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

Megha Chakraborty[1,4], Wei Li[1], Johannes Faber[1,2], Georg Rümpker[1,4], Horst Stoecker[1,2,3,5], and Nishtha Srivastava[1,4] *

[1]Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany
[2]Institute for Theoretical Physics, Goethe Universität, 60438 Frankfurt am Main, Germany
[3]Xidian-FIAS international Joint Research Center, Giersch Science Center, D-60438 Frankfurt am Main, Germany
[4]Institute of Geosciences, Goethe-University Frankfurt, 60438 Frankfurt am Main, Germany
[5]GSI Helmholtzzentrum für Schwerionenforschung GmbH, 64291 Darmstadt, Germany

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

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 have 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 s, 3 s, 10 s, 20 s 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 characterization (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 above-mentioned reasons and the empirical nature of majority of the magnitude scales, it is one of the most difficult parameters to estimate (Chung and Bernreuter, 1981; Ekström and Dziewonski, 1988). Some of the classical approaches to obtain first estimates of earthquake magnitude have used empirical relations for parameters such as predominant period τ^{p}_{max} (Nakamura, 1988; Allen and Kanamori, 2003), effective average period τ_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 seconds 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 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).

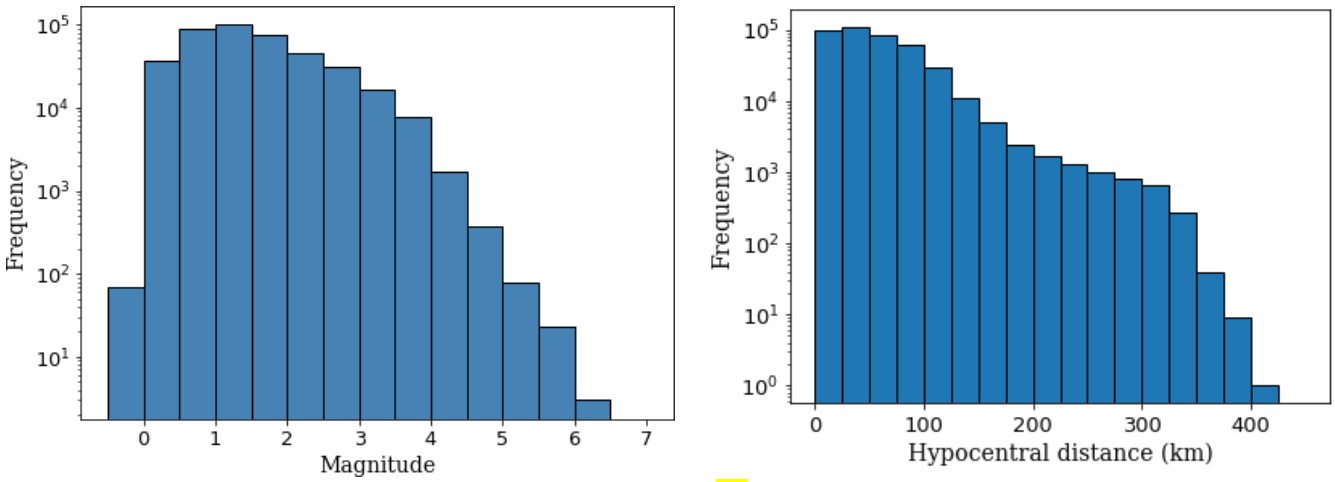
Considering that timeliness is of the essence in rapid earthquake characterisation, it becomes important to find an optimum duration for the input data, that can provide a reliable and statistically significant estimate for various earthquake parameters while using minimum amount of P-wave data. In this 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), which uses data from three seismic station 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 and Resources) to train and test our model. STEAD is a high-quality bench-marked dataset created for machine learning and deep learning applications and contains seismic event and noise waveforms of duration 1 minute recorded by over 2,500 seismic stations across the globe. The waveforms have been detrended and filtered with a bandpass filter between 1.0 to 40.0 Hz, followed by a resampling at 100Hz. A metadata consisting of 35 attributes for earthquake traces and 8 attributes for noise traces is provided by the authors.

To ensure consistency in magnitude we only use traces for which the magnitude is provided in 'ml' 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 s, 3 s, 10 s, 20 s and 30 s) on the performance of this classifier model. In each case the P-wave data is preceded by 2.8-3.0 seconds 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.



(a) (b)

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.

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 Figure 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). 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. 1/3rd of events with magnitude between 4.0 and 4.5 are used.
  3. 1/50th of events with magnitude between 2.0 and 4.5 are used.
  4. 1/100th of events with magnitude less than 2.0 are used.

  1/25th 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 Figure 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 is 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.

