Received: 25 July 2022 | Revised: 24 October 2022 | Accepted: 8 December 2022

DOI: 10.1002/eqe.3801

WILEY

SPECIAL ISSUE ARTICLE

Site classification using deep-learning-based image recognition techniques

Kun Ji[1] | **Chuanbin Zhu**[2] | **Saman Yaghmaei-Sabegh**[3] | **Jianqi Lu**[4,5] | **Yefei Ren**[4,5] | **Ruizhi Wen**[4,5]

[1]College of Civil and Transportation Engineering, Hohai University, Nanjing, China

[2]Department of Civil and Natural Resources Engineering, University of Canterbury, Christchurch, New Zealand

[3]Department of Civil Engineering, University of Tabriz, Tabriz, Iran

[4]Key Laboratory of Earthquake Engineering and Engineering Vibration, Institute of Engineering Mechanics, China Earthquake Administration, Harbin, China

[5]Key Laboratory of Earthquake Disaster Mitigation, Ministry of Emergency Management, Harbin, China

Correspondence
Chuanbin Zhu, Department of Civil and Natural Resources Engineering, University of Canterbury, Christchurch, New Zealand, Ilam, Christchurch 8041, New Zealand.
Email: chuanbin.zhu@canterbury.ac.nz

Funding information
Chinese National Natural Science Fund, Grant/Award Number: 51908518

Abstract

Classification of local soil conditions is important for the interpretation of structural seismic damage, which also plays a vital role in site-specific seismic hazard analyses. In this study, we propose to classify sites as an image recognition task using a deep convolutional neural network (DCNN)-based technique. We design the input image as a combination of the topographic slope and the mean horizontal-to-vertical spectral ratio (HVSR) of earthquake recordings. A DCNN model with five convolutional layers is trained using 1649 sites in Japan. The recall rates for site classes C, D, and E using our DCNN classifier for Japanese sites are 82%, 70%, and 60%, respectively. When compared with existing site classification schemes relying on predefined standard HVSR curves, our proposed method achieves the highest total accuracy rate (between 73% and 75%). The generality and applicability of our trained classifier are further validated using sites in Europe with a total accuracy between 64% and 66%. The proposed data-driven approach could be extended to other types of site amplification functions in the future.

KEYWORDS

deep convolutional neural network (DCNN), horizontal-to-vertical spectral ratio (HVSR), image recognition, site classification, topographic slope

1 | INTRODUCTION

Average shear-wave velocity down to 30 m depth (V_{S30}) is widely utilized in soil condition classification, for example, in the National Earthquake Hazard Reduction Program (NEHRP) and EuroCode8 (EC8). However, due to time and budgetary constraints, site classification using measured V_{S30} is not always feasible, especially for small projects (like family houses and small condominiums) or temporary strong-motion stations.[1]

Considering the limitations of V_{S30}-based site classification, many (e.g., ref. 2–4) used other proxies which are correlated with V_{S30}, for example, geological or topographic parameters which can be readily extracted from digital elevation models available on a global scale. Another popular method in site classification is the use of horizontal-to-vertical spectral ratio (HVSR) of seismic recordings or ambient noise data.[5] The HVSR approach does not need a nearby outcrop or borehole rock site as reference, with the assumption that the site amplification of vertical component of ground motions is negligible. This method is widely applied for site classification in a variety of countries, including Japan, Italy, Iran, and China (e.g., ref. 6–10).

It is shown by many that V_{S30} alone cannot describe the site amplification over the whole frequency range of engineering interest (e.g., ref. 1, 11). HVSR curves carry much more site-specific features of site response.[12] Several approaches were thus proposed to utilize the whole HVSR curve, rather than just its peak parameters, in site classification. Zhao et al.[6] suggested to classify sites according to the degree of absolute similarity between reference HVSR curves and the target curve. Ghasemi et al.[9] proposed using Spearman's correlation coefficient to evaluate the relative similarity between HVSR curves. To improve the classification performance, Yaghmaei-Sabegh and Tsang[13] applied artificial neural networks (ANN) to extract the shape feature of HVSR curves, and the results were validated using several earthquakes in Iran and Taiwan.[13–15] The general regression neural network (GRNN) approach was further developed by Ji et al.[16] for sites in China.

However, in these ANN or GRNN methods, reference HVSR curves or patterns are required. The reference curves are usually constructed by averaging the HSVR curves of sites within the same class. Because between-site variability in seismic site responses is not negligible,[17] a significant portion of site-specific information is lost in the averaging operation. The efficacy of the empirical HVSR-based classification approach is heavily dependent on the reference curves. Furthermore, the specific site classification criteria vary across regions, and the reference HVSR curves and classification scheme should ideally be tailored for each target region (e.g., ref. 18).

To tackle the limitation of previous site classification methods, we apply image recognition techniques to capture the shape characteristic of different HVSR curves without predefined references. The convolutional neural network (CNN) is designed to deal with multi-dimensional data such as images (e.g., ref. 19, 20). The objective of the convolution operation is to extract high-level features such as edges, from the input image. Through the use of appropriate filters, CNN can successfully capture the spatial and temporal dependencies in an image. Because of the reduced number of involved parameters and the reusability of weights, the network could be trained to better understand the sophistication of images. As one of the widely used deep learning approaches, CNN is typically not limited to a single convolutional layer and is often referred to as the deep convolutional neural network (DCNN).[21]

In this work, HSVR curves and topographic slopes are utilized to train DCNN models to classify NEHRP site classes. To identify NEHRP site classes for sites without V_{S30} measurements, we train DCNN models using the images of HVSR curves of earthquake recordings and topographic slopes at 1649 sites in Japan. The performance and robustness of proposed DCNN models are cross-validated considering different forms of input images and parameters. Then DCNN classifier is compared with empirical HVSR-based classification methods on the same dataset. The trained DCNN classifier is finally applied to European sites to test its generality.

2 | DATASET

2.1 | Training dataset

Because DCNN is a data-driven methodology, it is important to use an open and transparent training dataset so that others can validate and build on our results. Zhu et al.[22,23] constructed a comprehensive open-source site database of strong-motion stations in Japan. The dataset contains detailed site information for 1045 K-net and 697 KiK-net sites, for example, average shear-wave velocity to a certain depth, average HVSR curve (over available events), and topographic slope (from a 30 arc-second digital elevation model) for each site. For more details on the site database, we refer interested readers to the paper by Zhu et al.[23]

In our study, only sites with intact borehole information and more than five recordings are selected. After the screening, 1649 sites remain in the database (Figure 1A). KiK-net velocity profiles reach or exceed 30 m, thus their V_{S30} values are derived directly from available velocity profiles.[23] However, the depths of K-net profiles are less than 30 m, thus their V_{S30} values are estimated in this study from shallower velocity properties using the extrapolation method proposed by Wang and Wang.[24] The distribution of V_{S30} against slope is shown in Figure 1B. The Pearson coefficient between the two parameters is 0.40, which indicates a moderate correlation.

Based on measured V_{S30}, these 1649 sites are assigned to different classes according to the NEHRP classification scheme, with 121, 808, 612, and 108 sites in classes A + B, C, D, and E, respectively (Figure 1C). The NEHRP site classes are used as labels in subsequent model training. The numbers of A + B and E sites in our database are much smaller than those of the C and D sites. The imbalance ratio, defined as the ratio of the maximum and minimum sample size of different classes, is approximately 1:8. Because a deep learning model may learn insufficient features from underrepresented classes,[25] we take this problem into account when evaluating the DCNN classifier's performance.

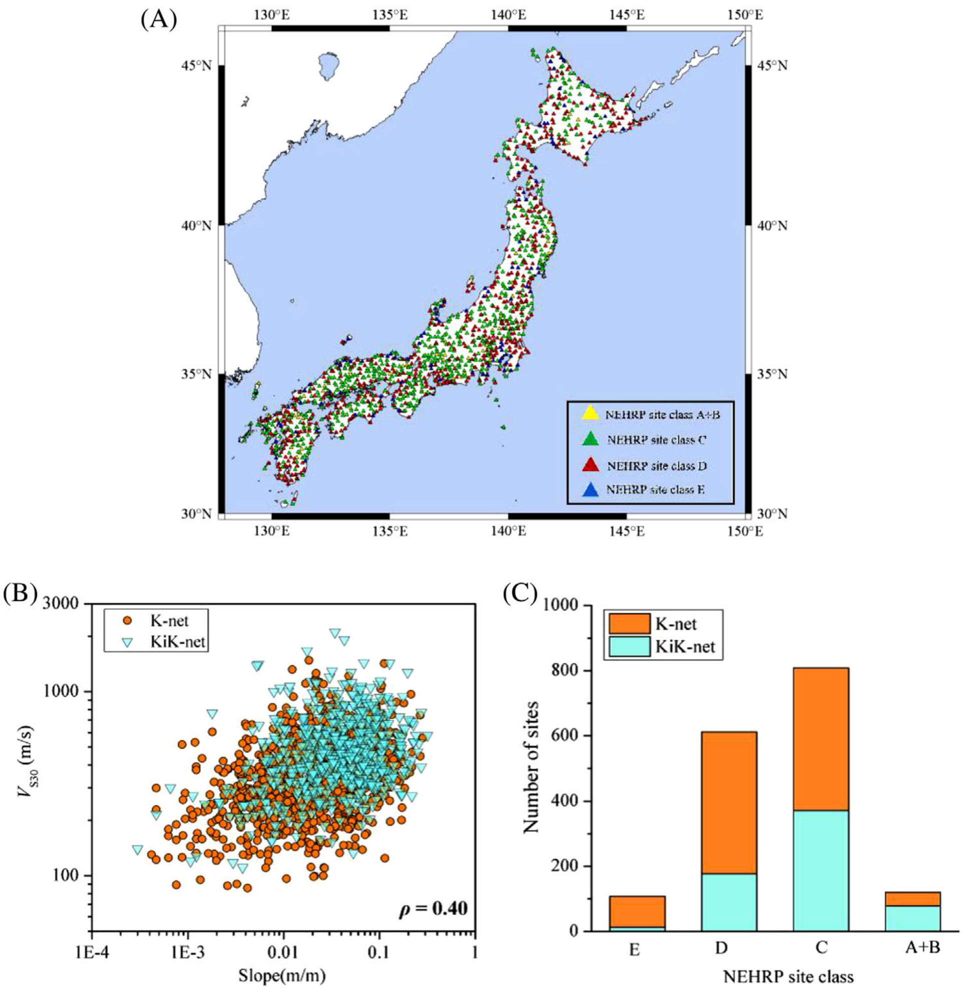

FIGURE 1 A total number of 1649 KiK-net and K-net sites, Japan, selected for model training. (A) Spatial distribution, (B) V_{S30} versus topographic slope where ρ represents the Pearson's correlation coefficient, and (C) Number of sites in different NEHRP site classes.

