Geoscience Letters

RESEARCH LETTER

Open Access

Multimodal deep learning network for fast seismic discrimination and magnitude classification

Jingbao Zhu[1,2], Shanyou Li[1,2] and Jindong Song[1,2*]

Abstract

Quickly and accurately identifying seismic events from background noise and classifying earthquake magnitudes is extremely important for improving the performance of earthquake early warning (EEW) systems. Microtremors are weak nonearthquake-induced vibrations that may trigger EEW systems, leading to false alarms and causing unnecessary public concern. Moreover, quickly predicting whether an earthquake event is of low or high magnitude is important for EEW systems to determine the potential earthquake damage and the alert area. Here, we develop a multimodal deep learning network (MDLNet) that can identify seismic events while determining whether an earthquake is of low magnitude ($M < 5.5$) or high magnitude ($M \geq 5.5$). MDLNet can handle multimodal data and uses time-domain and spectrum encoders to extract features. Then, the features extracted by these encoders are fused with ground-motion parameter data. We train MDLNet using multimodal data from seismic event signals and microtremor signals recorded by the Japanese Kyoshin Network. We demonstrate that using data from the 3-s period following the onset of *P*-waves, MDLNet can recognize 99.92% of microtremor signals, 96.65% of low-magnitude seismic signals and 90.68% of high-magnitude seismic signals, values higher than those for single-mode data. Multimodal deep learning techniques can provide new insight into seismology and EEW.

Keywords Earthquake, Multimodal, Deep learning, Seismic discrimination, Magnitude classification, Earthquake early warning

Introduction

Earthquake event identification and magnitude classification are fundamental tasks in seismic monitoring and earthquake early warnings (EEW) systems. Through the meticulous analysis and processing of seismic data, the prompt detection and issuance of alerts serve as pivotal mechanisms that can significantly mitigate casualties and property losses (Allen and Melgar 2019). Microtremors are subtle vibrations induced by nonseismic sources, and those proximate to seismic stations have the potential to trigger EEW systems, resulting in false alarms and unwarranted public apprehension (Liu et al. 2022). Moreover, for earthquakes with high magnitudes, an alert includes an estimated line (finite-fault) source (Kohler et al. 2020; Li et al. 2021). Rapid determination of whether an earthquake is of low or high magnitude can help us accurately determine the type, damage, and alert area of the earthquake, and effective warning measures and response strategies can be adopted (Li et al. 2021; Mittal et al. 2022; Zollo et al. 2010). When an earthquake occurs, based on signals recorded by stations within a finite timeframe, EEW systems need to promptly and accurately identify seismic events and classify magnitudes for alert dissemination

*Correspondence:
Jindong Song
jdsong@iem.ac.cn
[1] Key Laboratory of Earthquake Engineering and Engineering Vibration, Institute of Engineering Mechanics, China Earthquake Administration, Harbin 150080, China
[2] Key Laboratory of Earthquake Disaster Mitigation, Ministry of Emergency Management, Harbin 150080, China

(Kohler et al. 2020). Due to the limited seismic data recorded by stations within a matter of seconds, there exists the potential for missed alarms in cases with large earthquakes, leading to extensive casualties and economic losses (Minson et al. 2018). Therefore, the development of a method capable of comprehensively leveraging available data and swiftly and reliably identifying seismic events and classifying magnitudes is particularly critical for EEW systems.

Seismic event detection typically involves the identification of earthquake *P*-waves through the extraction of certain peak amplitudes and dominant frequencies from seismic waveforms (Allen 1978; Bose et al. 2009). Moreover, magnitude estimation usually involves establishing empirical magnitude prediction equations using certain ground-motion parameters related to amplitude and frequency obtained from *P*-waves (Kanamori 2005; Wu and Zhao 2006; Wu et al. 2023). Given the intricacies of seismic waveforms and the timeliness requirements of EEW systems, it is difficult to discern all features characterizing seismic events and predict magnitudes based on brief-duration *P*-waves. Consequently, these methods often manifest instances of both false alarms and missed alarms (Cochran et al. 2017). In the wake of advancements in computer science, a quest has emerged to harness artificial intelligence methodologies for the processing of attainable data, seismic event detection and magnitude estimation. Several studies have employed ground-motion parameter data extracted from seismic waveforms as inputs for machine learning models, facilitating magnitude estimation and discrimination between seismic and nonseismic signals (Kong et al. 2016; Zhu et al. 2022). Meier et al. (2019) employed data for 25 ground-motion parameters as inputs for a machine learning classifier and assessed the performance of diverse machine learning approaches in the task of distinguishing seismic signals from noise. Furthermore, some scholars have employed deep neural networks to identify features from time-domain seismic waveform data for seismic discrimination and magnitude estimation (Chen et al. 2019; Mousavi and Beroza 2020; Mousavi et al. 2020). Li et al. (2018) leveraged generative adversarial networks to extract features from time-domain seismic waveform data, employing a random forest classifier for the classification of seismic *P*-waves and impulse noise. Perol et al. (2018), based on inputs from time-domain seismic waveform data, utilized a convolutional neural network for the detection and localization of seismic events. Before inputting time-domain seismic waveforms into deep learning network, some studies typically normalize them for efficient training (Perol et al. 2018), which may remove amplitude information from the seismic waveforms. To address this problem, Lomax et al. (2019) proposed the ConvNetQuake–INGV model, which not only takes seismic waveforms as model inputs, but also inputs amplitude information. Simultaneously, certain scholars have applied deep learning networks to extract features within seismic spectrum data for the identification of seismic events (Njirjak et al. 2022; Trani et al. 2022). Linville et al. (2019) employed spectrum data recorded by stations as inputs for convolutional and recurrent neural networks, distinguishing between seismic events and explosions. Presently, the efficacy of deep learning methodologies in seismic event identification is primarily limited by the processing of single-mode data (such as time-domain data, spectrum data, and ground-motion parameters), and the potential to enhance the accuracy of seismic event identification and magnitude classification through the comprehensive utilization of available multimodal data remains largely untapped.

Maximizing the performance of deep learning models through the comprehensive utilization of multimodal data is the focus of various types of contemporary research (Baltrusaitis et al. 2018; Guzhov et al. 2022; Ngiam et al. 2011; Soleymani et al. 2011; Vinker et al. 2022). To capture more valuable and holistic information to solve intricate problems and perform various tasks, multimodal deep learning methods predominantly employ encoders to extract insights from diverse modal data, facilitating feature fusion to achieve synergistic information integration. The contrastive language-image pretraining (CLIP) model is currently one of the most commonly used multimodal models (Radford et al. 2021). Anchored in multimodal data training, the CLIP model employs a contrastive learning paradigm to discern the intricate relationship between images and text, thereby elevating its prowess in feature extraction. In the realm of medicine, Tiulpin et al. (2019) introduced a multimodal machine learning-based model for predicting the progression of knee osteoarthritis by intricately integrating raw radiological data, clinical examination results, and patients' medical histories. In addition, in the domain of meteorology, Boussioux et al. (2022) proposed a multimodal framework called Hurricast, which employs an encoder–decoder architecture to extract spatiotemporal data along with statistical information, thus enhancing predictions related to the intensity and trajectory of tropical cyclones.

Drawing inspiration from the multimodal machine learning model proposed by Tiulpin et al. (2019), the Hurricast multimodal framework presented by Boussioux et al. (2022), and the ConvNetQuake_INGV model proposed by Lomax et al. (2019), to integrate multimodal data (such as time-domain data, spectrum data, and ground-motion parameters) and improve the

accuracy of seismic event identification and magnitude classification, we develop a multimodal deep learning network (MDLNet) for seismic event identification and magnitude classification, and explore the feasibility of this method, using information from a short window around the *P*-wave arrival and relying on data from only a single seismometer. The tasks of earthquake event identification and magnitude classification are combined in a three-classification problem, and a multimodal deep learning network (MDLNet) that can identify seismic events while determining whether an earthquake is of low magnitude ($M<5.5$) or high magnitude ($M\geq5.5$) is established. Unlike the ConvNetQuake–INGV model proposed by Lomax et al. (2019), MDLNet employs time domain and spectrum encoders to extract features separately from seismic time-domain data and spectrum data. Subsequently, the features extracted by the encoders are fused with ground-motion parameter data (including amplitude information, energy information, and frequency information). A multilayer perceptron is then employed to recognize microtremor signals, low-magnitude ($M<5.5$) seismic signals, and high-magnitude ($M\geq5.5$) seismic signals. An extensive array of multimodal data from Japan's Kyoshin network (*K*-NET) is used for the training of MDLNet. Our investigation demonstrates that the multimodal deep learning method proposed in this paper can enhance the performance of seismic event identification and magnitude classification in EEW systems.

