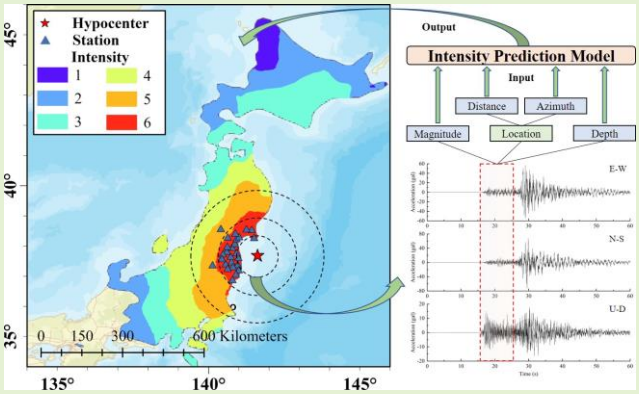
This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

Intensity Prediction Model Based on Machine Learning for Regional Earthquake Early Warning

Kaiwen Zhang, Fidel Lozano-Galant, Ye Xia, Jose Matos

***Abstract*—Seismic intensity plays a crucial role in influencing the decision-making process of users utilizing earthquake early warning systems (EEWs) upon receiving warning information. Improving intensity warnings' speed and accuracy is vital. We present a straightforward and dependable model for predicting intensity, which is based only on location and magnitude information. We use the catalog of intensity data from the Japan Meteorological Agency (JMA) released as a dataset, totaling 944,877 intensity instances. To address the issue of imbalanced dataset distribution, we employ the Synthetic Minority Over-Sampling Technique (SMOTE) as a means to improve this situation. Considering the distribution of high intensity data and the importance of features in the model, we construct and jointly apply intensity prediction models for magnitude below 5.7 and above 5.7, respectively. Further, we analyze the robustness of the model by adding random errors for magnitude and location information. We test the transfer capability of the proposed model with four earthquake events in China. Further, we use 466 seismic events (20,542 intensity instances) without published intensity data from the Kyoshin network (K-NET) as the application dataset. We simulate the phenomenon of underestimation of large earthquakes and overestimation of small earthquakes, which is used to analyze the possible application of the proposed model to EEWs. The findings indicate that the model achieves an accuracy of 97.77% when subjected to a magnitude error of 0.3 and a location error of 0.2°. Finally, we analyze the timeliness of the proposed model with a magnitude 7.4 event in 2022.**



***Index Terms*—Earthquake early warning, earthquake engineering, earthquake intensity, machine learning, seismic signal processing**

I. Introduction

EARTHQUAKE early warning systems (EEWs) leverage the velocity discrepancy between *P* wave and *S* wave, as well as between *S* wave and electromagnetic waves [1]-[3]. During an earthquake event, EEWs can provide alerts to the general public and specific users, such as high-speed railways, tall buildings, and slope construction projects, several seconds to approximately ten seconds ahead of time [4]-[8]. These alerts typically encompass information about the hypocenter's location, the earthquake's magnitude, predicted local intensity, and the expected arrival time of potentially destructive seismic waves [9]-[12]. Such data from EEWs furnish users with essential insights to inform potential emergency responses and risk-assessment decisions [13].

EEWs are categorized into regional and on-site warning systems [14]-[15]. On-site EEWs predict subsequent ground motion peaks based on local monitoring information for the first few seconds, thus providing warning information to users before destructive seismic waves arrive [16]. Conversely, regional EEWs dispatch alerts to distant users based on *P* wave data from stations proximate to the hypocenter [17].

Traditional methods for EEWs intensity prediction usually use empirical fitting. The initial *P* wave information is subjected to feature parameter extraction and a statistical relationship is established with the site motion amplitude. These *P* wave characteristic parameters include the peak displacement (P_d) [18]-[19] within the 3s time window of the initial *P* wave, the cumulative absolute velocity (CAV) [20], and so on. In addition to establishing relationships with waveform features, intensity prediction also establishes statistical relationships with magnitude, epicenter distance, and depth of epicenter to include diverse information to improve the accuracy of prediction [21]-[23]. However, the uncertainty of earthquake occurrence and the complexity of seismic wave propagation limit the accuracy and reliability of traditional empirical models. Therefore, it is necessary to propose more capable intensity prediction methods in order to obtain higher accuracy.

Machine learning algorithms have the advantage of being automated and adaptive, giving them the ability to build data mapping relationships without the aid of physical models [24]. This feature has led to the widespread use of machine learning algorithms in a variety of application areas, including earthquake engineering [25]-[27]. As machine learning

The paper is supported by the National Natural Science Foundation of China 52278313, and Project to Attract Foreign Experts G2023133018L. (*Corresponding author: Ye Xia*)

Kaiwen Zhang is with the Department of Bridge Engineering, Tongji University, Shanghai 200092, China (e-mail: zkevin259@163.com).

Fidel Lozano-Galant is with Department of Civil and Building Engineering, University of Castilla-La Mancha, Spain(e-mail: Fidel.LozanoGalant@uclm.es).

Ye Xia is Shanghai Qizhi Institute and Department of Bridge Engineering, Tongji University, Shanghai 200092, China(e-mail: yxia@tongji.edu.cn).

Jose Matos is with Department of Civil & Environmental Engineering, University of Minho, Portugal (e-mail: jmatos@civil.uminho.pt).

XXXX-XXXX © XXXX IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: U-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

continues to evolve, ensemble learning has excelled in different areas of application due to its efficient performance. Ensemble learning reduces the preference for a single model and further improves model adaptation by integrating different or identical machine learning models [28]. Currently, widely accepted models for ensemble learning include Bagging and Boosting.

Ensemble learning also currently has a wide range of applications in EEWs. For intensity prediction, Sarkar et al. [16] proposed a site peak prediction model using initial P-wave feature parameters based on multilayer perceptron. Hu et al. [29]proposed a peak site motion prediction model using initial P-wave feature parameters based on Extreme Gradient Boosting (XGBoost) algorithm. Abdalzaher et al. [30] used the initial time series data of seismic waves as input and intensity level as output and compared different machine learning models and identified the XGBoost model as the optimal model. In addition, some researchers have constructed local intensity prediction models using initial P-wave time series based on deep learning models [31]-[33]. However, the current machine learning or deep learning methods are based on local EEWs, and no further research has been carried out on the intensity prediction aspect of regional EEWs.

We statistically determine the time and epicentral distance of the maximum acceleration reaching other areas after the earthquake, with an average velocity of 3.17 km/s. This result is consistent with the velocity of destructive seismic *S* wave. According to our dataset, the epicentral distance of recorded intensity events averages at 155 km. Typically, the velocity of *P* wave is about 6.5 km/s, which also means that *P* wave takes an average of 23.84 s to trigger a local station after an earthquake. This also means that there is a certain lag in the use of on-site warning, but there is potential for further improvement in timeliness. High intensity may cause more damage to people's lives and property [29]. However, high intensity (≥ 5) is usually caused by a small number of large earthquakes, which suggests a serious imbalance between high and low intensity data. Nevertheless, no improvements have been made to address this problem in the current researches.

Addressing the issues of timeliness, intricate computation, and data imbalance in existing methods, this study introduces a streamlined yet dependable seismic intensity prediction model grounded in machine learning. The model hinges solely on magnitude and location data. We refer to the study of Abdalzaher [30] to construct and compare the intensity prediction models based on Bagging and Boosting algorithms, and compared the results of different algorithms on validation set and test set, and determine the best algorithm. Additionally, a random oversampling technique is employed to rectify the data imbalance, leading to enhanced performance with high-intensity data. The model's robustness and transfer capability are also assessed. Conclusively, simulations depict a decrease in the prediction error of seismic information for EEWs as time progresses, facilitating an analysis of the timeliness and precision of the model when integrated into EEWs.

II. Framework

The main work of this study includes main four steps, as shown in Fig. 1.

Step 1: constructing seismic intensity datasets

Collect and clean seismic intensity data from Japan, and complete the construction of datasets. Divide the test set in terms of training set, validation set and test set. Preprocess the seismic data in southwest China and Japan to construct a Chinese transfer dataset and Japanese application dataset (see in Dataset section).

Step 2: constructing intensity prediction model

Build the intensity prediction models based on Random Forest, Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) respectively, train the models and optimize the model parameters (see in Method section). Determine the final model based on the test results. Utilize Synthetic Minority Over-Sampling Technique (SMOTE) to improve data imbalance, distinguish between intensity prediction models of different magnitudes and apply them jointly. (see in Intensity prediction model section).