In previous site classification schemes, HSVR curves with no dominant peak are usually classified with a subjective threshold or directly excluded.[8,16,18] The significant peak of the HVSR curves at such sites is hard to recognize and, as a result, these sites tend to be misclassified in practice.[7] In the work by Zhu et al.,[23] an automatic HVSR peak picking scheme was proposed using the amplitude of the peak, height of the isolated peaks, and height/half-width of isolated peaks. After the exclusion of the HVSRs without predominant peaks, there are 1404 sites left. As illustrated in Figure 2A, A + B, C, D, and E site classes are assigned to 77, 684, 543, and 100 sites, respectively. The proportion of sites in each class does not differ significantly after the exclusion. The imbalance ratio is approximately 1:9, which is similar to the imbalance ratio (1:8) of the 1649 samples. The Pearson coefficient between slope and V_{S30} for the 1404 sites is 0.43 (Figure 2B), which is also similar to the value of 0.4 for the original dataset. Rock or stiff soil sites (A + B) tend to have a flat HVSR curve with no predominant peaks. However, due to the complex soil properties and variability of the site response, approximately 1/10 of the HVSRs of sites in C and D classes also have no significant peaks (Figure 2C). The amplitudes of HVSR curves are not used as a criterion for identifying the site class A + B because this would remove a large number of sites in classes C and D. Given that we do not need to compute the reference HVSR curve as a target in the DCNN scheme, the necessity of excluding HVSR curves with no significant peaks is investigated in this study.

2.2 | Test dataset

To test the applicability of the DCNN classifier in a region outside Japan, we select sites in the Pan-European region from the Engineering Strong Motion database (ESM, ref. 26). We apply the same site selection criteria as used in Japan, that is,

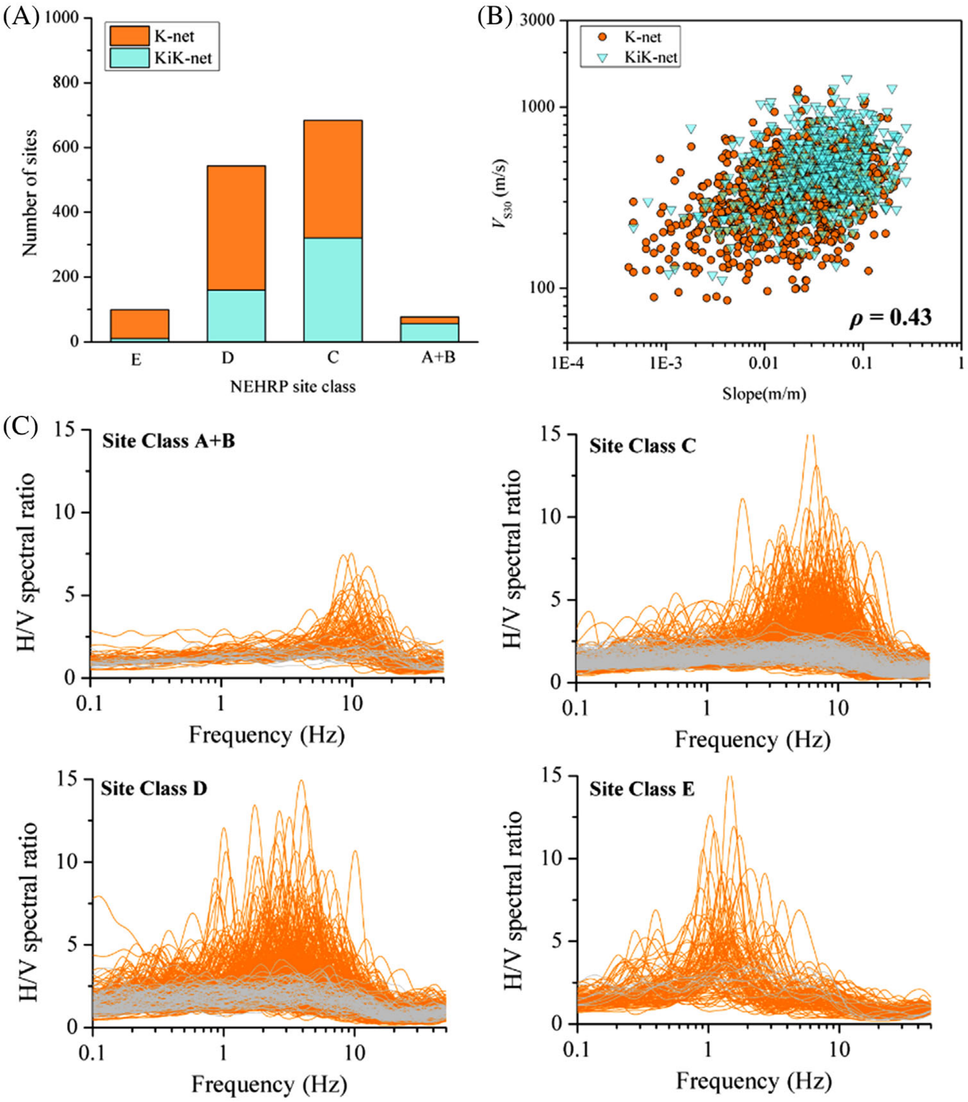

FIGURE 2 A total number of 1404 KiK-net and K-net training sites, Japan, with significant peaks identified on HVSR curves. (A) Number of sites in different NEHRP site classes, (B) V_{S30} versus topographic slope (ρ – Pearson's correlation coefficient), and (C) individual HVSR curves with gray lines indicating those without a significant peak.

sites with at least five earthquake recordings and with V_{S30} measurements. This leads to 217 sites in our testing dataset, as shown in Figure 3A. We then compute the average HVSR at each selected site following the same procedure as adopted in Japan. The topographic slope is collected from the ESM database. The Pearson correlation between V_{S30} and slope (Figure 3B) is stronger than that for the training data in Japan. The proportion of C and D sites is similar for the training and testing datasets (Figure 3C). However, in the testing set, there are only seven E sites, resulting in an imbalance ratio of 1:12. Among the 217 testing sites, only 118 sites have significant peaks on HVSR curves.

3 | METHOD

3.1 | DCNN model architecture

Firstly, we briefly review the basic concepts of the DCNN method. DCNN architecture is typically made of multiple convolutional layers, pooling (subsampling) layers, and fully connected layers.[21] Traditionally, the first convolutional layer is used to capture low-level feature information such as edges, colors, and so on. Convolution is used to move the filter bank across the input, resulting in activation at each receptive field, which combines to form a feature map. The

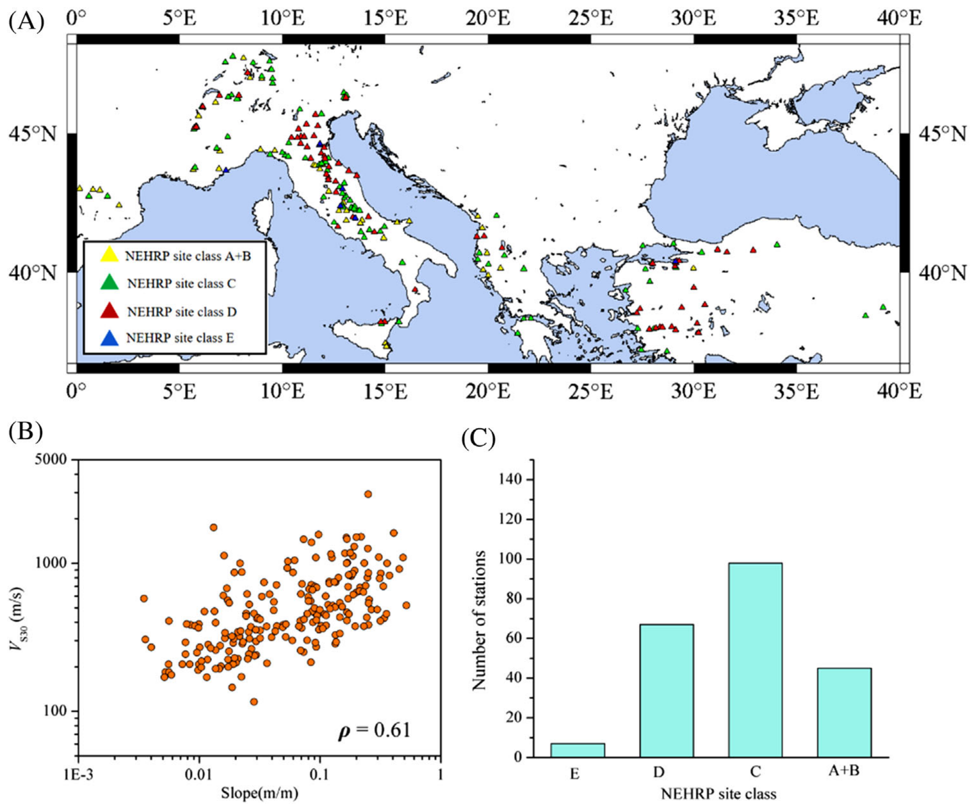

FIGURE 3 A total number of 217 ESM sites in Europe selected for model testing. (A) Spatial distribution, (B) V_{S30} versus topographic slope (ρ – Pearson's correlation coefficient), and (C) number of sites in different NEHRP site classes.

