IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 2, 2021, accepted November 15, 2021, date of publication December 6, 2021,
date of current version December 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132901

Augmenting Seismic Data Using Generative Adversarial Network for Low-Cost MEMS Sensors

AMING WU[1], JUYONG SHIN[1], JAE-KWANG AHN[2], AND YOUNG-WOO KWON[1]
[1]School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]Korea Meteorological Administration, Seoul 07062, South Korea

Corresponding author: Young-Woo Kwon (ywkwon@knu.ac.kr)

This study was supported by the Development of Earthquake, Tsunami, Volcano Monitoring and Prediction Technology (KMA-135002988) and by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394).

ABSTRACT The performance of a deep learning (DL) model depends on sufficient training datasets and its algorithmic structure. Even though seismological research using low-cost micro-electro-mechanical systems (MEMS) sensor received much attention recently, because of the lack of data recorded by such MEMS sensors whose data are usually polluted by different types of noise. Therefore, increasing seismic datasets is required by intelligently generating seismic data through data-augmentation techniques. However, it is difficult to characterize and measure the evolution process of seismic sequences, making the feature extraction and data generation of seismic sequences still a significant challenge. By combining the framework of Generative Adversarial Network (GAN) with long short-term memory (LSTM), attention mechanism and neural network (NN), a novel deep generation model (DGM) named EQGAN is developed to overcome the challenges, which can automatically capture the different time histories and dimension characteristics of seismic sequences, meanwhile stably generating high-quality seismic data. The reality of generated data is qualitatively clarified through the analysis of frequency domain and data autocorrelation distribution. Based on the High-throughput Screening (HTS) Theory, the quantitative evaluation index of statistical metrics is designed, and the generation performance of different machine learning models (standard GAN, LSTM, NN) is compared to prove the stability and effectiveness of EQGAN. The experimental results denote that the EQGAN has excellent stability and performance (up to 81%, much higher than that of other generation models), which provides a suitable data expansion approach for the field of seismological research.

INDEX TERMS Deep learning, generative adversarial network, data augmentation, Wasserstein distance.

I. INTRODUCTION

Earthquake detection [1]–[3] and earthquake early warning (EEW) [4], [5] are the main tasks of seismological research. Many data processing techniques used in traditional seismological research originated from small datasets and limited computing power. Low-cost MEMS acceleration sensors have been extensively used in the monitoring system of Internet of Things (IoT) over the last few years, because of their low installation and operation costs, the examples include a wireless sensor network (WSN) [6], [7], community seismic network (CSN) [8]. Although they have a great potential to replace the traditional expensive seismic networks whose coverage is hardly dense due to the high installation and operation costs, however, the large noises inherent in a low-cost MEMS acceleration sensor reduce the quality of data recorded [9], thus a novel approach is needed to adapt to the data with different signal-to-noise ratios (SNR).

In recent years, the machine learning (ML) has been widely applied to earthquake detection [10]–[13], including earthquake first arrival recognition [14]–[16] and source location [17], [18]. Compared with other time series (stock market price, WiFi signals), a high-dimensional seismic sequence has many implict features (evolution process of different

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

components and a single component, etc.) that are difficult to capture. Khan *et al.* [19] developed an artificial neural network (ANN) model [20], [21] to detect seismic events by artificially selecting labels. Also, many researchers developed different seismic detection models based on different convolutional neural network (CNN) methods [3], [22], [23]. Nevertheless, no matter which method is adopted, the above mentioned models are supervised. A real seismic waveform needs to be identified by subjectively selecting feature labels, which will affect the detection performance of the model. Therefore, seismic detection methods based on the ML gradually used to eliminate the influence of subjective factors.

Considering that the performance of the DL algorithm depends on the size and quality of the training dataset, in our work, however, we utilize low-cost MEMS sensors to record ground motion signals instead of smartphones, which are polluted by different noise levels (human activities or the sensor themselves), resulting in the lack of high-quality seismic data (High SNR). Too few training datasets easily leads to overfitting [5]. Therefore, to solve this problem, the current research is mainly based on the following three solutions:

- *To Use Transformation:* To overcome the problems of incomplete real seismic waveform and low SNR, Dokht *et al.* [24] designed a general deep convolution network model for the seismic event and phase detection based on time-frequency representation and convolution neural network. Saad and Chen [25] used automated unsupervised approaches to extract waveform signals from continuous microseismic data according to the time-frequency representation of microseismic trajectory, which was also applicable in an environment with a low SNR, confirming that the waveform-based inverse time migration method could be used in the model to improve the resolution of microseismic imaging.
- *To Train Model Using a Generalized Deep Learning Model Based on a Small Dataset:* Using a generalized deep learning architecture to extract the most representative features from limited/small training datasets, Saad *et al.* [26] successfully proposed the SCALODEEP model to detect ground demand signals. Similarly, Zhu *et al.* [27] proposed a CNN-based phase recognition classifier (CPIC) for phase detection and picked up from small and medium-sized training datasets. While Saad and Chen [28] used a capsule neural network (CapsNet) to identify and detect earthquakes automatically and confirmed that it could learn from small datasets with a good generalization performance.
- *To Develop a Data Augmentation-Approach:* Data augmentation is also an effective method to increase data samples. The conditional GAN [29] was used to generate the seismic dataset effectively. Wang *et al.* [30] developed the EarthquakeGen to generate a short seismic waveform and verified its rationality. Although seismic data can be generated by inferring the implicit and explicit characteristics of the seismic waveform, it is not easy to ensure the diversity or efficiency of the data generated.

In this paper, we propose a novel DGM for data augmentation. However, the design of the generation model depends on the measurement of the distribution pattern of the original data. The speech synthesis technology based on Hidden Markov Model (HMM) has been proven to be very effective in synthesizing an acceptable speech [31]. Due to the discrete limitation of HMM, it cannot represent a continuous space. The LSTM model [32] based on natural language is applied to capture and memorize the long-term and short-term features of the sequences, so as to generate realistic text information. However, compared with the traditional recurrent neural network (RNN) algorithm, the training difficulty also increases because of too many parameters. Zhao *et al.* [33] proposes the seq2seq based on LSTM together with an attention mechanism to improve the efficiency quality of text summarization, which, however, lack text information coherence. Although different DL models have been developed in these studies to achieve time series generation, they cannot fully represent the distribution of the original data. The performance of generation model cannot meet the best expectations. Therefore, we designed a variant of the GAN structure called EQGAN, automatically capturing different dimensional and time history features.

In addition, we mix the real seismic data and noise data as the data input layer, integrate Wasserstein Distance (WD) and spectral normalization (SN) to improve the stability of model training, overcome mode collapses, and generate high-quality earthquake data recorded by approximate acceleration sensors. Since there is no absolute one-to-one correspondence between the generated and real data, it is not easy to evaluate the quality of the generated data. Frechet Inception Distance (FID) [34], [35] have been proposed in previous studies to evaluate the similarity between generated images and real images. However, for non-image data, an accurate evaluation of the GAN model is still a challenge. Reagan's powerful seismic data generation ability is qualititatively analyzed in this study through visual representation, frequency domain and autocorrelation schemes, and a new quantitative error evaluation scheme is designed based on HTS theory, which proves the excellent stability and high efficiency of our model.

The rest of this article is organized as follows. In Section II, we describe the basic theory of the GAN, design and develop the DGM by analyzing the data distribution patterns as well as characteristics of real seismic data. Section III provides details of data collection and preprocessing techniques. We then present the experimental results and evaluate our model from different metrics in Section IV. Finally, Section V includes the conclusion and future work of the research.

II. THEORY AND MODEL DESIGN

In this section, we first present the basic theoretical framework of standard GAN in Part *A*, the characteristics of real seismic sequences are analyzed in Part *B* according to

TABLE 1. Overview of related work.

References	Feature used	Proposed model	Database	Category
[24]	Earthquake/Phase detection	ConvNet	incomplete/low SNR waveform	Using transformation method
[25]	Eaethquake detection	Unsupervised technique	Microseismic dataset	Using transformation method
[26]	Earthquake detection	SCALODEEP	Small training dataset	Using generalizad deep learning based on a small daraset
[27]	Earthquake detection	CPIC	Small-sized dataset	Using generalizad deep learning based on a small daraset
[28]	Earthquake detection	CapsNet	Small training dataset	Using generalizad deep learning based on a small daraset
[31]	Synthesize speech	HMM	Speech dataset	Developing a data augmentation approach
[32]	Text generation	LSTM	Text dataset	Developing a data augmentation approach
[33]	Text summarization	Seq2seq	Text dataset	Developing a data augmentation approach
[29]	Seismic data augmentation	Conditional GAN	Seismic dataset	Developing a data augmentation approach
[30]	Short seismic waveform generation	EarthuqakeGen	Seismic dataset	Developing a data augmentation approach

the evolution process of the seismic event. Finally, we also design the EQGAN model based on appropriate algorithms in Part *C*.