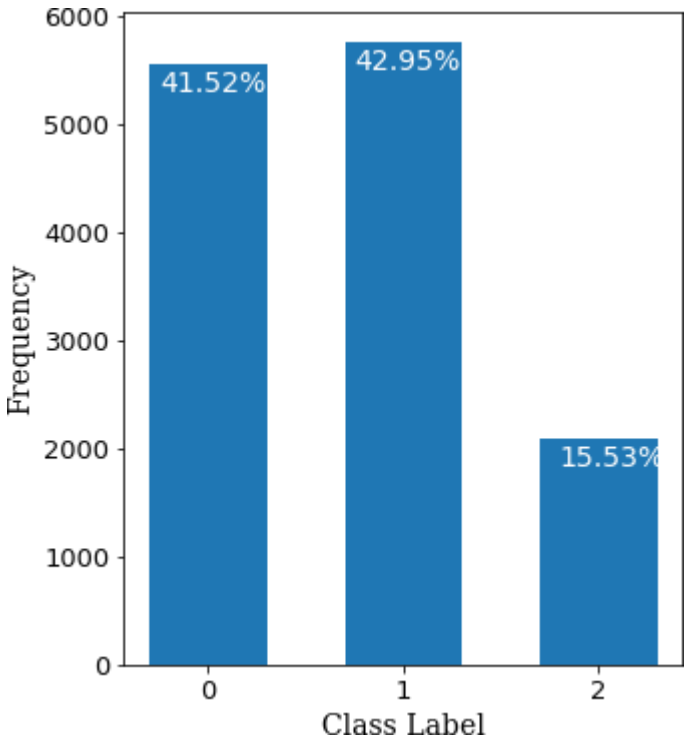


Figure 2: The distribution of classes in the training dataset obtained by under-sampling 'noise' (represented by class 0) and 'low-magnitude' (represented by class 1) data and applying data-augmentation to 'high-magnitude' (represented by class 2) events. A similar distribution of classes is seen in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

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 (Srivastava et al., 2014) and MaxPooling (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 Maxpooling 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 (Figure 3). A detailed description of the model architecture is provided in the caption for Figure 3.

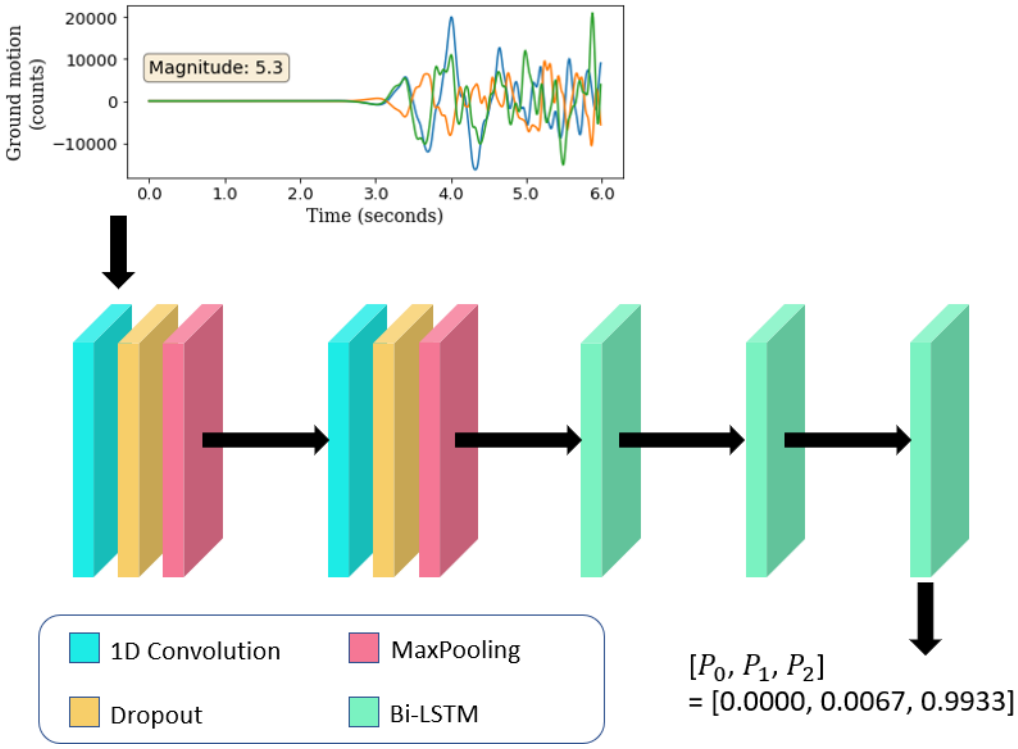


Figure 3: The architecture of the model used to perform the 3-class classification. The input to the model is 3-component seismic waveform data from a single station. The example shown here corresponds to the case where 3 seconds of P-wave data is used (the total length of data is, thus, 6 seconds). The 1D Convolution layers have a kernel size of 4 and 8 filters each; the drop rate for each Dropout layer is 0.2, and each MaxPooling 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 P0, P1 and P2 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.