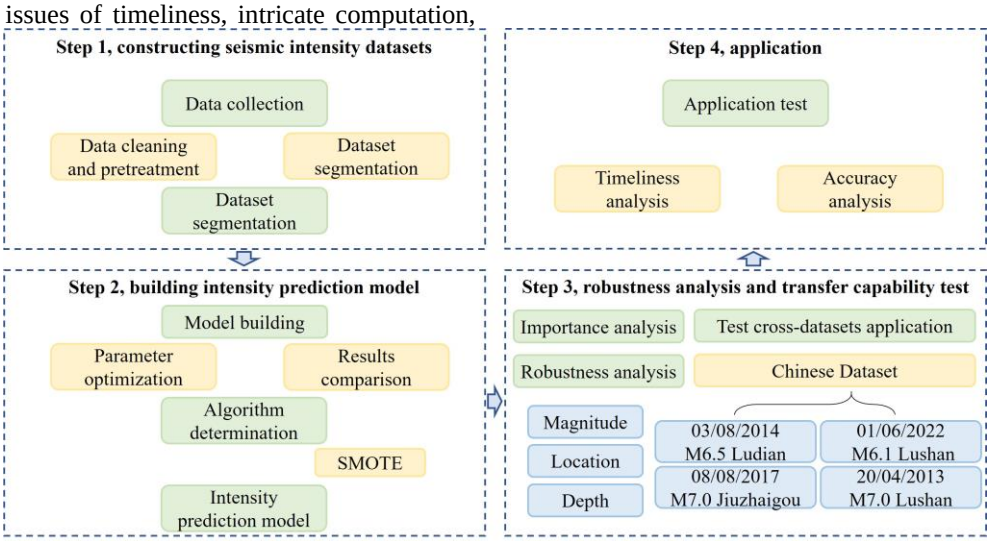


Fig. 1. Overall framework of the work.

Step 3: testing robustness analysis and cross-datasets validation

Add random errors to the magnitude and location information of the Japanese test set to analysis the robustness of the proposed model. Apply the intensity prediction model to the

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

Chinese transfer dataset to test the cross-datasets potential application of the model (see in DISCUSSION section).

Step 4: analyzing application possibilities

Simulate the reduction of the prediction error of seismic information for EEWs with time increasing, analyze the accuracy and timeliness by a real earthquake event (see in APPLICATION section).

III. DATASET

A. Japanese Dataset

A catalog of seismic intensity data published by Japan Meteorological Agency (JMA) is used as the dataset for this work [34]. The catalog stores data in hexadecimal files, so we convert the catalog to ASCII format for easy reading and manipulation. The catalog covers the period from 1919 to 2020, and the information has been calibrated. The catalog includes a total of 6,800 deactivated or active seismic intensity stations, distributed as shown in Fig. 2(a) Seismic intensity data announced by JMA are values observed using seismic intensity meters installed on the ground or on the first floor of low-rise buildings. We mainly use information from the catalog including (1) seismic intensity; (2) maximum acceleration observation time T_{max}; (3) earthquake origin time T_{origin}; (4) earthquake magnitude, longitude, latitude, and depth; (5) station longitude, latitude. JMA classifies seismic intensity as 0, 1, 2, 3, 4, 5 Lower, 5 Upper, 6 Lower, 6 Upper, and 7. Where 5 Lower, 5 Upper, 6 Lower, and 6 Upper are indicated in the catalog as A, B, C, and D.

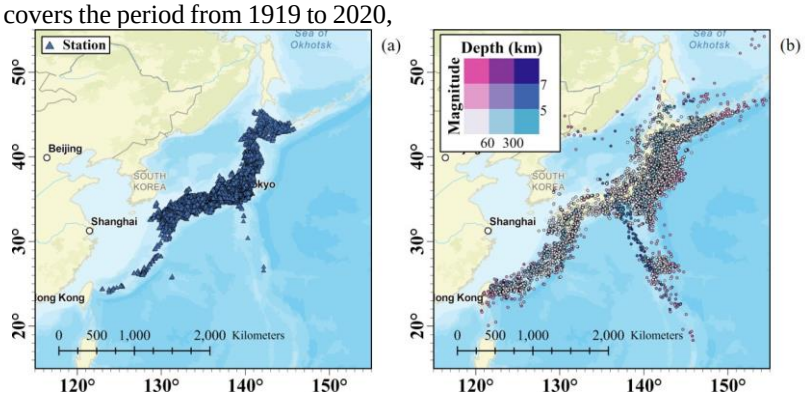


Fig. 2. Distribution of the stations (a) and epicenters (b), and the different color of the dots indicate the magnitude and depth of the earthquake, where shallow earthquakes are defined as those below 60km and deep earthquakes are defined as those above 300km.

We perform data cleaning and preprocessing of the catalog, which consists of the following:

1. **Temporal Selection:** We choose the time period from 1997 to 2020 for our catalog since T_{max} is not recorded in the 1996 and earlier files. Including data without Tmax would compromise the consistency and completeness of our dataset;

2. **Data Cleaning:** We remove intensity data that lacked T_{max} between 1997 and 2020. This step is crucial to avoid the introduction of null values that could skew our analysis and predictions;

3. **Data Validation:** Intensity data where T_{max} is less than T_{origin} are excluded. This criterion is based on the physical impossibility of such an occurrence and suggests a recording error. Removing such outliers helps prevent the model from learning from erroneous data;

4. **Data Encoding:** We replace the record intensity levels A, B, C, and D with numeric values 5, 5, 6, and 6, respectively. This encoding facilitates the mathematical manipulation of the data and the application of numerical methods. It also ensures uniformity in the treatment of categorical data;

5. **Magnitude Threshold:** We screen for seismic events with magnitudes greater than 4.0. Earthquakes with magnitudes below 4.0 are unlikely to cause significant ground shaking, and high intensity events are more important for EEWs.

6. **Time Difference Filtering:** We exclude intensity data where the time difference between T_{max} and T_{origin} is less than 1 second. Too short a time difference usually does not allow accurate estimation of basic parameters (e.g., magnitude, location, and depth), which are essential for reliable intensity prediction.

7. **Geospatial Calculations:** We calculate the epicentral distance and azimuth between the station and the hypocenter location using equations (1) to (6). This step is fundamental in establishing the spatial relationship between the earthquake source and the recording station, which is a significant factor in predicting the intensity of ground motions.

$$a = \sin^2(\Delta\text{lat}/2) + \cos(\text{lat}_1)\cdot\cos(\text{lat}_2)\cdot\sin^2(\Delta\text{lon}/2) \quad (1)$$

$$c = 2\cdot\arctan 2(\sqrt{a}, \sqrt{1-a}) \quad (2)$$

$$d = R\cdot c \quad (3)$$

$$x = \sin(\Delta\text{lon})\cdot\cos(\text{lat}_2) \quad (4)$$

$$y = \cos(\text{lat}_1)\cdot\sin(\text{lat}_2) - \sin(\text{lat}_1)\cdot\cos(\text{lat}_2)\cdot\cos(\Delta\text{lon}) \quad (5)$$

$$\theta = \arctan 2(x, y) \quad (6)$$

where, $\Delta\text{lat} = \text{lat}_2 - \text{lat}_1$, $\Delta\text{lon} = \text{lon}_2 - \text{lon}_1$, lat_1 and lon_1 are location of hypocenter, lat_2 and lon_2 are location of station, d and θ are distance and azimuth between the station and the hypocenter location, and R is the Earth's radius (mean radius = 6,371 km).

A total of 13,266 earthquake events (distributed as shown in Fig. 2(b)) and 944,877 intensity instances are obtained as a dataset after selection. The distribution of epicentral distance, magnitude and number of records is shown in Fig. 3(a), and the distribution of epicentral distance, intensity and number of records is shown in Fig. 3(b). In Fig. 4, the relationship between magnitude, depth, epicentral distance and intensity are given for all high intensity events. The results show that the intensity

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

greater than or equal to 5 has a strong relationship with the magnitude. Over 90% of high intensity are caused by earthquakes with magnitudes equal to or greater than 5.7.

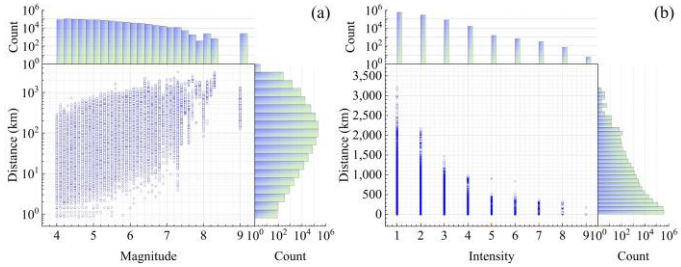


Fig. 3. (a) Data distribution versus catalog magnitude and epicentral distance used in this study. (b) Data distribution versus catalog intensity and epicentral distance used in this study.

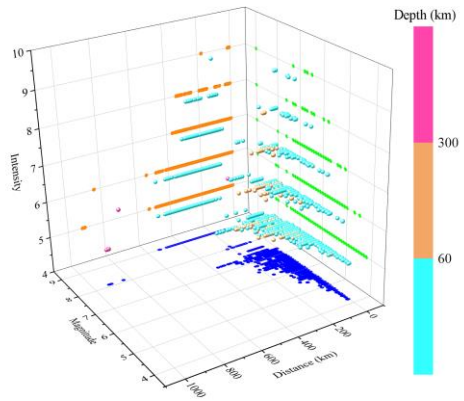


Fig. 4. Statistical scatterplot of magnitude, epicentral distance, and epicentral depth for high intensity data, with the color of the dots representing the depth of the epicenter.