A. THEORETICAL BASIS

3-component earthquake data is a series of discrete measurements captured in a continuous time series. Acceleration components change in different degrees at different time or in different dimensions at the same time. To automatically extract data features through the ML model instead of the traditional ANN model, so as to infer implicit data features, we introduce the GAN framework to build a generation model by fitting real data distribution.

In 2014, Goodfellow [36] proposed the concept of GAN, an epoch-making unsupervised learning algorithm framework (Fig.1c), in which only backpropagation is used to train the network, avoiding the use of Markov chain, and making a deep learning breakthrough. GAN's basic idea comes from the 0-1 Game Theory, which is mainly composed of a generator and a discriminator. The purpose of the generator is to learn and capture the potential distribution of real data as much as possible meanwhile generating new data. At the same time, the essence of the discriminator is a binary classifier with the purpose of identifying whether the incoming data is from real data or generated data as accurately as possible. This learning optimization process is a maximin game to optimize and improve their generation or discrimination ability continuously, whose purpose is to find a Nash equilibrium between the two sides. The performance of the generation task depends on the design of the GAN adversarial mechanism. The optimal objective function can be expressed as follows:

$$\min_G \max_D \ \mathrm{V(D,}\, G) = E_{x\sim p_{\mathrm{r}}(x)}[\log(D(x))] + E_{z\sim p_g(z)}[\log(1 - D(G(z)))] \quad (1)$$

where $D(*)$ is a discriminator, and $G(*)$ is a generator. The distribution probability of generated data $g(z)$ corresponding to random variable matrix z is $p_{g(z)}$, and the data distribution of the real data x is p_r, $D(g(z))$ is the probability of fast data $g(x)$ estimated by the discriminator, whose output value is 0-1. When the generator $g(*)$ is optimized, the discriminator $D(*)$ is fixed. The purpose of generator optimization is to cheat discriminator $D(*)$, so that $D(g(z))$ tends to be 1 and $1 - D(g(z))$ tends to be 0, so the optimization generator is to minimize $V(g, d)$; on the contrary, the optimized discriminator is to recognize no matter how realistic the generated samples are, that is, $D(g(z))$ tends to be 0 and $1-D(g(z))$ tends to be 1, so that the optimal discriminator is to maximize $V(D, G)$. Compared with other generation models, more real samples can be generated with only backpropagation through GAN. At the same time, no complex Markov chain is needed.

B. ANALYSIS OF SEISMIC DATA FEATURES

Due to the complex high-dimensional data structure of seismic data, it is impossible to directly represent its specific distribution pattern. The evolution of the seismic waveform is mapped into a controllable observation sequence, showing multiple multi-dimensional time-series correlations. To ensure the performance of the model training and generate various seismic waves, we systematically analyze the evolution of earthquake waves: (i) Fully extract the implicit and explicit characteristics of the seismic sequence's time evolution and spatial dimension. (ii) A realistic earthquake sequence generation model is established, and appropriate countermeasures are designed and adjusted to meet the dependence of high-quality generation tasks. Different time series distribution patterns lead to different construction methods of mapping space. To complete the task of seismic data generation, we need to find the spatial distribution pattern P according to the discrete real seismic data points and get the optimal parameter combination of continuous data space:

$$\theta^* = \arg\max_\theta \prod_{n=1}^{N} P\left(x^n, \theta\right) \quad (2)$$

where N is the data size.

Since earthquake sequence x in the spatial dimension $i = \{1,\ 2, 3\}$ (i represents three different dimensions, namely east-west, north-south, and up-down) and time series $N = \{1, 2, \ldots, n\}$, θ is the parameter space that satisfies the mapping relation, the maximum likelihood function Eq 3 is used to get the optimal parameter combination:

$$\theta^* = \arg\max_\theta \Pi_{n=1}^{N} \prod_{i=1}^{3} \underbrace{P\left(x_i^N \mid x_i^{1\to n-1}, \theta_f\right)}_{\text{feature extract}} \cdot \underbrace{P\left(y_i^n \mid x_i^n, \theta_g\right)}_{\text{data generation}}, \quad \theta = \left\{\theta_f, \theta_g\right\} \quad (3)$$

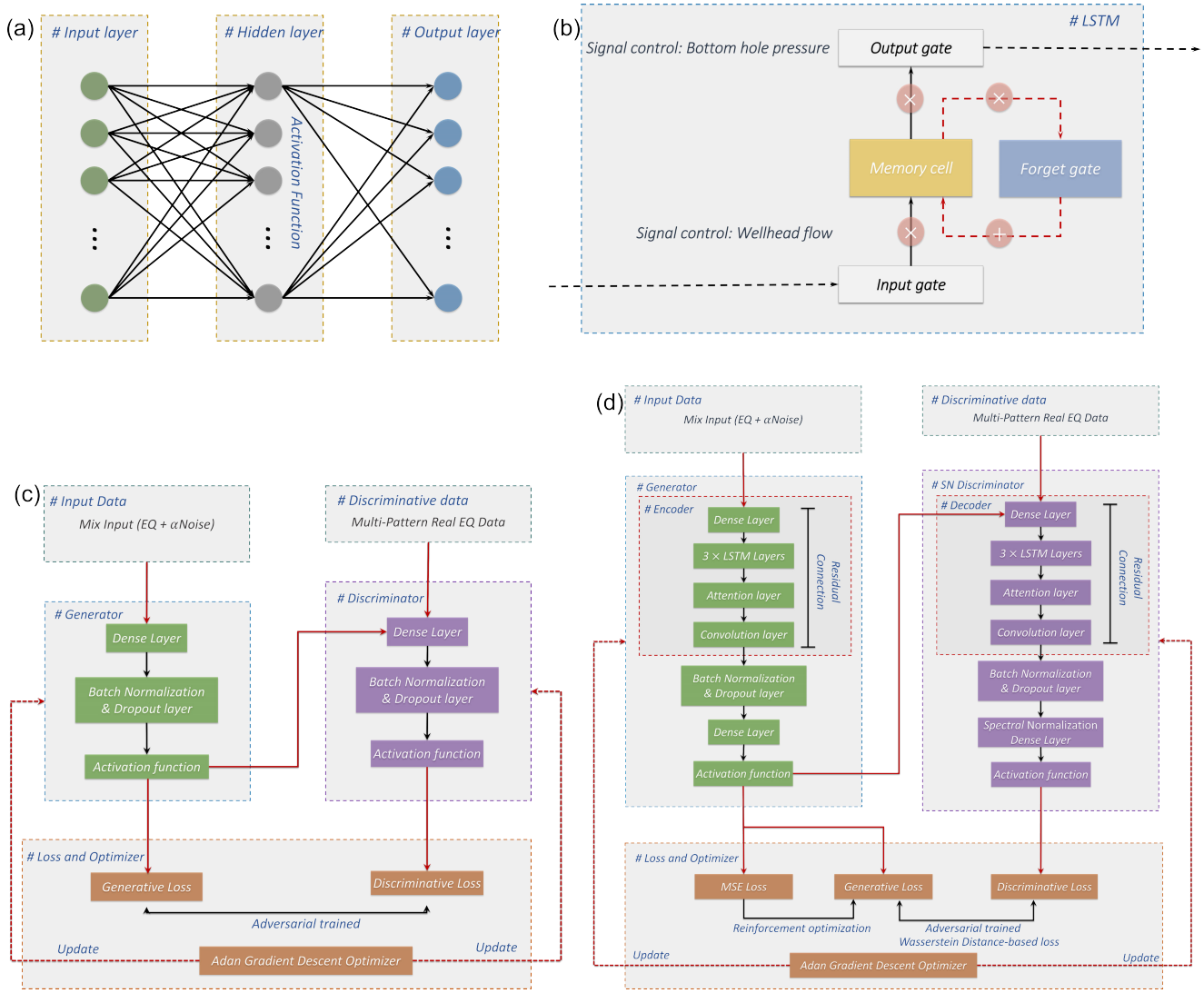


FIGURE 1. System structure of different generation models. (a) NN, (b) LSTM, (c) Normal GAN, (d) EQGAN.