convolved features are then reduced in dimensionality by applying "valid padding", or remain the same using the "same padding" approach. With added layers, the architecture adapts to the high-level features and has a wholesome understanding of the images in the dataset. By mixing many filter banks into a single convolutional layer, the layer can learn to recognize multiple features in the input, and the resulting feature maps become the input of the next layer. Pooling layers are inserted after one or more convolutional layers to blend semantically similar features and reduce dimensionality.[27] Following the convolutional and pooling layers, the multi-dimensional output is flattened and passed to fully-connected layers for classification.

Figure 4 depicts the input image samples used in our study. The input is a 64 × 64 pixel image that is comprised of two parts: mean HVSR curve of the site and its topographic slope. The slope value is normalized using the minimum and maximum values of the dataset and then represented in the upper left corner of the image by a line segment with a fixed width (3 pixels) and varying length (1 to 60 pixels), as shown in Figure 4A. Parts of the sample training images for four NEHRP classes are shown in Figure 4B.

We do not directly use decimated vector of HVSR values at various frequencies and the scalar value of slope as input for two reasons: (1) In previous study,[16] the HVSR curves were used as input for GRNN models. It turns out the results are sensitive to the location of the predominant peak of curves. For curves without predominant peaks, it will lead to misleading results. In other words, the characteristics of "flat" curves could not be learned well by GRNN given limited number of the training samples. (2) Secondly, the single scalar slope value would be easily merged with the HVSR values at multiple frequencies. It is hard to decide the corresponding weight of the slope value in the input layer comparing with the HVSR curves. The bottom-up saliency-maps are plotted to detect the interesting points or areas in our input images, as shown in Figure 5. The saliency-maps are constructed using the algorithm based on natural image statistics as proposed by Kanan and Cottrell.[28] It is called a "bottom-up saliency map" because it does not depend on the specific classification target and is determined by the sparse visual features of the image itself. The HVSR curves and the slope segments are both recognized as the interesting features in our input images.

The DCNN model used in this study has five convolutional layers, four max pooling layers, one fully connected layer, and one output layer that corresponds to four target NEHRP site classes (Figure 6). The first convolutional layer has four

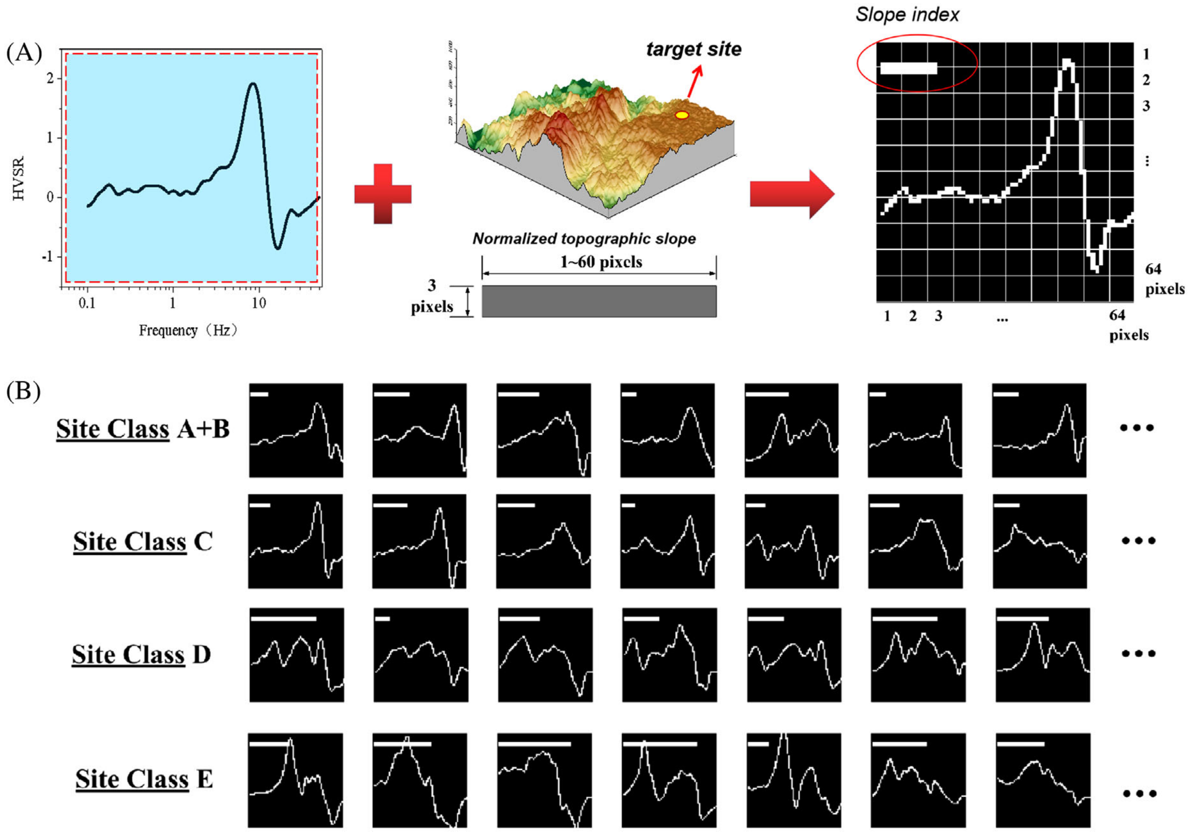

FIGURE 4 Input representation. (A) Construction of the 64 × 64 pixel image as input, which consists of HVSR curves and normalized topographic slope value, and (B) example training images for different NEHRP classes.

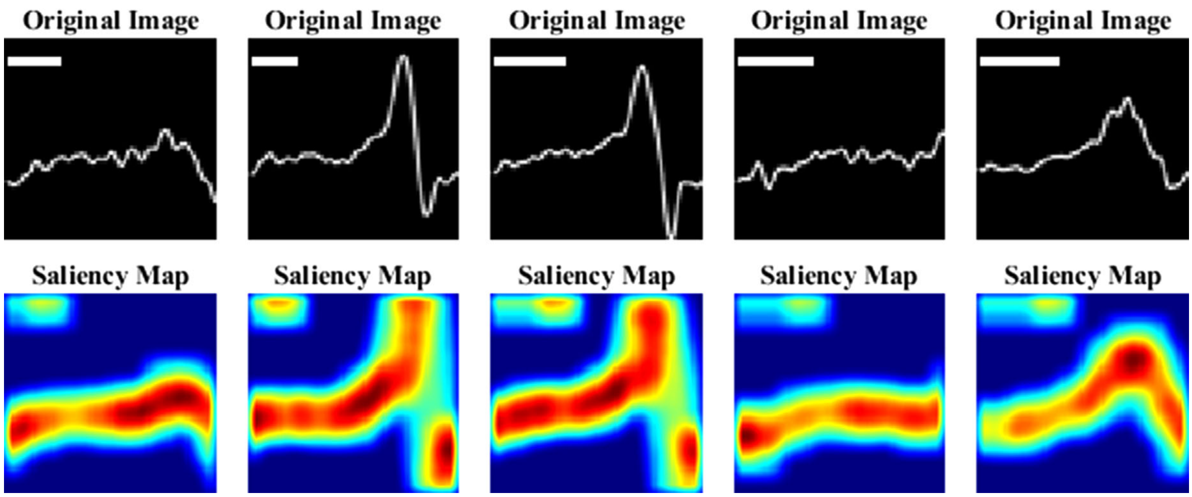

FIGURE 5 Samples of input images and their corresponding bottom-up saliency maps. The high and low salience values, indicating the interesting features in an image, are presented using red and blue colors, respectively.

filters and a single kernel. In the other four convolutional layers, the number of filters is 4, 8, 16, and 32, respectively. The kernel size of the filter in each convolutional layer is set to 3, the stride step is 2, and the padding type is "same". Except for the first one, each convolutional layer is followed by a batch normalization layer.[29] After the batch normalization layer, it is a max pooling layer with a max pooling size and stride of 2, and the padding type is "same". The final pooling layer is flattened and then fed to the fully connected layer, generating a depth of 4. The four neurons in the last layer correspond to four NEHRP site classes: A + B, C, D, and E, respectively. To avoid possible overfitting and improve generalizability, we