The dataset is divided into training set, validation set and test set. To avoid the overfitting problem of the training model, which the duplication of data in the validation set and the training set, we divide the dataset in terms of time of occurrence of earthquake. Intensity instances from 1997 to 2016 are used as the training set and from 2017 to 2018 are used as the validation set. Intensity instances from 2019 to 2020 are selected as a test set used to evaluate the generalization capability. The distribution of the number of intensity instances for the training, test and validation sets is shown in Fig. 5.

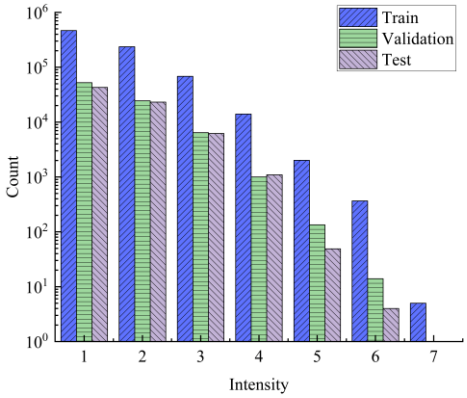


Fig. 5. Histogram of intensity statistic for different datasets.

The model constructed in this work is a parameter less intensity prediction model, which it only includes depth, magnitude, epicentral distance and azimuth. The output of the proposed model is the intensity. The inputs to the model are shown in (7),

$$Input_i = \{d_i, M_i, R_i, \theta_i\} \tag{7}$$

where, d_i, M_i, R_i and θ_i are depth, magnitude, epicentral distance and azimuth, respectively.

B. Chinese Transfer Dataset

To test the cross-datasets potential application of the proposed model, four earthquake events from southwest China are selected for testing. These four earthquake events are the 2022 Lushan M 6.1 earthquake, the 2013 Lushan M 7.0 earthquake, the 2014 Ludian M 6.5 earthquake, and the 2017 Jiuzhaigou M 7.0 earthquake, for a total of 319 recorded intensity instances [35]. The distribution of the hypocenters and stations of these earthquake events is shown in Fig. 6(a). It should be noted that the magnitude published in Japan is the Japan Meteorological Agency Magnitude (M_{JMA}), while the magnitude published in China is the Moment Magnitude (M_W). According to Katsumata's study [36], when the magnitude is 5~7 and the epicenter depth is less than 60km, the calculation methods of these two magnitudes are basically equal. The magnitude and epicenter depth of these four Chinese seismic events are eligible. These two magnitudes are considered the same here and are uniformly labeled with M in the text. Since the intensity instances are not directly given in the Chinese dataset, we first performed the intensity conversion according to the JMA [37] . It consists of two main steps: 1) Combining the three- component acceleration, which is shown in (8); 2) Calculating the intensity, which is shown in (9)

$$a(t_i) = \sqrt{a^2(t_i)_{E-W} + a^2(t_i)_{N-S} + a^2(t_i)_{U-D}} \tag{8}$$

$$I = 2 \times \log_{10}(a_c) + 0.94 \tag{9}$$

where, $a(t)_{E\text{-}W}$ is the east-west component of acceleration, $a(t)_{N\text{-}S}$ is the north-south component of acceleration, $a(t)_{U\text{-}D}$ is the vertical acceleration, and $a(t)$ is the acceleration synthesized from the three components, a_c is the acceleration corresponding to an exceedance probability of 0.3s, and I is the JMA seismic intensity. After the above calculation, the intensity statistics of earthquake events in southwest China is shown in Fig. 6(b).

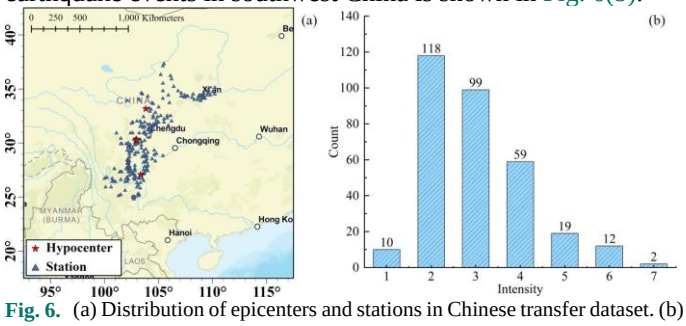


Fig. 6. (a) Distribution of epicenters and stations in Chinese transfer dataset. (b) Statistics on the number of intensity labels in Chinese transfer dataset.

C. Japanese Application Dataset

To further test the possibility of model application to EEWs, we construct a Japanese application dataset for testing the accuracy and timeliness of the proposed model application to EEWs. The Japanese application dataset is from the Kyoshin network (K-NET) in Japan for the period 2021 to 2022 [38]. Japan K-NET updates the seismic events occurring in the Japanese region on a daily basis and publishes waveform data from different stations. However, since there is no intensity data involved in the waveform files. Therefore, we calculate the

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

JMA intensity based on (8) to (9). We screen the data and the magnitude should not be less than 4.0. In addition, we calculate the trigger time of the station by using short-term average/long-term average (STA/LTA) algorithm [39]-[41] , which is shown in (10) to (12),

$$\mathrm{STA}(i)=\frac{1}{N}\sum_{j=i-N+1}^{i}X(j) \tag{10}$$

$$\mathrm{LTA}(i)=\frac{1}{M}\sum_{j=i-M+1}^{i}X(j) \tag{11}$$

$$R_i=\frac{\mathrm{STA}_i}{\mathrm{LTA}_i} \tag{12}$$

where, M and N, respectively, represent the length of the long term and the short term and X represents the raw data. We apply an STA/LTA operator on the filtered data and the length of the short term is 50 samples (0.5 s in K-NET data), while the length of the long term is 3000 samples (30 s in K-NET data). And the threshold of R_i is set to 10. The number of records from stations triggered by the STA/LTA method should be no less than five. After screening, 466 earthquake events and 20542 intensity instances are obtained. The distribution of hypocenters and stations is shown in Fig. 7(a), and the statistics of the intensity are shown in Fig. 7(b).

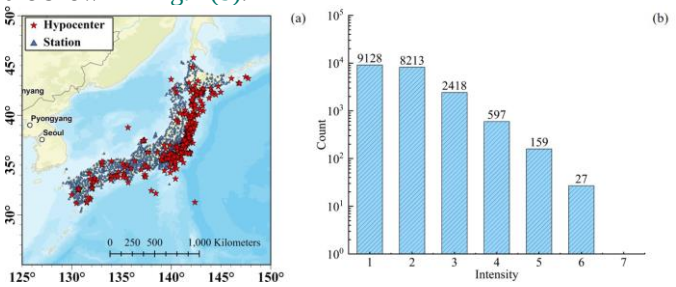


Fig. 7. (a) Distribution of epicenters and stations in Japanese application dataset. (b) Statistics on the number of intensity labels in Japanese application dataset.

IV. Method

A. Overview

Decision Tree is a supervised learning algorithm in machine learning, mainly used for classification and regression tasks. The decision tree model is a tree-like structure, where each non-leaf node represents a test on a feature attribute, each branch signifies the output of this feature attribute within a certain range, and each leaf node stands for a predicted classification (for classification trees) or a continuous value (for regression trees). The process of constructing a decision tree usually involves feature selection, decision tree generation and pruning [42].

When evaluating the performance of a machine learning model, there are two main aspects that are often considered: bias and variance. Bagging [43] and Boosting [44] are both ensemble learning strategies designed to improve overall model performance by combining multiple models. However, they focus on different things: while Bagging focuses on reducing variance, and Boosting focuses on reducing bias [45].

Bagging reduces model variance by generating multiple models and integrating their predictions. Random Forest is a classic application of Bagging that uses a decision tree as the base learner and introduces random selection of features to increase the diversity of the model. Boosting reduces bias and improves overall model performance by sequentially building learners and correcting errors in the previous learner. AdaBoost is a preliminary Boosting method, and in subsequent iterations of updates, XGBoost and LightGBM further improve the performance of the Boosting algorithm.

B. Random forest

A random forest is composed of multiple decision trees, each of which is independently sampled (shown in Fig. 8(a)). Random forests randomly draw (with put-back) samples of the same size from the original dataset when constructing each decision tree. Instead of considering all features when splitting each node of the tree, a subset of features is randomly selected and the best split is found on this subset. For the classification task, each tree votes and then the category with the most votes is chosen as the final prediction [46].

The main function of the random forest model is shown as (13),

$$\bar{H}(X)=(\frac{1}{K})\sum\nolimits_{i=1}^{K}H(X;\theta_i) \tag{13}$$

where, $H(X;\ \theta_i)$ denotes the kth predictor tree, and θ represents the random vector of the random forest. The supervised learning is achieved by minimizing the following loss function shown as (14),

$$\sum\nolimits_{n=1}^{N}(O_n-\bar{H}(X)_n)^2 \tag{14}$$

where, O denotes the seismic intensity and N denotes the number of training earthquakes. Random forest model optimization parameters include maximum depth of the decision tree (M_{dep}), number of decision trees (M_{tree}) and minimum number of samples a node can be divided (S), parameter meanings and selection ranges are shown in Table I.

