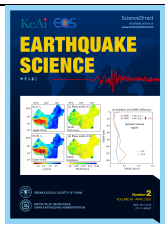
Earthquake Science 36 (2023) 84–94

Contents lists available at ScienceDirect

Earthquake Science

Journal homepage: https://www.keaipublishing.com/en/journals/earthquake-science/



Article

DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology

Ming Zhao[1,2,✉], Zhuowei Xiao[3,✉], Shi Chen[1,2] and Lihua Fang[1,4]

[1] *Institute of Geophysics*, *China Earthquake Administration*, *Beijing* 100081, *China*
[2] *Beijing Baijiatuan Earth Sciences National Observation and Research Station*, *Beijing* 100095, *China*
[3] *Institute of Geology and Geophysics*, *Chinese Academy of Sciences*, *Beijing* 100029, *China*
[4] *Key Laboratory of Earthquake Source Physics*, *China Earthquake Administration*, *Beijing* 100081, *China*

Key points:

- The first large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology is proposed.
- The dataset is of great significance for training high generalization models for seismic data processing applicable to Chinese mainland.
- The dataset serves a high-quality benchmark for various kinds of machine learning model development and data-driven seismological research.

A B S T R A C T

In recent years, artificial intelligence technology has exhibited great potential in seismic signal recognition, setting off a new wave of research. Vast amounts of high-quality labeled data are required to develop and apply artificial intelligence in seismology research. In this study, based on the 2013–2020 seismic cataloging reports of the China Earthquake Networks Center, we constructed an artificial intelligence seismological training dataset ("DiTing") with the largest known total time length. Data were recorded using broadband and short-period seismometers. The obtained dataset included 2,734,748 three-component waveform traces from 787,010 regional seismic events, the corresponding P- and S-phase arrival time labels, and 641,025 P-wave first-motion polarity labels. All waveforms were sampled at 50 Hz and cut to a time length of 180 s starting from a random number of seconds before the occurrence of an earthquake. Each three-component waveform contained a considerable amount of descriptive information, such as the epicentral distance, back azimuth, and signal-to-noise ratios. The magnitudes of seismic events, epicentral distance, signal-to-noise ratio of P-wave data, and signal-to-noise ratio of S-wave data ranged from 0 to 7.7, 0 to 330 km, –0.05 to 5.31 dB, and –0.05 to 4.73 dB, respectively. The dataset compiled in this study can serve as a high-quality benchmark for machine learning model development and data-driven seismological research on earthquake detection, seismic phase picking, first-motion polarity determination, earthquake magnitude prediction, early warning systems, and strong ground-motion prediction. Such research will further promote the development and application of artificial intelligence in seismology.

Production and Hosting by Elsevier on behalf of KeAI

✉ **Corresponding author.** Zhao M, email: mzhao@cea-igp.ac.cn; Xiao ZW, email: xiaozhuowei@mail.iggcas.ac.cn

Article history:
Received 30 December 2021
Received in revised form 3 February 2022
Accepted 14 February 2022
Available online 30 March 2022

https://doi.org/10.1016/j.eqs.2022.01.022

Keywords: artificial intelligence; benchmark dataset; earthquake detection; seismic phase identification; first-motion polarity

Citation: Zhao M, Xiao ZW, Chen S and Fang LH (2023). DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology. Earthq Sci **36**(2): 84–94, doi: 10.1016/j.eqs.2022.01.022.

1. Introduction

Seismology is an observation-based subject. With the continuous densification of seismic observation networks, large amounts of observation data have been accumulated. To this end, there is an urgent need for more methods to mine and analyze the information associated with observed waveform data in a timely manner. In recent years, big data and machine learning techniques have developed rapidly, enabling seismic data processing to develop toward efficient, accurate, and automated intelligent processing. For instance, many studies have focused on applying machine learning methods to the critical steps of seismic data processing, especially earthquake detection and seismic phase identification. Among them, deep learning-based seismic phase identification is the most progressed (Perol et al., 2018; Ross et al., 2018a; Xiao ZW et al., 2021; Wang J et al., 2019; Zhou YJ et al., 2019; Zhu LJ et al., 2019; Zhao M et al., 2019a, b; Zhou BW et al., 2020).

As deep learning models usually involve a large number of parameters, it is often necessary to train models using large-scale datasets to avoid "overfitting." In recent years, scholars worldwide have established large-scale training datasets while developing algorithms. For instance, Ross et al. (2018b) trained and tested a convolutional neural network (CNN) using up to 4,847,248 manually picked seismic waveform traces from the Southern California Seismic Network. The authors obtained an average standard deviation of only 0.023 s between the test data and manually picked arrival times, and 95% accuracy for the P-wave first-motion polarity classification. Mousavi et al. (2020) collected approximately 1.2×10^6 earthquake waveform traces from publicly available global data to build the STEAD dataset, which was then trained using the Transformer network with an attention mechanism. The resulting EQTransformer model has been applied to the data recorded by the Japanese Hi-net stations. In addition, Michelini et al. (2021) collected nearly 1.2×10^6 waveform traces from approximately 50,000 earthquakes listed in the Italian Seismic Bulletin from 2005 to 2020 to construct the INSTANCE dataset specifically for machine learning research.

Benefiting from the collection and collation of vast amounts of high-quality seismic waveform data and labels, some deep learning-based earthquake detection and phase picking models have achieved high generalization abilities, showing preliminary prospects for practical application. For example, Zhu WQ and Beroza (2019) trained the PhaseNet algorithm with approximately 700,000 waveform samples from previous earthquakes over 30 years recorded by 889 stations in the Northern California Seismic Network. The authors found that the PhaseNet algorithm could automatically pick up P- and S-phases and exhibited a desirable generalization ability in the phase picking process of regional earthquake events with an epicentral distance within 120 km. This approach has been used to study the induced seismicity of the 2010 Guy-Greenbrier earthquake sequence (Park et al., 2020), the 2019 Ridgecrest, California, earthquake sequence (Liu M et al., 2020), and induced earthquakes recorded by short-period dense arrays (Wang RJ et al., 2020). It has also been used to process real-time data from regional seismic networks (Liao SR et al., 2021).