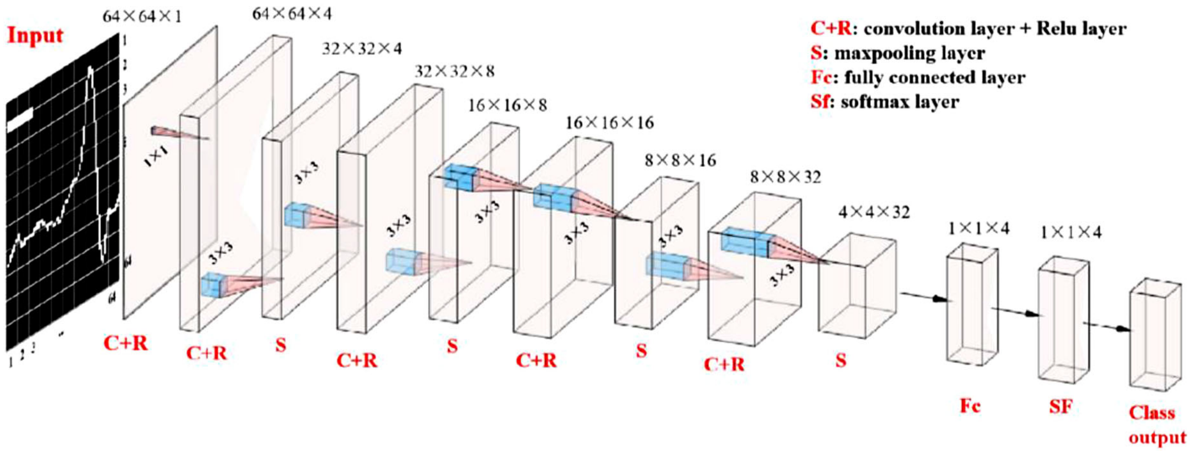

FIGURE 6 Architecture of the DCNN classification model. The size of the output in each layer is expressed as "height × width × depth" and is labeled in the figure.

TABLE 1 Four different datasets used for training the DCNN classifier

Case code	Including HVSR curves without significant peaks?	Including the slope index?	Number of sites
No. 1	Yes	No	1649
No. 2	Yes	Yes	1649
No. 3	No	No	1404
No. 4	No	Yes	1404

use dropout operation with a dropout rate of 0.5 before the fully connected layer.[30] Furthermore, the rectified linear unit (ReLU) activation function is applied to each pooling layer and fully connected layer.[31] Because a larger batch size usually results in poorer generalization,[29] a batch size of 50 with 50 epochs is utilized according to our experiments, which reveals a tradeoff between efficiency and generalizability. The DCNN model is trained using the Deep Learning Toolbox in Matlab and requires approximately 4 min to train (Intel[R] Core[TM] i7-8550U CPU with 18 GB memory). The corresponding train/validation example cases and script for the DCNN model are made available on Github (see data and resources).

3.2 | Training and validation

Before applying the DCNN approach for site classification in different regions, it is vital to validate the classifier's stability and robustness. If variability within sample observations is not negligible and the class in the training dataset is unbalanced, like in our classification problem, we must ensure that the model is not overfitted.[25] In this study, we use five-fold cross validation to evaluate the performance of the trained classifier. The entire dataset is divided into five equal-sized blocks. The classifier is then trained five times on four blocks before being tested on the fifth block. As illustrated in Figure 1–3, the number of sites in different site classes is unbalanced in our training and test datasets. There are more sites in classes C and D than in A + B and E. Therefore, we utilize stratified random sampling to ensure that each training and validation dataset has roughly the same representation. N-fold cross validation has a benefit over random subsampling in that the testing sets are distinct (non-overlapping) and can provide an objective evaluation of the trained classifier.

The DCNN model is trained using the stochastic gradient descent with momentum (SGDM) optimizer with an initial learning rate of 0.01 by optimizing a loss function defined as the cross entropy loss function.[32] The training process is terminated when the loss on the validation set exceeds or equals the previously smallest loss for a predefined number of times. After approximately 10 to 15 epochs, the accuracy rates on the training and validation datasets are stable, and the corresponding loss values do not fluctuate.

We design four cases to investigate whether including the HVSR curves without significant peaks is necessary and whether including the slope index improves classification performance, as shown in Table 1. The DCNN architecture and hyperparameter settings for these four cases are the same as described in Section 3.1.

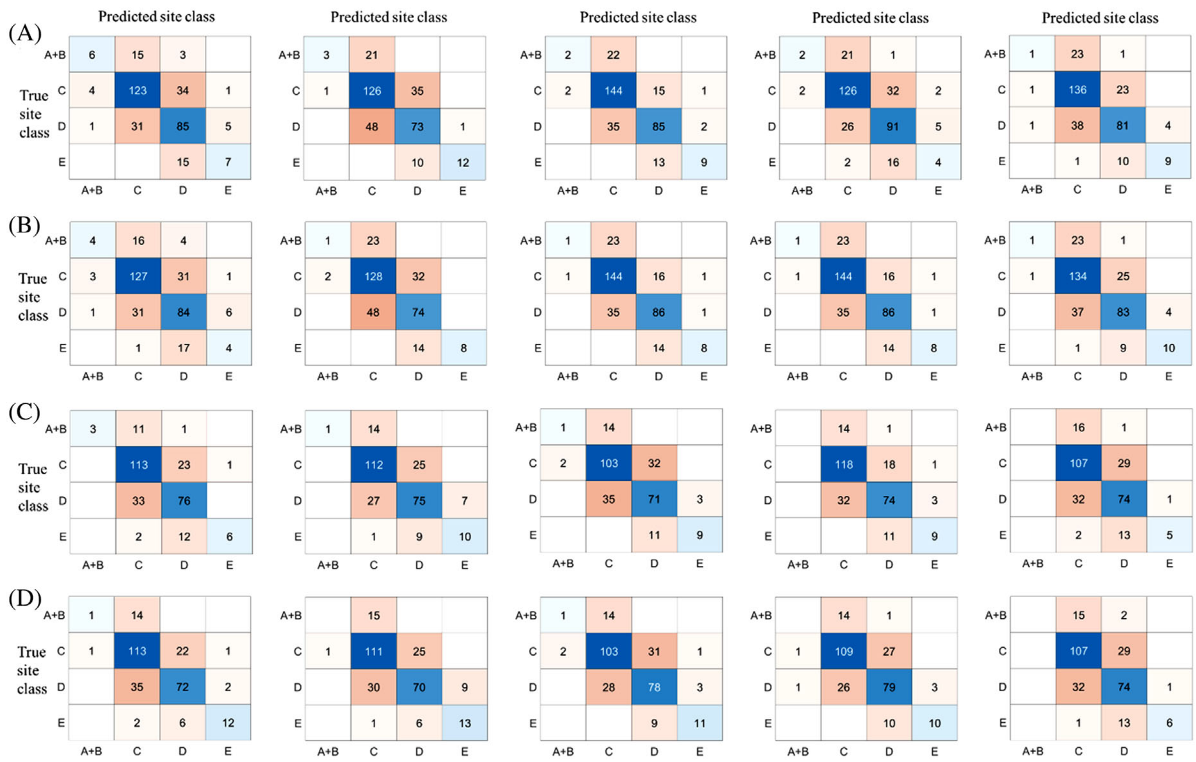

FIGURE 7 The confusion matrices for five-fold cross validation in four different cases in Table 1. (A) case no. 01, (B) case no. 02, (C) case no. 03, and (D) case no. 04.

More than 70% of the training sites are in class C or D. The success rate of the majority site class (i.e., C and D) may have a significant impact on the overall site classification accuracy rate. Therefore, we utilize the confusion matrix to objectively describe the classifier's performance for each site class. The confusion matrices of the validation folds in each case (Table 1) are illustrated in Figure 7 where x_{ij} indicates the number of class j sites that are recognized as class i. Hence, diagonal elements are those classified correctly, whereas others are misclassified. The number of correctly classified sites does not significantly vary between five-folds cross validation, indicating that our DCNN model is relatively stable. Overall, the DCNN model performs best for C and D sites and does not wrongly assign them to adjacent class A + B or E sites. In other words, most of the misclassified C and D sites are identified as D and C sites, respectively. This holds true for all the four studied cases in Table 1. The majority of A + B sites are incorrectly classified as class C. Almost half of the E sites are misclassified as class D.

To qualify the classification performance, we calculate the precision rate and recall rate for each validation fold. The precision rate quantifies the proportion of sites that are correctly classified. The recall rate for a certain site class is obtained by dividing the total number of sites in that class by the number of sites correctly classified. The total number of correct classifications among all sites is the overall accuracy rate. According to the definitions of precision rate and recall rate, there is a trade-off between these two metrics. Precision rate is sensitive to class imbalance because it considers the number of negative samples incorrectly labeled as positive. In contrast, the recall rate is insensitive to the class imbalance because it only depends on the positive classification results. Because the five-fold cross-validation is utilized, five DCNN models are trained for each case. Then the average and standard deviation of the precision/recall rate for five-folds are computed as the qualification metrics (Table 2). The recall rate for class C is around 70% to 80% in four studied cases (Table 2), which is approximately 10% higher than the corresponding precision rate. For class D, the precision rate and recall rate are similar, ranging from 65% to 68%. The similarity between recall rate and precision rate indicates that the classification performance for site class C and D is not significantly influenced by the class imbalance of the dataset. For classes A + B and E, however, the precision rate is much higher than the recall rate. The results are consistent with the previous observation that the classification performance for classes C and D is much better than A + B, regardless of the training subset. When A + B sites are excluded, the total accuracy rate increases from around 68% to 74%.