C. XGBoost

XGBoost (shown in Fig. 8(b)) is an optimized distributed gradient boosting algorithm with parallel processing, support for custom optimization objectives and evaluation criteria [47]. XGBoost controls the splitting nodes by the maximum depth of the tree and then continues to split through the child nodes, effectively preventing overfitting of the model. In addition, XGBoost can perform cross-validation in each round of boosting iterations to get the rounds where the optimal iteration stops. The main function of XGBoost is shown as below,

$$O_{b,j}=\sum\nolimits_i l(y_i,\hat{y}_i)+\sum\nolimits_k\Omega(f_k) \tag{15}$$

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\,\|w\|^2\ \omega \tag{16}$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ $\hat{y_i}$, and the target y_i, Ω is the regularization term function and f_k is the kth tree model, T and w are the number and weights of leaf nodes, respectively. XGBoost adds a regularization term to the objective function as compared to GBDT (Gradient Boosting Decision Tree), which can prevent overfitting of the model. XGBoost optimization parameters include learning rate (*Eta*), M_{dep}, M_{tree}, S and proportion of subsamples of features used in building each tree (C), parameter meanings and selection ranges are shown in Table I.

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

D. LightGBM

The LightGBM (shown in Fig. 8(c)) model is obtained by improving on the XGBoost model. Compared with the XGBoost model, the LightGBM model has four main optimizations [48]. (1) The LightGBM model uses the leaf-wise algorithm instead of the level-wise algorithm as the tree growth strategy. (2) The LightGBM model uses the histogram algorithm. The histogram algorithm can save only the bin values after feature discretization, which reduces memory consumption and thus increases the computational speed. (3) LightGBM uses a Gradient-based One-Side Sampling (GOSS) algorithm. (4). LightGBM optimization parameters include *Eta*, M_{dep}, M_{tree}, *S*, *C*, parameter descriptions and selection ranges are shown in Table I.

TABLE I
DESCRIPTIONS AND RANGE OF VALUES OF DIFFERENT MODEL PARAMETERS

Model	Parameter	Meaning	Range and interval
Random Forest	M_{dep}	Prevent overfitting and simplify model	[6, 2, 30]
	M_{tree}	Improve performance	[50, 50, 1000]
	S	Increase computational efficiency	[0.1, 0.1, 1]
XGBoost	*Eta*	Increase computational efficiency and performance	[0.05, 0.05, 0.5]
	M_{dep}	Prevent overfitting and simplify model	[6, 2, 30]
	M_{tree}	Improve performance	[50, 50, 1000]
	S	Increase computational efficiency	[0.1, 0.1, 1]
	C	Improve the robustness of the model	[0.1, 0.1, 1]
LightGBM	*Eta*	Increase computational efficiency and performance	[0.05, 0.05, 0.5]
	M_{dep}	Prevent overfitting and simplify model	[6, 2, 30]
	M_{tree}	Improve performance	[50, 50, 1000]
	S	Increase computational efficiency	[0.1, 0.1, 1]
	C	Improve the robustness of the model	[0.1, 0.1, 1]

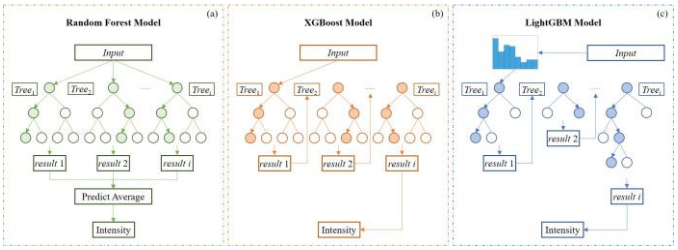


Fig. 8. (a) Random forest model. (b) XGBoost model. (c) LightGBM model.

E. Parameter Optimization

Suitable model parameters can effectively improve the accuracy and computational speed of the model. We use grid search and *k*-fold cross-validation (*k*=5) to determine the model parameters. The range of values and intervals of the parameters are shown in Table I (e.g. [6, 2, 30] represents an initial value of 6, an interval of 2, and an end value of 30). The order in which the model parameters are determined is sequential in the order of the table.

F. Evaluation Indicators

We refer to the methods of other researchers and scholars [17] [29], and predicting the intensity at ± 1 level is widely recognized as accurate, calculated as shown in (17). In the JMA intensity scale, intensity equal and higher to 5 may cause huge damage, since we regard intensity ≥ 5 as high else low. We consider high intensity accurate prediction as TP, high intensity false prediction as FP, low intensity accurate prediction as TN, and low intensity false prediction as FN (shown in Fig. 9). For users of EEWs, the cost of receiving inaccurate high intensity warning signals is more expensive than the cost of receiving inaccurate low intensity warning signals. Thus, we further introduce *Precision*, which responds to the accuracy of predicting high intensity, calculated as shown in (12). In addition, we introduced mean absolute error (*MAE*) and standard deviation (*STD*) to measure the accuracy of the predictions, which are calculated as shown below.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{19}$$

$$STD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{20}$$

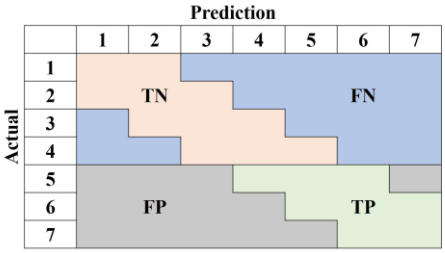


Fig. 9. Definitions of TP, FP, TN, and FN.

V. INTENSITY PREDICTION MODEL

The suitable parameters for different models are tested and the parameters for each model are shown in Table II. In Table III, the test results of different models are given. The results show that the *Accuracy*, *Precision*, *MAE*, and *STD* of the XGBoost model reach 98.97%, 93.96%, 0.308, and 0.484, respectively, which are better than those of the RF and LightGBM models. The results of the strengths and weaknesses of the test model were also the same as Abdalzaher's study [30] despite the different dataset and content used. Therefore, the subsequent work will further analyze and discuss based on the XGBoost model. In Fig. 10(a), the confusion matrix of the results on the validation set using the XGBoost model is given. The results show that the maximum error in XGBoost prediction is 3 levels of intensity, and the number of prediction errors greater than or equal to 3 levels is only 7 times. To further evaluate the effectiveness of XGBoost for testing, we save the model and test it on the test set. The results of the test set are shown in Table III and the confusion matrix for the test set is given in Fig. 10(b). The results of the test set are on par with the validation set in terms of the *Accuracy*, however the *Precision* is much lower than the validation set. This result is related to the imbalanced distribution of seismic intensity data.

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

In the full dataset, there are a total of only 2,584 events (0.27%) with an intensity equal or above 5. The imbalanced data distribution leads to the inability of the model to use the small amount of information from the high intensity data to learn the features of high intensity data. We adopt SMOTE to improve this problem. SMOTE balances the category distribution by synthesizing new instances for the minority class in the feature space, based on the difference between the minority class instances and their neighbors and random interpolation. The specific steps are as follows:

1) for each minority class sample, select the k nearest neighbors (after testing, k was chosen to be 4);

2) for each selected minority class sample, a random sample is selected from its k nearest neighbors;

3) new synthetic samples are generated in the feature space using linear interpolation between selected minority class samples and randomly selected samples;

4) add these synthetic samples to the original minority class samples to expand the dataset.

The formula for generating new samples for SMOTE is shown below,

$$x_{new} = x + rand(0,1) \times (\tilde{x} - x) \tag{21}$$

where, x is real value, $\tilde{x}$ is approach value, and x_{new} is new data.

After SMOTE, the number of each intensity in the training set is equal to that of intensity 1 at 466,233. We retrain the model based on these regenerated samples, named $\text{XGBoost}_{\text{SMOTE}}$. In Table III, the test results of $\text{XGBoost}_{\text{SMOTE}}$ on validation set and test set are given. The results show that $\text{XGBoost}_{\text{SMOTE}}$ performs better in the *Precision* metric of the test set. After the introduction of SMOTE, the *Precision* metric increased from 52.83% to 92.45%, which substantially improved the prediction accuracy of high intensity strength. While the *Accuracy* metric decreased from 98.79% to 96.23%. This reason is importantly related to the process of oversampling that leads to an excessive distribution of high intensity data, resulting in generally high predicted intensities for $\text{XGBoost}_{\text{SMOTE}}$.

These test results show that XGBoost, $\text{XGBoost}_{\text{SMOTE}}$ models have better performance in predicting low intensity instances and high intensity instances respectively. According to our statistics, more than 90% of the high-intensity instances are caused by earthquakes with $M \geq 5.7$ (shown as in Fig. 4). In addition, the magnitude has the highest percentage of importance in XGBoost at 0.4871. Therefore, we jointly apply XGBoost and $\text{XGBoost}_{\text{SMOTE}}$ models based on different magnitude intensities. This is done by using the XGBoost model to predict the intensity when the magnitude is less than 5.7, and the $\text{XGBoost}_{\text{SMOTE}}$ model to predict the intensity when the magnitude is $\geq$ 5.7. We call the model after the joint application $\text{XGBoost}_{\text{Intensity}}$ and re-tested the results on the test set and the results are shown in Table III. The results show that the *Accuracy* metric of $\text{XGBoost}_{\text{Intensity}}$ decreases by 0.77% but the *Precision* metric improves by 28.30% compared to XGBoost.