Data

In this work, to enable MDLNet to simultaneously identify earthquake events and classify earthquake magnitudes, the seismic signals are delineated into signals from low-magnitude seismic events and those stemming from high-magnitude seismic events. Previous researches (Böse et al. 2012; Kohler et al. 2020; Li et al. 2021) considered the rupture of earthquake events with $M<5.5$ as a point source with low magnitude, and the rupture of earthquake events with $M\geq5.5$ as a line source with high magnitude. Therefore, in this work, we define a magnitude of 5.5 as the boundary for both low-magnitude and high-magnitude seismic events. In addition, the data set utilized in this work predominantly comprises three categories of signals: low-magnitude ($M<5.5$) seismic signals, high-magnitude ($M\geq5.5$) seismic signals and subtle nonseismic microtremor signals (noise). Moreover, we selected earthquakes with magnitudes ranging from 3 to 8 and focal depths within 30 km recorded by the Japanese *K*-NET network (Aoi et al. 2011) from 2007 to 2017 (Table S1) (Huang et al. 2015). We did not impose restrictions on the epicentral distance or signal-to-noise ratio (SNR) for the strong-motion records. Finally, we have collected a total of 2774 earthquake events, including 81,819 three-component strong-motion records. The sampling rate for the strong-motion waveforms is 100 Hz. We employed the algorithm proposed by Allen (1978) to automatically determine the arrival times of *P*-waves, followed by meticulous manual verification of the results. Meanwhile, in this study, we extracted the data before the arrival of *P*-waves as the microtremor data (Liu et al. 2022). Figure 1a, b illustrates the spatial distribution of the seismic events and the *K*-NET stations utilized in this study, respectively.

We preprocessed the data set through the following steps: (1) removal of the mean value; (2) application of a fourth-order high-pass Butterworth filter with 0.075 Hz; (3) segmentation of waveforms from 1 s before the arrival of the *P*-wave to 3 s after the *P*-wave arrival, constituting the time-domain data for a seismic event; (4) random selection of 4 s waveforms starting from the onset of recording to the *P*-wave onset, constituting the microtremor time-domain data (Liu et al. 2022); and (5) Fourier spectrum data for seismic events and microtremors based on the amplitude derived from the fast Fourier transformation of seismic event time-domain data and microtremor time-domain data, respectively. Subsequently, we acquired 81,819 three-component seismic event time-domain data, 81,819 three-component seismic event spectrum data, 78,039 three-component microtremor time-domain data, and 78,039 three-component microtremor spectrum data. Figures S1 and S2 present examples of time-domain and spectrum data, respectively. Concurrently, from the 4 s waveform segments, we extracted nine ground-motion parameters, which are seismologically motivated and hitherto applied in discriminating seismic events, estimating magnitudes, and assessing seismic damage (Meier et al. 2019; Zhu et al. 2022, 2023; Zollo et al. 2010). These nine parameters encompass three amplitude parameters: peak displacement, peak velocity, and peak acceleration. Correspondingly, the three energy parameters include the squared velocity integral, cumulative absolute velocity, and Arias intensity. In addition, three frequency parameters, namely, the average period, peak ratio, and product parameter, are used. Detailed descriptions of these ground-motion parameters can be found in the literature and are not provided here. Furthermore, Text S1 provides descriptions of these nine ground-motion parameters. We collectively amassed 27,273 seismic event ground-motion parameter data and 26,013 microtremor ground-motion parameter data. Each ground-motion parameter data are a one-dimensional vector composed of nine ground-motion parameters. In this work, the microtremor signals are labeled 0, seismic signals from

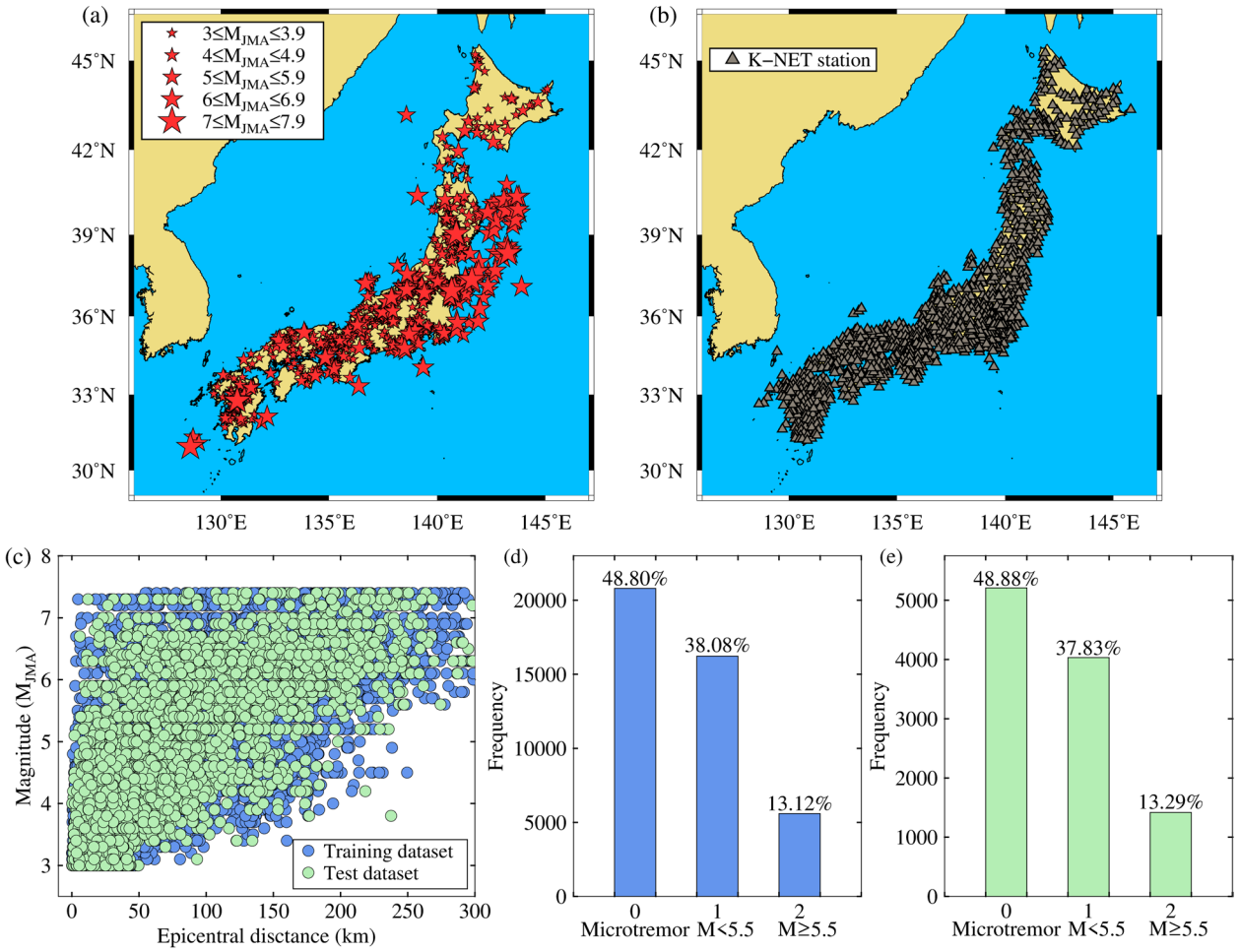

Fig. 1 **a** Spatial distribution of the seismic events utilized in this study. **b** Spatial distribution of the *K*-NET stations utilized in this study. **c** Relationship between the magnitude of earthquake events and epicentral distance. **d** Data distribution of different labels in the training data set. **e** Data distribution of different labels in the test data set

low-magnitude ($M < 5.5$) seismic events are labeled 1, and seismic signals from high-magnitude ($M \geq 5.5$) seismic events are labeled 2. In this study, according to previous studies (Mousavi and Beroza 2020), we randomly divided the data set into an independent training data set (80%) and a test data set (20%) by setting "random state = 30" (Pedregosa et al. 2011). Figure 1c delineates the relationship between the seismic event magnitude and epicentral distance for data from the training and test data sets. Figure 1d, e shows the distributions of various signals within the training and test data sets, respectively.