Whether a deep learning model can be generalized depends on two critical factors: the size of the training dataset and the feature distribution. In Chinese mainland, the crust-mantle velocity structure is complex, the spatial distribution of seismic stations is heterogeneous, and the environmental noise and instrumentation levels vary among stations. These aspects result in prominent differences in the characteristics of earthquake waveforms recorded by stations in different regions. Therefore, after being trained with foreign datasets, the performances of several deep learning models with good generalization abilities and practical effects have degraded to varying degrees when applied in China. E.g., Zhao M and Chen S (2021) tested the generalization ability of the PhaseNet model using data from the Capital Area Seismograph Network and Sichuan Seismograph Network. In addition, Jiang C et al. (2021) compared the earthquake detection performances of PhaseNet and EQTransformer by taking the aftershock sequences of the Yangbi and Maduo earthquakes as examples. Transfer learning can improve the detection performance of an original model in a new area to some extent, as demonstrated by Zhao M et al. (2021), who used PhaseNet to analyze the foreshock sequence of the M_S6.0 Changning earthquake in 2019. However, a key premise for successful transfer learning is that the dataset of the new region is roughly similar to the training dataset of the original model in terms of the statistical feature

distribution, which is difficult to fulfill in many areas of China. Therefore, it is necessary to establish a labeled dataset based on waveform traces from the China Seismograph Network to assist data-driven seismological research and further promote the application of machine learning methods to automated and intelligent seismic data processing in China.

After more than 50 years of construction and development, the China Seismograph Network now consists of more than 1,300 high-standard permanent stations and a professional team for analyzing seismic data. The China Earthquake Administration regularly holds technical training sessions for earthquake scientists and technicians, and conducts random checks and comparisons on seismic observation reports. Thus, the China Seismograph Network has unique advantages in the quantity of accumulated seismic data and the quality control on manually cataloged seismic phase reports. In this study, we compiled the "DiTing" dataset using the 2013–2020 seismic observation reports of the China Earthquake Networks Center and earthquake waveforms downloaded from the Data Backup Center of the China Seismograph Network (http://www.seisdmc.ac.cn) after data cleaning and desensitization. We included the waveform records of regional earthquakes with an epicentral distance within 330 km, mainly labeled with P- and S-phases and P-wave first-motion polarity, in the "DiTing" dataset. The magnitudes of seismic events, epicentral distance, back azimuth, signal-to-noise ratios, and first motion sign of each waveform segment were indexed in a descriptive information file. The primary aim of this study is to provide a standard benchmark for studying and processing Chinese seismic records based on machine learning methods. The dataset could be used to train and test various machine learning models, and then to evaluate model performance in P- and S-phase identification, P-wave first-motion polarity determination, and earthquake magnitude prediction. Compared with the existing STEAD and INSTANCE datasets, "DiTing" contains a larger number of samples, including more seismic records with an epicentral distance of >120 km. Moreover, the magnitude and signal-to-noise ratios in the DiTing dataset are more evenly distributed, thus compensating for the shortcomings of existing public datasets. Thus, the dataset is of great significance for training intelligent seismic data processing models applicable to Chinese mainland.

2. Overview of the DiTing dataset

The sources of data included in the DiTing dataset are shown in Figure 1. The dataset contains 2,734,748 three-component waveform traces corresponding to 787,010 seismic events recorded using broadband and short-period seismometers from 2013 to 2020 at >1,300 permanent stations distributed throughout China. The dataset also includes 2,734,748 labels of P-phases, 2,734,748 labels of S-phases, and 641,025 labels of P-wave first-motion polarities. In the dataset, earthquake magnitudes (99.2% as M_L) range from 0 to 7.7, the epicentral distance ranges from 0 to 330 km, the signal-to-noise ratio of P-wave data is predominantly between −0.05 dB and 5.31 dB, and the signal-to-noise ratio of S-wave data is mainly between −0.05 dB and 4.73 dB.

Figure 2 illustrates the frequency distribution histograms of the epicentral distance, magnitude, and signal-to-noise ratios of P- and S-waves in the DiTing dataset. Figure 2a shows that the number of events decreased as the epicentral distance increased. The epicentral distance distribution range of the DiTing dataset is better than that of the STEAD dataset (92% of labeled data have an epicentral distance within 120 km) and the INSTANCE dataset (the epicentral distance of most data is within 200 km). The magnitude distribution of our dataset follows the Gutenberg-Richter law because most seismic events had a magnitude of 1–2.5 and the number of events declined rapidly with increasing magnitude. In this study, the

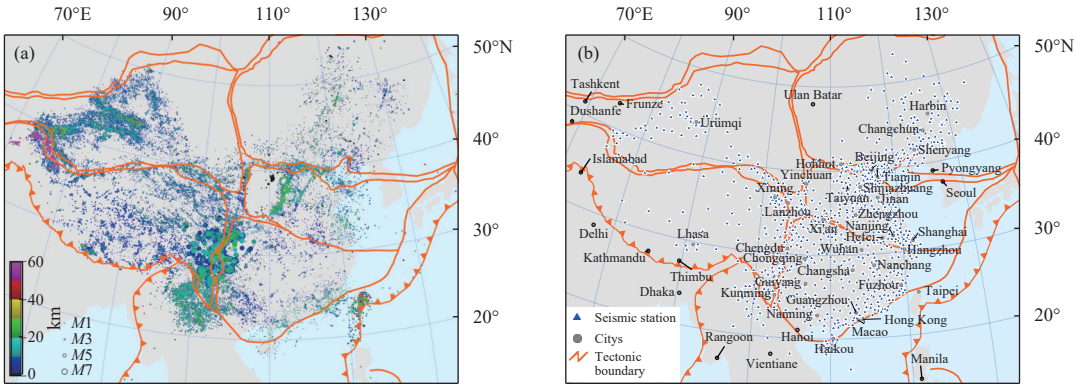


Figure 1. Distribution of the earthquakes (a) and seismic stations (b) included in the DiTing dataset. The dots denote earthquakes, with the color indicating the earthquake depth. The orange line is the tectonic line.

