TYPE Original Research
PUBLISHED 16 October 2023
DOI 10.3389/feart.2023.1267386

OPEN ACCESS

EDITED BY
Biswajeet Pradhan,
University of Technology Sydney,
Australia

REVIEWED BY
Pijush Samui,
National Institute of Technology Patna,
India
Xianyang Qiu,
Central South University, China

*CORRESPONDENCE
Byungmin Kim,
✉ byungmin.kim@unist.ac.kr

RECEIVED 26 July 2023
ACCEPTED 26 September 2023
PUBLISHED 16 October 2023

CITATION
Kim J, Kang J-D and Kim B (2023), Machine-learning models to predict P- and S-wave velocity profiles for Japan as an example.
Front. Earth Sci. 11:1267386.
doi: 10.3389/feart.2023.1267386

Machine-learning models to predict P- and S-wave velocity profiles for Japan as an example

Jisong Kim[1], Jae-Do Kang[2] and Byungmin Kim[1]*

[1]Department of Civil, Urban, Earth, and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, [2]Earthquake Disaster Mitigation Center, Seoul Institute of Technology, Seoul, Republic of Korea

Wave velocity profiles are significant for various fields, including rock engineering, petroleum engineering, and earthquake engineering. However, direct measurements of wave velocities are often constrained by time, cost, and site conditions. If wave velocity measurements are unavailable, they need to be estimated based on other known proxies. This paper proposes machine learning (ML) approaches to predict the compression and shear wave velocities (V_P and V_S, respectively) in Japan. We utilize borehole databases from two seismograph networks of Japan: Kyoshin Network (K-NET) and Kiban Kyoshin Network (KiK-net). We consider various factors such as depth, N-value, density, slope angle, elevation, geology, soil/rock type, and site coordinates. We use three ML techniques: Gradient Boosting (GB), Random Forest (RF), and Artificial Neural Network (ANN) to develop predictive models for both V_P and V_S and evaluate the performances of the models based on root mean squared errors and the five-fold cross-validation method. The GB-based model provides the best estimation of V_P and V_S for both seismograph networks. Among the considered factors, the depth, standard penetration test (SPT) N-value, and density have the strongest influence on the wave velocity estimation for K-NET. For KiK-net, the depth and site longitude have the strongest influence. The study confirms the applicability of commonly used machine-learning techniques in predicting wave velocities, and implies that exploring additional factors will enhance the performance.

KEYWORDS

shear wave velocity, compression wave velocity, machine learning, gradient boosting, random forest, artificial neural network, cross-validation

1 Introduction

Compression and shear wave velocities (V_P and V_S, respectively) are often employed to assess the properties of underground rock environments and to design geotechnical projects. V_P and V_S are significant in various fields such as rock mechanical property calculations (Chang et al., 2006; Ameen et al., 2009; Jamshidi et al., 2018; Rahman and Sarkar, 2021), pore structure identification (Eberli et al., 2003; Panza et al., 2019), lithology determination (Pickett, 1963; Deng et al., 2017), fluid saturation (Si et al., 2016; Roy et al., 2017; Ding et al., 2019), seismic liquefaction (Samui et al., 2011; Karthikeyan and Samui, 2014; Jena et al., 2023), seismic site responses, and ground motion predictions (Fiorentino et al., 2019; Harmon et al., 2019; Kim, 2019). Such wave velocities are measured by invasive tests such as down-hole test, cross-hole test, and suspension PS logging, as well as non-destructive tests such as Multichannel Analysis of Surface Wave (MASW), Spectral Analysis of Surface Wave (SASW), and Multichannel Simulation with One Receiver (MSOR). However, these tests are often constrained by time, cost, and site conditions (Hasancebi and Ulusay, 2007; Anemangely et al., 2019; Xiao et al., 2021).

To address the problems mentioned in the prior paragraph, numerous researchers have proposed indirect methods to estimate V_P or V_S. For instance, several studies have presented the relationships between V_P and the mechanical properties of rock materials, such as uniaxial compressive strength (Pappalardo, 2015), density (Yasar and Erdogan, 2004), and porosity (Sousa et al., 2005). Various correlation models between V_S and standard penetration test (SPT) resistance (N-value) have also been suggested (e.g., Ohta and Goto, 1978; Andrus et al., 2004; Akin et al., 2011; Sil and Haloi, 2017; Bajaj and Anbazhagan, 2019). For example, Kwak et al. (2015); Tsai et al. (2019) inferred V_S using empirical equations conditioned on the N-value and other independent variables such as vertical effective stress and soil type. Rahimi et al. (2020) presented the effect of soil aging on SPT-V_S correlations. Furthermore, researchers have predicted the time-averaged shear-wave velocity in the upper 30 m of soil deposits (V_{S30}) based on various proxies, such as topographic slope, surface geology, elevation, and terrain type (e.g., Kottke et al., 2012; Parker et al., 2017; Kwok et al., 2018; Heath et al., 2020).