The model is trained using Adam optimiser (Kingma and Ba, 2014), Categorical Crossentropy 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 positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

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

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + FN + TN + FP} \quad (1)$$

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

$$Precision = \frac{\mathrm{TP}}{TP + FP} \quad (2)$$

- **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:

$$Recall = \frac{\mathrm{TP}}{TP + FN} \quad (3)$$

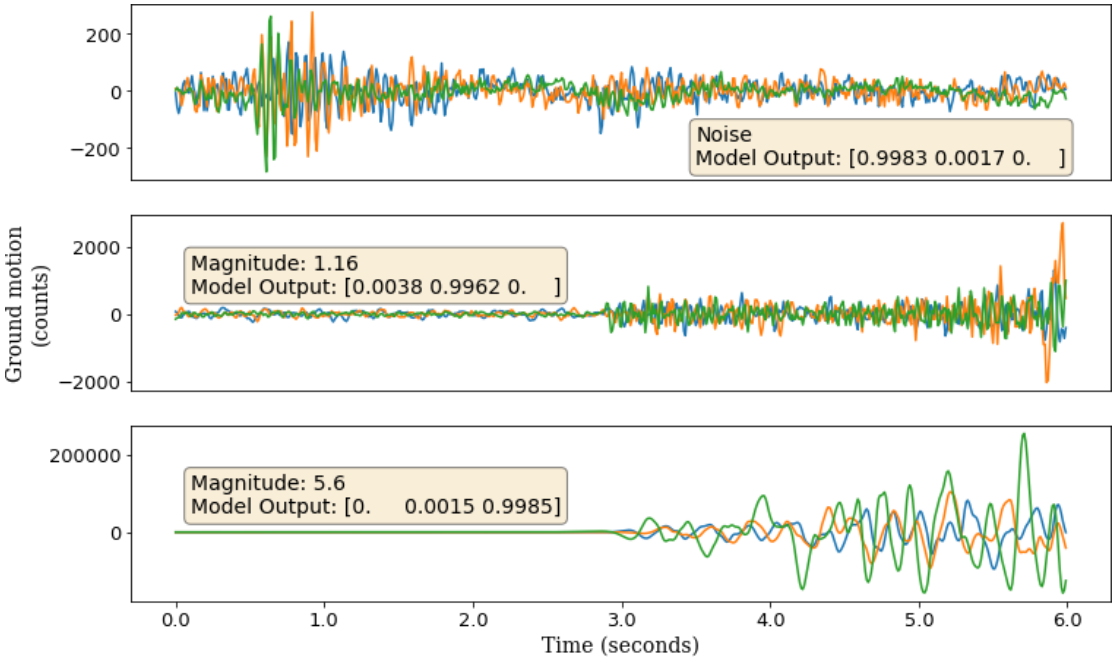


Figure 4: Examples of waveforms that have been correctly classified. In each case the highest probability corresponds to the respective class.

