnitude), represented here as  $P_0$ ,  $P_1$  and  $P_2$ , respectively. In this case a probability of 0.9933 is assigned to class 2 for an event with magnitude 5.3; thus, this is a case of correct classification.



**Figure 4.** Examples of waveforms that have been correctly classified. In each case the highest probability corresponds to the respective class.



**Figure 5.** Softmax probabilities for different input lengths of the same waveform, predicted by the models trained on the corresponding lengths of data. The waveform used here corresponds to an event of magnitude 2.8; although the maximum probability corresponds to class 1, the values of these probabilities are different for different data lengths, and there is no clear dependence between the length of the data and this probability.



**Figure 6.** (a) Variation in classifier model performance when different durations (1, 3, 10, 20, 30 s) of P-wave data are used; (b) variation in the classifier model performance when the same model is retrained on the same data (in this case 3 s of P-wave data used) five times. This shows that the variations in the two cases are comparable.

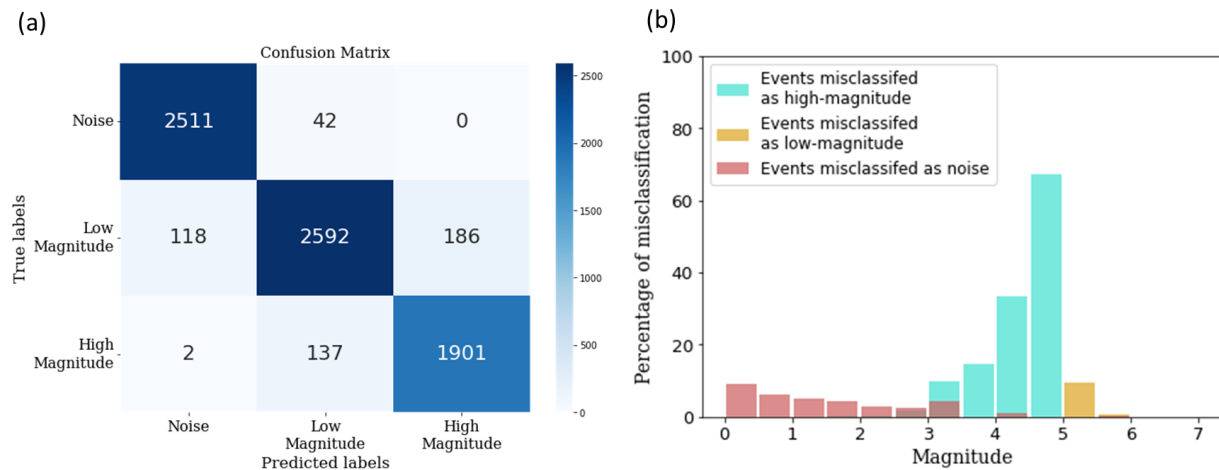
tude. The model obtains an overall accuracy ranging between 90.04 % and 93.86 %, which is comparable to the magnitude classification accuracy of 93.67 % achieved by Saad et al. (2020) using data from three seismic stations. This shows great potential in the area of single-station waveform analysis for earthquake early warning.

## 5 Conclusions

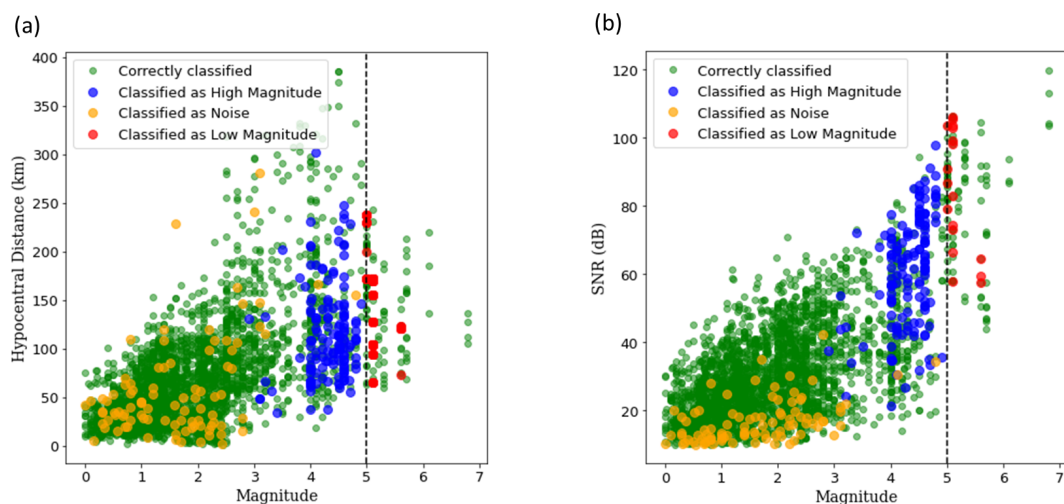
In this study, we present a deep learning model that classifies seismic waveform into three classes: noise, low-

magnitude events and high-magnitude events, with events of local magnitude equal to or above 5.0 categorised as “high-magnitude”. We investigate the effect of using different durations of P-wave data to perform the said task and demonstrate that changing the length of the waveform (1, 3, 10, 20 or 30 s after P arrival) has no significant effect on the model performance. We also find that the model classifies most of the data above a magnitude of 4.5 as high-magnitude, even though the decision boundary is chosen at 5.0, due to the higher class weight assigned to high-magnitude events. We obtain an overall accuracy of up to 93.86 %, and we expect