Methods

Multimodal deep learning network

Time-domain seismic waves and their spectrum and ground-motion parameters are not completely independent. Although time-domain signals and their Fourier transforms contain essentially equivalent information (aside from the loss of phase information), the difference in representation may lead deep learning models to learn different features (Abercrombie 1995; Boatwright and Fletcher 1984). In other words, the feature extracted and learned by deep learning networks from different dimensions and scales of data (such as time-domain seismic wave data, spectrum data, and ground-motion parameter data) may not be exactly the same. In addition, deep learning network is currently a black box, and the information extracted by deep learning networks from input data is not clear. Therefore, in this study, we established a multimodal deep learning network (MDLNet) with the aim of using it to extract features from multimodal data of different dimensions and scales (time-domain seismic wave data, spectrum data, ground-motion parameter data) to obtain more effective information and improve the accuracy of seismic event identification and magnitude classification.

MDLNet is mainly composed of three branches that fully utilize data from three different modalities, as shown in Fig. 2a. The multimodal data encompass time-domain data, spectrum data, and ground-motion parameter data. The inputs to the time-domain branch are three-component seismic waveform time-domain data with a structure of (400, 3), and time-domain features are captured using a time-domain encoder. The spectrum branch, on the other hand, uses three-component seismic spectrum data with a structure of (200, 3), utilizing a spectrum encoder to capture spectrum features. Due to the commendable performance of the combination of convolutional neural network (CNN) and recurrent neural network (RNN) in data processing in prior research (Mousavi and Beroza 2020; Song et al. 2023), the time-domain encoder (Fig. 2b) and spectrum encoder (Fig. 2c) employed in this study are primarily composed of CNN and RNN sub-blocks. The CNN

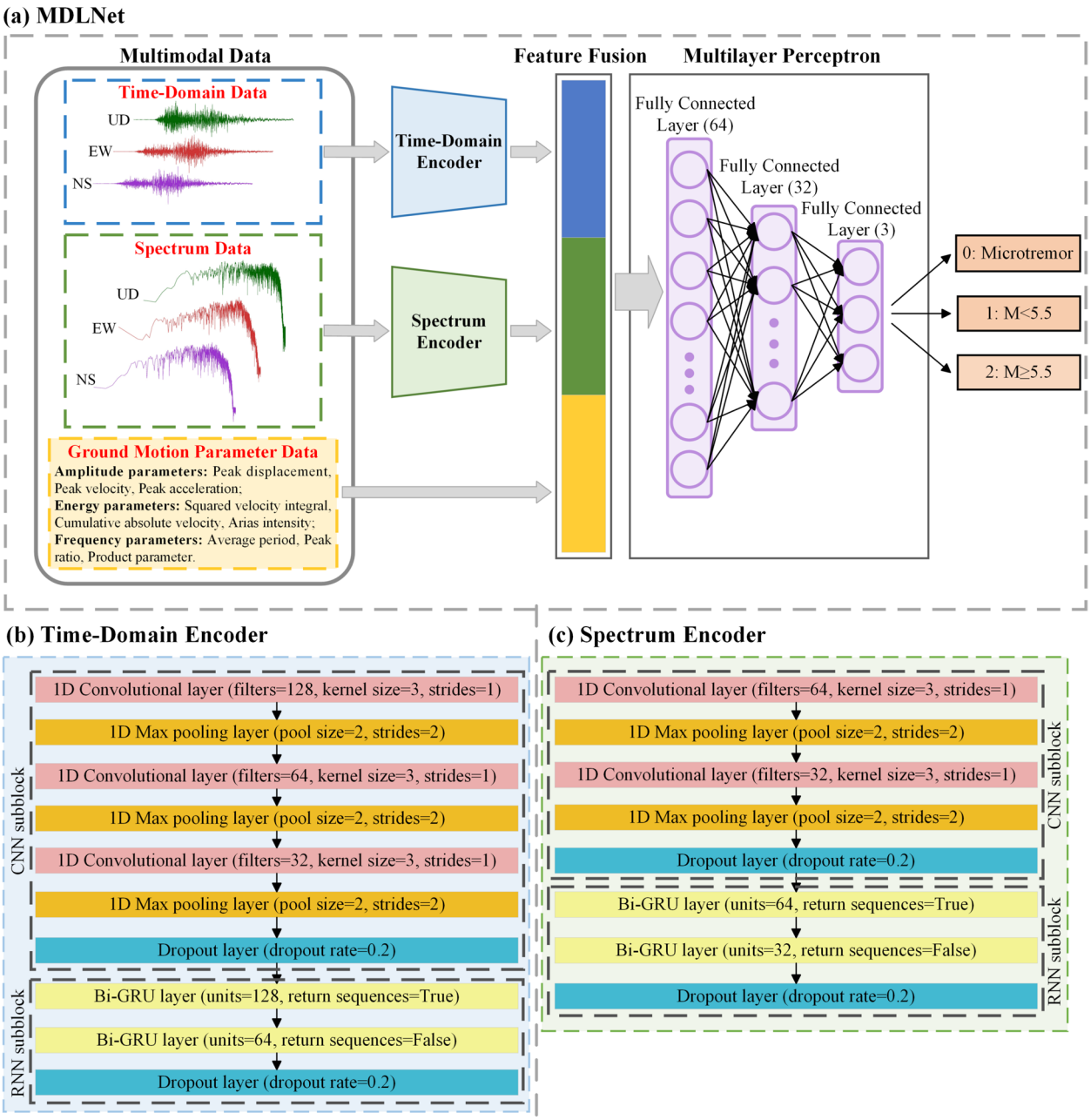

Fig. 2 **a** Architecture of MDLNet used in this work. **b** Architecture of the time-domain encoder. **c** Architecture of the spectrum encoder

sub-block is primarily composed of convolutional layers, max pooling layers, and dropout layers, and the RNN sub-block is primarily composed of bidirectional gated recurrent unit (Bi-GRU) layers and dropout layers. Simultaneously, the ground-motion parameter data branch involves nine ground-motion parameters with a (9, 1) structure. Based on the feature concatenation layer, we fuse the flattened ground-motion parameter data with the features extracted by the time-domain encoder and the spectrum encoder to integrate the multimodal data and improve the model's capability for feature extraction. Subsequently, these fused features are fed into a multilayer perceptron composed of three fully connected layers (FCLs). The number of neurons in each FCL is 64, 32, and 3. Notably, all convolutional layers and FCLs in this network employ a rectified linear unit (ReLU) activation function, except for the last FCL. The last FCL uses the Softmax activation function (Goodfellow et al. 2016), which is used to predict the probability of each category, and the total probability is 1. Meanwhile, select the class with the highest probability as the prediction result. In this work, the microtremor signals are labeled class 0, seismic signals from low-magnitude ($M<5.5$) seismic events are labeled class 1, and seismic signals from high-magnitude ($M\geq 5.5$) seismic events are labeled class 2. For example, if the output of the last FCL is [0.895, 0.095, 0.010], then it can be known that the value with the highest probability is 0.895 and corresponds to class 0, which means that the model predicts the sample to be a microtremor signal.