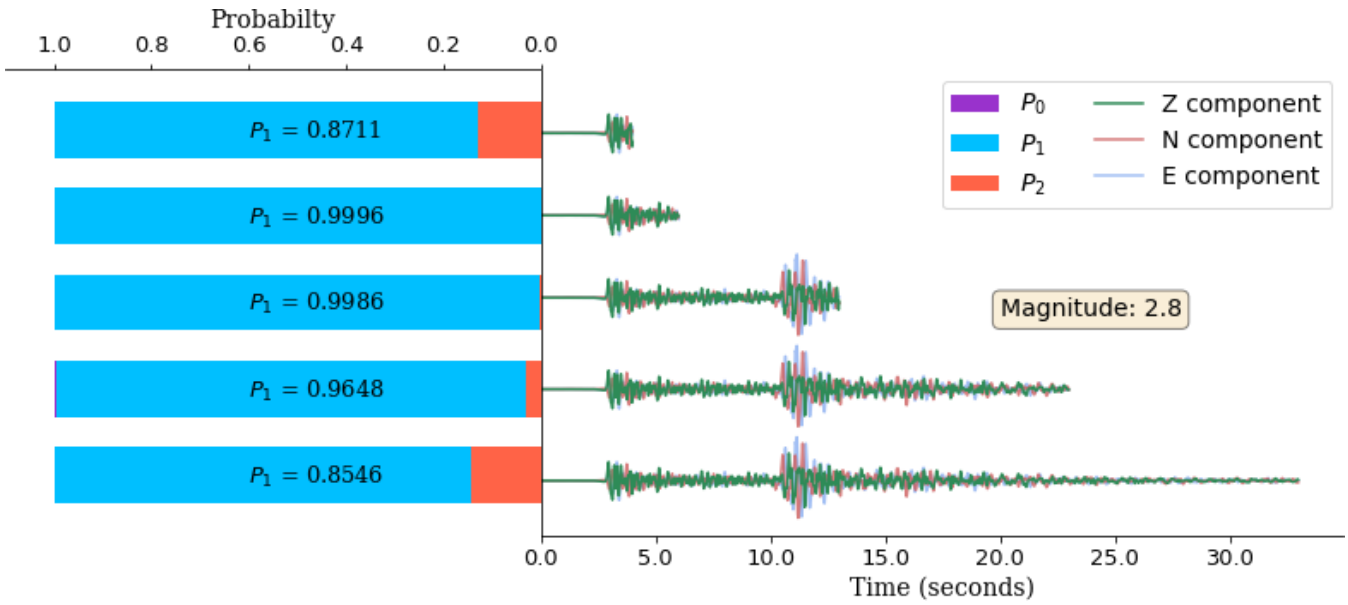
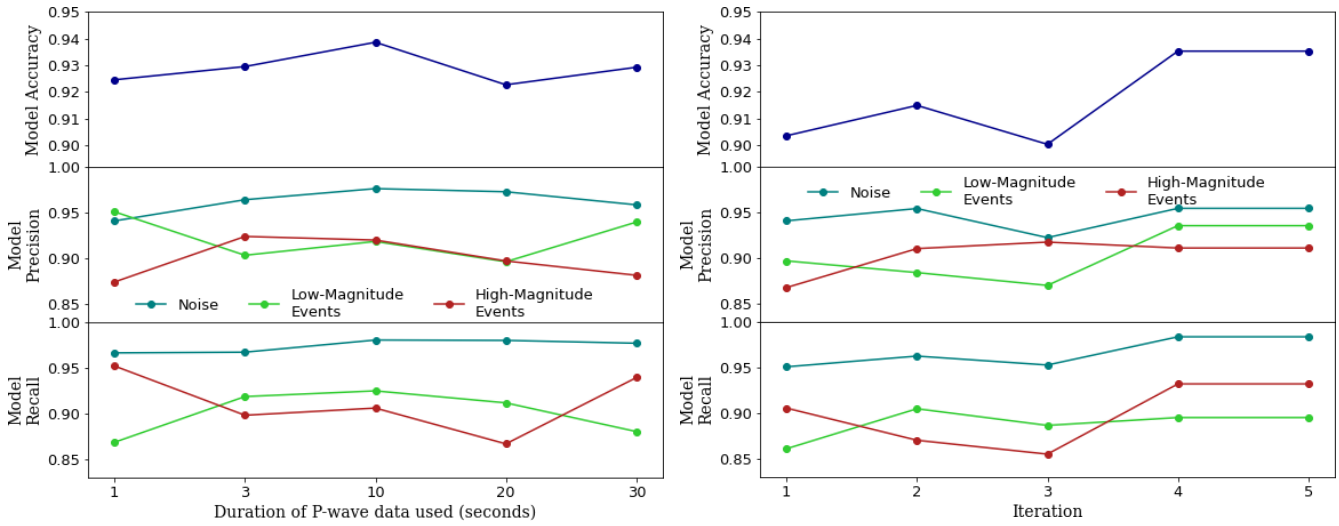


Figure 5: Softmax probabilities for different input lengths of the same waveform, predicted by the models trained on the corresponding lengths of data. The waveform used here corresponds to an event of magnitude 2.8, although the maximum probability corresponds to class 1, the values of these probabilities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

Figure 4 shows three waveforms, (one from each class) that has been correctly classified. The softmax probabilities, as described in section 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.