However, it is infeasible to directly solve the equation to measure the evolution process of time series, and we introduce Kullback-Leibler (KL) Divergence [37], [38]:

$$KL\left(\mathbb{P}_r \| \mathbb{P}_g\right) = \int_{\mathcal{X}} P_r(x) \log \frac{P_r(x)}{P_g(x)} \mathrm{d}x \tag{4}$$

The optimal parameter combination can be obtained by minimizing the KL divergence of the data distribution generated P_g and the real data distribution P_r.

C. ALGORITHM AND MODEL DESIGN

In order to expand the scope of seismic sequence analysis and extract the long-range as well as short-range spatiotemporal correlation characteristics of the earthquake wave evolution process, we introduce the LSTM [39] algorithm (Fig.1b) to capture the invisible evolution relationship among the earthquake sequences. Keep the crucial data and information that need to be memorized for a long time and forget the unimportant information. Input information x at time i is represented by Eq 5:

$$x = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{5}$$

where $\sigma(*)$ is a deterministic nonlinear sigmoid function, while W_i and b_i are the weight and deviation of output units, respectively, which has only one transfer state h_t compared with the traditional RNN, LSTM is calculated by changing the operation mode of neurons, based on the input x and the last time hidden layer transmission h, to control the transmission process of the data hidden state relying on the compling work among the three gating units, as is shown in the following Eq 6:

$$h_t = \underbrace{\sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)}_{\text{output gate}} \cdot \underbrace{\tanh\left(\underbrace{f_t \times C_{t-1}}_{\text{forget gate}} + \underbrace{i_t \times \tanh\left(\tilde{C}_t\right)}_{\text{input gate}}\right)}_{\text{state unit}} \tag{6}$$

Among them, f_t, i_t, and C_t represent forgetting gate, input gate, and new cell state, respectively. In Eq 6, the forgetting unit keeps the previously accumulated sequence information correct, and the information input of the current time is added by the input gate, which effectively captures the multi-range pattern of the time series. It is highlighted that the generation of seismic data is neither a noise version nor a copy of real data, and it is difficult to generate high-quality seismic data by retaining the correlation with earthquake data evolution process, considering that the seismic waveform results from a multi-dimensional time series. Therefore, to obtain the correlation among different dimensions, we compiled the multi-head attention mechanism [40], [41] and use the weight of the attention vector as the approximate value of the target (The attention mechanism is detailed in Appendix 2). In addition, to further enhance the local feature extraction of earthquake waveforms, the NN algorithm (Fig.1a) is developed.

In the earthquake data generation process, measuring the distance among the low-dimensional manifolds of the distribution patterns in a high-dimensional hidden space more accurately drives the generated sequences towards the objective function. On the other hand, the highly scattered data points in the training dataset lead to an irregular gradient transmission, which cannot guarantee the training's stability. Therefore, it is unreasonable to train GAN by minimizing KL divergence to make the two distributions approach each other. We introduce Wasserstein distance [42], [43] to describe the similarity between the distribution of generated and real data, so as to solve this problem, the specific theory on WD is explained in Appendix 3.

$$\mathrm{W}\left(P_r, P_g\right) = \inf_{\gamma \sim \Pi\left(P_r, P_g\right)} \mathbb{E}_{(x,y)\sim\gamma}[\|\mathrm{x} - \mathrm{y}\|] \tag{7}$$

where $\Pi\left(P_r, P_g\right)$ is the set of all possible joint distributions of real data P_r and generated data P_g, samples x and y are obtained from each possible joint distribution γ, and the distance between the two samples is calculated, so the expected value of the sample pair distance under joint distribution γ can be obtained. In all possible joint distributions, the expected value from the sample to distance can be obtained, and the lower bound of the expected value is obtained: $\inf_{\gamma\sim\Pi\left(P_r,P_g\right)} E_{(x,y)\sim\gamma}\left[||\mathrm{x}-\mathrm{y}||\right]$ and its WD is found.

Compared with KL divergence, WD can provide smooth and meaningful distance even if two distributions are located in low-dimensional manifolds with no or less overlap. It can effectively measure the pattern difference in the submanifolds of high-dimensional distributions and describe the similarity of the two distributions, which fundamentally solves the problem of vanishing gradients.

To further ensure the stability of EQGAN model training, the discriminator must satisfy the Lipschitz constraint. Therefore, we use the spectral normalization method to normalize the spectral normalization [44] of weight matrix W to satisfy Lipschitz constraint $\sigma(W) = 1$:

$$\bar{W}_{\mathrm{SN}}(W) := W/\sigma(W) \tag{8}$$

where W is the total weight of EQGAN and $\sigma(*)$ is the maximum singular value. Unlike the commonly used methods in the calculation of gradients (such as gradient penalty), SNR operation $W_{S_N}(*)$ is performed before the backpropagation of the discriminator. The weight of different time series features is intelligently adjusted to avoid overlearning EQGAN iterative training, which retains the time correlation of the carefully designed weights, thus ensuring the high stability of generation and training process.

The system framework of our EQGAN model is summarized in Fig.1d, in which LSTM, attention, NN, SN, and WD are combined to realize the extraction and generation of seismic data features; it is developed through the platform of Python 3.6 and TensorFlow v2.0.

III. DATA COLLECTION AND PREPROCESSING

Data collection and preprocessing are the premises of model training and analysis. Here we present the experimental data source and preprocessing to better explain the model workflow.

A. DATASET

The occurrence of earthquake events is accidental and irreplaceable. With this in mind, we need a special data acquisition scheme. The earthquake datasets used for model training are mainly from the National Research Institute for Earth Science and Disaster Resilience (NIED) [45] databases, we also integrated the seismic data recorded by our sensors into the dataset. The earthquake events with magnitudes ranging from 4 to 8 recorded from April 2009 to May 2019 were selected from the NIED database and preprocessed to convert them into units (g). Also, the data of 120 stations for the three earthquakes of Tottori (2000) (M6.61), Niigata (2004) (M6.63), and Chuetsuoki (2007) (M6.8) were downloaded from the United States Geological Survey (USGS) database [46]. In addition, the earthquake data also include small events (2020) (approximately M2.5) recorded by our MEMS sensor in Korea. The sampling rate of all earthquake events is 100 Hz. The data is shown in three channels titled X, Y, and Z, where X (east-west) and Y (north-south) are horizontal components, and Z (up and down) is the vertical component. Noise data is a time series of non-earthquake datasets recorded byour low-cost sensors for several hours. We used two kinds of non-earthquake data in the experiment, namely human activity data and noise data. Human activity data include cars (hands), rope skipping, running (hands, pockets), table shaking (when moving above), climbing stairs (up and down), walking (bags, hands, pockets), standing still, and working. Instead, noise data includes floor noise (for example, different elevation angles) and mechanical noise. These noise data are external source data.

B. DATA PREPROCESSING AND TRAINING DETAILS

To facilitate training and analysis, we preprocess the seismic data. Each data (they are all recorded at a sampling rate of 100 data points per second) only retains 3600 data points