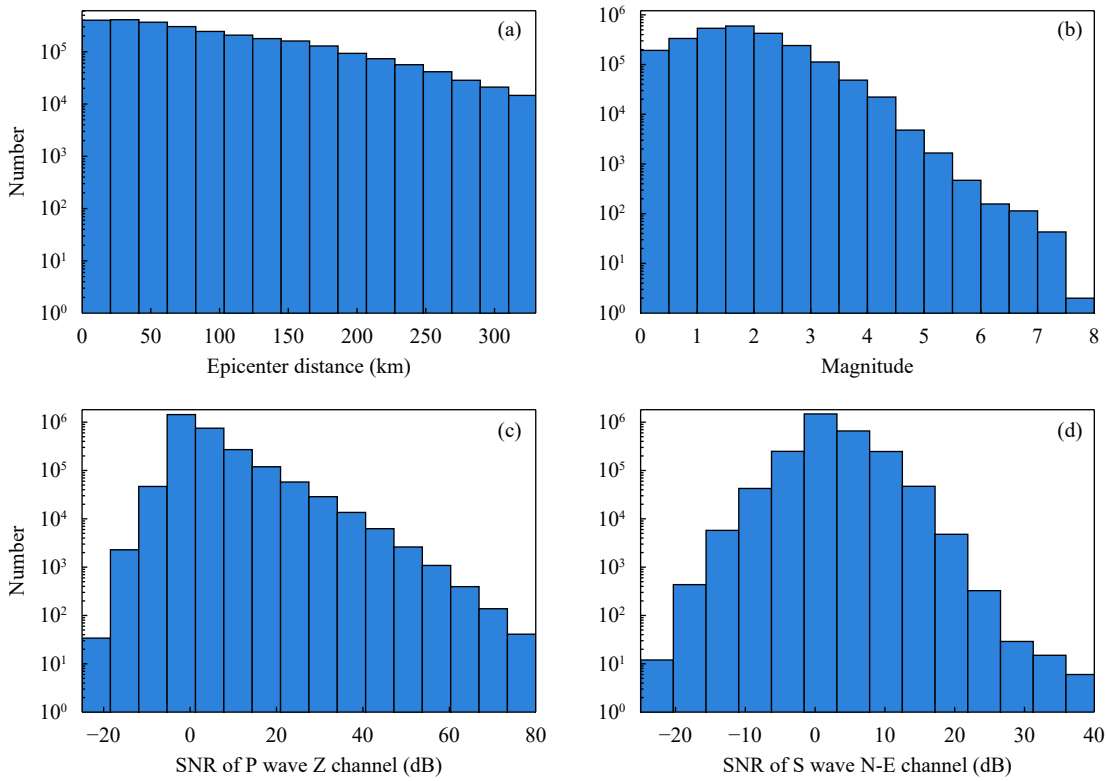


Figure 2. Histograms of the epicentral distance (a), magnitude (b), signal-to-noise ratio of P-waves (c), and signal-to-noise ratio of S-waves (d) in the DiTing dataset.

signal-to-noise ratio was calculated as follows:

$$\mathrm{SNR} = 10 \times \log_{10} \frac{\frac{1}{N}\sum_{i=1}^{N} x_{\mathrm{signal}}[i]^2}{\frac{1}{N}\sum_{i=1}^{N} x_{\mathrm{noise}}[i]^2}$$

The noises of the P- and S-phases are collected within 0.5 s and 1.5 s before their picked-up onsets, respectively, while their signals are recorded within 0.5 s and 1.5 s after their onsets, respectively. Thus, signal-to-noise ratios can effectively indicate the clarity of the labeled P- and S-phases. In addition, unlike the STEAD and INSTANCE datasets, the DiTing dataset contains large amounts of data with low signal-to-noise ratios. The lack of sufficient data with low signal-to-noise ratios in the STEAD and INSTANCE datasets partly explains why some models trained with foreign datasets (e.g., PhaseNet and EQTransformer) have not performed well when tested in certain regions in China. To this end, incorporating the DiTing dataset into the training process of machine learning algorithms will help improve their detection performance for data with low signal-to-noise ratios.

Figures 3a and 3b present the distributions of the back azimuth, first-motion polarity in the DiTing dataset. The distribution of the back azimuth was relatively even over the range of 0–360 owing to the large monitoring area of the China Seismograph Network and the wide distribution of earthquakes and seismic stations. As shown in Figure 3b, 641,025 P-wave onsets were labeled with polarities and divided into six categories according to their first-motion features (I, E, and -) and first-motion polarities (U and D). As shown in Figure 3c, the main magnitude type of previous earthquakes was the local magnitude (M_L), followed by surface wave (M_S) and body-wave magnitudes (m_b and m_B).

3. Dataset construction and evaluation

3.1. Dataset construction process

First, we retrieved the nationwide seismic phase reports of 2013–2020 from the China Earthquake Networks Center (China Earthquake Data Center) to collect the corresponding waveform records. The detailed method for processing the seismic phase reports is described in Dai GH et al. (2019). Then, we screened out the seismic records that fulfilled the following three conditions: (1) the epicentral distance was <330 km; (2) both P- and S-phases were manually labeled; (3) the waveform record was complete and the calculated signal-to-noise ratios did not contain “NaN” or “inf” values. To