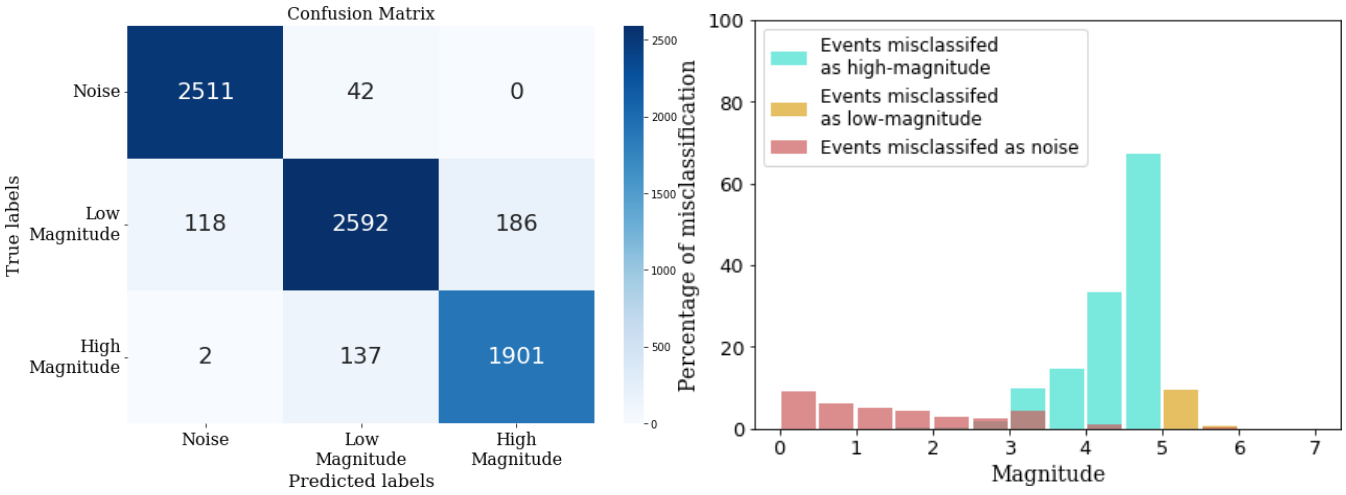
4 Discussions

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 (Figure 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 re-training the model on the same data.



(a) (b)

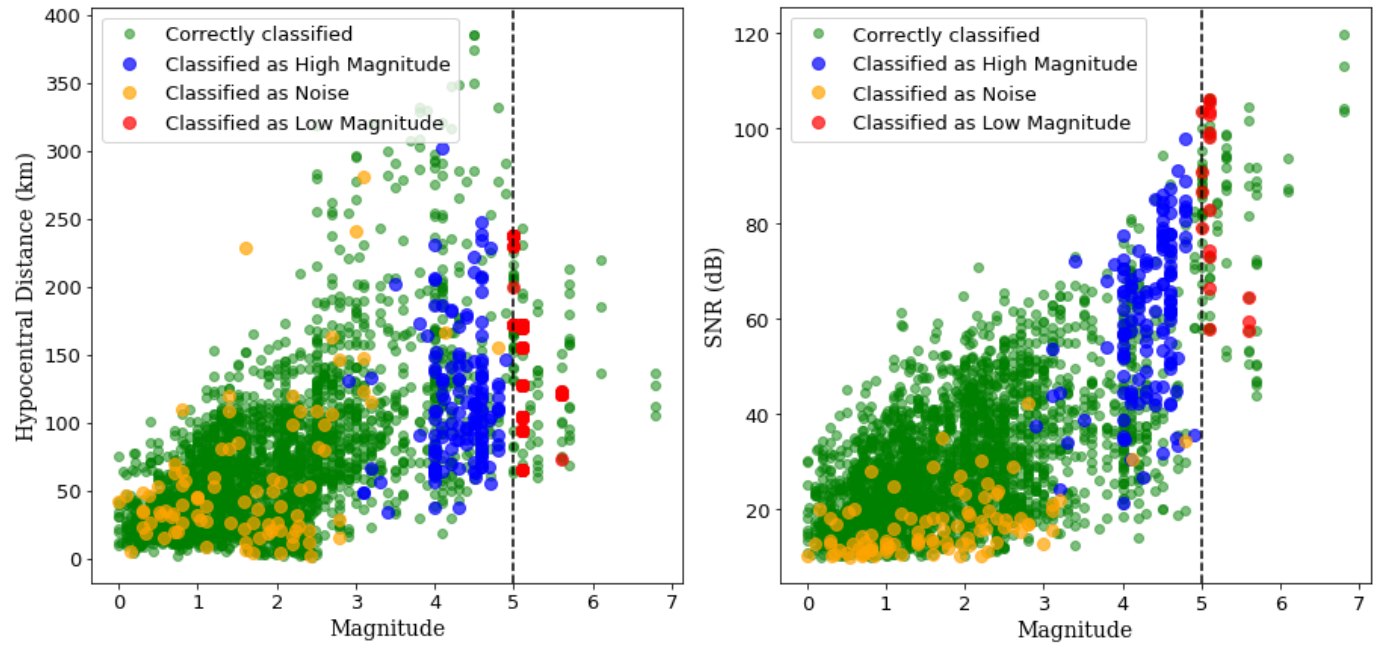
Figure 6: (a) Variation in classifier model performance when different duration (1 s, 3 s, 10 s, 20 s, 30 s) of P-wave data are used; (b) Variation in the classifier model performance when the same model is re-trained on the same data (in this case 3 seconds of P-wave data used) five times. This shows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



(a) (b)

Figure 7: The classification results for a model trained on the 3 second data. (a) The confusion matrix (Ting 2017) for a model trained and tested on the 3 second data. (b) The mis-classification statistics for the same model, for different magnitude values. Note how the highest degree of mis-classification 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 7 shows the classification statistics for one of the iterations of the model trained on the 3 second data. The events classified as noise tend to be of low magnitude, while the mis-classification 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 less 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.



(a) (b)

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 20dB.