The demand for Machine learning (ML) applications has been increasing as huge volumes of data are accessible over a computer network. ML algorithms are well suited for making regression models on complex data-driven problems. Researchers have studied V_P or V_S estimation based on ML (e.g., Singh and Kanli, 2016; Paul et al., 2018; Anemangely et al., 2019; Dumke and Berndt, 2019; Wang and Peng, 2019; Zhang et al., 2020). In particular, Dumke and Berndt (2019) used the Random Forest (RF) regression algorithm to estimate V_P as a function of depth on global marine locations. They used data from 333 boreholes and considered 38 geological variables, such as site coordinates, sediment thickness, and depth below the seafloor. They validate the ML model using 10-fold cross-validation (CV). Paul et al. (2018) used an Artificial Neural Network (ANN) algorithm on data from five wells in India to estimate the V_P. Singh and Kanli (2016) used an ANN to estimate V_S in an oil field located in southeastern Turkey. Anemangely et al. (2019) adopted the least square version of the support vector machine (LSSVM) algorithm combined with three optimization algorithms to predict V_S using data from two oilfields located in the southwest of Iran.

This study aims to train the three ML algorithms (i.e., gradient boosting, random forest, and artificial neural network) to estimate both V_P and V_S in Japan. We utilize borehole databases, covering all of Japan from two seismograph networks: Kyoshin Network (K-NET) and Kiban Kyoshin Network (KiK-net). We consider various factors such as depth, N-value, density, slope angle, elevation, geology, soil/rock type, and site coordinates. We quantitatively evaluate the prediction performances of the ML-based algorithms based on five-fold cross-validation and evaluate the relative importance of the factors.

2 Data

In this study, we obtained site data from two seismograph networks of Japan, Kyoshin Network (K-NET) and Kiban Kyoshin Network (KiK-net), where the National Research Institute for Earth Science and Disaster Resilience (2019) has operated since 1996. Each site of these two networks has profiles of V_P, V_S, and soil/rock types. In addition, the K-NET site has profiles of standard penetration test (SPT) resistance values (N-values) and density with a depth interval of 1 m. The energy efficiency is unknown for the borings at the K-NET sites (Kwak et al., 2015). Therefore, we utilized unnormalized N-values. Because of the inconsistent datasets between the two seismograph networks, we considered training the ML models for each network.

For the datasets, the velocity profile data were resampled to a depth interval of 1 m. For all of the K-NET sites, a minimum depth interval is 1 m. Furthermore, approximately 43% of KiK-net sites have minimum depth intervals of 1 m or shorter. Therefore, we consider that resampling the profile data into a depth interval of 1 m is reasonable. We also screen the suspicious profile data such as those with the velocity of zero. In addition to the depth-dependent variables provided by the networks, we also considered the following five depth-independent variables: site latitude, site longitude, geology, topographic slope angle, and elevation. The geology map was obtained from the seamless digital geological map of Japan (1: 200,000) (Geological Survey of Japan, 2015), and the slope angle and elevation were obtained from the digital elevation map (DEM) of the Shuttle Radar Topography Mission (SRTM) with a resolution of 30 m. We then used the nine independent variables (i.e., site longitude, site latitude, geology, slope angle, elevation, N-value, density, depth, soil/rock type) for the K-NET, and seven (i.e., site longitude, site latitude, geology, slope angle, elevation, depth, and soil/rock type) for the KiK-net sites, as summarized in Table 1.

We considered all sites where all of the variables were available: 996 K-NET sites with 15,253 data samples for each of V_P and V_S and 677 KiK-net sites with 136,315 data samples for V_P and 132,855 data samples for V_S. The dataset information is summarized in Table 1. The considered sites (i.e., recording stations) covering Japan are shown in Figure 1.