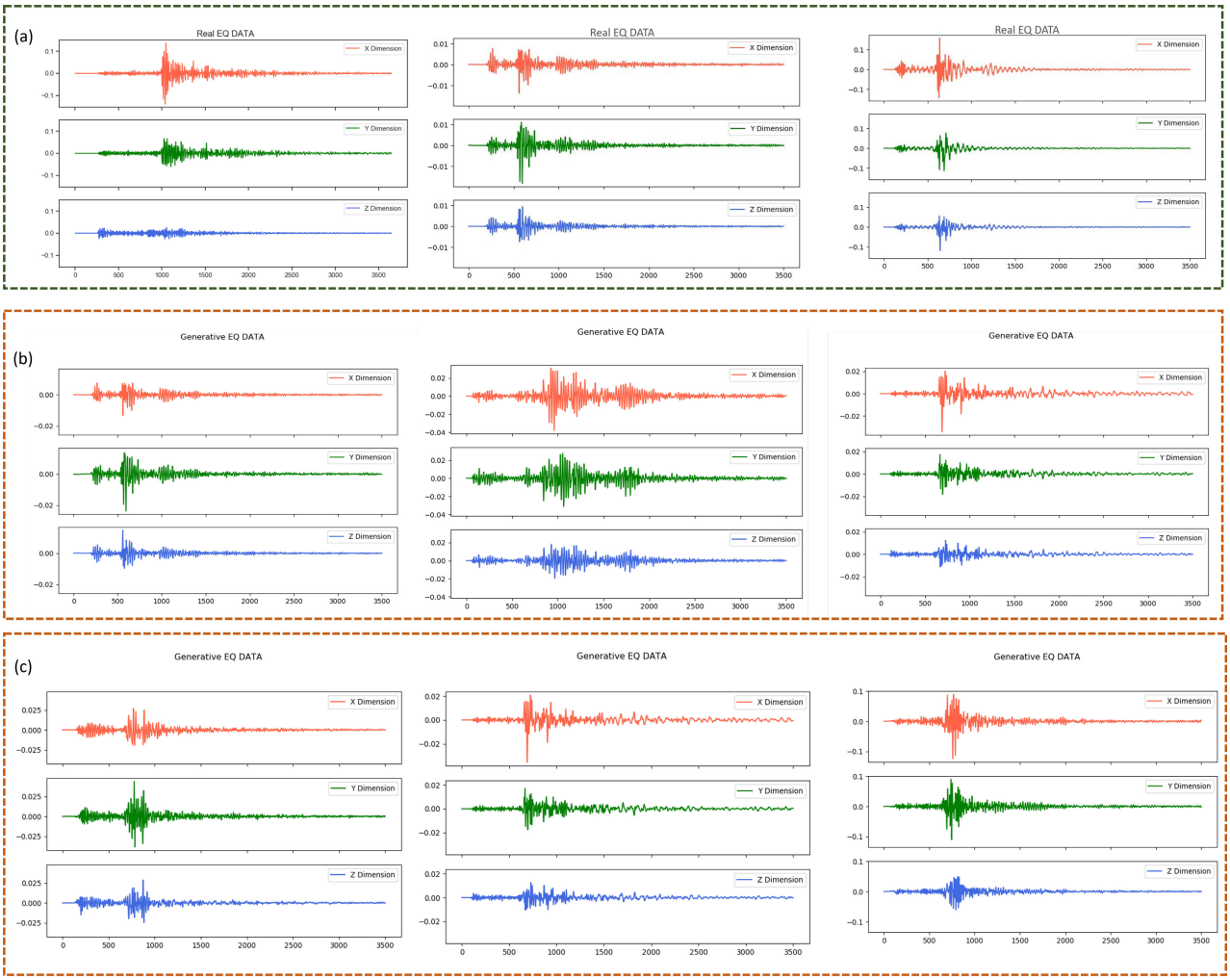


FIGURE 2. **The diversity and fidelity of the data generated by our EQGAN model, (a) The fundamental waveform of seismic data, (b) and (c) represent the waveform of generated data.**

included the P-wave and S-wave (including 50 abnormal data points at the two data endpoints), and only obtains the final length of 3500 data. The training and test dataset are divided according to the ratio of 7:3. The experiment was carried out on Ubuntu 18.04 operating system, and the learning rate was set to 10^{-5}, the batch size and epoch were set 64 and 1000, respectively. In order to improve the stability of model training, in addition to compiling the WD algorithm, in the Input layer, we have mixed the real earthquake data with random noise as input data (Appendix 4 for the schematic diagram) to improve the robustness of our model, and the data generated by our model is the same type of data length in output. We mainly use diagram to intuitively present the evolution process of seismic waveforms.

IV. RESULTS AND DISCUSSION

A more important fact is how to evaluate the quality of generated data. With all things considered, we design a variety of evaluation schemes in this section:

(1) Initially, we analyze the performance of the EQGAN model to generate data through visual appearance.

(2) Then, we compare and analyze the frequency domain of generated and real data.

(3) From the perspective of seismic data distribution pattern, we introduce a paired scatter plot to analyze the distribution of generated data points and real data or noise data and the correlation among different channels. At the same time, we also compare the performance of other generation models.

(4) Also, to make the evaluation more reliable, based on the statistical analysis index, we introduce the Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and WD error quantification index and use the High-throughput Screening Theory to design a novel generation model quantitative accuracy evaluation method.

(5) Finally, we compare the computational complexity of different models and discuss the robustness of our model from the perspective of computational cost.

Through an analysis of generated data, real data, and noise data, results of our model cast a new light on augmenting seismic data.

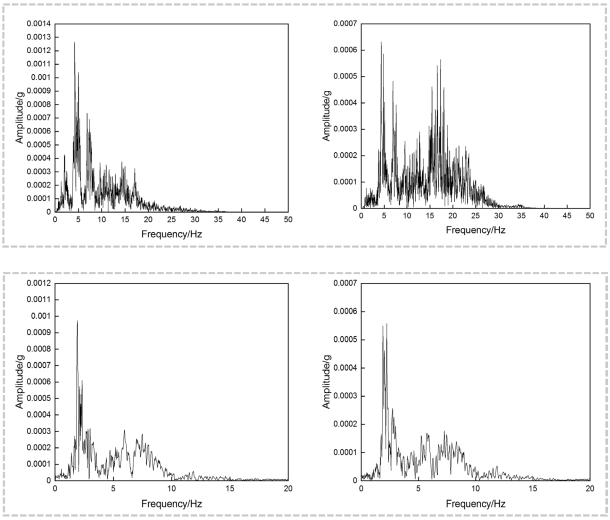


FIGURE 3. The comparision of frequency domain between real earthquake data and generated data, (a) Real data frequency domain (b) Generate data frequency domain.

A. VISUAL APPEARANCE

The accuracy of earthquake detection depends not only on the first arrival of earthquake waves but also the amplitude and frequency. Therefore, visual performance of generated data is one of the primary indicators to evaluate the quality of generated data. Fig.2b-c highlights the seismic data generated by our EQGAN model, which presents similar characteristics to that of the real seismic waveform. It is clear to observe the arrival time of P-wave and S-wave in different dimensions of generated data. It is worth noting that compared with Channel Z, the amplitude of S-wave after arrival in Channel X and Y is more prominent, which indicates that the EQGAN model we designed has taken into the essential statistical characteristics of real earthquake data and can generate seismic waveforms that are highly similar to the appearance of real data.

From the perspective of generating data diversity, in a real earthquake sequence, some earthquake waves contain small vibrations such as foreshocks or aftershocks, which is in line with the scientific nature of seismology. We can not only generate a single epicenter or aftershock, but also capture the characteristics of a single epicenter. Simultaneously, in Fig.2c, it is evident that the amplitude of generated data is also significantly different. The presentation of these different data features can prove the diversity of the data generated by our generation model, which is highly coincide with the real seismic data recorded by acceleration sensors.

B. FREQUENCY DOMAIN ANALYSIS

To further confirm the quality of generated data, we use the Fast Fourier Transform (FFT) to obtain its frequency domain [47] and the real data (Fig.3). Because of the frequency synchronization between generated and real data, it is not easy to represent the frequency domain of the whole dataset through an exact measurement in the frequency domain. However, we can evaluate the fluctuation in the frequency range by randomly selecting 100 real data samples and generating data for testing. As can be seen, the frequency domain of real data is maintained at 0-40 Hz. Similarly, most of the frequency fields of generated data are maintained in the same range, and there is no false data found beyond the real data frequency domain, which further suggests that real data and generated data are highly similar in the frequency range.

C. AUTOCORRELATION DISTRIBUTION ANALYSIS

Through further exploration, with another evaluation method, we randomly select a piece of data from the corresponding dataset and use the scatter matrix diagram to visualize data analysis. Fig.4 can be divided into two parts: the scatter diagram shows that in all kinds of data (real data, generated data and noise data), any two channels of X, Y and Z are paired to measure the correlation of them. The Kernel Density Estimation (KDE) represents the autocorrelation distribution of particular channel data, and the horizontal as well as vertical axes correspond to data points of the channels.