Figure 8 visualizes the classification of events across different hypocentral distance (Figure 8a) and signal-to-noise ratios (Figure 8b). We observe that there are instances of correct classification across a wide-range of hypocentral distances and SNRs, which means that the model is capable to learn 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 ratio.

Despite maximizing the amount of data on either size of the decision boundary between low and high magnitude we find some incorrect classifications, most of which lies within a range of 5.0 ± 0.5 as can be seen in Figure 8. However, considering that sometimes even magnitudes in 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 magnitude. 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 Conclusion

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 having local magnitude equal to or above 5.0 categorised as 'high-magnitude'. We investigate the effect of using different duration of P-wave data to perform the said task and demonstrate that changing the length of the waveform (1 s, 3 s, 10 s, 20 s or 30 s after P-arrival) has no significant effect on the model performance. We also find that the model classifies most 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 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 was downloaded from https://github.com/smousavi05/STEAD.

Author contributions

Megha chakraborty, Nishtha Srivastava, Georg Ruempker and Horst Stöcker contributed to conception and design of the study. Megha Chakraborty did the analysis with the help of Wei Li and Johannes Faber. Megha chakraborty wrote the first draft of the manuscript. Georg Ruempker and Nishtha Srivastava wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Declaration of Competing Interests

The authors declare no competing interests.

Acknowledgement

This research is supported by the “KINachwuchswissenschaftlerinnen" - 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 BMBF for the project Seismologie und Artifizielle Intelligenz (SAI).

References

Allen, R., P. Gasparini, O. Kamigaichi, and M. Böse (2009). “The Status of Earthquake Early Warning around the World: An Introductory Overview”. In: *Seismol. Res. Lett.* 80,pp. 682–693. DOI: 10.1785/gssrl.80.5.682.

Allen, R. and H. Kanamori (May 2003). “The Potential for Earthquake Early Warning in SouthernCalifornia”. In: *Science (New York, N.Y.)* 300, pp. 786–789. DOI: 10.1126/science.1080912. Allen, R. M. and D. Melgar (2019). “Earthquake Early Warning: Advances, Scientific Challenges,and Societal Needs”. In: *Annu. Rev. Earth Planet Sci.* 47.1, pp. 361–388. DOI: 10.1146/annurev-earth-053018-060457.

Aly, M. (2005). “Survey on multiclass classification methods”. In: *Neural Netw.* 19, pp. 1–9.

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard (June 2004). “A Study of the Behaviorof Several Methods for Balancing Machine Learning Training Data”. In: *SIGKDD Explorations Newsletter* 6.1, pp. 20–29. Issn: 1931-0145. DOI: 10.1145/1007730 .1007735. url: https://doi.org/10.1145/1007730.1007735.

Bengio, Y. (2012). “Practical Recommendations for Gradient-Based Training of Deep Architec- tures”. In: *Neural Networks: Tricks of the Trade: Second Edition*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 437–478. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_26. url: https://doi.org/10.1007/978-3-642-35289-8_26.

Chakraborty, M., G. Rümpker, H. Stöcker, W. Li, J. Faber, D. Fenner, K. Zhou, and N. Srivastava(2021). “Real Time Magnitude Classification of Earthquake Waveforms using Deep Learning”. In: *EGU General Assembly 2021, online* EGU21-15941. url: https://doi.org/10.5194/ egusphere-egu21-15941.

Chung, D. H. and D. L. Bernreuter (1981). “Regional relationships among earthquake magnitude scales”. In: *Rev. Geophys.*19.4, pp. 649–663. DOI: https://doi.org/10.1029/RG019i004p00649. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029RG019i004p00649. url: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/RG019i004p00649.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). “Natural language processing (almost) from scratch”. In: *J. Mach. Learn. Res.* 12.AR- TICLE, pp. 2493–2537. DOI: 10.5555/1953048.2078186.

Ekström, G. and A. Dziewonski (1988). “Evidence of bias in estimations of earthquake size.” In: *Nature* 332, pp. 319–323. DOI: https://doi.org/10.1038/332319a0.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press. url: http://www. deeplearningbook.org.

He, K., S. Ren, J. Sun, and X. Zhang (2016). “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”. In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97. DOI: 10.1109/MSP.2012.2205597.

Hochreiter, S. and J. Schmidhuber (Nov. 1997). “Long Short-Term Memory”. In: *Neural Com- put.* 9.8, pp. 1735–1780. Issn: 0899-7667. DOI: 101162/neco.1997.9.8.1735. url: https://doi.org/10.1162/neco.1997.9.8.1735.

Howell Jr, B.F. (1981). “On the saturation of earthquake magnitudes.” In: *Bull. Seismol. Soc. Am*71(5), pp. 1401–1422. url: https ://doi.org/10.1785/BSSA0710051401.