Network training and optimization

During the training process, we employed the "validation split = 0.1" command from the TensorFlow framework to randomly partition 10% of the training data set as the validation data set. Simultaneously, we utilized the sparse categorical cross-entropy loss function (Goodfellow et al. 2016) and the Adam optimizer (Kingma and Ba 2017). The initial learning rate was set at 0.001, the batch size was fixed at 256, and the maximum number of epochs was set to 500. To optimize the learning rate and mitigate model overfitting, we implemented both a learning rate scheduler and an early stopping mechanism (Prechelt 2012). If the validation accuracy remained unchanged over five epochs, the learning rate was multiplied by 0.1 to obtain the new learning rate. In the event of no improvement in validation accuracy within ten epochs, we terminated the training process, selecting the model with the highest accuracy as the optimal training model. Figure S3 shows the loss and accuracy curves of MDLNet for both the training and validation data sets. For the initialization of parameters during the training process of the model (such as initial value assignment for each neuron), this study sets "random seed" (Goodfellow et al. 2016). Dropout rate is set to 0.2. Through the above operations, ensure that the results of repeated training of the model are consistent to avoid the randomness of the output results each time. Furthermore, employing the grid search method, we explored the hyperparameters of MDLNet, prioritizing the overall precision, recall, F1 score, and accuracy as optimization metrics. Tables S2–S4 delineate the grid search process for the hyperparameters of the time-domain encoder, spectrum encoder, and multilayer perceptron within MDLNet, respectively. It can be observed from Tables S2–S4 that for grid search, the variation range of model performance indicators is only within 0.008. This to some extent indicates the stability of the MDLNet model proposed in this study, and the differences in hyperparameters and data segmentation methods have not had a significant impact on the MDLNet model. Figure 2c, d presents details regarding the time-domain encoder and spectrum encoder structures employed in this study, respectively. Concurrently, Fig. S4 offers additional insights into the network. Furthermore, we demonstrated through an ablation study that an encoder composed of CNN and RNN sub-blocks can improve the performance of MDLNet (Table S5). Moreover, we employed a multilayer perceptron-based encoder for preprocessing the ground-motion parameter data before feature fusion. The grid search process for the hyperparameters of the ground-motion parameter encoder based on a multilayer perceptron is detailed in Table S6. From the findings in Table S6, it is clear that MDLNet best identifies seismic events when the ground-motion parameters are not processed by the encoder.

Results and discussion

Based on the test data set, Table S7 reveals that MDLNet exhibits overall precision, recall, accuracy, and F1 score of 0.97458, 0.97457, 0.97457, and 0.97458, respectively, surpassing those of the baseline models based on single-mode data. Meanwhile, we used the same data set, the same training and testing methods, and the same model optimization methods to compare and analyze the performance of MDLNet and the baseline models. Further details on these performance metrics are provided in Text S2. In addition, for the evaluation of multiclass classification, metrics were computed individually for each category, as outlined in Text S2. Consequently, concerning the baseline models for different single-mode data, Fig. 3a-d shows the receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC) values for each category (Fawcett 2006). Text S3 provides a detailed introduction to the ROC curve in this study.

For all categories, MDLNet based on multimodal data consistently yields superior AUC values compared to the baseline models relying on single-mode data. Meanwhile, it can be seen from Fig. 3a that for microtremor signal recognition, using only the baseline model based on time-domain waveform can achieve similar performance to MDLNet, and is superior to the performance of baseline model based on spectrum data and baseline model based on ground-motion parameter data. We infer that (1) time-domain waveform data contributes more to microtremor signal recognition in MDLNet than amplitude spectrum data and ground parameter data and (2) due to the loss of phase information in amplitude spectrum data, this may also be one of the reasons why baseline model based on time-domain waveform can achieve higher performance than baseline model based on amplitude spectrum data. In addition, for a more insightful analysis, Fig. 3e-h illustrates confusion matrices for different modalities. As shown in Fig. 3e-h, MDLNet identifies 99.92% of microtremor signals, 96.65% of low-magnitude ($M<5.5$) seismic signals, and 90.68% of high-magnitude ($M\geq5.5$) seismic signals, surpassing the accuracy of baseline models based on single-mode data. Meanwhile, by comparing Fig. 3e, f, it can be found that compared with the baseline model based on time-domain data, the MDLNet improves the accuracy of low-magnitude ($M<5.5$) seismic signal recognition by 0.64%, and improves the accuracy of high-magnitude ($M\geq5.5$) seismic signal recognition by about 3%. By comparing Fig. 3e, h, it can be found that compared with the baseline model based on ground-motion parameter data, the MDLNet has improved the accuracy of microtremor signal recognition by 3.03%, low-magnitude ($M<5.5$) seismic signal recognition by 5.65%, and high-magnitude ($M\geq5.5$) seismic signal recognition by about 14.9%.

As shown in Fig. 3e, only two low-magnitude ($M<5.5$) seismic signals and six high-magnitude ($M\geq5.5$) seismic signals are misclassified as microtremor signals. Furthermore, Fig. 4a, b depicts MDLNet's seismic recognition and magnitude classification results concerning different epicentral distances and SNRs, respectively. Notably, from Fig. 4a, b, low magnitudes, low SNRs, and long epicentral distances emerge as the predominant factors contributing to the misclassification of seismic signals as microtremors. Simultaneously, data indicating misclassifications of $M<5.5$ signals as $M\geq5.5$ signals and vice versa are predominantly clustered near decision boundaries ($M5.5\pm0.5$). This occurrence may be attributed to the initial stages of seismic events, where the rupture characteristics near a decision boundary are similar (Ide 2019; Melgar and Hayes 2019). In addition, in Fig. 4c–e, we analyze the dependency of seismic event recognition and magnitude classification accuracy on magnitude, SNR, and epicentral distance under different modalities. Across various magnitude ranges, SNRs, and epicentral distances, MDLNet based on multimodal data exhibits superior accuracy in seismic event recognition and magnitude classification compared to the baseline models reliant on single-mode data. Meanwhile, it can be found from Fig. 4c that within the magnitude range of 5–6, compared with the baseline model based on single-mode data, the MDLNet based on multimodal data significantly improves the accuracy of magnitude classification within the magnitude range of 5–6, which also shows that the MDLNet can learn and extract more effective information and features from multimodal data. When the magnitude range is between 5 and 6, the MDLNet improves the accuracy of magnitude classification by approximately 5% compared to the baseline model based on time-domain data. Besides, based on the above results, we can infer that although time-domain seismic waves, their spectrum, and ground-motion parameters are not completely independent, the MDLNet model can extract more effective information and features from multimodal data at different dimensions and scales (time-domain seismic waves, spectrum data, and ground-motion parameter data), and inputting non independent and different dimensional multimodal data into deep learning network has certain advantages. Furthermore, Fig. 4c–e indicates that in ranges distant from the decision boundary, high SNR ranges, and ranges proximate to the epicenter, MDLNet demonstrates excellent performance in seismic event recognition and magnitude classification.

Based on different amounts of training data sets to train the model and for the same test data set, Fig. 5a–d, respectively, shows the performance of the model about the balanced accuracy (BACC) metric (Nakano

(See figure on next page.)