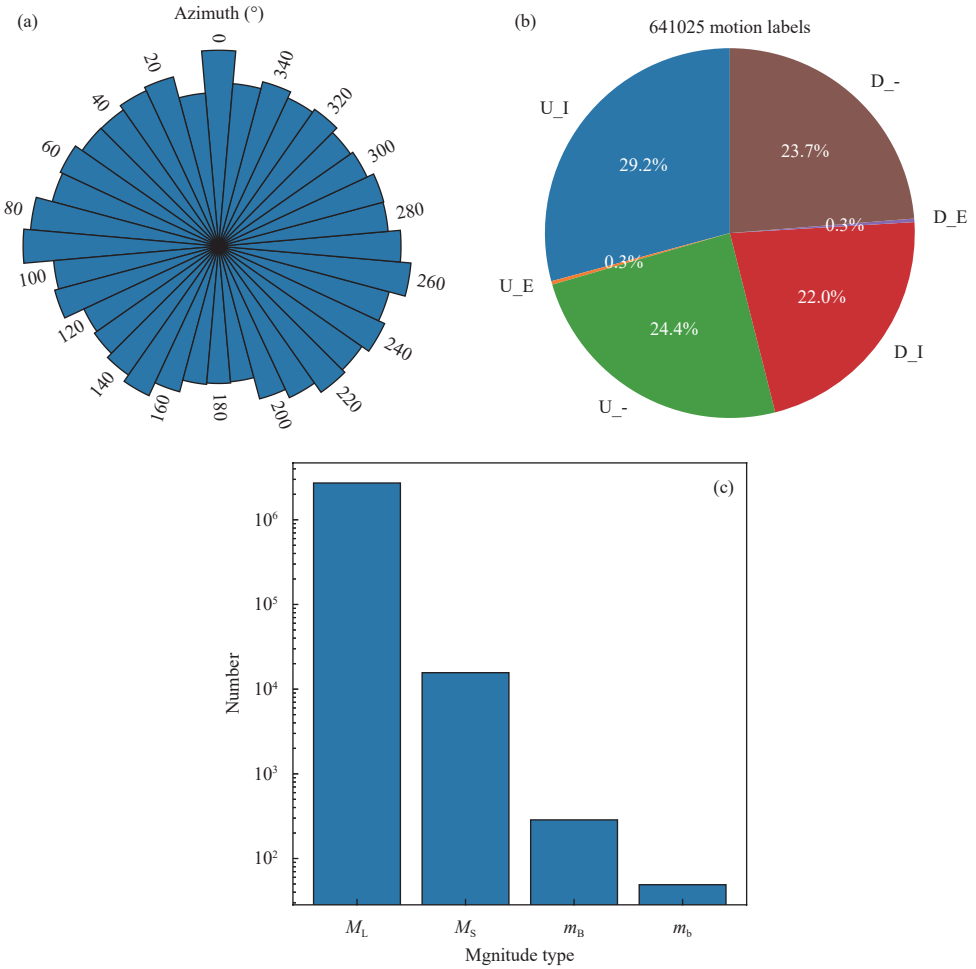


Figure 3. Distributions of back azimuth (a), P-wave first-motion polarity (b) (where "U" represents upward, "D" represents downward, and "I," "E," and "-" represent impluse, emergent, and uncertain first-motion labeling features, respectively), and magnitude type (c) for the DiTing dataset.

ensure data security and facilitate data sharing, we desensitized the following information: (1) the station location information was removed, the name of the seismic network was deleted, and the station name was mapped to a serial number; (2) the earthquake location and time of occurrence were removed and mapped to serial numbers; (3) the downloaded earthquake waveforms were uniformly sampled at 50 Hz; (4) the P- and S-wave arrival times used in the dataset were the number of sampling points relative to the starting time of the waveform segment, and the starting time was randomly selected from a specific time range before the time of earthquake occurrence. As the earthquake occurrence time was removed, the absolute P- and S-wave arrival times were also deleted.

3.2. Introduction to the metadata of the dataset

Owing to the large data volume, the entire dataset was stored in multiple hdf5 files. The metainformation of the dataset was stored in a single csv file containing the following terms to refer to different information: (1) "part" refers to the corresponding hdf5 file number; (2) "key" refers to the index of the waveform data, comprising the seismic event number and station number; (3) "ev_id" refers to the desensitized serial number of the earthquake corresponding to the waveform file; (4) "ev_mag" refers to the magnitude of a seismic event; (5) "mag_type" refers to the corresponding type of seismic magnitude; (6) "p_pick" refers to the P-wave arrival time; (7) "p_sharpness" refers to the sharpness of the P-wave first motion; (8) "p_motion" refers to the P-wave first-motion polarity; (9) "s_pick" refers to the S-wave arrival time; (10) "net_id" refers to the serial number of the seismic network, uniformly desensitized as "AA"; (11) "sta_id" refers to the desensitized serial number of the station; (12) "dis" refers to the epicentral distance in the waveform file; (13) "st_mag"refers to the single-station magnitude in the waveform file; (14) "baz"refers to the back azimuth in the waveform file; (15) "Z_P_amplitude_snr",

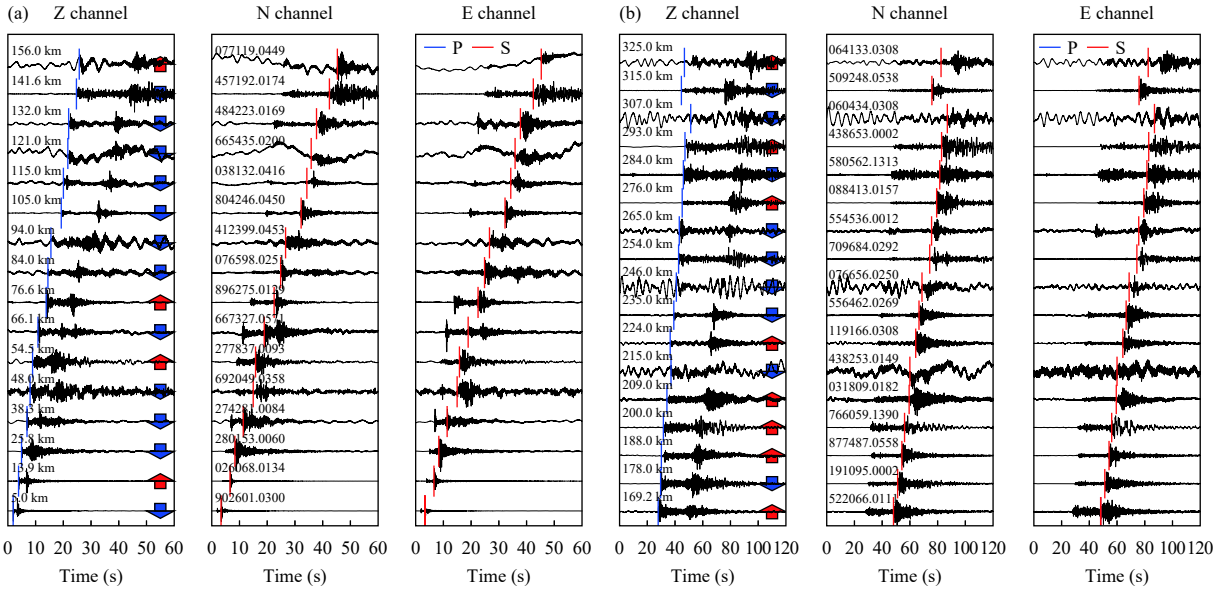


Figure 4. Examples of labeled waveforms at epicentral distances of (a) 0–160 km and (b) 160–330 km. The blue lines are labels of P-wave arrival times; the red lines are labels of S-wave arrival times; the upward and downward arrows denote the "positive" and "negative" P-wave first-motion polarities, respectively; the waveforms are arranged by epicentral distance; the serial number is in the "seismic event number.station number" format, which is the unique number of the waveform in the DiTing dataset.