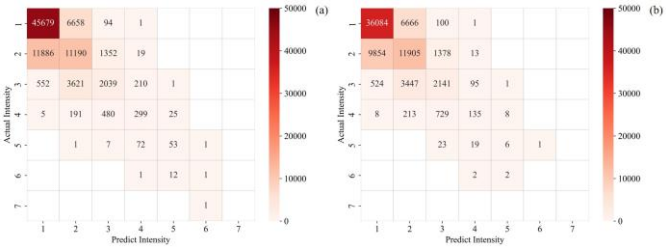


Fig. 10. Confusion matrix for XGBoost results in validation (a) and test (b) sets.

TABLE II
OPTIMAL PARAMETERS FOR DIFFERENT ALGORITHMIC MODELS

Model	Parameter	Value
Random Forest	M_{dep}	10
	M_{tree}	150
	S	0.1
XGBoost	*Eta*	0.15
	M_{dep}	6
	M_{tree}	100
	S	0.1
	C	1
LightGBM	*Eta*	0.05
	M_{dep}	22
	M_{tree}	550
	S	0.5
	C	1

VI. DISCUSSION

A. Robustness Analysis

In EEWs, seismic parameters (magnitude, location, etc.) are obtained by analyzing the information contained within the initial P wave time window [50]-[51]. As the length of the *P* waves time window increases, more and more information are contained and the seismic parameters are estimated more accurately. Although the estimation becomes more accurate as the amount of information increases, it is still inevitable that estimation errors will still exist in actual applications. Therefore, this section analyzes the effect of seismic parameter errors on the prediction accuracy of the proposed model. On the one hand, it simulates the change in the accuracy of the proposed model as the error gradually decreases; on the other hand, it tests the robustness of the model. We add random different errors to magnitude, latitude, longitude, and depth, and choose the test set for robustness analysis. The specific robustness analysis consists of three main components, magnitude, location and depth.

We introduce uniformly distributed random errors in hypocenters' magnitude, location (latitude and longitude), and depth to assess the robustness of the model. We use *Accuracy* and *Precision* evaluation metrics to quantify the impact of different features on model predictions. In order to simulate the increase in the accuracy of earthquake parameters over time, we sequentially reduce the range of the uniform distribution of errors in the contaminated dataset. In addition, we add errors at magnitude, location, and depth in turn to better rate the effect of different features on the prediction results. Eventually we obtained nine contaminated datasets as shown in Table IV. The serial number of each contaminated dataset represents a different combination of magnitude, location (epicentral distance and azimuth), and depth error. In addition, we have

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

recalculated the epicentral distance and azimuth as inputs to the new contaminated dataset after the introduction of location errors. For depth, we restrict its value to be non-negative to reflect the physical reality that seismic depth cannot be negative.

TABLE IV
RANDOM ERRORS ADDED TO DIFFERENT CONTAMINATED DATASETS.

NO.	Magnitude	Location (°)	Depth (km)
1	[-0.5, 0.5]	/	/
2	[-0.3, 0.3]	/	/
3	[-0.1, 0.1]	/	/
4	[-0.5, 0.5]	[-0.3, 0.3]	/
5	[-0.3, 0.3]	[-0.2, 0.2]	/
6	[-0.1, 0.1]	[-0.1, 0.1]	/
7	[-0.5, 0.5]	[-0.3, 0.3]	[-10, 10]
8	[-0.3, 0.3]	[-0.2, 0.2]	[-5, 5]
9	[-0.1, 0.1]	[-0.1, 0.1]	[-2, 2]

In Fig. 11 histograms of the test results for different contaminated datasets are given. The *Accuracy* of No.7 contaminated dataset with the addition of random maximum is 94.91%. As the error decreases, the *Accuracy* of the prediction is further improved. In the contaminated dataset No. 9, the model achieved an *Accuracy* of 97.65% and a *Precision* of 83.02%, achieving results that are essentially equal to those of the origin test set. Furthermore, when comparing the contaminated dataset vertically in conjunction with Fig. 12, the accuracy drops by just less than 1% when errors in position and depth are added sequentially. When comparing the contaminated dataset horizontally, the reduction in magnitude error leads to a significant increase in both *Accuracy* and *Precision*. The results show that the model's predictions are most sensitive to magnitude errors, followed by location and depth errors, which is consistent with the ordering of feature importance in our model. These results are significant because they show that our model can provide reliable and robust intensity predictions despite initial uncertainties in the estimation of seismic parameters. This robustness is critical for the practical deployment of EEWs, where timely and accurate warnings can save lives and reduce economic losses.

TABLE III
RESULTS OF DIFFERENT MODELS ON DIFFERENT DATASETS

Model	Dataset	Total	TP	TN	FP	FN	*Accuracy*	*Precision*	*MAE*	*STD*
Random Forest			118	83334	31	968	98.82%	79.19%	0.312	0.488
XGBoost	Validation	84451	140	83439	9	863	98.97%	93.96%	0.308	0.484
LightGBM			139	82991	10	1311	98.44%	93.29%	0.333	0.513
XGBoost	Test	73355	28	72442	25	860	98.79%	52.83%	0.326	0.494
$XGBoost_{SMOTE}$	Validation	84451	147	82112	2	2190	97.40%	98.66%	0.398	0.541
	Test	73355	49	70539	4	2763	96.23%	92.45%	0.543	0.607
$XGBoost_{Intensity}$	Test	73355	43	71904	10	612	98.02%	81.13%	0.353	0.520

B. Cross-datasets Validation

Seismic waves produce different ground motions under different geologic formations and velocity models. Nonetheless, a common observation is that seismic intensity increases with higher magnitude and lower epicentral distance. Therefore, the intensity prediction model constructed in this work may be applicable to intensity prediction in other different regions. To further test the cross-datasets validation of the proposed model, we select four earthquake events from southwest China for evaluation. Due to the small number of datasets in China, we do not work on transfer learning training. The model remains the $XGBoost_{Intensity}$ model trained by using the Japanese earthquake data. Since the model is not further trained, this also means that the results may be poorer. Notably, the results show that the *Accuracy* of the Chinese transfer dataset reaches 90.28% and the *Precision* reaches 93.93% (shown as in Fig. 11). In addition, the *MAE* and *STD* of the Chinese transfer dataset are 0.736 and 0.624, respectively. These test results are lower than the performance of the Japanese dataset. However, these results are without further training. The performance of the model is expected to be further optimized if further training is performed using large scale datasets from China.

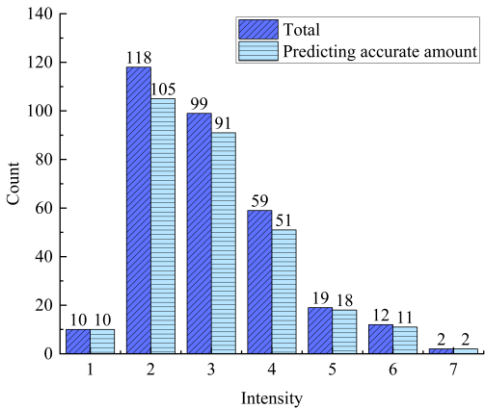


Fig. 11. Histogram of predicted outcomes for the Chinese transfer dataset.

C. Limitations

The models proposed in this paper are all based on the intensity prediction with existing seismic parameters. Although we test the robustness of the proposed models under different errors. However, due to the highly random nature of earthquakes, the estimation results of earthquake parameters may still have some bias. In addition, all the tests in this paper are done under computer simulation, and the effects of errors due to signal transmission, system delay, etc. are not considered in this paper in terms of timeliness analysis. These factors will have a certain impact on the results of the proposed model.