When the slope index is included in the input image, the mean value of recall rate of class C for case no. 02 increases by 11% compared with case no. 01 (without slope), and the overall accuracy rate increases by 5%. The largest standard

TABLE 2 The average and standard deviation of recall rate and precision rate among five trained folds

NEHRP site class	Case no. 01 Mean recall rate (std.[a])	Case no. 01 Mean precision rate (std.)	Case no. 02 Mean recall rate (std.)	Case no. 02 Mean precision rate (std.)	Case no. 03 Mean recall rate (std.)	Case no. 03 Mean precision rate (std.)	Case no. 04 Mean recall rate (std.)	Case no. 04 Mean precision rate (std.)
A + B	11 (8)	52 (22)	6 (5)	43 (3)	6 (8)	46 (3)	3 (3)	16 (2)
C	72 (16)	59 (22)	83 (5)	69 (3)	80 (4)	68 (3)	79 (3)	70 (2)
D	68 (5)	66 (5)	67 (4)	68 (6)	66 (4)	66 (6)	68 (4)	67 (3)
E	38 (13)	64 (21)	36 (11)	73 (23)	36 (9)	82 (11)	51 (15)	73 (9)
Overall accuracy	64 (5)		69 (3)		67 (2)		68 (1)	
Overall accuracy (without A + B)	69 (4)		74 (4)		71 (2)		73 (1)	

[a]std. refers to the standard deviation of recall/precision rate values of five training folds.

deviation of recall rates in case no. 01 also indicates that the stability of classifier could not be guaranteed. It suggests that including the slope index could improve classification performance when the HVSR curves are not screened by significant peaks.

For cases no. 03 and 04, in which the HVSR curves without predominant peaks are excluded, the inclusion of the slope index does not significantly influence the classification results of class C or D, and the overall accuracy rate is also similar to that in no. 02. For class E, the corresponding mean recall rate increases from 36% to 51% after the slope index is included. However, since the number of class E sites in the validation fold is small (approximately 20), we are inconclusive about the impact of including slope on class E sites.

No matter which strategy is utilized, the recall rate for site class A + B is not satisfactory, and the underlying reason will be discussed in Section 3.3. The comparison shows that implementation of the slope index is necessary when HVSR curves are not screened by significant peaks. If the HVSR curves with no significant peaks are already excluded, the slope index is not required in the input image according to Occam's Razor,[33] because the performance would not be improved significantly.

3.3 | DCNN ensemble

As shown in Figure 7 and Table 2, the DCNN classifier has good classification performance for site classes C and D, but accuracy for A + B and E sites needs to be improved further. It is observed that a large number of wrongly classified A + B and E sites are assigned to adjacent C and D sites, respectively. In other words, the A + B sites are not identified as D or E sites, and the E sites are not wrongly identified as A + B and C sites. Therefore, the multi-classification problem could be converted into a binary classification problem. Two extra classifiers are designed to distinguish A + B from C sites, and E from D sites, respectively. After the whole dataset is used to train the DCNN model (called classifier 1 hereafter), the classified A + B and C sites are used to train a new DCNN model, which is referred to as classifier 2. Then the D and E sites that are recognized using classifier 1 are utilized to construct classifier 3. This voting-like strategy is illustrated in Figure 8. It is worth noting that the training dataset is the only difference between these three classifiers. The DCNN architecture and hyperparameters are the same as described in Section 3.1.

The confusion matrices in Figure 8 show that the number of successfully classified A + B and E sites increases after applying classifiers 2 and 3. The recall rate for site classes A + B increases from 13.2% to 23.7%, while the recall rate for site class E increases from 44% to more than 60%. The recall rate for C and D sites is nearly identical, which suggests that using DCNN ensembles could improve classification performance for A + B and E sites without sacrificing accuracy for C and D sites.

To investigate the underlying reasons for the difficulty in identifying the A + B and E classes, the HVSR curves at correctly classified and misclassified sites by classifier 1 are compared in Figure 9. Flat HVSR curves are likely to be successfully classified as A + B sites by classifier 1, while most of the sites with predominant peaks at high frequency (around 10 Hz) are recognized as site class C (Figure 9A). This is also the characteristic of many HVSR curves at C sites (Figure 9B). The HVSR curves of C sites cover those of A + B sites (Figure 9C). This explains why the DCNN models do not distinguish A + B sites from C sites well using only the HVSR curves. Although a number of E sites are wrongly classified

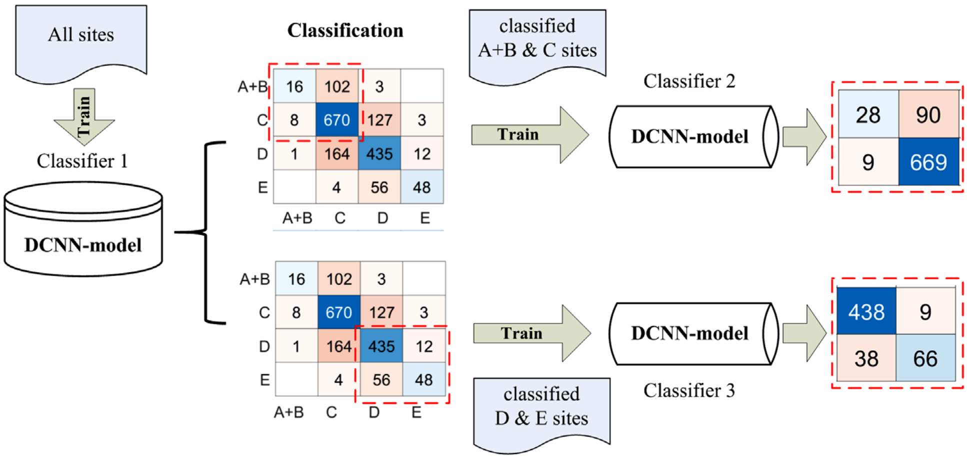

FIGURE 8 An ensemble of DCNN classifiers designed to distinguish A + B and E sites from C and D sites, respectively. The classifier 1 is the DCNN classifier trained using the whole training set. Classifier 2 and 3 are trained using the classified sites from classifier 1. The confusion matrices of these three classifiers are also presented.

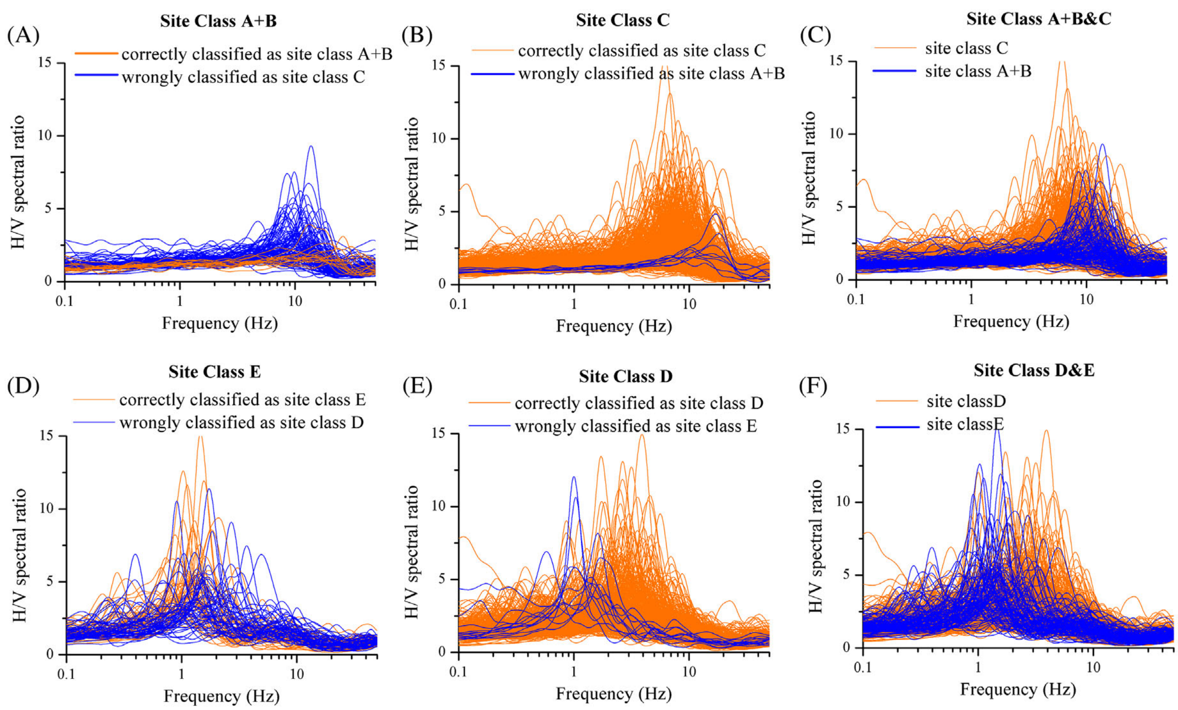

FIGURE 9 Comparison between HVSR curves of correctly classified sites and misclassified sites by classifier 1 regarding (A) class A + B, (B) class C, (D) class E, and (E) class D. The HSVR curves of site class A + B and C, and D and E recognized by classifier 1, are shown in (C) and (F), respectively.

as class D because of the similar curve shapes (Figure 9D,E), the predominant frequencies of site class E (around 1.0 Hz) tend to be lower than those of site class D (around 3.0 Hz) (Figure 9F). Therefore, a number of HVSR curves of E sites are not covered by the D sites, which may explain why we have a better classification for E sites than for A + B sites.