"Z_P_power_snr", "N_S_amplitude_snr", "N_S_power_snr", "E_S_amplitude_snr", and "E_S_power_snr" refer to the signal-to-noise ratios of P- and S-waves calculated by amplitude and power corresponding to the waveforms recorded on Z, N, and E channels, respectively; (16) "P_residual" refers to the residual of the manually labeled P-wave arrival time against the theoretical arrival time; (17) "S_residual" refers to the residual of the manually labeled S-wave arrival time against the theoretical arrival time.

3.3. Examples of labeled waveforms

Figure 4 shows three-component waveforms with different epicentral distances that were randomly selected from the DiTing dataset at epicentral distance intervals of approximately 10 km. These were labeled with their respective P- and S-wave arrival times and P-wave first-motion polarities. Each waveform was assigned with an ID in the "seismic event number.station number" format; for example, the serial number "064122.0308" indicates that the waveform was from seismic event 064122 recorded at station 0308. The IDs correspond to the "key" values in Table 1 and are convenient for information retrieval.

3.4. Benchmark test for the dataset

To further evaluate the quality of the DiTing dataset and its importance for improving the generalization abilities of deep learning algorithms for Chinese data, we conducted benchmark tests for the dataset using the EQTransformer and PhaseNet algorithms (Figure 5). Then, we randomly inspected the data with relatively large differences between the results of automatic identification and manual labeling (Figure 6). As EQTransformer and PhaseNet were trained with data sampled at 100 Hz, the data used in the benchmark tests were resampled at 100 Hz. For the P- and S-wave data, the test results with manual labeling and automatic detection errors within 0.5 s and 1 s were recorded as true positives (TPs). If there was no automatic detection result within the corresponding time range, the result was recorded as false negatives (FNs). Other detection results that were within the 10 s time window around the manual labels and did not fulfill the TP requirements were recorded as false positives (FPs). Figure 5 shows that both models maintained high precision for the Chinese dataset, verifying the reliability of the labeling in the DiTing dataset. However, all the recall rates were relatively low, which may have been mainly caused by the differences in their training data, especially the lack of data from China. It is worth noting that the definitions of a TP and FP in the present study differed slightly from those in the original literature of EQTransformer and PhaseNet. These test indices were only used to analyze the quality and necessity of our dataset and do not reflect the merits of the two methods. This situation was also observed by Jiang C et al. (2021) when constructing the

Table 1. Example of the metadata of the DiTing Dataset.

part	key	ev_id	ev_mag	mag_type	p_pick	p_sharpness	p_motion	s_pick	net_id	sta_id
0	000002.0006	2	1	M_L	1.54	I	R	2.64	AA	6

dis	st_mag	baz	Z_P_amplitude_snr	Z_P_power_snr	N_S_amplitude_snr	N_S_power_snr	E_S_amplitude_snr	E_S_power_snr	P_residual	S_residual
3.7	0.5	256.1	16.712	16.508	0.359	−3.852	0.359	−3.852	−0.01	0.03

aftershock sequences of the Yangbi and Maduo earthquakes using EQTransformer and PhaseNet. As shown in the random inspection results (Figure 6), most of the false and missing identifications were generated by the judgement errors of the deep learning models because the training datasets of the two models did not contain any data (or only included a small amount of data) for China. Therefore, the DiTing dataset is of great significance for enriching the diversity of the current artificial intelligence dataset in seismology, and improving the generalization abilities of intelligent processing models for seismic data from Chinese mainland.

3.5. Advantages of the DiTing dataset

We compared the DiTing dataset with other publicly available datasets in terms of the earthquake seismogram, noise seismogram, epicentral distance, waveform length, collection region, and first-motion polarity data. The detailed comparison results are listed in Table 2, which shows that the DiTing dataset ranks second only to the SCEDC dataset with respect to the earthquake seismogram data, whereas it ranks above the SCEDC dataset with respect to epicentral distance. Among the considered datasets, P-phase, S-phase, and first-motion polarity labels are only included in the DiTing dataset, which also has the longest waveform length (180 s). Thus, the DiTing dataset can be used to train various machine learning models. In contrast, the SCEDC dataset can only be used to train P-wave arrival picking and first-motion classification because its waveform length is only 6 s. In terms of epicentral distance, only the INSTANCE and NEIC datasets have larger data collection ranges than the DiTing dataset. However, if the sample size is not large enough, a wide sample distribution range may not be an advantage for machine learning. Overall, the comparison indicated that the DiTing dataset is better than other publicly available datasets in terms of either data size or versatility, except for the fact that it lacks a separate noise seismogram category (in fact, the 180-s-long waveform already contains a considerable amount of noise).

Another critical aspect is the labeling quality of the dataset, which is not easy to compare among different datasets due to the strong subjectivity of manual labeling. The labels in the DiTing dataset were obtained based on the seismic phase reports from the China Earthquake

Table 2. Comparisons of the DiTing dataset with other publicly available datasets.