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

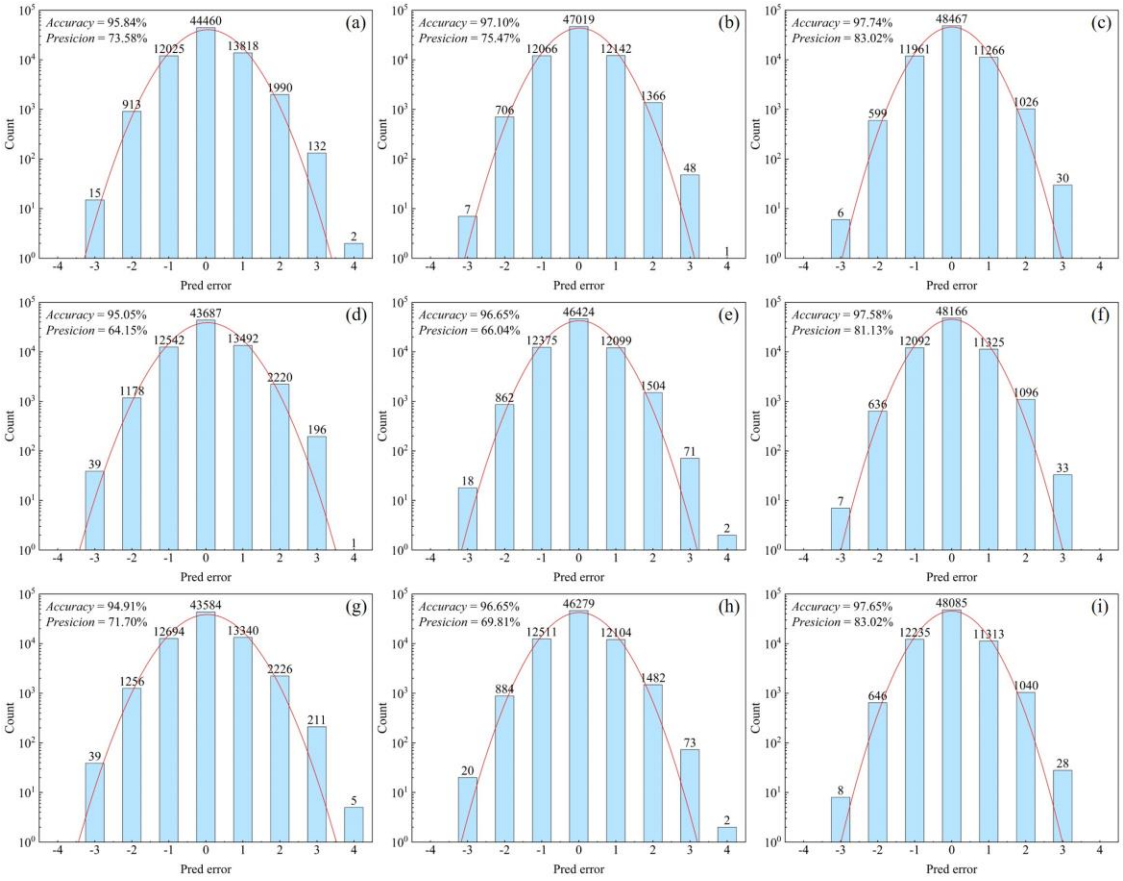


Fig. 12. (a) to (i) are the histograms of contaminated test sets No.1 to No.9, respectively.

VII. Application

In EEWs, the usual way of working is trigger, localization, estimation, and warning. Among them, the seismic information is mainly estimated by using the *P* wave monitoring information from stations near the epicenter. Usually, the seismic information is also predicted more accurately with time increasing. We simulate this process of gradually decreasing prediction error and test it on Japanese Application Dataset. It is worth noting that we more realistically reproduce the effect of magnitude estimation, which the underestimation of large earthquakes and overestimation of small earthquakes [50]-[51]. We add a biased small error for earthquakes with magnitude greater than 5.7 and a biased large error for earthquakes with magnitude less than 5.7. The random error added to each earthquake event is the same. The dataset for our simulation of the process of random error reduction is shown in TABLE V. The errors we add for position and depth are guaranteed to be added in absolute values (e.g. [-0.3, -0.2]/ [0.2, 0.3] represents the addition of random errors of -0.3° to -0.2° or 0.2° to 0.3°).

In Table VI, the overall test results of the applied dataset are given. The results show that the prediction error of the intensity keeps decreasing as the error decreases, which is basically consistent with the previous test results. In addition, these test results further show that the proposed method has strong robustness and generalization ability.

TABLE V
Applying the Dataset to Simulate the Process of Error Reduction

NO.	Magnitude		Location (°)	Depth (km)
1	M≥5.7	[-0.5, -0.4]	[-0.3, -0.2]/	[-10, -5]/
	M<5.7	[0.4, 0.5]	[0.2, 0.3]	[5, 10]
2	M≥5.7	[-0.3, -0.2]	[-0.2, -0.1]/	[-5, -2]/
	M<5.7	[0.2, 0.3]	[0.1, 0.2]	[2, 5]
3	M≥5.7	[-0.1, 0]	[-0.1, 0.1]	[-2, 2]
	M<5.7	[0, 0.1]		

TABLE VI
Results of Different Models on Different Datasets.

Dataset	TP	TN	FP	FN	*Accuracy*	*Precision*
No.1	135	19355	51	1001	94.88%	72.58%
No.2	169	19915	17	441	97.77%	90.86%
No.3	183	19930	3	426	97.91%	98.39%

In addition to the overall testing of the entire Japanese Application Dataset, we also analyze the effect of testing each seismic event and epicentral distance. Fig. 13(a) shows the results of the *Accuracy* metric for each seismic event. The results show that there are 53, 14, and 3 seismic events with Accuracy below 90% for datasets No.1, No.2, and No.3, respectively. For the case of dataset No.1, the distribution of magnitude errors for seismic events is more spread out, and this reason is related to the addition of too high magnitude random errors. As the error decreases, the estimation accuracy continues to improve. Fig. 13(b) shows the estimation effect of the proposed model at different epicenter distances, and Accuracy is counted every 50km. The results show that the *Accuracy* of the proposed model have a significant decreasing trend at 250-300 km, 500-550 km, and 700-750 km. This reason

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

may be related to the uneven distribution of data. Overall, the proposed model has good performance at different magnitudes and epicentral distances.

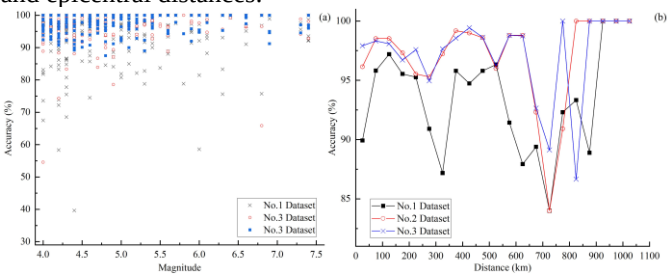


Fig. 13. Distribution of Accuracy with (a) magnitude and (b) epicentral distance.

In addition, we use a real earthquake event of *M* 7.4 on March 16, 2022 as an example for testing the timeliness of the proposed model for intensity prediction. In Fig. 14(d), we give the distribution of true intensity and the distribution of predicted intensity from the model. In Fig. 14, in combination with the proposed *Accuracy* metric, stations with intensity prediction error ≤ 1 are regarded as true intensity. The results show that the predicted intensity maps converge with the true intensity maps as the error in the seismic parameters decreases.

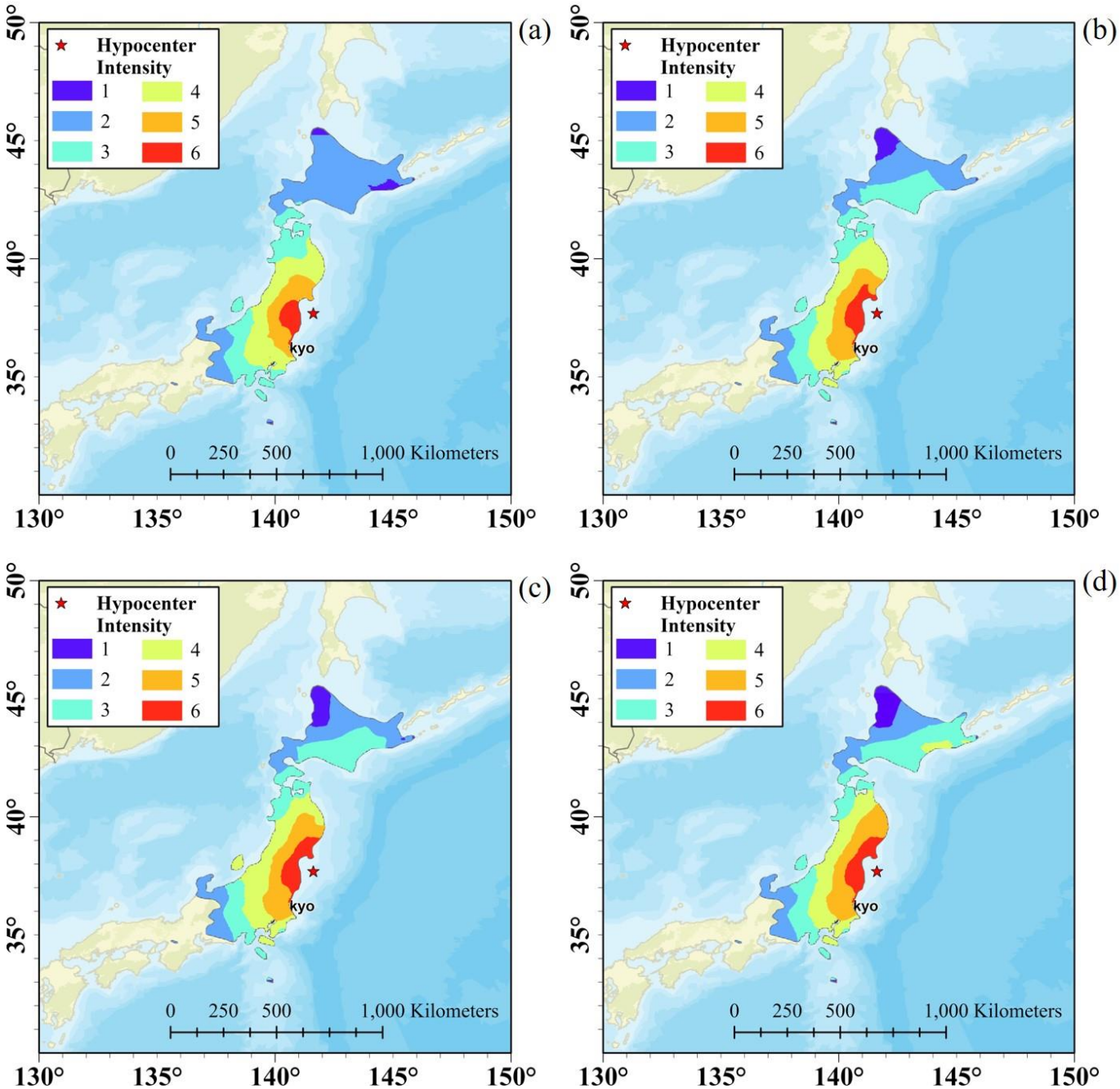


Fig. 14. Intensity distribution of the M 7.4 event. (a), (b) and (c) are the predicted intensities under errors No.1, No.2 and No.3, respectively, and (d) is the true intensity.