Ismail Fawaz, H., G. Forestier, J. Weber, L. Idoumghar, and P.A. Muller (2019). “Deep learning for time series classification: a review.” In: *Data Mining and Knowledge Discovery* 33, pp. 917–963. DOI: https://doi.org/10.1007/s10618-019-00619-1.

Jin, X., H. Zhang, J. Li, Y. Wei, and Q. Ma (2013). “Earthquake magnitude estimation using theτ_c and P_d method for earthquake early warning systems”. In: *Earthq. Sci.* 26.es-26-1-23, pp. 23–31. Issn: 1674-4519. DOI: 10.1007/s11589-013-0005-4.

Kanamori, H. (1983). “Magnitude scale and quantification of earthquakes.” In: *Tectonophysics* 93(3-4), pp. 185–199. url: https://doi.org/10.1016/0040-1951(83)90273-1.

Kanamori, H. (2005). “REAL-TIME SEISMOLOGY AND EARTHQUAKE DAMAGE MITIGATION”. In: *Annu. Rev. Earth Planet Sci.* 33.1, pp. 195–214. DOI: 10.1146/annurev.earth.33.092203.122626.

Kanamori, H. and G. S. Stewart (1978). “Seismological aspects of the Guatemala Earthquake ofFebruary 4, 1976”. In: *J. Geophys. Res. Solid Earth* 83.B7, pp. 3427–3434. DOI:https://doi.org/10.1029/JB083iB07p03427.

Kingma, D. P. and J. Ba (Dec. 2014). *Adam: A Method for Stochastic Optimization*.

Kiranyaz, S., O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman (2021). “1D convolu-tional neural networks and applications: A survey”. In: *Mech. Sys. Signal Process.* 151, p. 107398. Issn: 0888-3270. url: https://doi.org/10.1016/j.ymssp.2020.107398.

Kong, Q., D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft (2018). “Machine Learning in Seismology: Turning Data into Insights”. In: *Seismol. Res. Lett.* 90.1, pp. 3–14. url: https://doi.org/10.1785/0220180259.

Krawczyk, B. (2016). “Learning from imbalanced data: open challenges and future directions”. In: *Prog. Artif.* 5.4, pp. 221–232. url: https://doi.org/10.1007/s13748- 016-0094-0.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2017). “ImageNet Classification with Deep Convo- lutional Neural Networks”. In: *Communications of the ACM* 60.6, pp. 84–90. Issn: 0001-0782. url: https://doi.org/10.1145/3065386.

Kuyuk, H. S. and O. Susumu (2018). “Real-time classification of earthquake using deep learning”. In: *Proc. Comput. Sci* 140, pp. 298–305. url: https://doi.org/10.1016/j.procs.2018.10. 316.

LeCun, Y., Y. Bengio, and G. Hinton (2015). “Deep learning.” In: *Nature* 521, pp. 436–444. url: https://doi.org/10.1038/nature14539.

Li, W., M. Chakraborty, D. Fenner, J. Faber, K. Zhou, G. Ruempker, H. Stoecker, and N. Srivastava (2021). *EPick: Multi-Class Attention-based U-shaped Neural Network for Earthquake Detection and Seismic Phase Picking.* url: https://arxiv.org/abs/2109.02567.

Liao, W.Y., E.J. Lee, D. Mu, P. Chen, and R.J. Rau (Mar. 2021). “ARRU Phase Picker: Atten- tion Recurrent-Residual U-Net for Picking Seismic P- and S-Phase Arrivals”. In: *Seismol. Res. Lett.* 92.4, pp. 2410–2428. Issn: 0895-0695. DOI: 10.1785/0220200382. eprint: https://pubs.geoscienceworld.org/ssa/srl/article- pdf/92/4/2410/5351037/srl- 2020382.1.pdf. url: https://doi.org/10.1785/0220200382.

Lomax, A., A. Michelini, and D. Jozinović (Feb. 2019). “An Investigation of Rapid Earthquake Characterization Using Single-Station Waveforms and a Convolutional Neural Network”. In: *Seismol. Res. Lett.* 90, pp. 517–529. DOI: 10.1785/0220180311.

Madhyastha, P. and R. Jain (2019). *On Model Stability as a Function of Random Seed.* arXiv: 1909.10447 [cs.LG].

Meier, M.A., Z. E. Ross, A. Ramachandran, A. Balakrishna, S. Nair, P. Kundzicz, Z. Li, J. Andrews,E. Hauksson, and Y. Yue (2019). “Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning”. In: *J. Geophys. Res. Solid Earth* 124.1, pp.

788–800. DOI: https://doi.org/10.1029/2018JB016661. url: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JB016661.

Mikolov, T., A. Deoras, D. Povey, L. Burget, and J. Černocký (2011). "Strategies for training large scale neural network language models". In: *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 196–201. DOI: 10.1109/ASRU.2011.6163930.