Since the distribution of data points is a relatively scattered and weak correlation due to the difference in the amplitude of P-wave and S-wave, in contrast, the distribution of non-seismic data points is uniform and concentrated. It is evident in the autocorrelation distribution map that both real and generated data present an approximate Gaussian distribution pattern on Channel X, Y and Z, which is quite different from the distribution of non-seismic data. No matter it is a scatter diagram or a distribution diagram, we can undoubtedly find that data generated by our model is similar to the real data, which is consistent with what has been found in the above analysis. Also, we dissect a standard GAN model, LSTM, and NN model with paired scatter diagram (Fig.4d-f). As we expected, the distribution of data generated by the standard GAN model is very similar to the real data, nonetheless, from the perspective of the sequence evolution process, which does not conform to the natural characteristics of the P-wave and S-wave. On the contrary, the autocorrelation distribution of data generated by the LSTM model is not different from the real data; the correlation scatter diagram of data generated by the LSTM model presents an obvious positive correlation among three channels, which is contrary to that of real seismic data. The data of correlation scatter diagram and autocorrelation scatter diagram generated by the NN model is different from real data. There are significant differences between the correlation distribution map and the real data.

Even if this method can reflect the excellent quality of the data generated by EQGAN model, it is difficult to distinguish the false positive data generated by standard GAN and NN model. Moreover, it can only be used to analyze randomly selected single data instead of evaluating a dataset as a whole.

D. COMPARATIVE ANALYSIS OF DIFFERENT GENERATION MODELS

Although previous evaluation methods can be used to verify the potential of EQGAN in earthquake sequence generation,

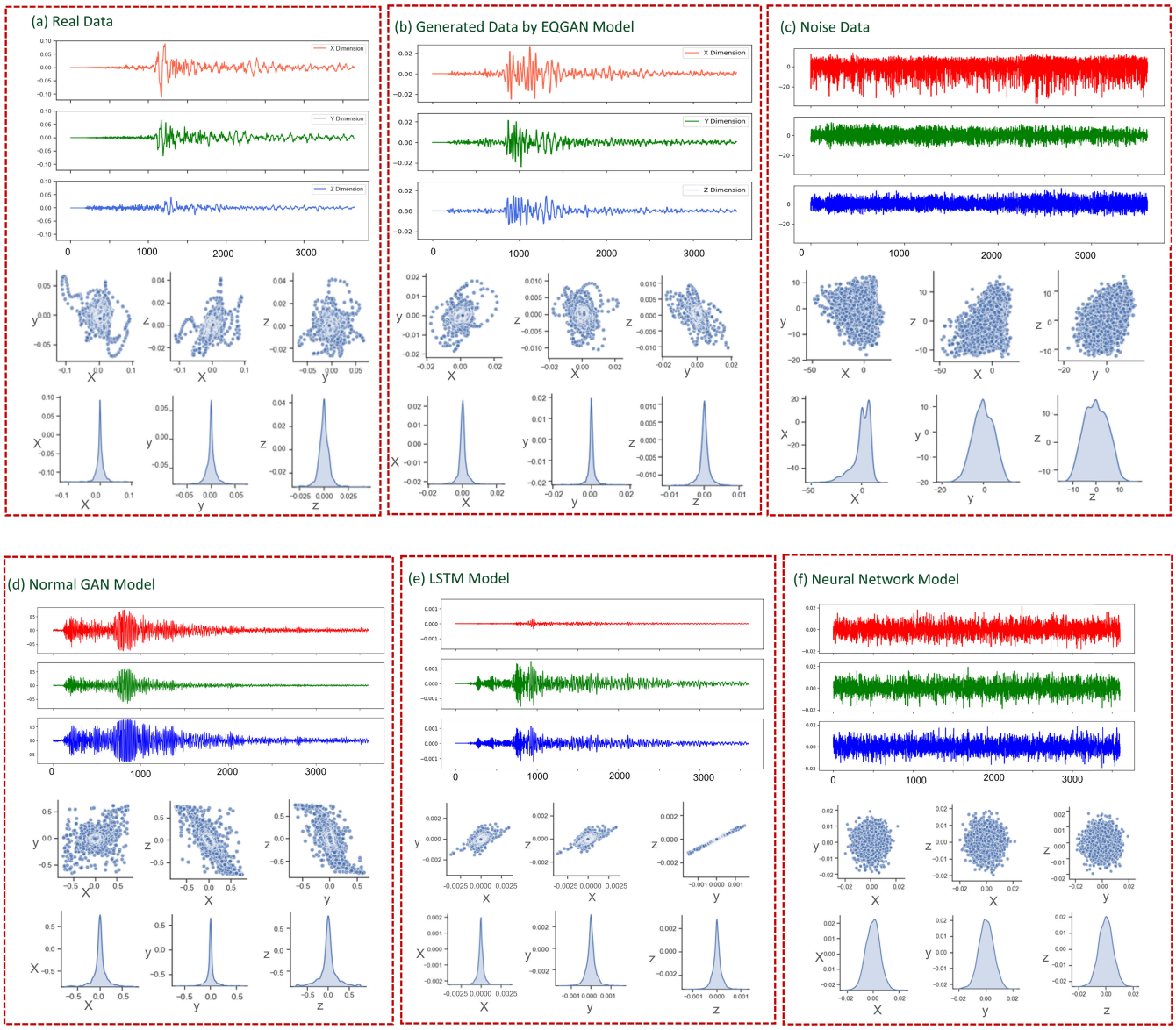


FIGURE 4. **Different data distribution patterns including real data, noise data, and generated data by different generation models.**

one limitation of our implementation is that they all be used to qualitatively evaluate the quality of individual generated data, which is also a common weakness in the evaluation of many ML model. In this research, to further clarify excellent performance of our EQGAN model, we design a scheme to quantitatively verify the generated data based on MSE, MAPE, and WD meanwhile evaluating the performance of different generation models. (i) Firstly, eight samples of representative seismic data is selected from all real datasets as the standard dataset.

(ii) Then, calculate the MSE of sample dataset and standard dataset:

$$MSE = \sum_{i=1}^{N} \frac{(c_i - r_i)^2}{N} \tag{9}$$

where c_i represents the sample dataset, r_i is the standard dataset, and N represents the data length. We use real data, noise data, and data generated by different models as sample datasets to obtain MSE (Eq 9) corresponding to standard datasets.

(iii) The mean value, minimum value, and standard deviation of the MSE vector are extracted as characteristic parameters for experimental confirmation. Different characteristic parameters show similar distribution patterns. Fig.5a displays the distribution difference between each sample dataset and the real dataset when MSE is minimized. It can be seen that the distribution of the dataset generated by EQGAN has the highest similarity with the real dataset, and there are cliff-like differences between the distribution of other sample datasets and the real dataset.

FIGURE 5. Distribution diagram of error quantification index. (a) Minimizing MSE, (b) Minimizing MAPE, (c) Minimizing WD. Here, the y-axis (Vertical) represents the distribution probability, and the x-axis (Horizontal) shows the value of the corresponding error-index.

Comparing the models generated by different ML algorithms, the results reveal that the overall similarity between different sample datasets and the real dataset is as follows: EQGAN > GAN > LSTM > NN.

Although MSE strongly indicates the actual situation of the error between generated and real data, it is not convincing to judge the generation ability from the value of MSE alone. Consequently, we handle the same scheme to calculate the mean absolute percentage error (MAPE) of the sample dataset and the standard dataset, respectively (Eq 10):

$$MAPE = \frac{\sum_{i=1}^{N} \frac{|c_i - r_i|}{r_i}}{N} \times 100\% \tag{10}$$

MAPE is a statistical index to measure the accuracy of prediction, which considers the error between predicted and actual value as well as the ratio between the error and actual value. It is generally believed that the closer the MAPE is to 0, the higher the similarity between the two groups of

data will be. By calculating the MAPE of different generated, real, noise and standard datasets, the results explain that the distribution pattern of real and generated datasets is very similar. Fig.5b confirms the similarity between the dataset generated by different models and the real dataset: EQGAN > GAN > LSTM > NN. From the statistics perspective, both MSE and MAPE are widely applicable to the quantitative error analysis of data. WD is a special indicator to measure the difference of probability distribution between two pieces of high-dimensional data. Accordingly, we work the corresponding scheme to calculate the WD (Eq 11) and further analyze the quality of generated data. The results are exhibited in Fig.5c.

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^p d\gamma(x, y) \right)^{1/p} \quad (11)$$

where γ represents the joint distribution of real data x and generated data y, which is called coupling and requires that the edge distribution is μ and ν.