**Figure 7.** The classification results for a model trained on the 3 s data. **(a)** The confusion matrix (Ting, 2017) for a model trained and tested on the 3 s data. **(b)** The misclassification statistics for the same model, for different magnitude values. Note how the highest degree of misclassification happens close to the decision boundary. The percentage of low-magnitude events classified as high-magnitude is much higher than the percentage of high-magnitude events classified as low-magnitude. This is a result of the class weights we used while training the model.



**Figure 8.** Classification of events with different **(a)** hypocentral distance and **(b)** signal-to-noise ratio (SNR). It is observed that the model can correctly classify traces over a range of hypocentral distance and SNR, which exhibits its ability to learn from the frequency characteristics and does not directly learn from amplitude or SNR to some extent. There seems some visible clustering of misclassification of low-magnitude events as noise for SNR below 20 dB.

this to be very useful in the fast classification of seismological data.

**Data availability.** The seismic waveforms used in our research are a part of the Stanford EArthquake Dataset (STEAD) (Mousavi et al., 2019), and the dataset was downloaded from <https://github.com/smousavi05/STEAD>.

**Author contributions.** MC, NS, GR and HS contributed to the conception and design of the study. MC did the analysis with the help

of WL and JF. MC wrote the first draft of the manuscript. GR and NS wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

**Competing interests.** The contact author has declared that none of the authors has any competing interests.

**Disclaimer.** Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Acknowledgements.** This research is supported by the “KINach-wuchswissenschaftlerinnen” (grant SAI 01IS20059) by the Bundesministerium für Bildung und Forschung (BMBF). Calculations were performed at the Frankfurt Institute for Advanced Studies’ new GPU cluster, funded by the BMBF for the project Seismologie und Künstliche Intelligenz (SAI).

**Financial support.** This research has been supported by the Bundesministerium für Bildung und Forschung (grant no. SAI 01IS20059).