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

We further analyze the timeliness of the model applied to real earthquake event. After an earthquake occurs, EEWs first proceed to identify the seismic phase and issue an alarm. We define the moment of the first station triggering the alarm as the original moment T_0, and the time difference between the i th station triggering the alarm afterward and the first station triggering the alarm as T_i (i > 1). In this earthquake event, trigger difference T_{159} for the last intensity 5 station is 28.74 s.

We refer to Wang's study [52] to test the magnitude estimation results. Wang [52] constructed a magnitude estimation method based on traditional parameters using waveform data from Japan. In this paper, the fitting model of displacement peak, magnitude and epicenter distance after the first arrival of seismic wave for 3s proposed by Wang is used for magnitude estimation. The magnitude estimates are averaged over multiple stations. The formula is shown below,

$$\log_{10}(P_{d10}) = \log_{10}(P_d) - c \times \log_{10}(R/10) \quad (22)$$

$$\log_{10}(P_{d10}) = a \times M + b \quad (23)$$

where P_d is the displacement amplitude 3s after the initial arrival of the seismic wave, P_{d10} is the displacement amplitude normalized to 10km, R is the epicenter distance, M is the magnitude, and a, b and c are the fitting coefficients. After Wang's test [52], a, b and c are obtained as 0.65, -5.15, and -0.70, respectively.

First, we use the real epicentral distance for the test, and the magnitude estimation results are M 6.93, M 7.15, and M 7.31 when 3, 4, and 6 stations are used, respectively. This result corresponds to the simulated dataset with reduced error. In the No.1 dataset, the mean epicentral distance error is 8.91 km, and the magnitude estimation results are M 6.85, M 7.07, and M 7.24 when using 3, 4, and 6 stations, respectively. According to Saad's study [12], when three triggering stations are used for hypocenter localization, the average error of hypocenter location is about 2.5km. This implies that the error of the epicentral distance will be lower when a more advanced epicentral localization model is used. Therefore, we used the No.3 dataset for the timeliness analysis. In No.3 dataset, when 6 stations are used, the magnitude estimation results are M 7.29 with an estimation error of -0.11 magnitude units, respectively. The time T_6 for the 6th triggering station is 4.47 s. Since the calculation of P_d requires the use of a time window of 3 s, the time needed to complete the magnitude estimation is 7.47 s. This means that the test results for the dataset of No.3 can be achieved 7.47 s after the first station is triggered.

A total of 16 stations were triggered 7.47s after the first station was triggered. From the 17th station to the 159th station, a total of 62 stations are high intensity areas. The average trigger T_i of these 62 stations is 16.18s. This also means that the application of the proposed model can provide an average warning time of 8.71s for these high intensity areas. The predicted *Precision* of these 62 stations reaches 98.38% (61/62), and the *MAE* and *STD* reach 0.70 and 0.48 intensity units, respectively. As for the whole earthquake event, the average triggering T_i of the stations from 17th to 447th (431 stations) is 41.96 s, which means that the proposed model can provide an average warning time of 34.49 s if the whole seismic impact area is considered.

In terms of the computation time of the model, we evaluate the time to run the Japanese application dataset and get the test results. The model is run on a personal computer with an Intel i7-10700K processor. The results show that the run included three processes of loading data, loading model and predicting results, which took a total time of 1.02 s. This also means that the estimation time of the model proposed in this work for a single intensity is about 0.00005 s, which can satisfy the timeliness requirement of EEWs.

VIII. Conclusion

In this study, we introduce a streamlined machine learning intensity prediction model that solely depends on magnitude and location data for its predictions. The primary findings of our research are as follows:

1. We compare the results of three machine learning algorithms, Random Forest, XGBoost and LightGBM, for intensity prediction. The results show that the XGBoost algorithm achieves 98.93% and 93.96% in *Accuracy* and *Precision* in the validation set, which is better than Random Forest and LightGBM, respectively.

2. We adopt SMOTE to improve the problem of data imbalance. Considering the magnitude distribution of high intensity, we construct the intensity prediction model $\text{XGBoost}_{\text{Intensity}}$ under different magnitudes. $\text{XGBoost}_{\text{Intensity}}$ achieves 98.02% *Accuracy* and 81.13% *Precision* in the test set. After improving the results on the test set, we sacrifice 0.77% of the *Accuracy* and increase the *Precision* by 28.30%.

3. We analyze the robustness of the model by adding random errors to each intensity instance. When the random errors of magnitude, location and depth are ±0.5, ±0.3° and ±10km, the *Accuracy* of the test set still reaches 94.91%. As the random errors decrease, the *Accuracy* of the test set reaches up to 97.91%, which is basically the same as the original data. In addition, we directly apply $\text{XGBoost}_{\text{Intensity}}$ to four seismic events in the Chinese transfer dataset. The *Accuracy* and *Precision* of the Chinese transfer dataset reached 90.28% and 93.93%, respectively. This indicates the cross-datasets potential application of the proposed model.

4. We construct a Japanese application dataset using 466 seismic events from the Japanese K-NET network for 2021-2022 and analyze the timeliness and accuracy of the proposed model. We simulate the overestimation of large earthquakes and underestimation of small earthquakes in the magnitude estimation of EEWs and predict the intensity of the Japanese application dataset. The results show that in the worst case simulated (magnitude 0.5, location 0.3°, depth 10 km), the model achieves *Accuracy* and *Precision* of 94.88% and 72.58%, respectively, which agrees with the results of the robustness analysis. We test the timeliness of the proposed model using an occurrence of a real M 7.4 earthquake as an example. We tested the time required for magnitude estimation, and the magnitude estimation error reached -0.11 units after 7.47 s after the first station was triggered. The proposed model can provide an average warning time of 7.47 s for other high intensity areas compared with on-site EEWs. In addition, for the whole earthquake event, the proposed model can provide an average warning time of 34.49 s for 447 stations on average.

5. The proposed model improves the timeliness of prediction compared with the existing intensity prediction methods. This also means that if the proposed model is applied, it is expected

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

to further improve people's escape time during earthquakes and better protect people's lives and properties.

ACKNOWLEDGMENT

Acknowledgement for the data support from China Earthquake Networks Center, National Earthquake Data Center. (http://data.earthquake.cn). The strong ground motion data are recorded by the Kyoshin network (K-NET), which can be downloaded from the National Research Institute for Earth Science and Disaster Prevention (NIED K-NET, KiK-net, https://www.doi.org/10.17598/NIED.0004).

REFERENCES

[1] H. Kanamori, E. Hauksson, and T. Heaton, "Real-time seismology and earthquake hazard mitigation," *Nature,* vol. 390, no. 6659, pp. 461-464, Dec 4, 1997.

[2] R. M. Allen, "The ElarmS earthquake early warning methodology and application across california," *Earthquake Early Warning Systems*, pp. 21-43, 2007.

[3] J. Fayaz, and C. Galasso, "A deep neural network framework for real-time on-site estimation of acceleration response spectra of seismic ground motions," *Comput-Aided Civ Inf,* vol. 38, no. 1, pp. 87-103, Jan, 2023.

[4] X. M. Lei, L. M. Sun, and Y. Xia, "Seismic fragility assessment and maintenance management on regional bridges using bayesian multi-parameter estimation," *B Earthq Eng,* vol. 19, no. 15, pp. 6693-6717, Dec, 2021.

[5] J. P. Yang, Y. Xia, X. M. Lei *et al.*, "Hysteretic parameters identification of RC frame structure with Takeda model based on modified CKF method," *B Earthq Eng,* vol. 20, no. 9, pp. 4673-4696, Jul, 2022.

[6] E. Ahmadi, M. R. Salami, R. De Risi *et al.*, "Multi-pulse decomposition for nonlinear seismic analysis of structural systems," *Soil Dyn Earthq Eng,* vol. 163, Dec, 2022.

[7] C. W. Yang, X. H. Tong, G. P. Chen *et al.*, "Assessment of seismic landslide susceptibility of bedrock and overburden layer slope based on shaking table tests," *Eng Geol,* vol. 323, Sep 20, 2023.

[8] C. W. Yang, X. H. Tong, D. S. Wu *et al.*, "A new model for mechanical calculation of h-type anti-slide piles," *Structures,* vol. 56, Oct, 2023.

[9] X. T. Zhang, W. Reichard-Flynn, M. Zhang *et al.*, "Spatiotemporal Graph Convolutional Networks for Earthquake Source Characterization," *J Geophys Res-Sol Ea,* vol. 127, no. 11, Nov, 2022.