The DCNN model has a stable classification performance for C and D sites with recall rates greater than 70%. After applying an ensemble of classifiers, the recall rate for site class E reaches more than 60%. However, due to data limitations and similarities between the shapes of HVSR curves of A + B and C sites, A + B sites with a high predominant frequency are still not well recognized by the ensemble DCNN classifier. This is mainly due to the limitation of the V_{S30}-based NEHRP site classification criterion used to label the instances. V_{S30} is only one proxy and cannot capture all the features of the site amplification over the whole frequency range. Many alternative proxies have been proposed to give more refined site

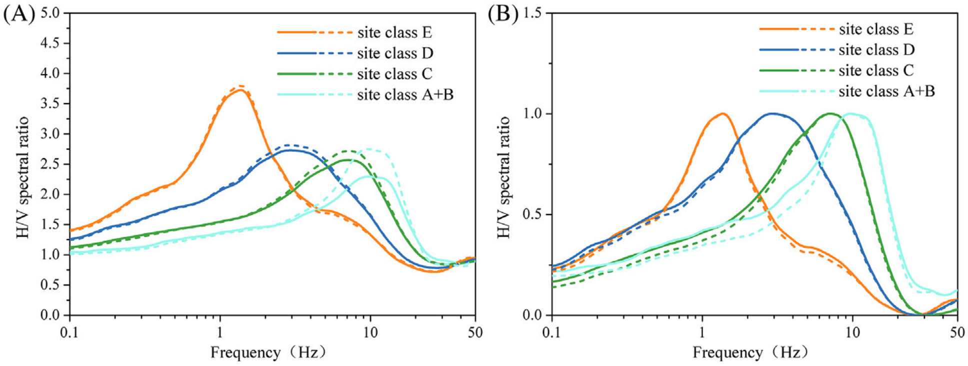

FIGURE 10 Reference HVSR curves (A) without normalization, and (B) with normalization, for site class A + B, C, D, and E. The solid lines represent the mean HVSR curves calculated using 1649 sites, and the dash lines represent those using 1404 sites (exclusion of HVSR curves without significant peaks).

classification scheme, which includes the shallow impedance information, the overall sedimentary thickness, and other parameters (e.g., ref. 34, 35). Kotha et al.[36] directly classified sites according to the site amplification functions. With a more reasonable site classification criterion or scheme, the classification performance of the DCNN model would be further improved. In current common practice, HVSR is generally treated as a supplement to V_{S30} index rather than as a replacement. But taking advantage of the available seismic waveforms to explore local site characteristics using empirical and theoretical approaches is a promising direction with accumulation of observation data. New methods based on seismicity data are promising due to cheaper sensors and more seismicity data (e.g., more small earthquakes detected by artificial intelligence techniques).

4 | RESULTS

4.1 | Comparing with other classification methods

Next, we compare the classification performance of our DCNN model with existing empirical classification schemes. The empirical methods proposed by ref., 6, 9 identifying site classes according to the similarity in the shape of HVSR curves, are referred to as Zhao06 and Ghasemi09 methods hereafter. Another method is the GRNN model developed by Ji et al.[16] based on the work of Yaghmaei-Sabegh and Tsang.[13]

The first step of all these methods, including the GRNN model, is to derive the mean HVSR curves as targets or reference patterns for classification. Given that the datasets and site classification criteria used in these studies are not identical, we recalculate the average mean HVSR curves using our dataset in Japan. The reference HVSR curves are built using HVSR curves from 1649 and 1404 sites respectively, as shown in Figure 10A. After excluding HVSR curves without significant peaks, the amplitude of the mean HVSR curve of site classes A + B increases significantly. This means that many relatively flat HVSR curves are left out of the 1404-site dataset. The mean HVSR curves of the other three classes are nearly identical before and after exclusion. As shown in Figure 10B, if the mean HVSR curves are normalized, the shape difference is not obvious before and after exclusion, including the site class A + B.

We compare the classification results of existing methods with our DCNN model on the 1649 sites (Figure 11). The overall accuracy rates for Zhao06 and Ghasemi09 methods are approximately the same, reaching 44% to 45%. They have a better classification performance for site class A + B and site class E with a recall rate nearly 60% to 80%, while more than half of sites in class C and D are misclassified. According to Zhao et al.,[6] the HVSR curves of rock and hard soil sites are emphasized at about 7 Hz and have a nearly uniform shape, thus soft soil can be easily identified due to its high peaks at low frequency. This partially explains the good performance for site classes A + B and E. Since C and D sites dominate the dataset, the total accuracy rate is relatively low. The total accuracy rate of the GRNN model using the workflow recommended by Ji et al.[16] is 54%, with recall rates of 73%, 44%, 57%, and 78% for site classes A + B, C, D, and E, respectively. Though the GRNN model achieves a higher total accuracy rate than Zhao06 and Ghasemi09, its lead on C and D sites is rather limited. A large number of C and D sites are still wrongly classified by GRNN as A + B and E sites.

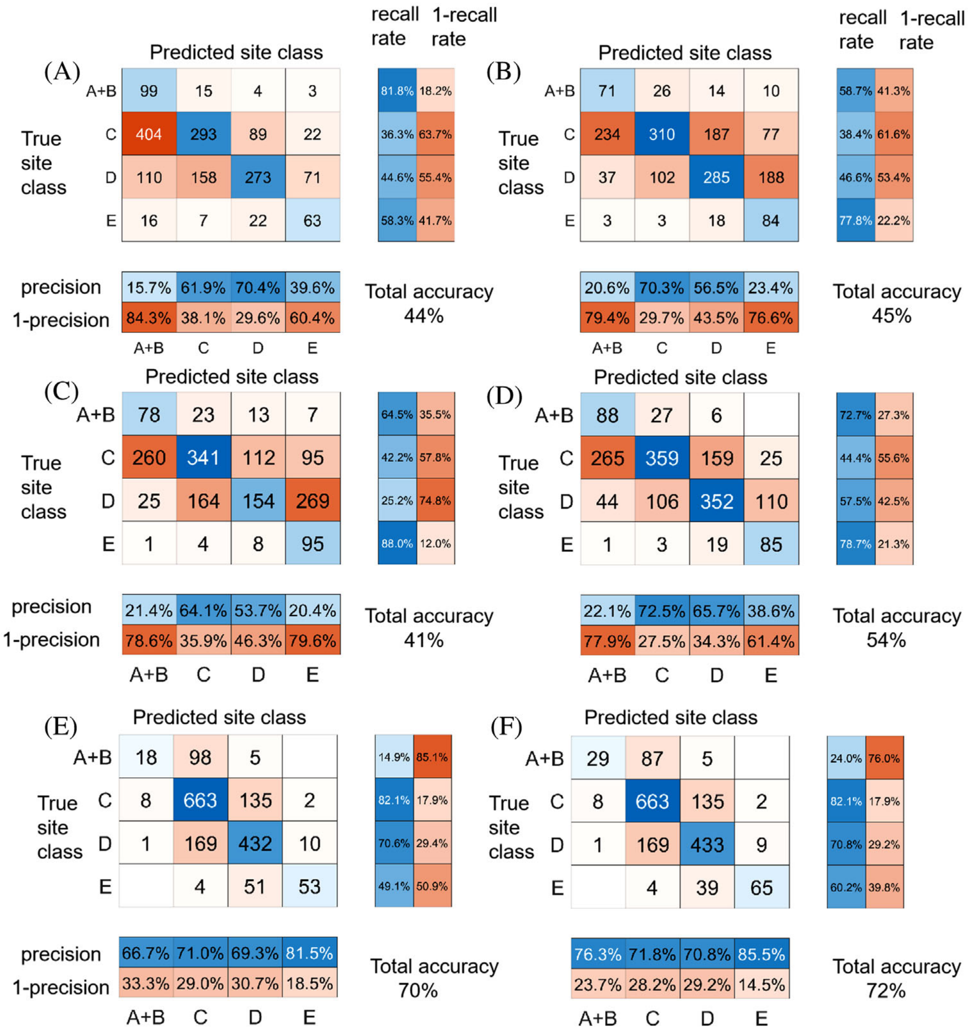

FIGURE 11 The confusion matrices regarding classification results on the 1649 sites using (A) Zhao06 method, (B) Ghasemi09 method, (C) GRNN method using normalized HVSR curves as reference patterns, (D) GRNN method using non-normalized HVSR curves as reference patterns, (E) DCNN classifier 1, and (F) DCNN ensemble.

Compared with the existing methods on the 1649 sites, the DCNN classifier performs best for site classes C and D. The recall and precision rates are both nearly 70%, and the overall accuracy rate is 73%. When the DCNN ensemble is utilized, the recall rates of A + B and E classes increase to 24% and 60%, respectively. The overall accuracy increases from 70% into 72%. The majority of misclassified sites are identified as the adjacent site classes, which outperforms other methods. In this way, sites can potentially avoid being classified into a category that is completely different from the true category. On the 1404 sites, of which the HVSR curves without significant peaks are removed, the classification results are similar to those on the 1649 sites.