**Review statement.** This paper was edited by Irene Bianchi and reviewed by Filippo Gatti and one anonymous referee.

## References

- Allen, R., Gasparini, P., Kamigaichi, O., and Böse, M.: The Status of Earthquake Early Warning around the World: An Introductory Overview, *Seismol. Res. Lett.* 80, 682–693, <https://doi.org/10.1785/gssrl.80.5.682>, 2009.
- Allen, R. and Kanamori, H.: The Potential for Earthquake Early Warning in Southern California, *Science*, 300, 786–789, <https://doi.org/10.1126/science.1080912>, 2003.
- Allen, R. M. and Melgar, D.: Earthquake Early Warning: Advances, Scientific Challenges, and Societal Needs, *Annu. Rev. Earth Planet Sc.*, 47, 361–388, <https://doi.org/10.1146/annurev-earth-053018-060457>, 2019.
- Aly, M.: Survey on multiclass classification methods, *Neural Netw.*, 19, 1–9, 2005.
- Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, *SIGKDD Explorations Newsletter*, 6, 20–29, <https://doi.org/10.1145/1007730.1007735>, 2004.
- Bengio, Y.: Practical Recommendations for Gradient-Based Training of Deep Architectures, *Neural Networks: Tricks of the Trade: Second Edition*, Berlin, Heidelberg: Springer Berlin Heidelberg, 437–478, [https://doi.org/10.1007/978-3-642-35289-8\\_26](https://doi.org/10.1007/978-3-642-35289-8_26), 2012.
- Chakraborty, M., Rümpler, G., Stöcker, H., Li, W., Faber, J., Fenner, D., Zhou, K., and Srivastava, N.: Real Time Magnitude Classification of Earthquake Waveforms using Deep Learning, *EGU General Assembly 2021*, online, 19–30 Apr 2021, EGU21-15941, <https://doi.org/10.5194/egusphere-egu21-15941>, 2021.
- Chung, D. H. and Bernreuter, D. L.: Regional relationships among earthquake magnitude scales, *Rev. Geophys.*, 19, 649–663, <https://doi.org/10.1029/RG019i004p00649>, 1981.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P.: Natural language processing (almost) from scratch, *J. Mach. Learn. Res.*, 12, 2493–2537, <https://doi.org/10.5555/1953048.2078186>, 2011.
- Ekström, G. and Dziewonski, A.: Evidence of bias in estimations of earthquake size, *Nature*, 332, 319–323, <https://doi.org/10.1038/332319a0>, 1988.
- Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. A.: Deep learning for time series classification: a review, *Data Min. Knowl. Disc.*, 33, 917–963, <https://doi.org/10.1007/s10618-019-00619-1>, 2019.
- Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT press., <http://www.deeplearningbook.org> (last access: 20 October 2022), 2016.
- He, K., Ren, S., Sun, J., and Zhang, X.: Deep Residual Learning for Image Recognition, *Proc. Cvprr IEEE*, 770–778, <https://doi.org/10.1109/CVPR.2016.90>, 2016.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Proc. Mag.*, 29, 82–97, <https://doi.org/10.1109/MSP.2012.2205597>, 2012.
- Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Comput.*, 9, 1735–1780, 1997.
- Howell, Jr., B. F.: On the saturation of earthquake magnitudes, *B. Seismol. Soc. Am.*, 71, 1401–1422, <https://doi.org/10.1785/BSSA0710051401>, 1981.
- Jin, X., Zhang, H., Li, J., Wei, Y., and Ma, Q.: Earthquake magnitude estimation using the  $\tau_c$  and  $P_d$  method for earthquake early warning systems, *Earthq. Sci.*, 26, 23–31, <https://doi.org/10.1007/s11589-013-0005-4>, 2013.
- Kanamori, H.: Magnitude scale and quantification of earthquakes, *Tectonophysics*, 93, 185–199, [https://doi.org/10.1016/0040-1951\(83\)90273-1](https://doi.org/10.1016/0040-1951(83)90273-1), 1983.
- Kanamori, H.: Real-time seismology and earthquake damage mitigation, *Annu. Rev. Earth Planet Sc.*, 33, 195–214, <https://doi.org/10.1146/annurev.earth.33.092203.122626>, 2005.
- Kanamori, H. and Stewart, G. S.: Seismological aspects of the Guatemala Earthquake of February 4, 1976, *J. Geophys. Res.-Sol. Ea.*, 83, 3427–3434, <https://doi.org/10.1029/JB083iB07p03427>, 1978.
- Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings, 2015.
- Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J.: 1D convolutional neural networks and applications: A survey, *Mech. Sys. Signal Proc.*, 151, 107398, <https://doi.org/10.1016/j.ymssp.2020.107398>, 2021.
- Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P.: Machine Learning in Seismology: Turning Data into Insights, *Seismol. Res. Lett.*, 90, 3–14, <https://doi.org/10.1785/0220180259>, 2018.
- Krawczyk, B.: Learning from imbalanced data: open challenges and future directions, *Prog. Artif.*, 5, 221–232, <https://doi.org/10.1007/s13748-016-0094-0>, 2016.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Communications of the ACM*, 60, 84–90, <https://doi.org/10.1145/3065386>, 2017.
- Kuyuk, H. S. and Susumu, O.: Real-time classification of earthquake using deep learning, *Proc. Comput. Sci.*, 140, 298–305, <https://doi.org/10.1016/j.procs.2018.10.316>, 2018.
- LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, *Nature*, 521, 436–444, <https://doi.org/10.1038/nature14539>, 2015.
- Li, W., Chakraborty, M., Fenner, D., Faber, J., Zhou, K., Ruempker, G., Stoecker, H., and Srivastava, N.: EPick: Multi-Class Attention-based U-shaped Neural Network for Earth-