[10] M. Yamada, and J. Mori, "P-wave picking for earthquake early warning: refinement of a T-pd method," *Geophys J Int,* vol. 228, no. 1, pp. 387-395, Jan, 2022.

[11] S. M. Mousavi, and G. C. Beroza, "A Machine-Learning Approach for Earthquake Magnitude Estimation," *Geophys Res Lett,* vol. 47, no. 1, Jan 16, 2020.

[12] O. M. Saad, Y. F. Chen, A. Savvaidis *et al.*, "Real-Time Earthquake Detection and Magnitude Estimation Using Vision Transformer," *J Geophys Res-Sol Ea,* vol. 127, no. 5, May, 2022.

[13] Y. P. Fan, Z. J. Chen, and T. Su, "Abnormal Electromagnetic Wave Detection Method Before Earthquake Based on Feature Extraction of Radio Frequency I/Q Signal," *Ieee Sens J,* vol. 23, no. 11, pp. 11796-11805, Jun 1, 2023.

[14] D. T. Trugman, M. T. Page, S. E. Minson *et al.*, "Peak Ground Displacement Saturates Exactly When Expected: Implications for Earthquake Early Warning," *J Geophys Res-Sol Ea,* vol. 124, no. 5, pp. 4642-4653, May, 2019.

[15] A. Wang, S. Y. Li, J. Q. Lu *et al.*, "Prediction of PGA in earthquake early warning using a long short-term memory neural network," *Geophys J Int,* vol. 234, no. 1, pp. 12-24, Feb 24, 2023.

[16] S. Sarkar, A. Roy, S. Kumar *et al.*, "Seismic Intensity Estimation Using Multilayer Perceptron for Onsite Earthquake Early Warning," *Ieee Sens J,* vol. 22, no. 3, pp. 2553-2563, Feb 1, 2022.

[17] J. B. Zhu, S. Y. Li, and J. D. Song, "Hybrid Deep-Learning Network for Rapid On-Site Peak Ground Velocity Prediction," *Ieee T Geosci Remote,* vol. 60, 2022.

[18] S. Colombelli, A. Caruso, A. Zollo et al., "A P wave-based, on-site method for earthquake early warning," *Geophys Res Lett,* vol. 42, no. 5, pp. 1390-1398, Mar 16, 2015.

[19] Y. M. Wu, and H. Kanamori, "Rapid assessment of damage potential of earthquakes in Taiwan from the beginning of waves," *B Seismol Soc Am,* vol. 95, no. 3, pp. 1181-1185, Jun, 2005.

[20] S. Parolai, D. Bindi, T. Boxberger et al., "On-Site Early Warning and Rapid Damage Forecasting Using Single Stations: Outcomes from the REAKT Project," *Seismol Res Lett,* vol. 86, no. 5, pp. 1393-1404, Sep-Oct, 2015.

[21] D. Spallarossa, S. R. Kotha, M. Picozzi et al., "On-site earthquake early warning: a partially non-ergodic perspective from the site effects point of view," *Geophys J Int,* vol. 216, no. 2, pp. 919-934, Feb, 2019.

[22] A. Zollo, O. Amoroso, M. Lancieri et al., "A threshold-based earthquake early warning using dense accelerometer networks," *Geophys J Int,* vol. 183, no. 2, pp. 963-974, Nov, 2010.

[23] S. Nazeri, Z. H. Shomali, S. Colombelli et al., "Magnitude Estimation Based on Integrated Amplitude and Frequency Content of the Initial Wave in Earthquake Early Warning Applied to Tehran, Iran," *B Seismol Soc Am,* vol. 107, no. 3, pp. 1432-1438, Jun, 2017.

[24] Denilson, and Barbosa, "Understanding machine learning: from theory to algorithms," *Computing reviews*, 2016.

[25] V. S. Chaudhary, D. Kumar, B. P. Pandey et al., "Advances in Photonic Crystal Fiber-Based Sensor for Detection of Physical and Biochemical Parameters-A Review," *Ieee Sens J,* vol. 23, no. 2, pp. 1012-1023, Jan 15, 2023.

[26] A. Shadab, S. K. Raghuwanshi, and S. Kumar, "Advances in Micro-Fabricated Fiber Bragg Grating for Detection of Physical, Chemical, and Biological Parameters-A Review," *Ieee Sens J,* vol. 22, no. 16, pp. 15650-15660, Aug 15, 2022.

[27] Z. F. Li, "A Generic Model of Global Earthquake Rupture Characteristics Revealed by Machine Learning," *Geophys Res Lett,* vol. 49, no. 8, Apr 28, 2022.

[28] O. Sagi, and L. Rokach, "Ensemble learning: A survey," *WIRES DATA MIN KNOW,* vol. 8, no. 4, Jul-Aug, 2018.

[29] J. J. Hu, Y. T. Ding, H. Zhang *et al.*, "A Real-Time Seismic Intensity Prediction Framework Based on Interpretable Ensemble Learning," *Seismol Res Lett,* vol. 94, no. 3, pp. 1579-1602, May, 2023.

[30] M. S. Abdalzaher, M. S. Soliman, and S. M. El-Hady, "Seismic Intensity Estimation for Earthquake Early Warning Using Optimized Machine Learning Model," *Ieee T Geosci Remote,* vol. 61, 2023.

[31] H. C. Zhang, D. Melgar, V. Sahakian *et al.*, "Learning source, path and site effects: CNN-based on-site intensity prediction for earthquake early warning," *Geophys J Int,* vol. 231, no. 3, pp. 2186-2204, Sep 12, 2022.

[32] A. Datta, D. J. Wu, W. Q. Zhu *et al.*, "DeepShake: Shaking Intensity Prediction Using Deep Spatiotemporal RNNs for Earthquake Early Warning," *Seismol Res Lett,* vol. 93, no. 3, pp. 1636-1649, May, 2022.

[33] M. P. Gonzalez, and J. L. Zapico, "Seismic damage identification in buildings using neural networks and modal data," *Comput*

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3354857

Struct, vol. 86, no. 3-5, pp. 416-426, Feb, 2008.

[34] The data set is provided by Japan Meteorological Agency (https://www.data.jma.go.jp/svd/eqev/data/bulletin/shindo.html).

[35] The data set is provided by China Earthquake Networks Center, National Earthquake Data Center. (http://data.earthquake.cn).

[36] A. Katsumata, "Comparison of magnitudes estimated by the Japan Meteorological Agency with moment magnitudes for intermediate and deep earthquakes," *B Seismol Soc Am*, vol. 86, no. 3, pp. 832-842, Jun, 1996.

[37] T. Kunugi, S. Aoi, H. Nakamura *et al.*, "An Improved Approximating Filter for Real-Time Calculation of Seismic Intensity," *Journal of the Seismological Society of Japan*, vol. 65, pp. 223-230, 2013.

[38] National Research Institute for Earth Science and Disaster Resilience, NIED K-NET, National Research Institute for Earth Science and Disaster Resilience, doi:10.17598/NIED.0004

[39] R. Allen, "Automatic Phase Pickers - Their Present Use and Future-Prospects," *B Seismol Soc Am*, vol. 72, no. 6, pp. S225-S242, 1982.

[40] R. V. Allen, "Automatic Earthquake Recognition and Timing from Single Traces," *B Seismol Soc Am*, vol. 68, no. 5, pp. 1521-1532, 1978.

[41] C. W. Yang, K. W. Zhang, D. S. Wu et al., "Fractal Slope-Based Seismic Wave Detection Method," *B Seismol Soc Am*, vol. 113, no. 6, pp. 2311-2322, Dec, 2023.

[42] Z. H. Zhou, "When Semi-supervised Learning Meets Ensemble Learning," *Lect Notes Comput Sc*, vol. 5519, pp. 529-538, 2009.

[43] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123-140, Aug, 1996.

[44] M. Kearns, and L. G. Valiant, "Learning Boolean formulae or finite automata is as hard as factoring," 1988.

[45] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data," *Ieee T Syst Man Cy A*, vol. 41, no. 3, pp. 552-568, May, 2011.

[46] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5-32, Oct, 2001.

[47] T. Q. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

[48] G. L. Ke, Q. Meng, T. Finley *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv Neur In*, vol. 30, 2017.

[49] J. B. Zhu, S. Y. Li, Y. X. Wei et al., "On-site instrumental seismic intensity prediction for China via recurrent neural network and transfer learning," *J Asian Earth Sci*, vol. 248, Jun 1, 2023.

[50] G. Q. Zhang, D. N. Li, Y. Gao *et al.*, "A Magnitude Estimation Approach and Application to the Yunnan Earthquake Early Warning Network," *Seismol Res Lett*, vol. 94, no. 1, pp. 234-242, Jan, 2023.

[51] Y. M. Wu, and L. Zhao, "Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning," *Geophys Res Lett*, vol. 33, no. 16, Aug 24, 2006.

[52] Y. W. Wang, X. J. Li, L. Li et al., "New Magnitude Proxy for Earthquake Early Warning Based on Initial Time Series and Frequency," *Seismol Res Lett*, vol. 93, no. 1, pp. 216-225, Jan, 2022.

Authorized licensed use limited to: b-on: UNIVERSIDADE DO MINHO. Downloaded on February 26,2024 at 14:26:49 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.