Despite the fact that the previous evaluation method has been used to fully explain and examine the quality of generated data, to prove the robustness and stability of EQGAN, we propose a High Throughput Screening (HTS) Theory to analyze the performance of different generation models. The HTS technology is an essential means of drug research and development based on experimental methods at the molecular and cellular level, in which microplate is used as the experimental tool carrier to screen high-quality data, so as to meet the needs automatically [48]–[50]. The quality of data screening depends on the design of the microplate, and some data will show false-positive results with different types of microplate screening [51], which is entirely consistent with the error evaluation scheme of MSE, MAPE and WD designed by us. Fig.6 presents our basic screening process of generated data. In view of the significant difference in data generated by different models, in order to eliminate the dimensional influence among different error indexes and obtain the comparability among seismic datasets, we normalized all datasets under different error quantitative indexes. It is worth mentioning that normalization will reduce the differences among the data generated by different models and change their distributions. Hence, in this paper, we try different normalization methods for MSE, MAPE and WD, choosing the best method to process the data (Fig.7).

Compared with Fig.5 and Fig.7, it stands to reason that we can find that the generated datasets with normalization have a higher similarity with the real dataset, which does not mean that their distribution pattern is changed in the normalization process. Still, differences in the data are reduced, which does not affect the scientific nature of the statistical analysis. Therefore, through Fig.5 and Fig.6, we can conclude that the GAN framework training generation task is better than a single algorithm model, and generation performance of our EQGAN model is better than that of standard GAN.

Furthermore, based on MSE, MAPE and WD, we calculate the correlation between the sample dataset generated by

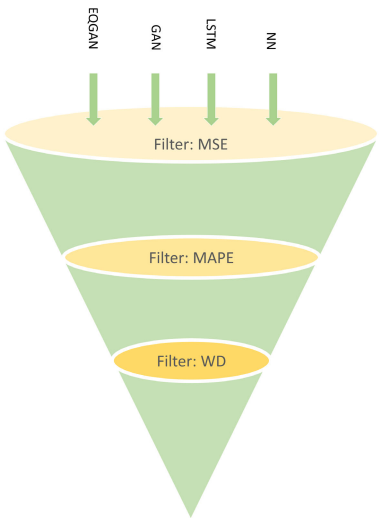


FIGURE 6. High-throughput screening process.

different models and the real dataset. The scatter plot matrix directly reveals the correlation between the datasets generated by different models and the real datasets under different quantitative indexes meanwhile the incidence matrix is used to quantify and summarize the linear strength relationship between the datasets (Fig.8a). It is observed that the correlation coefficient between the dataset generated by EQGAN and the real dataset is 0.11. Although it looks deficient, it is much higher than that of other generation models, which shows that the generation performance of the EQGAN model is not only high but also fully reflects that the data generated by EQGAN is not a copy of the real data.

It's noteworthy that based on the above analysis results, we use each generation model to complete 10 consecutive generation tasks, generating 2,000 data samples each time, and further verify the performance of different models. The error bars diagram (Fig.8b) shows the datasets generated based on different models and gives the maximum, minimum, and average value of MSE, MAPE and WD, respectively. At the same time, it allows us to master the efficiency and stability of the models. Through EQGAN, the generation task can be completed more stably under MSE, but the generation performance of GAN is the highest. The reason is that there will be false-positive data after the normalization of the data generated by GAN, which improves the efficiency. Simultaneously, in a more complex evaluation index map under MAPE and WD, the results show that the stability and accuracy of the data generated by the model with LSTM and NN algorithm are the best.

Finally, through the HTS method, we filter 2000 data samples generated by each model respectively according to the increasing complexity of MSE, MAPE and WD. Fig.8c shows that the generation performance of different models is: EQGAN (81%) > GAN (74%) > NN (21%) > LSTM (2%), implying that EQGAN possesses a strong generalization and stability to deal with distinct diversity earthquake series intelligently.

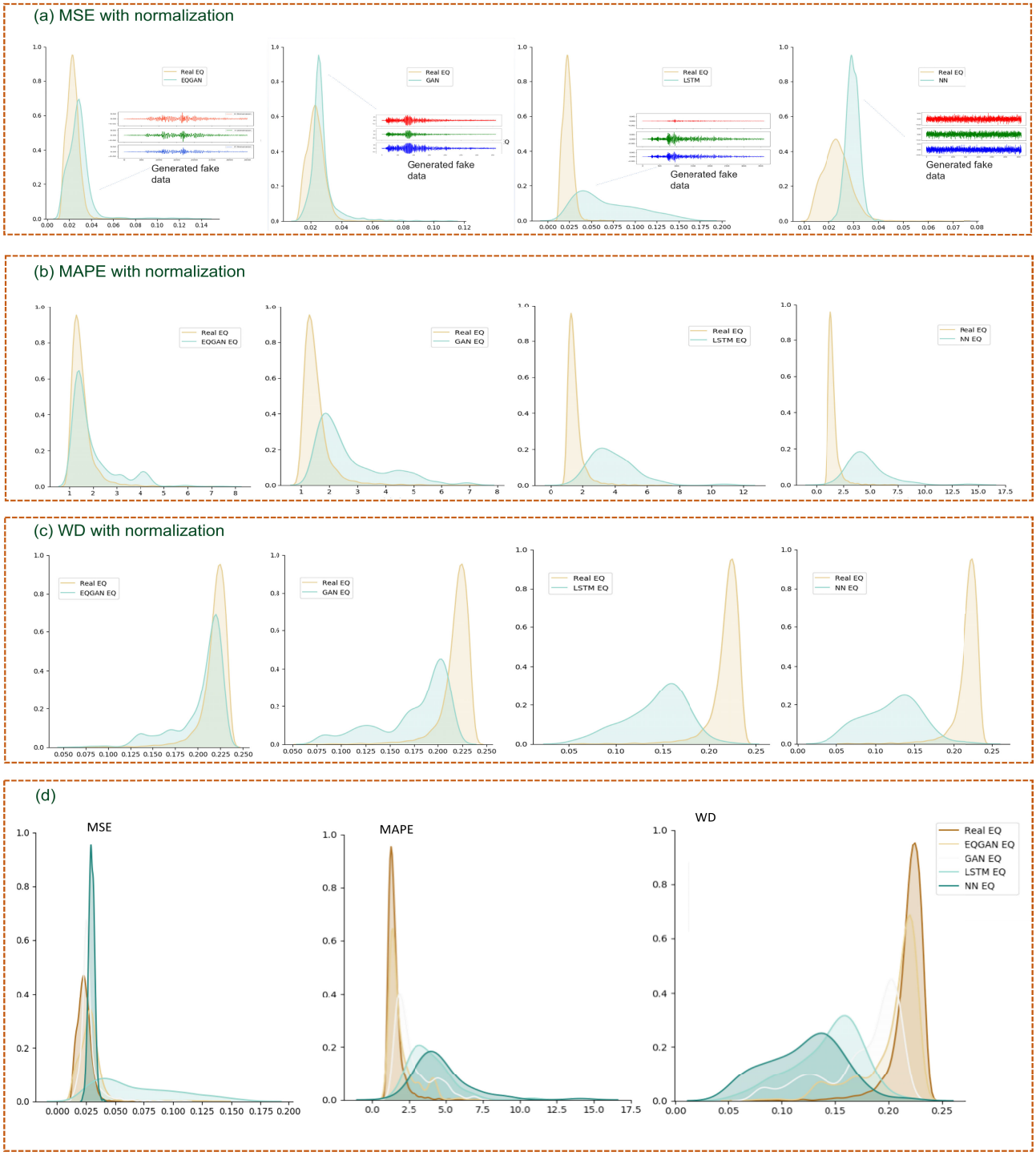


FIGURE 7. Distribution of different error quantification indexes with normalization. Here, the y-axis (Vertical) represents the distribution probability, and the x-axis (Horizontal) shows the value of the corresponding error-index.