Fig. 3 **a** ROC curves for microtremors in the three-classification task based on different modal data. **b** ROC curves for the class of low-magnitude ($M<5.5$) seismic signals in the three-classification task based on different modal data. **c** ROC curves for the class of high-magnitude ($M\geq5.5$) seismic signals in the three-classification task based on different modal data. **d** Macroaverage represents the arithmetic mean of the metrics for multiple classes. **e** Confusion matrix of the three-classification task for MDLNet based on multimodal data. **f** Confusion matrix of the three-classification task using time-domain data. **g** Confusion matrix of the three-classification task using spectrum data. **h** Confusion matrix of the three-classification task using ground-motion parameter data

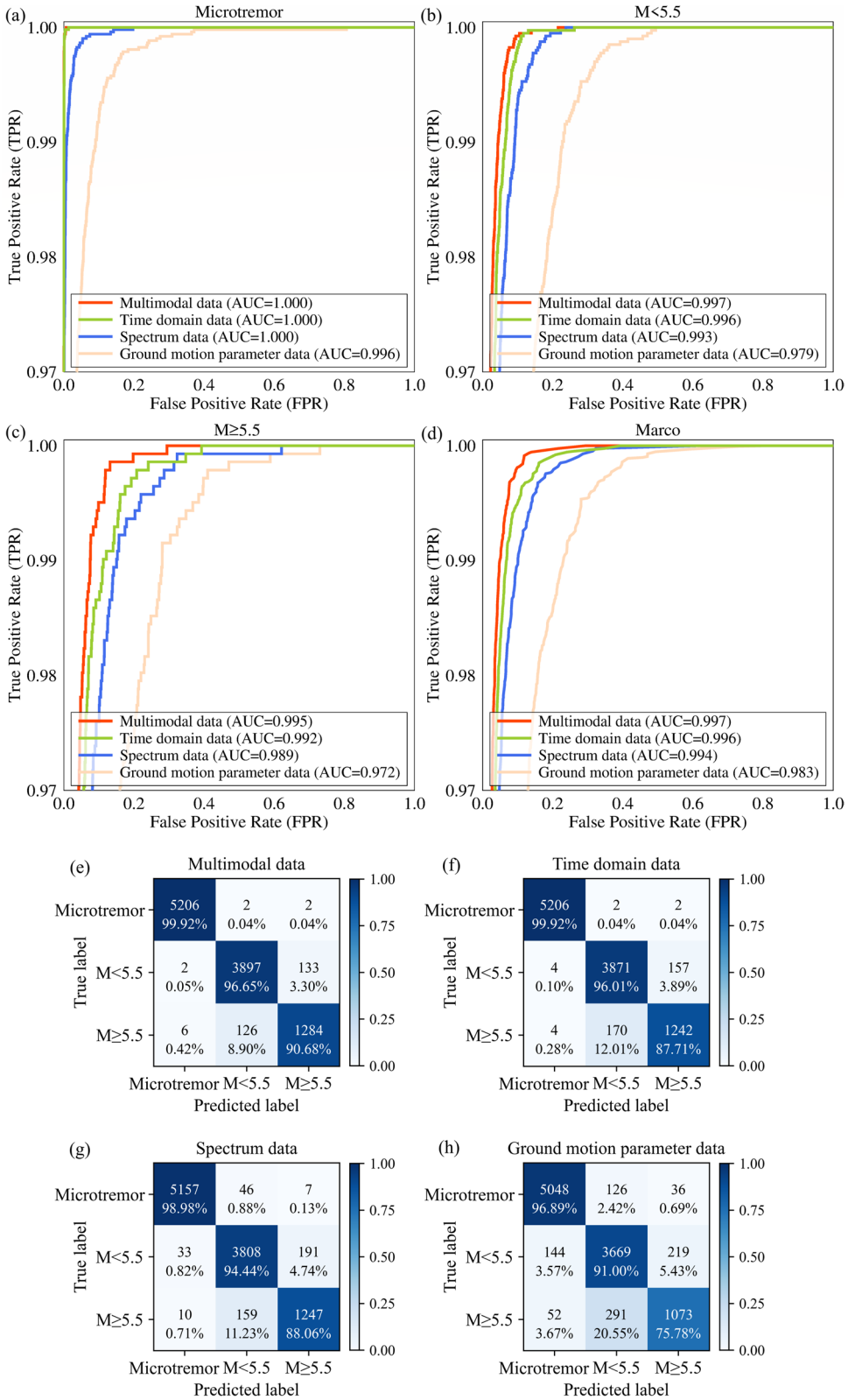

Fig. 3 (See legend on previous page.)

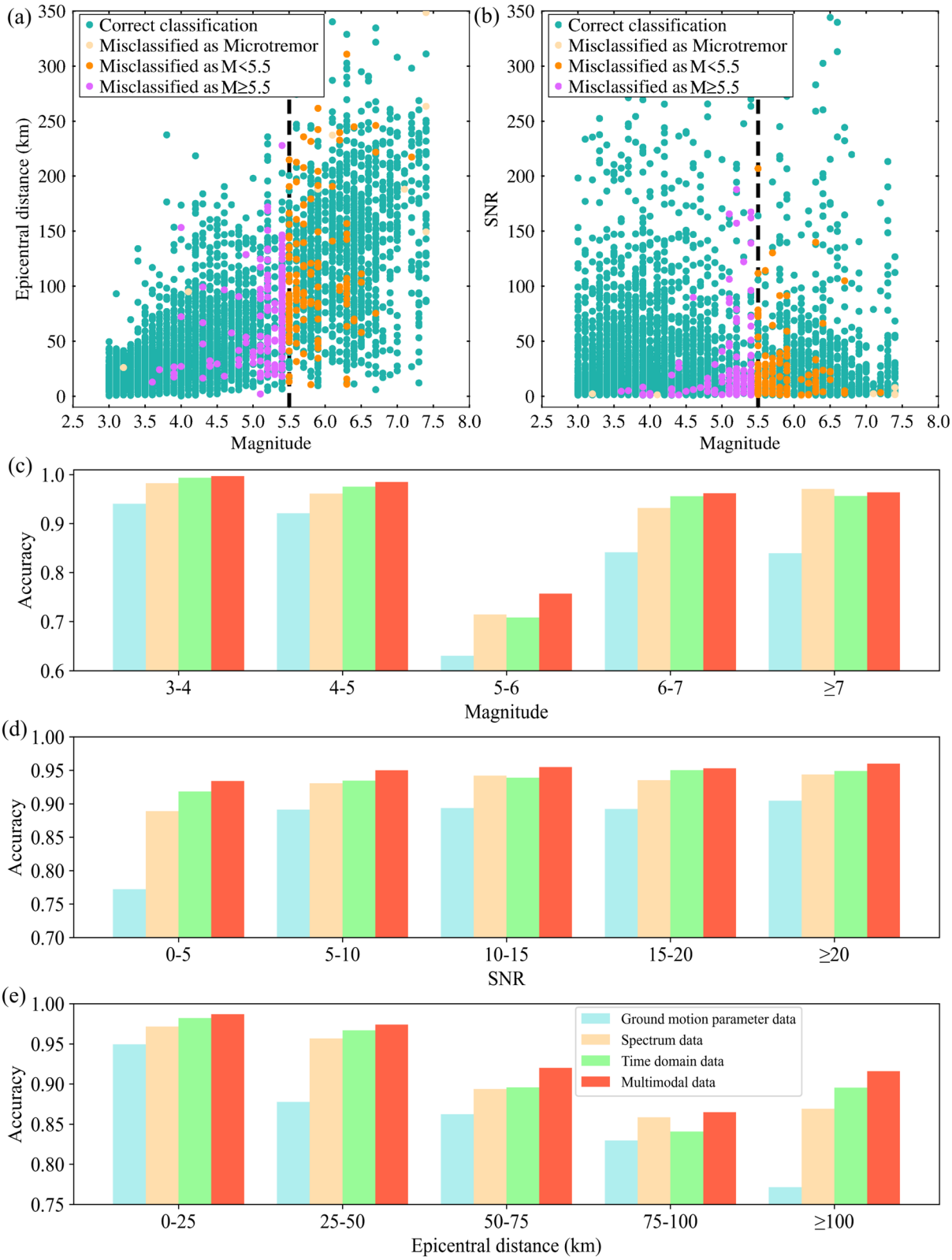

Fig. 4 **a** Distribution of classification results with epicentral distance using MDLNet. **b** Distribution of classification results with SNR using MDLNet. **c** Classification accuracy for earthquakes with different magnitude ranges under different modalities. **d** Classification accuracy for earthquakes with different SNR ranges under different modalities. **e** Classification accuracy for earthquakes with different epicentral distance ranges under different modalities