The distributions of the numerical variables for the K-NET and KiK-net datasets are shown in Figure 2 and Figure 3, respectively. The depth to the bottom of the borehole (D^{bh}) at the K-NET sites ranges from 5 to 20 m, with 83% concentration at 10 m and 20 m (Figure 2A). The elevation ranges from −3 m to 1,502 m, 75% of which are positioned under 179 m, as shown in the boxplot above the histogram (Figure 2B). The slope angle ranges from 0° to 30.87°, with 75% below 5.26° (Figure 2C). The 96 outliers are observed as circular forms in each boxplot (Figures 2B, C). The N-value with depth ranges from 0 to 500 with 69% below 90, where the four outliers are observed: three of which are 375, and one is 500 (Figure 2D). The density with depth is distributed from 0.69 g/cm^3 to 2.82 g/cm^3 with 75% under 1.98 g/cm^3, in which 306 outliers are detected (Figure 2E). The V_S is distributed from 37 m/s to 2,350 m/s with 75% slower than 450 m/s (Figure 2F). The V_P ranges from 140 m/s to 5,270 m/s with 75% slower than 1,800 m/s (Figure 2G). For categorical variables in K-NET sites in our dataset, 12 unique soil/rock types according to depth and 110 unique types of geology are observed.

Figure 3A depicts the D^{bh} of the KiK-net sites ranging from 92 m to 2,000 m, with 75% under 199 m, where 46 outliers are observed. Figure 3B presents the site elevation ranging from −5 m to 1,302 m with 75% below 330 m, where 34 outliers are detected. Figure 3C shows the slope angle that ranges from 0° to 36.23° with 75% under 10.42°, where 12 outliers are observed. Figure 3D shows the V_S

TABLE 1 Datasets used in this study.

Seismograph network	Dependent variables	Independent variables	Description
K-NET	V_P	Site longitude	1) 996 sites (V_P)
	V_S	Site latitude	2) 996 sites (V_S)
		Geology	3) 15,253 data samples (V_P)
		Slope angle	4) 15,253 data samples (V_S)
		Elevation	5) Velocity profiles were sampled at every 1-m depth interval
		N-value	
		Density	
		Depth	
		Soil/rock type	
KiK-net	V_P	Site longitude	1) 677 sites (V_P)
	V_S	Site latitude	2) 675 sites (V_S)
		Geology	3) 136,315 data samples (V_P)
		Slope angle	4) 132,855 data samples (V_S)
		Elevation	5) Velocity profiles were sampled at every 1-m depth interval
		Depth	
		Soil/rock type	

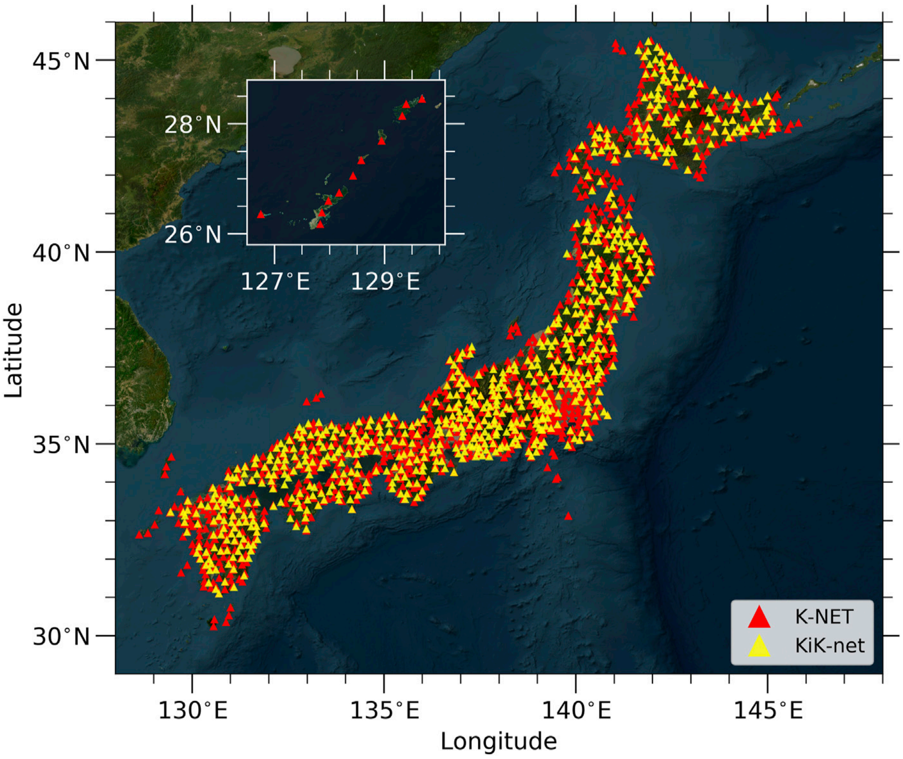

FIGURE 1
Locations of K-NET and KiK-net recording sites used in this study.