E. COMPUTATIONAL COMPLEXITY

One may expect that our EQGAN model with different algorithms would have high computational complexity. This is, however, not the case. To measure the quality of an algorithm, there are usually three considerations: (i) The time consumed in the execution of the algorithm (ii) The number of resources occupied during execution, such as the amount of memory space occupied (iii) The algorithm is easy to understand, implement and verify Therefore, different algorithms need to be selected in different cases. Due to a large amount of data processing, we first consider the difficulty of data processing in LSTM or GAN, which is not different from that

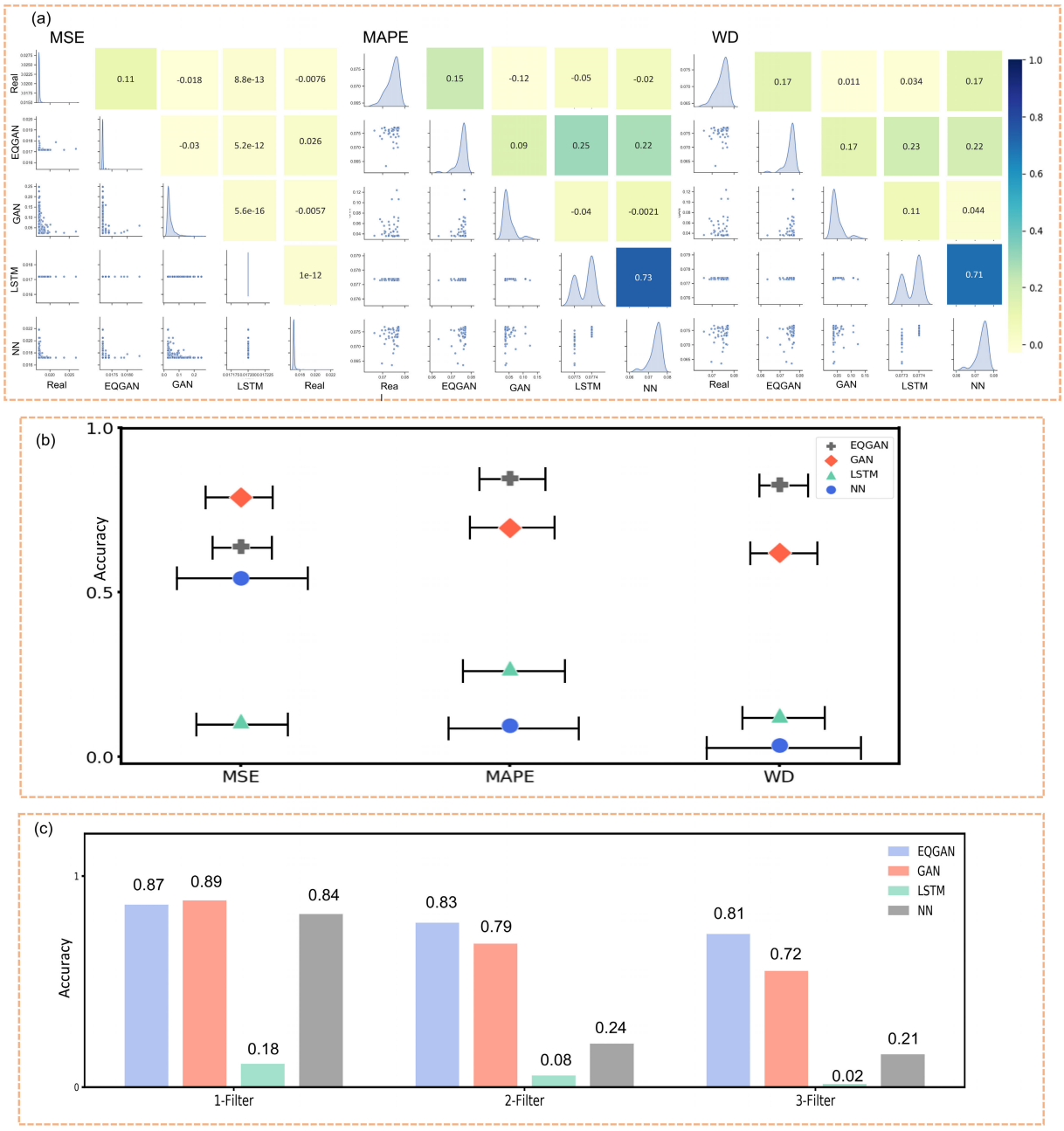


FIGURE 8. Stability and performance analysis from different models. (a) shows the correlation between datasets generated by different models and real datasets, (b) Comparison of performance and stability of different models under different filter screening conditions, (c) Accuracy analysis of the same amount of data from different generation models filtered by different filters (1-filter represents MSE, 2-filter denotes MSE+MAPE, 3-filter denotes MSE + MAPE + WD).

in EQGAN. Still, EQGAN has apparent advantages in time complexity and easy implementation of the algorithm. This paper mainly discusses the time complexity of the algorithm and its feasibility. Appendix 5 gives the measurement indexes of different algorithm complexity.

V. CONCLUSION

A new DGM called EQGAN is proposed in this research to capture the multi-dimensional temporal evolution of seismic sequences and generate high-quality seismic sequences containing P-waves and S-waves. In order to verify the performance of the EQGAN model, by comparing standard GAN, NN and LSTM model, we not only qualitatively evaluate the quality of generated data from the distribution pattern and frequency domain, but also quantitatively analyze the similarity between generated and real data by fusing statistical indexes of MSE, MAPE and WD with seismic data. Experimental results show that the efficiency of data generated by our EQGAN model reaches 81% (The generation performance of standard GAN, LSTM and NN models

are 72%, 2%, and 21%, respectively), which further demonstrate that our generation model has excellent performance and stability.

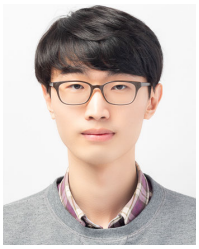
Even if the current discussion is not as easy to be explained as the traditional supervised training models, the data screening and evaluation scheme based on the HTS theory and techniques are highly consistent with the distribution pattern of seismic data. There is no apparent defect to prevent the expansion of the EQGAN model, which also promotes the application and innovation of ML algorithms in seismology. These findings provide a potential mechanism for data augmentation. We also assume that the EQGAN algorithm may generate seismic sequences similar to that recorded by a specific position sensor, which provides a more convenient data support scheme for earthquake prediction. Looking forward, the proposed EQGAN model provides a more convenient dataset support scheme for seismic prediction. Based on generated data, we will further develop and train the earthquake detection model to improve the accuracy and robustness of the EEW system. Moreover, with fault-detection/identification techniques as a future research direction [52], we will design fault and detection equipment for regulating and maintaining sensors under abnormal conditions to improve the quality of the data recorded by our sensors.