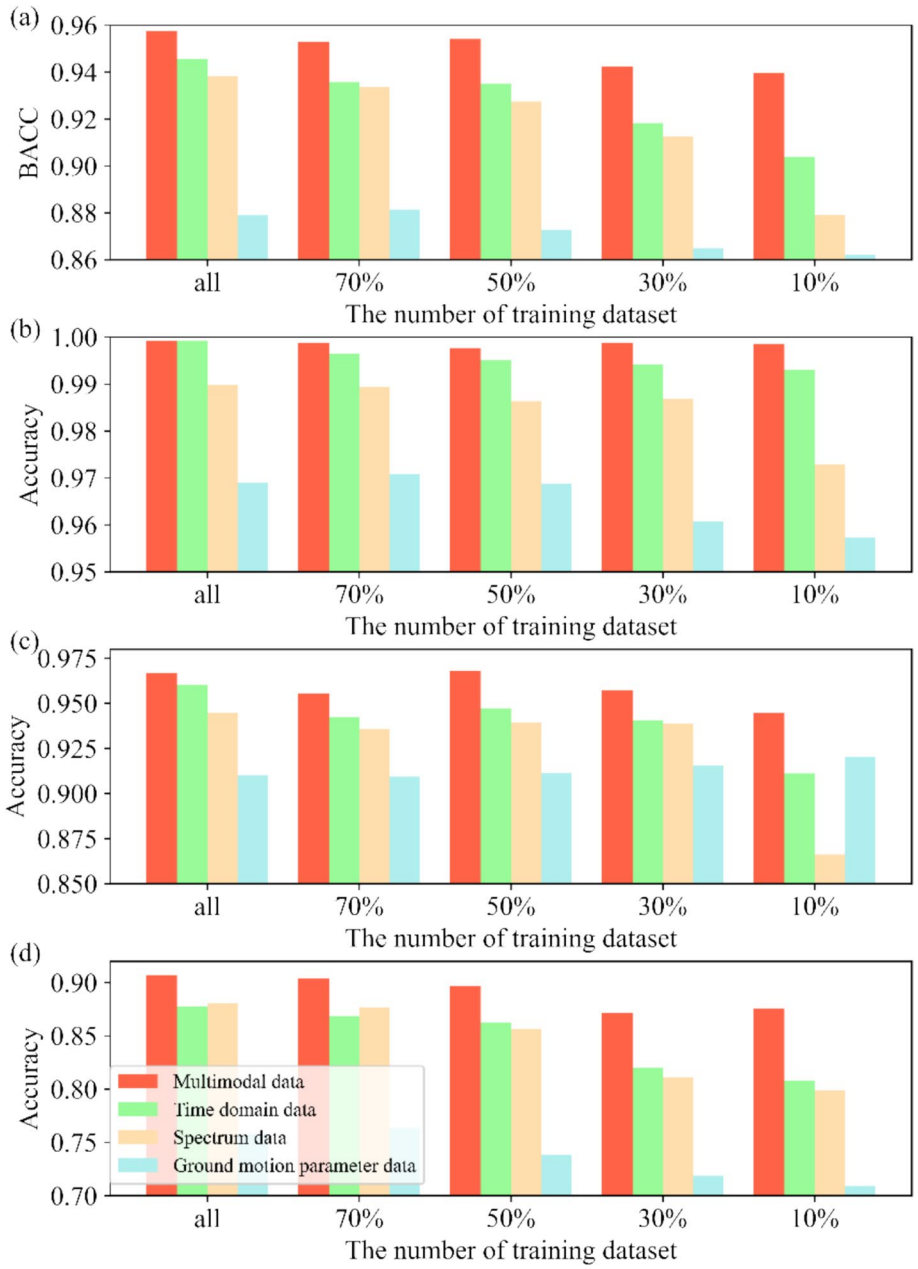

Fig. 5 **a** Balanced accuracy (BACC) metric of models based on different modal data for different numbers of training data. **b** Accuracy of microtremor signal recognition of models based on different modal data for different amounts of training data. **c** Accuracy of low-magnitude ($M < 5.5$) seismic signal recognition of models based on different modal data for different amounts of training data. **d** Accuracy of high-magnitude ($M \geq 5.5$) seismic signal recognition of models based on different modal data for different amounts of training data. "all" represents training the model using all the training data sets introduced in the 2 data section, "70%" represents training the model using the 70% training data set introduced in the 2 data section, "50%" represents training the model using the 50% training data set introduced in the 2 data section, "30%" represents training the model using the 30% training data set introduced in the 2 data section, and "10%" represents training the model using the 10% training data set introduced in the 2 data section

and Sugiyama 2022), the accuracy of the model for microtremor signal recognition, the accuracy of low-magnitude ($M < 5.5$) seismic signal recognition, and the accuracy of high-magnitude ($M \geq 5.5$) seismic signal recognition. As can be seen from Fig. 5a, the BACC of MDLNet based on multimodal data is higher than that of baseline model based on single-mode data. Meanwhile, compared with the baseline model based on single-mode

data, the MDLNet has more obvious advantages in performance with fewer training data sets. In addition, the BACC of the baseline model based on time-domain data and the baseline model based on spectrum data are also relatively close, and the difference in BACC between the two is mainly distributed between 0.01 and 0.02, which is consistent with the research results of Nakano and Sugiyama (2022). From Fig. 5b–d, it can be observed that when training the model using the "all" training data set, although the accuracy of the MDLNet model for microtremor signal recognition is similar to that of the baseline model based on time-domain data, the MDLNet model shows higher accuracy for low-magnitude ($M < 5.5$) seismic signal recognition and high-magnitude ($M \geq 5.5$) seismic signal recognition. Meanwhile, compared with the baseline model based on single-mode data, with the reduction of the number of training data sets, the MDLNet has more obvious advantages in performance. This also indicates that using the MDLNet can obtain effective information from time-domain and spectrum data, and has certain advantages, which can improve the accuracy of earthquake event identification and magnitude classification to a certain extent.

In this study, we used TensorFlow 2.3 framework, Python 3.6, and an NVIDIA Quadro T1000 GPU for data processing, model training, and testing. Table S8 shows the computational costs of MDLNet and baseline models based on single-mode data. From Table S8, it can be observed that (1) the processing time for obtaining spectrum data of one sample in this study is approximately 0.06 ms, and the processing time for obtaining ground parameter data of one sample is approximately 4.78 ms; (2) compared to baseline models based on single-mode data, although the MDLNet has an increased processing time for obtaining a single sample, the processing time is only within 5 ms; (3) the MDLNet takes approximately 29 s per epoch during training, which is higher than the baseline models based on single-mode data in terms of the duration of each epoch during training; and (4) compared with baseline model based on time-domain data, baseline model based on spectrum data and the baseline model based on ground-motion parameter data, the MDLNet increased the time required to predict a sample by approximately 10.14 ms, 12.72 ms, and 24.79 ms. In this study, compared with the baseline models based on single-mode data, we consider the increased computational cost of the MDLNet to be acceptable and negligible.