- quake Detection and Seismic Phase Picking, arXiv [preprint], <https://doi.org/10.48550/arXiv.2109.02567>, 6 September 2021.
- Liao, W. Y., Lee, E. J., Mu, D., Chen, P., and Rau, R. J.: ARRU Phase Picker: Attention Recurrent-Residual U-Net for Picking Seismic P- and S-Phase Arrivals, *Seismol. Res. Lett.*, 92, 2410–2428, <https://doi.org/10.1785/0220200382>, 2021.
- Lomax, A., Michelini, A., and Jozinović, D.: An Investigation of Rapid Earthquake Characterization Using Single-Station Waveforms and a Convolutional Neural Network, *Seismol. Res. Lett.*, 90, 517–529, <https://doi.org/10.1785/0220180311>, 2019.
- Madhyastha, P. and Jain, R.: On Model Stability as a Function of Random Seed, arXiv [preprint], <https://doi.org/10.48550/arXiv.1909.10447>, 2019.
- Meier, M. A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., Li, Z., Andrews, J., Hauksson, E., and Yue, Y.: Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning, *J. Geophys. Res. Sol.-Ea.*, 124, 788–800, <https://doi.org/10.1029/2018JB016661>, 2019.
- Mikolov, T., Deoras, A., Povey, D., Burget, L., and Černocký, J.: Strategies for training large scale neural network language models, 2011 IEEE Workshop on Automatic Speech Recognition Understanding, 196–201, <https://doi.org/10.1109/ASRU.2011.6163930>, 2011.
- Mousavi, S. M. and Beroza, G. C.: A machine-learning approach for earthquake magnitude estimation, *Geophys. Res. Lett.*, 47, e2019GL085976, <https://doi.org/10.1029/2019GL085976>, 2020.
- Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C.: STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, *IEEE Access*, 7, 179464–179476, <https://doi.org/10.1109/ACCESS.2019.2947848>, 2019 (data available at: <https://github.com/smousavi05/STEAD>, last access: 20 October 2022).
- Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C.: Earthquake transformer – an attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nat. Commun.*, 11, 3952, <https://doi.org/10.1038/s41467-020-17591-w>, 2020.
- Münchmeyer, J., Bindi, D., Leser, U., and Tilmann, F.: The transformer earthquake alerting model: a new versatile approach to earthquake early warning, *Geophys. J. Int.*, 225, 646–656, <https://doi.org/10.1093/gji/ggaa609>, 2020.
- Murphy, K. P.: Machine learning: a probabilistic perspective, MIT press, ISBN 9780262018029, 2012.
- Nagi, J., Ducatelle, F., Di Caro, G. A., Cireşan, D., Meier, U., Giusti, A., Nagi, F., Schmidhuber, J., and Gambardella, L. M.: Max-pooling convolutional neural networks for vision-based hand gesture recognition, 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 342–347, <https://doi.org/10.1109/ICSIPA.2011.6144164>, 2011.
- Nakamura, Y.: On the Urgent Earthquake Detection and Alarm System (UrEDAS), 9th world conference on earthquake engineering, VII, B7, 673–678, 1988.
- Panakkat, A. and Adeli, H.: Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators, *Comput.-Aided Civ. Infrastruct. Eng.*, 24, 280–292, <https://doi.org/10.1111/j.1467-8667.2009.00595.x>, 2009.
- Perol, T., Gharbi, M., and Denolle, M.: Convolutional neural network for earthquake detection and location, *Sci. Adv.*, 4, e1700578, <https://doi.org/10.1126/sciadv.1700578>, 2018.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L.: Deep contextualized word representations, *Association for Computational Linguistics*, 2227–2237, <https://doi.org/10.18653/v1/N18-1202>, 2018.
- Prechelt, L.: Early Stopping – But When?, *Neural Networks: Tricks of the Trade: Second Edition*, Springer Berlin Heidelberg, 53–67, [https://doi.org/10.1007/978-3-642-35289-8\\_5](https://doi.org/10.1007/978-3-642-35289-8_5), 2012.
- Ross, Z. E., Meier, M. A., and Hauksson, E.: P wave arrival picking and first-motion polarity determination with deep learning, *J. Geophys. Res. Sol.-Ea.*, 123, 5120–5129, <https://doi.org/10.1029/2017JB015251>, 2018.
- Saad, O. M., Hafez, A. G., and Soliman, M. S.: Deep Learning Approach for Earthquake Parameters Classification in Earthquake Early Warning System, *IEEE Geosci. Remote Sens. Lett.*, 18, 1293–1297, <https://doi.org/10.1109/LGRS.2020.2998580>, 2020.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *J. Mach. Learn. Res.*, 15, 1929–1958, 2014.
- Ting, K. M.: Confusion Matrix, *Encyclopedia of Machine Learning and Data Mining*, Boston, MA: Springer US, 260–260, [https://doi.org/10.1007/978-1-4899-7687-1\\_50](https://doi.org/10.1007/978-1-4899-7687-1_50), 2017.
- Wang, J. and Teng, T. L.: Artificial neural network-based seismic detector, *B. Seismol. Soc. Am.*, 85, 308–319, <https://doi.org/10.1785/BSSA0850010308>, 1995.
- Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., and Xu, H.: Time Series Data Augmentation for Deep Learning: A Survey, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, <https://doi.org/10.24963/ijcai.2021/631>, 2021.
- Wu, Y. M. and Zhao, L.: Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning, *Geophys. Res. Lett.*, 33, L16312, <https://doi.org/10.1029/2006GL026871>, 2006.
- Zhu, W. and Beroza, G. C.: PhaseNet: a deep-neural-network-based seismic arrival-time picking method, *Geophys. J. Int.*, 216, 261–273, <https://doi.org/10.1093/gji/ggy423>, 2019.
- Ziv, A.: New frequency-based real-time magnitude proxy for earthquake early warning, *Geophys. Res. Lett.*, 41, 7035–7040, <https://doi.org/10.1002/2014GL061564>, 2014.