4.2 | Testing in Europe

In this section, we test the classification performance of the trained DCNN classifiers on the European sites. For the test dataset of 217 European sites, we choose the classifier that is trained on the 1649 Japanese sites using input images that include the slope index. After the HVSR curves without significant peaks are excluded from the testing dataset, 118 European sites remain. The classifier trained on the 1404 Japanese sites is used for these 118 sites. The confusion matrices between classification results using classifier 1 and the ensemble are compared in Figure 12.

For the case of 217 sites, the results indicate that the recall rate of classification results of site class C, D, and E are similar to the results for the dataset in Japan, reaching approximately 70%. Although the recall rates for E sites are as high as 70%,

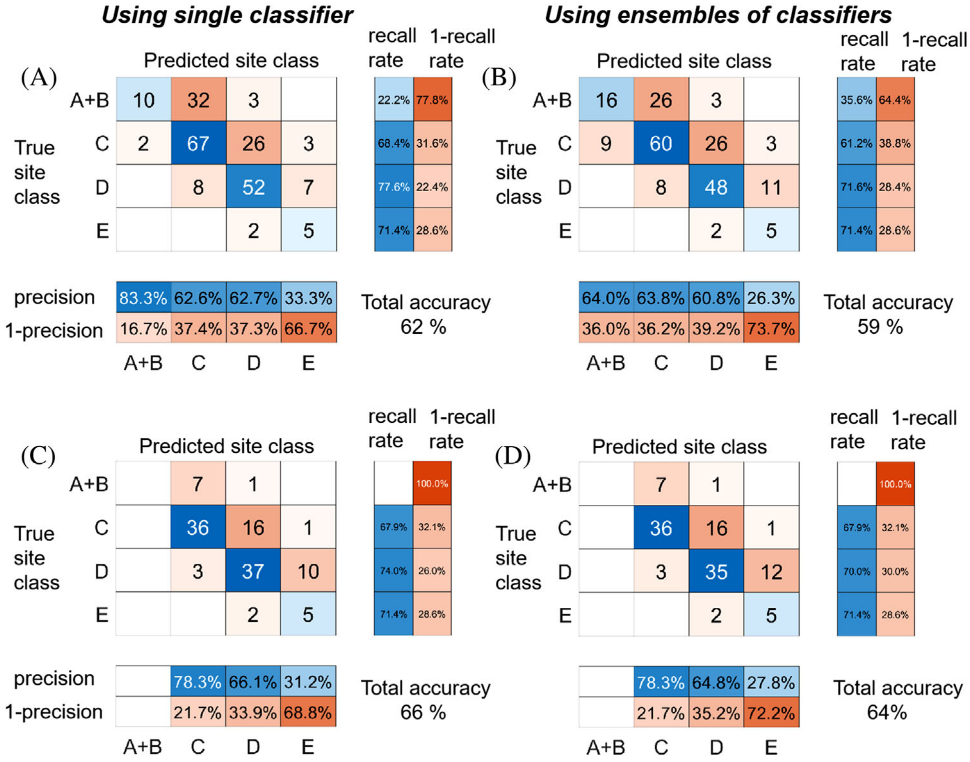

FIGURE 12 The confusion matrices regarding classification results of the DCNN classifier 1 and ensemble on testing sites in Europe. (A,B) Correspond to the dataset of 217 sites whose HVSR curves are not excluded by the predominant peaks. (C,D) Correspond to the dataset of 118 sites of which the HVSR curves without predominant peaks are excluded.

this is mainly related to the small total number of E sites (i.e., seven). Therefore, we would not expect our classifiers' recall rate for E sites to be as good as C and D sites. Because the number of E sites is too small, utilizing an ensemble of classifiers does not show significant improvement compared with using DCNN classifier 1. But using the ensemble still has merits in practice because we often do not know the proportion of different site classes for a given region a priori. It is illustrated that the ensembles at least do not adversely affect the classification performance of C and D sites. For the testing on the 118 sites, no A + B sites are recognized because the flat HVSR curves are removed, leaving only the HVSR curves with predominant peaks at high frequency. Overall, the constructed DCNN classifier has the same level of classification accuracy on the testing sites in Europe as on the training sites in Japan. This implies that our DCNN model can be applied to European sites without retraining, suggesting its generality.

5 | DISCUSSION

We utilize the Grad-CAM (Gradient-weighted Class Activation Map) to explain our DCNN classification results.[37,38] The Grad-CAM can detect the parts that our trained network focused on to give the corresponding prediction. The Grad-CAM technique uses the gradients of the classification score concerning the final convolutional feature map. As shown in Figure 13, the parts of an image with a large value for the Grad-CAM map are those that significantly impact the network scores for the identified class.

For the identification of site classes E and D, the Grad-CAM map shows that the neural network focused on some peaks of the HVSR curves. No obvious relation is found between the results and specific frequency band. In addition, the slope segment is not recognized as featuring parts in classification of these two site classes. For site class C, the Grad-CAM map shows that the neural network focused on the empty space outlined by the HVSR curves and slope segment. The map suggests that the neural network used the empty space in the image to detect the site class C. For some samples in site class A + B, with relatively flat curves without predominant peaks, the value of Grad-CAM map is zero, indicating that no parts of the image are recognized. Site class A + B is not directly classified but is predicted by exclusion of other classes. This also partially explains why the performance of site classification of A + B is not as good as other site classes.

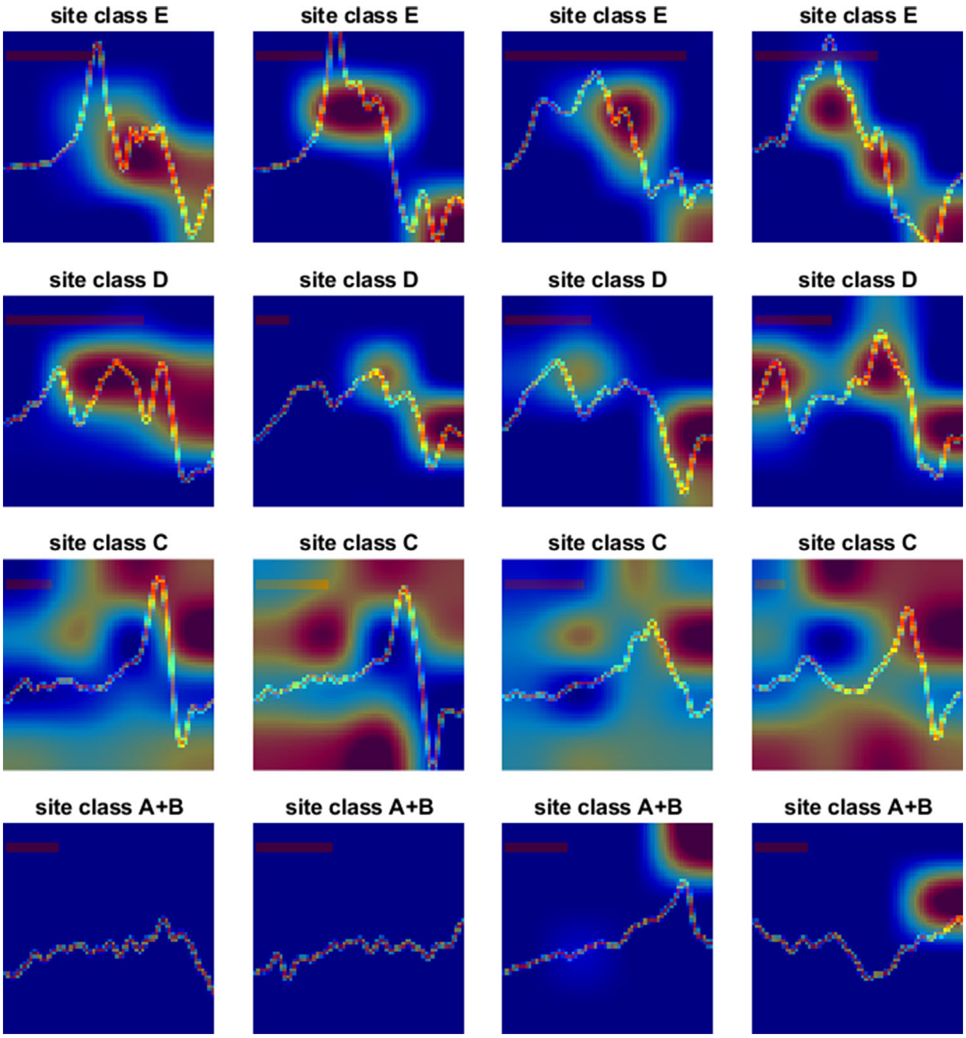

FIGURE 13 The Grad-CAM map for our trained DCNN classifier regarding sample images of different site classes. The image parts with the highest values are highlighted in red, as they have the greatest impact on the network scores for the predicted class.

Although we use HVSR curves in DCNN in this study, other types of site amplification functions can also be used as input to the data-driven framework, such as site terms of empirical ground motion model[39] and site responses from the generalized inversion technique.[40] As a data-driven approach, the DCNN model can be retrained by replacing the input images with those of other site amplification curves. In addition, HVSR curves of microtremor recordings could be utilized as input to retrain the model, which is appealing for cases without earthquake recordings.

6 | CONCLUSIONS

In this study, we used deep convolutional neural networks (DCNN) to classify sites into NEHRP classes from images of earthquake-based HVSR curves and topographic information. We trained the DCNN model on 1649 strong ground motion recording sites (instances) in Japan and tested it on 217 sites in Europe.