This study mainly proposes a multimodal deep learning network (MDLNet) for earthquake event identification and magnitude classification, which has not yet been applied to actual EEW systems. Moreover, we mainly collected earthquakes with magnitudes ranging from 3 to 8 recorded by the Japanese *K*-NET network to train and test MDLNet. Based on the test data set of this study, the results show that compared with the baseline models based on single-mode data, the MDLNet model can improve the accuracy of earthquake event identification and magnitude classification to a certain extent. We infer that the MDLNet model has certain applicability and feasibility for earthquake events occurring in Japan with magnitudes ranging from 3 to 8. In actual EEW application, earthquakes with $M < 3$ and earthquakes with $M > 8$ may occur, and the applicability of the model needs further verification for earthquake events that occur outside the magnitude range of the training data set. The MDLNet model may misidentify seismic event signals with $M < 3$ as microtremors (noise). Meanwhile, before applying the MDLNet to actual EEW systems, we also need to conduct offline and online testing of the MDLNet using a large amount of data. In addition, the applicability and feasibility of the MDLNet model for regions other than Japan still need further validation and analysis. Meanwhile, further in-depth research and exploration are needed in the future to address the limitations of the MDLNet model proposed in this study. We may consider taking the following measures and work: (1) collecting more earthquake data (including seismic events below magnitude 3 and above magnitude 8, etc.) to enrich our training data set, and retraining the MDLNet model to improve its generalization. (2) Through transfer learning, the MDLNet model is fine-tuned using earthquake events from different regions and earthquakes outside the magnitude range of the training data set to improve its generalization and robustness. Besides, the MDLNet proposed in this study has not yet reached the operational and practical EEW system level of Japan Meteorological Agency's Earthquake Early Warning System. Therefore, it is hard to achieve unbiased comparison between the results of this study and the predicted results of Japan Meteorological Agency's Earthquake Early Warning System. In future research, we will also consider collaborating with relevant researchers and technicians from Japan Meteorological Agency's Earthquake Early Warning System, attempting to highlight the advantages of the MDLNet proposed in this study in earthquake event identification and magnitude classification, as well as combining the advantages of Japan Meteorological Agency's Earthquake Early Warning System to further explore the potential for improvement of the MDLNet proposed in this study.

Meanwhile, the seismic data used in this study was selected from already cataloged events (downloaded from website https://www.kyoshin.bosai.go.jp/kyoshin/quake/index_en.html), rather than from all triggered waveforms at each *K*-NET station, and all continuous waveform data

are not used. This may have introduced an undesirable bias in data selection. *K*-NET system is made by the Japanese National Research Institute for Earth Science and Disaster Resilience (NIED). The time history of ground motion is recorded by an event-triggered system (ETS). Data on ground motion is retained solely when the motion satisfies the pre-established triggering criteria. The ETS of the data recorder is controlled using surface signals (*K*-NET). Normally, the recording of an event initiates at a threshold acceleration of 2 gal and concludes with a sustained 30 s signal that is at or below 2 gal. Upon triggering, the communication module promptly initiates a connection to the Data Management Center, employing a dial-up router that operates through a digital telephone line. For more detailed information on how the *K*-NET system selects or triggers seismic event data, please refer to previous studies (Aoi et al. 2004, 2011; Okada et al. 2004). Meanwhile, in future research, we will also consider collaborating with relevant researchers and technicians from the Japanese National Research Institute for Earth Science and Disaster Resilience (NIED), collect all triggered waveforms at each *K*-NET station, then train and test the model, and optimize the performance of the model to avoid introducing undesirable bias in data selection.

Conclusions

To apply multimodal data to the problem of seismic discrimination and magnitude classification, we develop a multimodal deep learning network (MDLNet), which considers the tasks of identifying earthquake events and classifying earthquake magnitudes as a three-classification problem. MDLNet employs a time-domain encoder and a spectrum encoder to extract features from time-domain and spectrum data, respectively. Subsequently, MDLNet fuses the features extracted by the time-domain encoder and spectrum encoder with the ground-motion parameter data, and the multilayer perceptron is used to identify microtremor signals, low-magnitude ($M < 5.5$) seismic signals, and high-magnitude ($M \geq 5.5$) seismic signals. We employ multimodal seismic and microtremor signals from the Japanese Kyoshin Network to train and test MDLNet. We demonstrate that during the 3 s following the *P*-wave onset, MDLNet can recognize 99.92% of microtremor signals, 96.65% of low-magnitude seismic signals and 90.68% of high-magnitude seismic signals, values higher than those for single-mode data. Furthermore, for varying ranges of magnitudes, SNRs, and epicentral distances, MDLNet outperforms the baseline models based on single-mode data. MDLNet has the potential to improve the performance of EEW systems. Multimodal deep learning can process multimodal data and integrate important information from different modal data, thus providing a basis for further work in the field of artificial intelligence. Here, we preliminarily explored the application of multimodal deep learning in seismic discrimination and magnitude classification, with the goal of extending the methodology to various other areas in the field of seismology in the future.

Abbreviations

EEW	Earthquake early warning
MDLNet	Multimodal deep learning network
CLIP	Contrastive Language-Image Pretraining
K-NET	Kyoshin Network
SNR	Signal-to-noise ratio
CNN	Convolutional neural network
RNN	Recurrent neural network
Bi-GRU	Bidirectional gated recurrent unit
FCL	Fully connected layer
ReLU	Rectified linear unit
ROC	Receiver operating characteristic
AUC	Area under the curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40562-025-00412-7.

Additional file 1.
Additional file 2.
Additional file 3.
Additional file 4.
Additional file 5.
Additional file 6.
Additional file 7.
Additional file 8.
Additional file 9.
Additional file 10.
Additional file 11.
Additional file 12.
Additional file 13.
Additional file 14.
Additional file 15.
Additional file 16.
Additional file 17.

Acknowledgements

We would like to thank the Editor Yuichiro Tanioka, and two anonymous reviewers for their constructive comments and suggestions, which are very helpful for improving our manuscript. We wish to thank the Japanese National Research Institute for Earth Science and Disaster Resilience (NIED) for providing the K-NET seismic data to the public. We are grateful to the Python community for making everything publicly available.

Author contributions

JZ and JS contributed to conceptualization; JZ and JS contributed to methodology; JZ provided software and were involved in formal analysis; JZ, JS and SL were responsible for validation, writing—original draft preparation, and writing—review and editing; JZ, JS and SL were involved in investigation; JS provided resources and performed supervision, project administration, and funding acquisition; JZ, JS and SL were involved in data curation;

JZ contributed to visualization. All authors have read and agreed to the published version of the manuscript. All authors contributed to all sections until the final revision of the manuscripts.

Funding
This study was supported by the Scientific Research Fund of Institute of Engineering Mechanics, China Earthquake Administration with Grant Number 2024B08, and the National Natural Science Foundation of China with Grant Number 42304074.

Data availability
Table S1 provides the seismic event information, and the website for the Japanese strong-motion records from K-NET used in this study is https://doi.org/10.17598/NIED.0004. The deep learning framework used in this work is TensorFlow (Abadi et al. 2016). The code for the MDLNet and ROC curve calculation in this work is available from GitHub (https://github.com/Jingbaozhu1996/MDLNet).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 20 December 2023 Accepted: 19 August 2025
Published online: 25 August 2025

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. (2016) TensorFlow: a system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16). Berkeley, CA, US: USENIX Association, pp 265–283. Retrieved from https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

Abercrombie RE (1995) Earthquake source scaling relationships from− 1 to 5 ML using seismograms recorded at 2.5-km depth. J Geophys Res Solid Earth 100(B12):24015–24036. https://doi.org/10.1029/95JB02397

Allen RV (1978) Automatic earthquake recognition and timing from single traces. Bull Seismol Soc Am 68(5):1521–1532. https://doi.org/10.1785/bssa0680051521

Allen RM, Melgar D (2019) Earthquake early warning: advances, scientific challenges, and societal needs. Annu Rev Earth Planet Sci 47(1):361–388. https://doi.org/10.1146/annurev-earth-053018-060457

Aoi S, Kunugi T, Fujiwara H (2004) Strong-motion seismograph network operated by NIED: K-NET and KiK-net. J Japan Assoc Earthquake Eng 4(3):65–74. https://doi.org/10.5610/jaee.4.3_65

Aoi S, Kunugi T, Nakamura H, Fujiwara H (2011) Deployment of new strong motion seismographs of K-NET and KiK-net. Earthquake data in engineering seismology: predictive models, data management and networks, pp 167–186. https://doi.org/10.1007/978-94-007-0152-6_12

Baltrusaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443. https://doi.org/10.1109/tpami.2018.2798607

Boatwright J, Fletcher JB (1984) The partition of radiated energy between P and S waves. Bull Seismol Soc Am 74(2):361–376. https://doi.org/10.1785/BSSA0740020361

Bose M, Hauksson E, Solanki K, Kanamori H, Wu YM, Heaton TH (2009) A new trigger criterion for improved real-time performance of onsite earthquake early warning in Southern California. Bull Seismol Soc Am 99(2A):897–905. https://doi.org/10.1785/0120080034