Dataset	Earthquake seismogram (×10^6)	Noise seismogram (×10^6)	Epicentral distance	Waveform length (s)	Region	First-motion Polarity
DiTing-330 km (this study)	2.74	0	0–3°	180	China	√
STEAD (Mousavi et al., 2019)	1.05	0.10	0–3°	60	Global	×
INSTANCE (Michelini et al., 2021)	1.20	0.13	0–6°	120	Italy	√
LEN-DB (Magrini et al., 2020)	0.63	0.62	0–1°	27	Global	×
NEIC (Yeck et al., 2021)	1.30	0	0–180°	60	Global	×
SCEDC-Phase (Ross et al., 2018a)	3.50	1.50	0–1°	6	U.S.	×
SCEDC-Motion (Ross et al., 2018b)	2.53	2.32	0–1°	6	U.S.	√

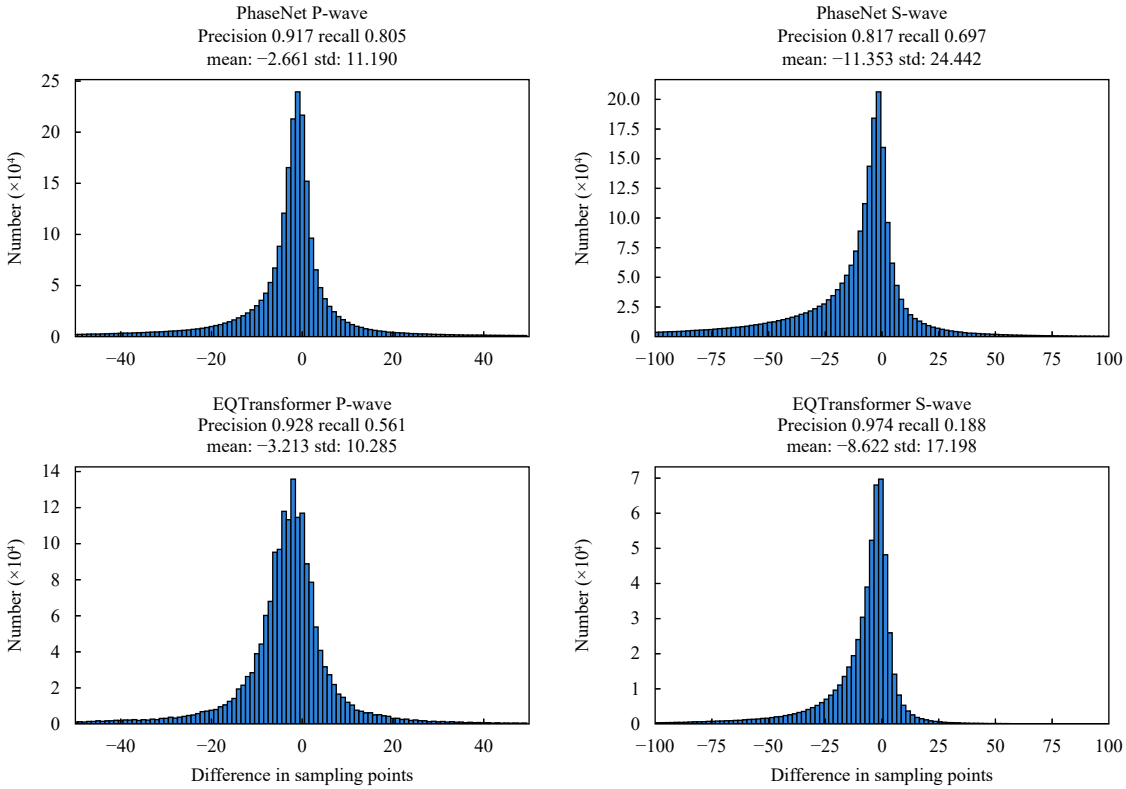


Figure 5. Results of benchmark tests for the DiTing dataset using the EQTransformer and PhaseNet algorithms. The horizontal axis is the difference between manually labeled sampling points and the sampling points picked up by the deep learning model.

Networks Center. Traveltime residuals against the theoretical traveltimes are calculated for the arrival times recorded at each station involved in recording earthquakes. This proportion of arrival time records accounts for 91% of the entire dataset. With the triple verification of manual picking, manual reviewing, and theoretical traveltime residual calculation stages, the labeling quality of the DiTing dataset can be generally considered more reliable. In comparison, the P- and S-phase arrival times in the STEAD and SCEDC datasets were labeled both manually and automatically. Figure 7 displays the histograms of the theoretical traveltime residuals for the manually picked P- and S-wave arrival times in the DiTing dataset, in which 76.04% of the P-wave data were observed to have residuals within ±0.5 s and S-wave data accounted for 61.16%.

4. Dataset acquisition method and user guide

Researchers interested in the DiTing dataset can visit the Earthquake Science website (https://www.equsci.org.cn) for further information. The DiTing dataset is shared online through the website of the China Earthquake Data Center (https://data.earthquake.cn), allowing users to retrieve the metadata information of the dataset online. To obtain the complete dataset, users need to register online, upgrade to an advanced user, and sign the "Data Sharing Agreements of the China Earthquake Data Center" before accessing the data. We prepared a Jupyter notebook manual named "ReadDiTingExample_50Hz.ipynb" in the dataset, which provides guidance on how to read the waveform and metadata information in the dataset and plot data. It can also be downloaded from: https://github.com/mingzhaochina/readDiTingExample. When using this dataset for publishable research, users must cite the data source. In addition, to help us keep track of the application of this dataset, users are requested to provide timely feedback on their results.

5. Discussion and conclusions

In this study, we compiled the DiTing dataset, which can be used in the supervised learning of phase-picking algorithms, training and validating the determination of first-motion polarity, and comparing model performance. The unique advantages of this dataset over the existing ones are as follows.

(1) Large data size: Compared with the STEAD dataset, which is the largest existing dataset of the same