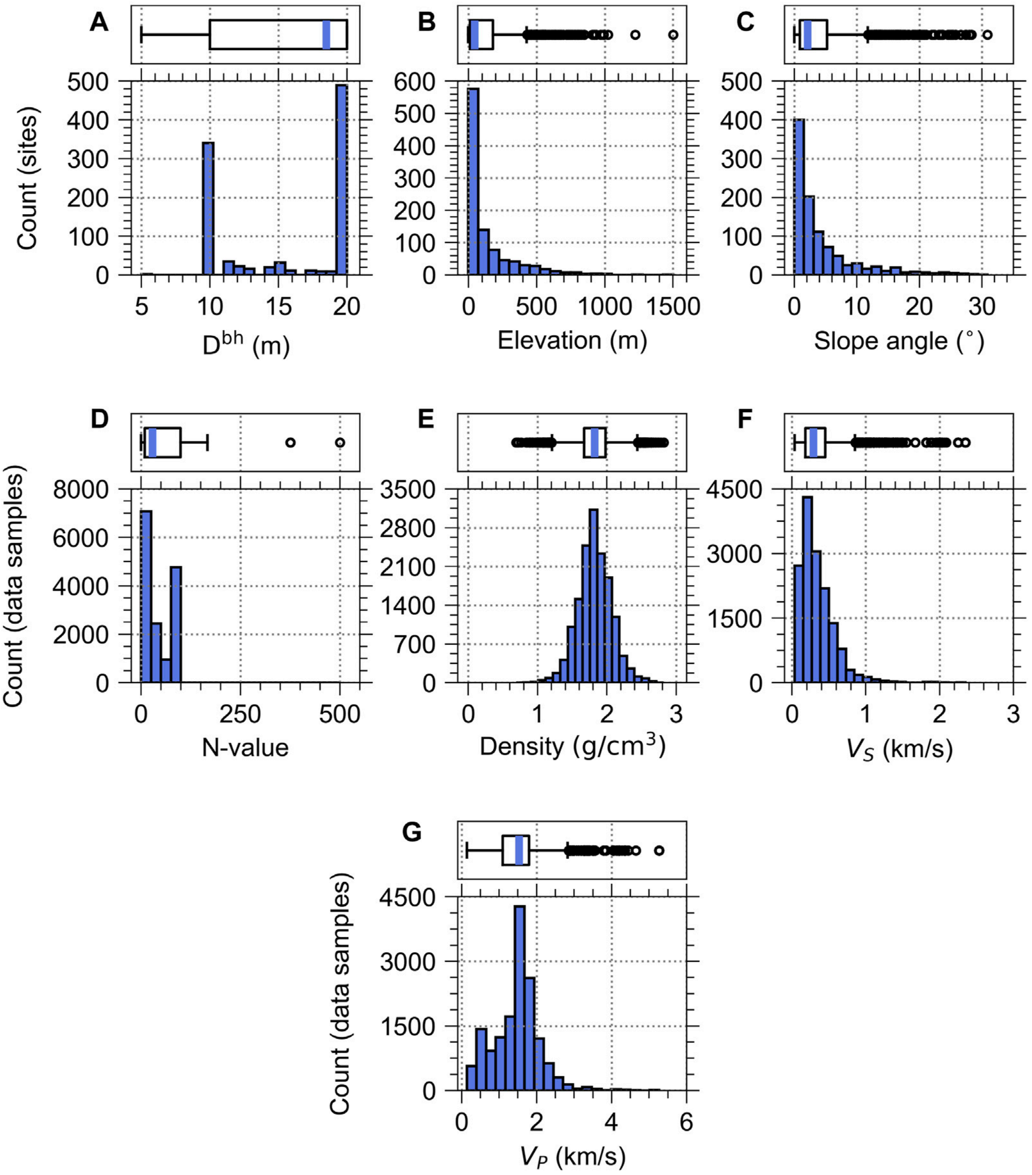

FIGURE 2
Variables distribution for K-NET dataset used in this study: **(A)** depth to the bottom of the borehole (D^{bh}), **(B)** elevation, **(C)** slope angle for the sites (i.e., 996 sites), and **(D)** N-value, **(E)** density, **(F)** V_S, **(G)** V_P for data samples (i.e., 15,253 data samples). In the boxplot above the histogram, the blue line represents a median value, and the box represents 25 and 75 percentiles of the data.

ranging from 20 m/s to 3,500 m/s with 75% slower than 1,720 m/s. Figure 3E presents the V_P that ranges from 50 m/s to 6,100 m/s with 75% slower than 3,830 m/s. For categorical variables, soil/rock type with depth has 588 unique features, among which the features include the various combinations of several soil types, such as 'sand and gravel', 'shale with gravel', and 'sandstone and mudstone'. Furthermore, 119 unique geological classes are observed.