Böse M, Heaton TH, Hauksson E (2012) Real-time finite fault rupture detector (FinDer) for large earthquakes. Geophys J Int 191(2):803–812. https://doi.org/10.1111/j.1365-246x.2012.05657.x

Boussioux L, Zeng C, Guénais T, Bertsimas D (2022) Hurricane forecasting: a novel multimodal machine learning framework. Weather Forecasting 37(6):817–831. https://doi.org/10.1175/waf-d-21-0091.1

Chen Y, Zhang G, Bai M, Zu S, Guan Z, Zhang M (2019) Automatic waveform classification and arrival picking based on convolutional neural network. Earth Space Sci 6(7):1244–1261. https://doi.org/10.1029/2018ea000466

Cochran ES, Kohler MD, Given DD, Guiwits S, Andrews J, Meier MA et al (2017) Earthquake early warning ShakeAlert system: testing and certification platform. Seismol Res Lett 89(1):108–117. https://doi.org/10.1785/0220170138

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge

Guzhov A, Raue F, Hees J, Dengel A (2022) Audioclip: extending clip to image, text and audio. ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Singapore, pp 976–980

Huang PL, Lin TL, Wu YM (2015) Application of τc* Pd in earthquake early warning. Geophys Res Lett 42(5):1403–1410. https://doi.org/10.1002/2014gl063020

Ide S (2019) Frequent observations of identical onsets of large and small earthquakes. Nature 573(7772):112–116. https://doi.org/10.1038/s41586-019-1508-5

Kanamori H (2005) Real-time seismology and earthquake damage mitigation. Annu Rev Earth Planet Sci 33(1):195–214. https://doi.org/10.1146/annurev.earth.33.092203.122626

Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. arXiv:1412.6980

Kohler MD, Smith DE, Andrews J, Chung AI, Hartog R, Henson I et al (2020) Earthquake early warning shakealert 2.0: public rollout. Seismol Res Lett 91(3):1763–1775. https://doi.org/10.1785/0220190245

Kong Q, Allen RM, Schreier L, Kwon YW (2016) MyShake: a smartphone seismic network for earthquake early warning and beyond. Sci Adv 2(2):e1501055. https://doi.org/10.1126/sciadv.1501055

Li Z, Meier MA, Hauksson E, Zhan Z, Andrews J (2018) Machine learning seismic wave discrimination: application to earthquake early warning. Geophys Res Lett 45(10):4773–4779. https://doi.org/10.1029/2018gl077870

Li J, Böse M, Feng Y, Yang C (2021) Real-time characterization of finite rupture and its implication for earthquake early warning: application of finder to existing and planned stations in Southwest China. Front Earth Sci 9:699560. https://doi.org/10.3389/feart.2021.699560

Linville L, Pankow K, Draelos T (2019) Deep learning models augment analyst decisions for event discrimination. Geophys Res Lett 46(7):3643–3651. https://doi.org/10.1029/2018gl081119

Liu H, Li S, Song J (2022) Discrimination between earthquake p waves and microtremors via a generative adversarial network. Bull Seismol Soc Am 112(2):669–679. https://doi.org/10.1785/0120210231

Lomax A, Michelini A, Jozinović D (2019) An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network. Seismol Res Lett 90(2A):517–529. https://doi.org/10.1785/0220180311

Meier MA, Ross ZE, Ramachandran A, Balakrishna A, Nair S, Kundzicz P et al (2019) Reliable real-time seismic signal/noise discrimination with machine learning. J Geophys Res Solid Earth 124:788–800. https://doi.org/10.1029/2018JB016661

Melgar D, Hayes GP (2019) Characterizing large earthquakes before rupture is complete. Sci Adv 5(5):eaav2032. https://doi.org/10.1126/sciadv.aav2032

Minson SE, Meier MA, Baltay AS, Hanks TC, Cochran ES (2018) The limits of earthquake early warning: timeliness of ground motion estimates. Sci Adv 4(3):eaaq0504. https://doi.org/10.1126/sciadv.aaq0504

Mittal H, Yang BM, Wu YM (2022) Progress on the earthquake early warning and shakemaps system using low-cost sensors in Taiwan. Geosci Lett 9:42. https://doi.org/10.1186/s40562-022-00251-w

Mousavi SM, Beroza GC (2020) A machine-learning approach for earthquake magnitude estimation. Geophys Res Lett 47(1):e2019GL085976. https://doi.org/10.1029/2019gl085976

Mousavi SM, Ellsworth WL, Zhu W, Chuang LY, 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
Nakano M, Sugiyama D (2022) Discriminating seismic events using 1D and 2D CNNs: applications to volcanic and tectonic datasets. Earth Planets Space 74(1):134. https://doi.org/10.1186/s40623-022-01696-1
Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696.
Njirjak M, Otović E, Jozinović D, Lerga J, Mauša G, Michelini A, Štajduhar I (2022) The choice of time–frequency representations of non-stationary signals affects machine learning model accuracy: a case study on earthquake detection from LEN-DB data. Mathematics 10(6):965. https://doi.org/10.3390/math10060965
Okada Y, Kasahara K, Hori S, Obara K, Sekiguchi S, Fujiwara H, Yamamoto A (2004) Recent progress of seismic observation networks in Japan—Hi-net, F-net, K-NET and KiK-net—. Earth Planets Space 56(8):xv–xxviii. https://doi.org/10.1186/BF03353076
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Perol T, Gharbi M, Denolle M (2018) Convolutional neural network for earthquake detection and location. Sci Adv 4(2):e1700578. https://doi.org/10.1126/sciadv.1700578
Prechelt L (2012) Early stopping—but when? In: Montavon G, Orr GB, Müller KR (eds) Neural networks: tricks of the trade, 2nd edn. Springer, Berlin, Heidelberg, pp 53–67
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning: PMLR, pp 8748–8763
Soleymani M, Pantic M, Pun T (2011) Multimodal emotion recognition in response to videos. IEEE Trans Affect Comput 3(2):211–223. https://doi.org/10.1109/t-affc.2011.37
Song J, Zhu J, Li S (2023) MEANet: Magnitude estimation via physics-based features time series, an attention mechanism, and neural networks. Geophysics 88(1):V33–V43. https://doi.org/10.1190/geo2022-0196.1
Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, Meurs JV et al (2019) Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep 9(1):20038. https://doi.org/10.1038/s41598-019-56527-3
Trani L, Pagani GA, Zanetti JPP, Chapeland C, Evers L (2022) Deepquake—an application of CNN for seismo-acoustic event classification in the Netherlands. Comput Geosci 159:104980. https://doi.org/10.1016/j.cageo.2021.104980
Vinker Y, Pajouheshgar E, Bo JY, Bachmann RC, Bermano AH, Cohen-Or D et al (2022) CLIPasso: semantically-aware object sketching. ACM Trans Graph 41(4):1–11. https://doi.org/10.1145/3528223.3530068
Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning. Geophys Res Lett 33(16):L16312. https://doi.org/10.1029/2006gl026871
Wu YM, Mittal H, Lin YH, Chang YH (2023) Magnitude determination using cumulative absolute absement for earthquake early warning. Geosci Lett 10(1):59
Zhu J, Li S, Song J (2022) Magnitude estimation for earthquake early warning with multiple parameter inputs and a support vector machine. Seismol Res Lett 93(1):126–136. https://doi.org/10.1785/0220210144
Zhu J, Li S, Wei Y, Song J (2023) On-site instrumental seismic intensity prediction for China via recurrent neural network and transfer learning. J Asian Earth Sci 248:105610. https://doi.org/10.1016/j.jseaes.2023.105610
Zollo A, Amoroso O, Lancieri M, Wu YM, Kanamori H (2010) A threshold-based earthquake early warning using dense accelerometer networks. Geophys J Int 183(2):963–974. https://doi.org/10.1111/j.1365-246x.2010.04765.x

Publisher's Note