REFERENCES

[1] G. Madureira and A. E. Ruano, "A neural network seismic detector," *IFAC Proc. Volumes*, vol. 42, no. 19, pp. 304–309, 2009.
[2] S. M. Mousavi, W. Zhu, Y. Sheng, and G. C. Beroza, "CRED: A deep residual network of convolutional and recurrent units for earthquake signal detection," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, Dec. 2019.
[3] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *Sci. Adv.*, vol. 4, no. 2, Feb. 2018, Art. no. e1700578.
[4] R. M. Allen and H. Kanamori, "The potential for earthquake early warning in southern California," *Science*, vol. 300, no. 5620, pp. 786–789, 2003.
[5] Z. Li, M.-A. Meier, E. Hauksson, Z. Zhan, and J. Andrews, "Machine learning seismic wave discrimination: Application to earthquake early warning," *Geophys. Res. Lett.*, vol. 45, no. 10, pp. 4773–4779, May 2018.
[6] J. Botero-valencia, L. Castano-Londono, D. Marquez-Viloria, and M. Rico-Garcia, "Data reduction in a low-cost environmental monitoring system based on LoRa for WSN," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3024–3030, Apr. 2018.
[7] D. Ciuonzo, S. H. Javadi, A. Mohammadi, and P. S. Rossi, "Bandwidth-constrained decentralized detection of an unknown vector signal via multi-sensor fusion," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 744–758, 2020.
[8] M. Faulkner, M. Olson, R. Chandy, J. Krause, K. M. Chandy, and A. Krause, "The next big one: Detecting earthquakes and other rare events from community-based sensors," in *Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2011, pp. 13–24.
[9] D. Li, R. Landry, and P. Lavoie, "Low-cost MEMS sensor-based attitude determination system by integration of magnetometers and GPS: A real-data test and performance evaluation," in *Proc. IEEE/ION Position, Location Navigat. Symp.*, May 2008, pp. 1190–1198.
[10] C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," *Sci. Adv.*, vol. 1, no. 11, Dec. 2015, Art. no. e1501057.
[11] K. Rong, C. E. Yoon, K. J. Bergen, H. Elezabi, P. Bailis, P. Levis, and G. C. Beroza, "Locality-sensitive hashing for earthquake detection: A case study of scaling data-driven science," 2018, *arXiv:1803.09835*.
[12] X. Huang, J. Lee, Y.-W. Kwon, and C.-H. Lee, "CrowdQuake: A networked system of low-cost sensors for earthquake detection via deep learning," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3261–3271.
[13] Y. Chen, "Automatic microseismic event picking via unsupervised machine learning," *Geophys. J. Int.*, vol. 222, no. 3, pp. 1750–1764, 2020.
[14] Y. Yu, J. Lin, L. Zhang, G. Liu, J. Hu, Y. Tan, and H. Zhang, "Identification of seismic wave first arrivals from earthquake records via deep learning," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Cham, Switzerland: Springer, 2018, pp. 274–282.
[15] W. Zhu and G. C. Beroza, "PhaseNet: A deep-neural-network-based seismic arrival-time picking method," *Geophys. J. Int.*, vol. 216, no. 1, pp. 261–273, 2019.
[16] J. Wang, Z. Xiao, C. Liu, D. Zhao, and Z. Yao, "Deep learning for picking seismic arrival times," *J. Geophys. Res., Solid Earth*, vol. 124, no. 7, pp. 6612–6624, 2019.
[17] L. Küperkoch, T. Meier, J. Lee, and W. Friederich, "Automated determination of P-phase arrival times at regional and local distances using higher order statistics," *Geophys. J. Int.*, vol. 181, no. 2, pp. 1159–1170, May 2010.
[18] S. M. Mousavi and G. C. Beroza, "Bayesian-deep-learning estimation of earthquake location from single-station observations," 2019, *arXiv:1912.01144*.
[19] I. Khan, S. Choi, and Y.-W. Kwon, "Earthquake detection in a static and dynamic environment using supervised machine learning and a novel feature extraction method," *Sensors*, vol. 20, no. 3, p. 800, Feb. 2020.
[20] J. Wang and T.-L. Teng, "Identification and picking of s phase using an artificial neural network," *Bull. Seismol. Soc. Amer.*, vol. 87, no. 5, pp. 1140–1149, Oct. 1997.
[21] Q. Kong, R. M. Allen, L. Schreier, and Y.-W. Kwon, "MyShake: A smartphone seismic network for earthquake early warning and beyond," *Sci. Adv.*, vol. 2, no. 2, Feb. 2016, Art. no. e1501055.
[22] Z. E. Ross, M.-A. Meier, E. Hauksson, and T. H. Heaton, "Generalized seismic phase detection with deep learning," *Bull. Seismol. Soc. Amer.*, vol. 108, no. 5A, pp. 2894–2901, 2018.
[23] Y. Wu, Y. Lin, Z. Zhou, D. C. Bolton, J. Liu, and P. Johnson, "DeepDetect: A cascaded region-based densely connected network for seismic event detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 62–75, Jan. 2018.
[24] R. M. Dokht, H. Kao, R. Visser, and B. Smith, "Seismic event and phase detection using time–frequency representation and convolutional neural networks," *Seismol. Res. Lett.*, vol. 90, no. 2A, pp. 481–490, 2019.
[25] O. M. Saad and Y. Chen, "Automatic waveform-based source-location imaging using deep learning extracted microseismic signals," *Geophysics*, vol. 85, no. 6, pp. KS171–KS183, Nov. 2020.
[26] O. M. Saad, G. Huang, Y. Chen, A. Savvaidis, S. Fomel, N. Pham, and Y. Chen, "SCALODEEP: A highly generalized deep learning framework for real-time earthquake detection," *J. Geophys. Res., Solid Earth*, vol. 126, no. 4, Apr. 2021, Art. no. e2020JB021473.
[27] L. Zhu, Z. Peng, J. McClellan, C. Li, D. Yao, Z. Li, and L. Fang, "Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M_w7. 9 Wenchuan earthquake," *Phys. Earth Planet. Interiors*, vol. 293, Aug. 2019, Art. no. 106261.
[28] O. M. Saad and Y. Chen, "Earthquake detection and P-wave arrival time picking using capsule neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6234–6243, Jul. 2021.
[29] Y. Li, B. Ku, S. Zhang, J.-K. Ahn, and H. Ko, "Seismic data augmentation based on conditional generative adversarial networks," *Sensors*, vol. 20, no. 23, p. 6850, Nov. 2020.
[30] T. Wang, Z. Zhang, and Y. Li, "EarthquakeGen: Earthquake generator using generative adversarial networks," in *Proc. SPIE, SEG Tech. Program Expanded Abstr.*, 2019, pp. 2674–2678.
[31] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 4, 2007, pp. 4–1229.
[32] S. Santhanam, "Context based text-generation using LSTM networks," 2020, *arXiv:2005.00048*.
[33] S. Zhao, E. Deng, M. Liao, W. Liu, and W. Mao, "Generating summary using sequence to sequence model," in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 1102–1106.
[34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*.
[35] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? A large-scale study," 2017, *arXiv:1711.10337*.
[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[37] D. I. Belov and R. D. Armstrong, "Distributions of the Kullback–Leibler divergence with applications," *Brit. J. Math. Stat. Psychol.*, vol. 64, no. 2, pp. 291–309, May 2011.
[38] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. IV–317.
[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[40] W. Schneider and A. D. Fisk, "Attention theory and mechanisms for skilled performance," in *Proc. Adv. Psychol.*, vol. 12, 1983, pp. 119–143.
[41] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.
[42] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio, "Learning with a wasserstein loss," 2015, *arXiv:1506.05439*.
[43] S. S. Vallender, "Calculation of the Wasserstein distance between probability distributions on the line," *Theory Probab. Appl.*, vol. 18, no. 4, pp. 784–786, 1974.
[44] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
[45] *National Research Institute for Earth Science and Disaster Prevention*. Accessed: Sep. 31, 2021. [Online]. Available: https://www.kyoshin.bosai.go.jp
[46] *Peer Ground Motion Database, Pacific Earthquake Engineering Research Center*. Accessed: Sep. 31, 2021. [Online]. Available: https://www.kyoshin.bosai.go.jp
[47] H. Katukura, S. Ohno, and M. Izumi, "Symmetrical fft technique and its applications to earthquake engineering," *Earthq. Eng. Struct. Dyn.*, vol. 18, no. 5, pp. 717–725, Jul. 1989.
[48] P. Gribbon, R. Lyons, P. Laflin, J. Bradley, C. Chambers, B. S. Williams, W. Keighley, and A. Sewing, "Evaluating real-life high-throughput screening data," *J. Biomolecular Screening*, vol. 10, no. 2, pp. 99–107, Mar. 2005.
[49] N. Malo, J. A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon, "Statistical practice in high-throughput screening data analysis," *Nature Biotechnol.*, vol. 24, no. 2, pp. 167–175, Feb. 2006.
[50] L. M. Mayr and D. Bojanic, "Novel trends in high-throughput screening," *Current Opinion Pharmacol.*, vol. 9, no. 5, pp. 580–588, Oct. 2009.
[51] Z.-Y. Yang, J.-H. He, A.-P. Lu, T.-J. Hou, and D.-S. Cao, "Frequent hitters: Nuisance artifacts in high-throughput screening," *Drug Discovery Today*, vol. 25, no. 4, pp. 657–667, Apr. 2020.
[52] H. Darvishi, D. Ciuonzo, E. R. Eide, and P. S. Rossi, "Sensor-fault detection, isolation and accommodation for digital twins via modular data-driven architecture," *IEEE Sensors J.*, vol. 21, no. 4, pp. 4827–4838, Feb. 2020.

AMING WU received the M.S. degree in mathematics and technology from The Education University of Hong Kong, China, in 2019. He is currently pursuing the Ph.D. degree in computer science and engineering with Kyungpook National University, South Korea. His research interests include deep learning and big data processing.



JUYONG SHIN received the B.S. degree in computer science and engineering from Kyungpook National University, South Korea, in 2021, where he is currently pursuing the M.S. degree in computer science and engineering. His research interests include distributed systems, system monitoring, and the Internet of Things.

JAE-KWANG AHN received the B.S., M.S., and Ph.D. degrees from Hanyang University, South Korea, in 2006, 2008, and 2017, respectively. He is currently a research officer at Korea Meteorological Administration. His research interests include Earthquake, Ground motion, GANs, EEW, Seismic Engineering, Sensors, and Liquefaction.

YOUNG-WOO KWON received the Ph.D. degree in computer science from Virginia Tech, in 2014. Prior to coming to KNU, he was an Assistant Professor with the Department of Computer Science, Utah State University. He is currently an Associate Professor with the School of Computer Science and Engineering, Kyungpook National University. His research interests include span mobile computing, cloud-based systems, the Internet of Things, and software engineering.