Using DCNN based classifier for Japanese sites, the recall rates for C, D, and E sites are 82%, 70%, and 60%, respectively. Three existing HVSR site classification schemes have a better performance for site classes A + B and E with a recall rate nearly 60% to 80% while more than half of C and D sites are misclassified. The total accuracy rate of DCNN classifier (72%) is much higher than previous schemes (lower than 55%). When tested on sites in Europe, the DCNN classifier achieves a recall rate of more than 70% for site classes C, D, and E. The classification performance for E sites needs to be further validated with more collected data. Another potential advantage of our DCNN classifier is that the majority of misclassified sites are assigned to adjacent site classes. In this way, sites can potentially avoid being classified into a category that is far different from the true site condition. Sites in class A and B with a high predominant frequency could not be distinguished from those in class C by the DCNN model, which holds true for sites in both Japan and Europe. This is because the NEHRP site classes (classification labels) are based on V_{S30}, which alone cannot sufficiently distinguish different sites.

The proposed DCNN-based site classification scheme is a promising method for characterizing site conditions using image recognition techniques. It is demonstrated that the trained DCNN classifier can be applied to different regions without retraining. It is worth noting that, as a data-driven method, the classification accuracy of the DCNN model could be improved with more training data in specific regions. Other types of site amplification functions can also be used as input to the data-driven framework, such as site-response curves from multi-station-based empirical approaches and HVSR curves of microtremor recordings.

ACKNOWLEDGMENTS

This work is financially supported by the Chinese National Natural Science Fund (grant number. 51908518). We would like to thank the Editor Dr. Henry Burton and two anonymous reviewers for providing very constructive feedback. We thank National Research Institute for Earth Science and Disaster Prevention (NIED), and Istituto Nazionale di Geofisica e Vulcanologia (INGV) for making their strong ground motion data publicly available.

DATA AVAILABILITY STATEMENT

The Deep Learning Toolbox in Matlab was used to construct the DCNN model. The Matlab script and data for model development, as well as example cases are made available on Github: https://github.com/JIKUN1990/DCNN-site-classification. Site data in Japan were collected from an open-source site database (v1.0.0) (Zhu et al.[23]) via https://doi.org/10.5880/GFZ.2.1.2020.006. Data in Europe were derived from the ESM strong-motion flatfile 2018 (Lanzano et al., 2018[41]): https://doi.org/10.13127/esm/flatfile.1.0. The Generic Mapping Tools (GMT) was utilized to produce some of the figures.

ORCID

Kun Ji https://orcid.org/0000-0003-0314-5216
Saman Yaghmaei-Sabegh https://orcid.org/0000-0003-0285-3528

REFERENCES

1. Gallipoli MR, Mucciarelli M. Comparison of site classification from VS 30, VS 10, and HVSR in Italy. *Bull Seismol Soc Am*. 2009;99(1):340-351.
2. Wald DJ, Allen TI. Topographic slope as a proxy for seismic site conditions and amplification. *Bull Seismol Soc Am*. 2007;97(5):1379-1395.
3. Wills CJ, Petersen M, Bryant WA, et al. A site-conditions map for California based on geology and shear-wave velocity. *Bull Seismol Soc Am*. 2000;90(6B):S187-S208.
4. Lee CT, Cheng CT, Liao CW, Tsai YB. Site classification of Taiwan free-field strong-motion stations. *Bull Seismol Soc Am*. 2011;91(5):1283-1297.
5. Nakamura Y. *A Method for Dynamic Characteristics Estimation of Subsurface Using Microtremor on the Ground Surface*. Quarterly Reports 30. Railway Technical Research Institute; 1989:25-33.
6. Zhao JX, Irikura K, Zhang J, et al. An empirical site-classification method for strong-motion stations in Japan using H/V response ratio. *Bull Seismol Soc Am*. 2006;96:914-925.
7. Fukushima Y, Bonilla LF, Scotti O, Douglas J. Site classification using horizontal-to-vertical response spectral ratios and its impact when deriving empirical ground-motion prediction equations. *J Earthquake Eng*. 2007;11(5):712-724.
8. Di Alessandro C, Bonilla LF, Boore DM, Rovelli A, Scotti O. Predominant-period site classification for response spectra prediction equations in Italy. *Bull Seismol Soc Am*. 2012;102(2):680-695.
9. Ghasemi H, Zare M, Fukushima Y, Sinaeian F. Applying empirical methods in site classification, using response spectral ratio (H/V): a case study on Iranian strong motion network (ISMN). *Soil Dyn Earthquake Eng*. 2009;29(1):121-132.
10. Wen R, Ren Y, Zhou Z, Li X. Temporary strong-motion observation network for Wenchuan aftershocks and site classification. *Eng Geol*. 2014;180:130-144.
11. Lee VW, Trifunac MD. Should average shear-wave velocity in the top 30m of soil be used to describe seismic amplification? *Soil Dyn Earthquake Eng*. 2010;30(11):1250-1258.
12. Zhu C, Cotton F, Kawase H, Haendel A, Pilz M, Nakano K. How well can we predict earthquake site response so far? *Earthq Spectra*. 2022;38(2):1047-1075.
13. Yaghmaei-Sabegh S, Tsang HH. A new site classification approach based on neural networks. *Soil Dyn Earthquake Eng*. 2011;31(7):974-981.
14. Yaghmaei-Sabegh S, Tsang HH. Site class mapping based on earthquake ground motion data recorded by regional seismographic network. *Nat Hazards*. 2014;73(3):2067-2087.
15. Yaghmaei-Sabegh S, Rupakhety R. A new method of seismic site classification using HVSR curves: a case study of the 12 November 2017 Mw 7.3 Ezgeleh earthquake in Iran. *Eng Geol*. 2020;270:105574.

16. Ji K, Ren Y, Wen R, Zhu C, Liu Y, Zhou B. HVSR-based site classification approach using general regression neural network (GRNN): case study for china strong motion stations. *J Earthquake Eng*. 2021;26:8423–8445.
17. Zhu C, Cotton F, Kwak DY, Ji K, Kawase H, Pilz M. Within-site variability in earthquake site response. *Geophys J Int*. 2022;229(2):1268-1281.
18. Ji K, Ren Y, Wen R. Site classification for National Strong Motion Observation Network System (NSMONS) stations in China using an empirical H/V spectral ratio method. *J Asian Earth Sci*. 2017;147:79-94.
19. LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput*. 1989;1(4):541-551.
20. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324.
21. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.
22. Zhu C, Weatherill G, Cotton F, Pilz M, Kwak DY, Kawase H. An open-source site database of strong-motion stations in Japan: K-NET and KiK-net. V. 1.0.0. *GFZ Data Services*. 2020. https://doi.org/10.5880/GFZ.2.1.2020.006
23. Zhu C, Weatherill G, Cotton F, Pilz M, Kwak DY, Kawase H. An open-source site database of strong-motion stations in Japan: K-NET and KiK-net (v1. 0.0). *Earthq Spectra*. 2021;37(3):2126-2149.
24. Wang HY, Wang SY. A new method for estimating VS (30) from a shallow shear-wave velocity profile (depth< 30m). *Bull Seismol Soc Am*. 2015;105(3):1359-1370.
25. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data*. 2019;6(1):1-54.
26. Lanzano G, Luzi L, Russo E, et al., Engineering Strong Motion Database (ESM) flatfile [Data set]. Istituto Nazionale di Geofisica e Vulcanologia (INGV). 2018. https://doi.org/10.13127/esm/flatfile.1.0
27. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.
28. Kanan C, Cottrell GW. Robust classification of objects, faces, and flowers using natural image Statistics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010; 2010.
29. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR. 2015.
30. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. 2012.
31. Agarap AF, Deep learning using rectified linear units (relu). arXiv: 1803.08375. 2018.
32. Bishop CM. *Pattern Recognition and Machine Learning*. Springer; 2006.
33. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Occam's razor. *Inf Process Lett*. 1987;24(6):377-380.
34. Luzi L, Puglia R, Pacor F, Gallipoli M, Bindi D, Mucciarelli M. Proposal for a soil classification based on parameters alternative or complementary to Vs, 30. *Bull Earthquake Eng*. 2011;9(6):1877-1898.
35. Pitilakis K, Riga E, Anastasiadis A. New code site classification, amplification factors and normalized response spectra based on a worldwide ground-motion database. *Bull Earthquake Eng*. 2013;11(4):925-966.
36. Kotha SR, Cotton F, Bindi D. A new approach to site classification: mixed-effects ground motion prediction equation with spectral clustering of site amplification functions. *Soil Dyn Earthquake Eng*. 2018;110:318-329.
37. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision. 2017;618-626.
38. Vinogradova K, Dibrov A, Myers G. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. In Proceedings of the AAAI conference on artificial intelligence. 2020;34(10):13943-13944.
39. Atkinson GM. Single-station sigma. *Bull Seismol Soc Am*. 2000;105(96):446-455.
40. Andrews DJ. Objective determination of source parameters and similarity of earthquakes of different size. In: Das S, Boatwright J, Scholz CH, eds. *Earthquake Source Mechanics*. American Geophysical Union; 1986: 259-268.
41. Lanzano G, Sgobba S, Luzi L, et al., The pan-European Engineering Strong Motion (ESM) flatfile: compilation criteria and data statistics. *Bull Earthquake Eng*. 2018;17:561-582

How to cite this article: Ji K, Zhu C, Yaghmaei-Sabegh S, Lu J, Ren Y, Wen R. Site classification using deep-learning-based image recognition techniques. *Earthquake Engng Struct Dyn*. 2022;1-16.
https://doi.org/10.1002/eqe.3801