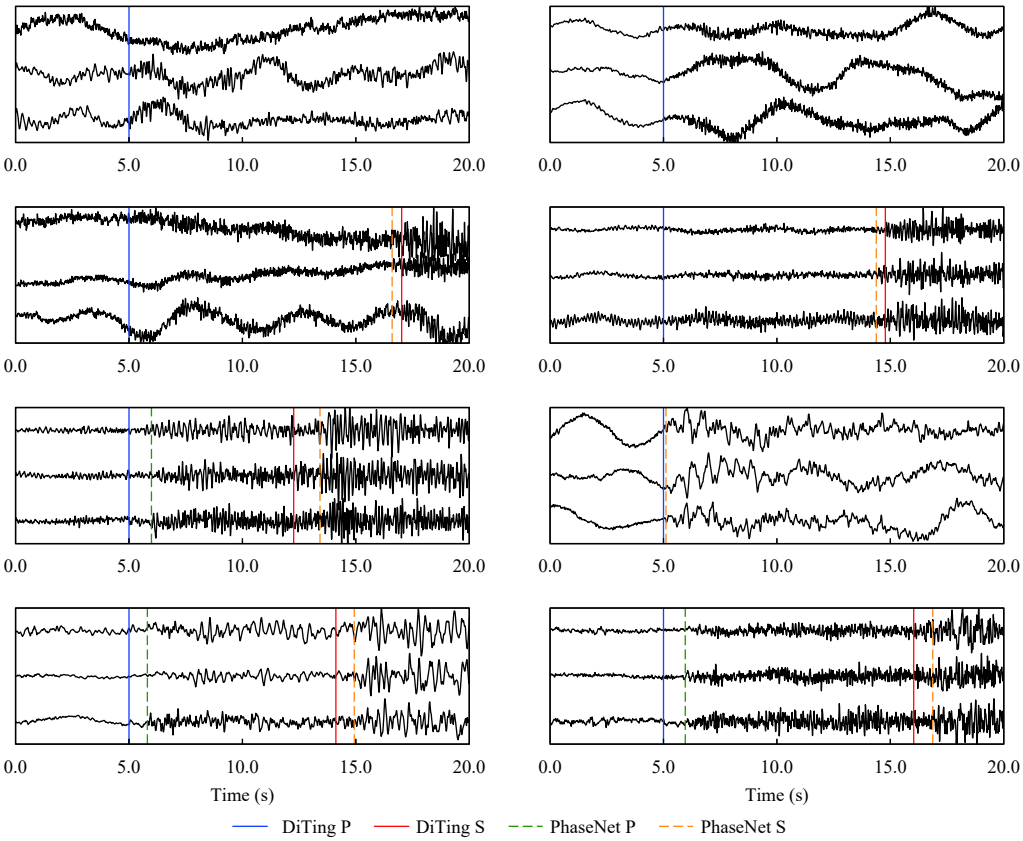


Figure 6. Examples of randomly inspected samples with large differences between the manual labeling of the DiTing dataset and the automatic detection results of PhaseNet.

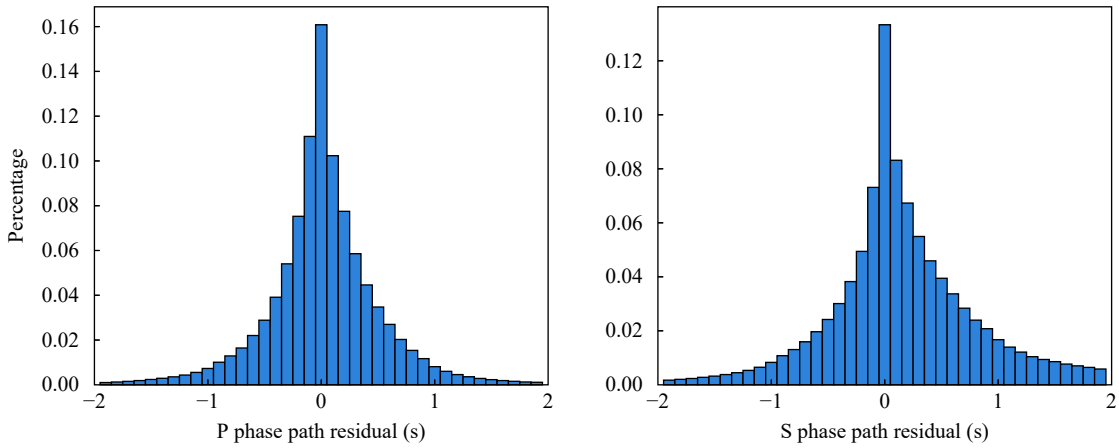


Figure 7. Histograms of the theoretical traveltime residuals for manually picked P- and S-wave arrival times.

type, the DiTing dataset contains ~3 times more manually labeled earthquake records.

(2) Reliable labels: The labels in the DiTing dataset were obtained based on the seismic phase reports of the China Earthquake Networks Center (China Earthquake Data Center). As 91% of the seismic phases were involved in detecting previous earthquakes and calculations were performed using theoretical traveltime residuals, the labels are highly reliable.

(3) Distinct waveform recording area: The waveforms included in the DiTing dataset were recorded by stations belonging to the China Seismograph Network and distributed within Chinese territory, thus compensating for the shortcoming of existing datasets that mainly include waveform records from other countries. This aspect is of great significance for training seismic event detection and phase identification algorithms applicable to China.

(4) Desirable epicentral distance coverage: The epicentral distance range of the DiTing dataset covers 0–330 km, including a large amount of data in the range of 110–330

km. This coverage is important for further expanding the applicable epicentral distance range of seismic event detection and phase identification algorithms.

(5) Wide application range: The DiTing dataset includes not only arrival time information but also many first-motion polarity labels, which can be used for training seismic phase identification algorithms and determining first-motion polarity.

In summary, the DiTing dataset serves as a high-quality benchmark for developing artificial intelligence-based earthquake detection techniques, picking seismic phases, determining first-motion polarity, predicting earthquake magnitudes, implementing early earthquake warning systems, and developing algorithms for strong ground-motion simulation in China, especially in key monitoring areas such as the China Seismic Experimental Site. The dataset provides basic information for data-driven seismological research, which will further promote the application of artificial intelligence methods in seismology and advance the level of intelligent seismic data processing in China.

The current DiTing dataset only includes the first arrival times of the P- and S-wave phases with epicentral distances of up to 330 km. However, for most areas in China, various phases, such as Pn, Pg, Sn, and Sg, are observed when the epicentral distance exceeds 200 km. Accordingly, in the future, an updated version of the dataset will be prepared to expand the epicentral distance range and include information on more types of seismic phases.

Acknowledgements

This research was jointly funded by the National Natural Science Foundation of China (Nos. 41804047 and 42111540260), Fundamental Research Funds of the Institute of Geophysics, China Earthquake Administration (NO. DQJB19A0114); and the Key Research Program of the Institute of Geology and Geophysics, Chinese Academy of Sciences (No. IGGCAS-201904). We would like to thank the Data Backup Center of the China Seismograph Network and China Earthquake Networks Center (China Earthquake Data Center) for providing data for this study. We would like to thank the Beijing Baijiatuan Earth Sciences National Observation and Research Station for providing the EarthStack server cluster. We are grateful to Libo Han, Shuguang Wang, and Leiyu Mou at the Institute of Geophysics, China Earthquake Administration, for their assistance and support in obtaining seismic waveform data from permanent stations throughout China. We would also like to thank Dr. Bei Zhang, Dr. Jiancheng Han, and Dr. Wen Shi for their technical support, and Guofeng Zhao for his help in checking and sharing data.