Mousavi, S. M. and G. C. Beroza (2020). "A machine-learning approach for earthquake magnitude estimation." In: *Geophys. Res. Lett.* 47, e2019GL085976. url: https://doi.org/ 10.1029/2019GL085976.

Mousavi, S.M., W.L. Ellsworth, W. Zhu, L.Y. Chuang, and G.C. Beroza (2020). "Earthquake trans- former—an attentive deep-learning model for simultaneous earthquake detection and phase picking". In: *Nat. Commun.* 11.3952. url: https://doi.org/10.1038/s41467-020- 17591-w.

Mousavi, S. M., Y. Sheng, W. Zhu, and G. C. Beroza (2019). "STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI." In: *IEEE Access* 7, pp. 179464– 179476. url: https://doi.org/10.1109/ACCESS.2019.2947848.

Münchmeyer, J., D. Bindi, U. Leser, and F. Tilmann (2020). "The transformer earthquake alerting model: a new versatile approach to earthquake early warning". In: *Geophys. J. Int.* 225.1, pp. 646–656. DOI: 10.1093/gji/ggaa609. url: https://doi.org/10.1093/gji/ggaa609.

Murphy, K.P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nagi, J., F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella (2011). "Max-pooling convolutional neural networks for vision-based hand gesture recognition". In: *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 342–347. DOI: 10.1109/ICSIPA.2011.6144164.

Nakamura, Y. (1988). "On the Urgent Earthquake Detection and Alarm System (UrEDAS)". In: *9th world conference on earthquake engineering* VII.B7, pp. 673–678.

Panakkat, A. and H. Adeli (2009). "Recurrent neural network for approximate earthquake timeand location prediction using multiple seismicity indicators". In: *Comput.-Aided Civ. Infrastruct. Eng.* 24.4, pp. 280–292. url: https://doi.org/10.1111/j.1467-8667.2009.00595.x.

Perol, T., M. Gharbi, and M. Denolle (2018). "Convolutional neural network for earthquake detec- tion and location". In: *Sci. Adv.* 4.2, e1700578. DOI: 10.1126/sciadv.1700578.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (June 2018). "Deep contextualized word representations". In: pp. 2227–2237. DOI: 10.18653/v1/N18-1202. url: https://aclanthology.org/N18-1202.

Prechelt, L (2012). "Early Stopping — But When?" In: *Neural Networks: Tricks of the Trade: Second Edition*. Springer Berlin Heidelberg, pp. 53–67. ISBN: 978-3-642-35289-8. DOI: 10.1007/ 978-3-642-35289-8_5. url: https://doi.org/10.1007/978-3-642-35289-8_5.

Ross, Z. E., M.A. Meier, and E. Hauksson (2018). "P wave arrival picking and first-motion polarity determination with deep learning". In: *J. Geophys. Res. Solid Earth*123.6, pp. 5120–5129. url: https://doi.org/10.1029/2017JB015251.

Saad, O. M., A. G. Hafez, and M. S. Soliman (2020). "Deep Learning Approach for EarthquakeParameters Classification in Earthquake Early Warning System." In: *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5. DOI: 10.1109/LGRS.2020.2998580.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). "Dropout: ASimple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* 15.56, pp. 1929–1958. url: http://jmlr.org/papers/v15/srivastava14a.html.

Ting, K. M. (2017). "Confusion Matrix". In: *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US, pp. 260–260. ISBN: 978-1-4899-7687-1. DOI: 10.1007/978-1-4899-7687-1_50. url: https://doi.org/10.1007/978-1-4899-7687-1_50.

Wang, J. and T.L. Teng (Feb. 1995). "Artificial neural network-based seismic detector". In: *Bull. Seismol. Soc. Am* 85.1, pp. 308–319. url: https://doi.org/10.1785/ BSSA0850010308.

Wen, Q., L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu (Aug. 2021). "Time Series Data Augmentation for Deep Learning: A Survey". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. DOI: 10.24963/ijcai.2021/631. url: http://dx.doi.org/10.24963/ijcai.2021/631.

Wu, Y.M. and L. Zhao (2006). "Magnitude estimation using the first three seconds P-wave am- plitude in earthquake early warning". In: *Geophys. Res. Lett.* 33.16. url: https://doi.org/10.1029/2006GL026871.

Zhu, W. and G. C. Beroza (2019). "PhaseNet: a deep-neural-network-based seismic arrival-time picking method". In: *Geophys. J. Int.* 216.1, pp. 261–273. url: https:// doi.org/10.1093/gji/ggy423.

Ziv, A. (2014). "New frequency-based real-time magnitude proxy for earthquake early warning".In: *Geophys. Res. Lett.* 41.16, pp. 7035–7040. url: https://doi.org/10.1002/2014GL061564.