Author contributions

Ming Zhao collected and preprocessed the data used in this study and wrote the first draft of this paper; Zhuowei Xiao produced the dataset and jointly contributed to writing the first draft; Shi Chen and Lihua Fang revised and proofread the paper.

References

Dai GH, Miao CL and Zhai LY (2019). Unified earthquake cataloging of China seismographic network. Earthq Res China **35**(1): 192–203 (in Chinese with English abstract).

Jiang C, Fang LH, Fan LP and Li BR (2021). Comparison of the earthquake detection abilities of PhaseNet and EQTransformer with the Yangbi and Maduo earthquakes. Earthq Sci **34** (5): 425–435 https://doi.org/10.29382/eqs-2021-0038.

Liao SR, Zhang HC, Fan LP, Li BR, Huang LZ, Fang LH and Qin M (2021). Development of a real-time intelligent seismic processing system and its application in the 2021 Yunnan Yangbi M_S 6.4 earthquake. Chin J Geophys **64**(10): 3632–3645 https://doi.org/10.6038/cjg2021O0532 (in Chinese with English abstract).

Liu M, Zhang M, Zhu WQ, Ellsworth WL and Li HY (2020). Rapid characterization of the July 2019 Ridgecrest, California, earthquake sequence from raw seismic data using machine-learning phase picker. Geophys Res Lett **47**(4): e2019GL086189 https://doi.org/10.1029/2019GL086189.

Magrini F, Jozinović D, Cammarano F, Michelini A and Boschi L (2020). Local earthquakes detection: a benchmark dataset of 3-component seismograms built on a global scale. Artif Intell Geosci **1**: 1–10 https://doi.org/10.1016/j.aiig.2020.04.001.

Michelini A, Cianetti S, Gaviano S, Giunchi C, Jozinović D and Lauciani V (2021). INSTANCE–the Italian seismic dataset for machine learning. Earth Syst Sci Data **13**(12): 5509–5544 https://doi.org/10.5194/essd-13-5509-2021.

Mousavi SM, Sheng YX, Zhu WQ and Beroza GC (2019). 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.

Mousavi SM, Ellsworth WL, Zhu WQ, Chuang LY and Beroza GC (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nat Commun **11**(1): 3952 https://doi.org/10.1038/s41467-020-17591-w.

Park Y, Mousavi SM, Zhu WQ, Ellsworth WL and Beroza GC (2020). Machine-learning-based analysis of the Guy-

Greenbrier, Arkansas earthquakes: a tale of two sequences. Geophys Res Lett **47**(6): e2020GL087032.

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

Ross ZE, Meier MA, Hauksson E and Heaton TH (2018a). Generalized seismic phase detection with deep learning. Bull Seismol Soc Am **108**(5A): 2894–2901 https://doi.org/10.1785/0120180080.

Ross ZE, Meier MA and Hauksson E (2018b). P wave arrival picking and first-motion polarity determination with deep learning. J Geophys Res: Solid Earth **123**(6): 5120–5129 https://doi.org/10.1029/2017JB015251.

Wang J, Xiao ZW, Liu C, Zhao DP and Yao ZX (2019). Deep learning for picking seismic arrival times. J Geophys Res:Solid Earth **124**(7): 6612–6624 https://doi.org/10.1029/2019JB017536.

Wang RJ, Schmandt B, Zhang M, Glasgow M, Kiser E, Rysanek S and Stairs R (2020). Injection-induced earthquakes on complex fault zones of the Raton Basin illuminated by machine-learning phase picker and dense nodal array. Geophys Res Lett **47**(14): e2020GL088168.

Xiao ZW, Wang J, Liu C, Li J, Zhao L and Yao ZX (2021). Siamese Earthquake transformer: a pair-input deep-learning model for earthquake detection and phase picking on a seismic array. J Geophys Res: Solid Earth **126**(5): e2020JB021444.

Yeck WL, Patton JM, Ross ZE, Hayes GP, Guy MR, Ambruz NB, Shelly DR, Benz HM and Earle PS (2021). Leveraging deep learning in global 24/7 real-time earthquake monitoring at the national earthquake information center. Seismol Res Lett **92**(1): 469–480 https://doi.org/10.1785/0220200178.

Zhao M, Chen S and Dave Y (2019a). Waveform classification and seismic recognition by convolution neural network. Chin J Geophys **62**(1): 374–382 https://doi.org/10.6038/cjg2019M0151 (in Chinese with English abstract).

Zhao M, Chen S, Fang LH and David AY (2019b). Earthquake phase arrival auto-picking based on U-shaped convolutional neural network. Chin J Geophys **62**(8): 3034–3042 https://doi.org/10.6038/cjg2019M0495 (in Chinese with English abstract).

Zhao M and Chen S (2021). The generalization ability research of deep learning algorithm in seismic phase detection of regional seismic network. Earthquake **41**(1): 166–179 (in Chinese with English abstract).

Zhao M, Tang L, Chen S, Su JR and Zhang M (2021). Machine learning based automatic foreshock catalog building for the 2019 M_S6.0 Changning, Sichuan earthquake. Chin J Geophys **64**(1): 54–66 https://doi.org/10.6038/cjg2021O0271 (in Chinese with English abstract).

Zhou BW, Fan LP, Zhang L, Li BR and Fang LH (2020). Earthquake detection using convolutional neural network and its optimization. Acta Seismol Sin **42**(6): 669–683 https://doi.org/10.11939/jass.20200045 (in Chinese with English abstract).

Zhou YJ, Yue H, Kong QK and Zhou SY (2019). Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks. Seismol Res Lett **90**(3): 1079–1087 https://doi.org/10.1785/0220180319.

Zhu LJ, Peng ZG, McClellan J, Li CY, Yao DD, Li ZF and Fang LH (2019). Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M_W 7.9 Wenchuan Earthquake. Phys Earth Planet Inter **293**: 106261 https://doi.org/10.1016/j.pepi.2019.05.004.

Zhu WQ and Beroza GC (2019). PhaseNet: a deep-neural-network-based seismic arrival-time picking method. Geophys J Int **216**(1): 261–273.