3 Machine learning (ML) models

The ML model uses the following variables: the depth and depth-related information (i.e., N-value, density, soil/rock type), and site information (i.e., coordinates, slope angle, elevation, geology) described in Table 1 to infer V_P or V_S on a specific depth (e.g., 15 m) of the site. This section describes the ML algorithms utilized for V_P and V_S prediction. We illustrated them using all K-NET data samples for V_S as an example. We used Scikit-learn (Pedregosa et al., 2011) for the implementation of Gradient Boosting (GB) and Random Forest (RF) algorithms and the Tensorflow (Abadi et al., 2016) for the Artificial Neural Network (ANN) algorithm. Note that comparing these three algorithms is a popular practice in the field of machine learning-based studies (e.g., Krauss et al., 2017; Kim et al., 2020; Jun, 2021; Seo et al., 2022). These methods represent different types of machine learning algorithms and have been proven effective in handling complicated relationships within various datasets. Given their proven reliability, we employed such methods to assess the

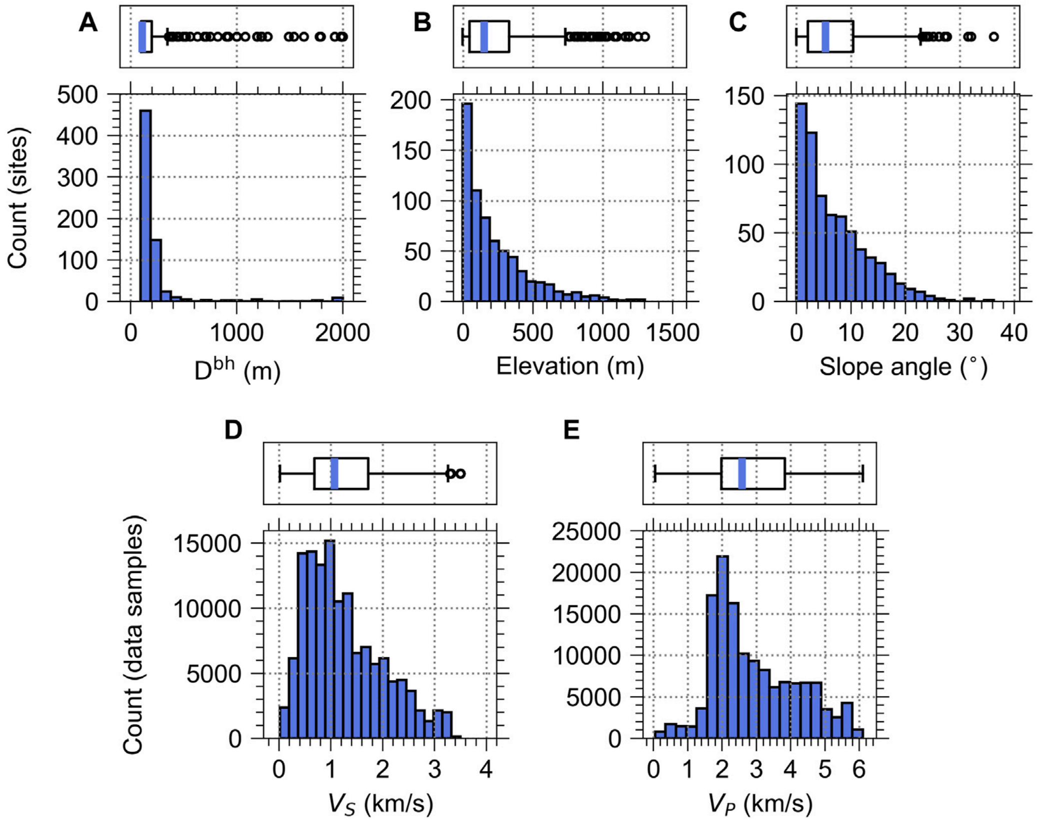

FIGURE 3
Variables distribution for KiK-net dataset used in this study: **(A)** depth to the bottom of the borehole (D^{bh}), **(B)** station elevation, **(C)** station slope angle for the sites of V_P dataset (i.e., 677 sites), and **(D)** V_S, **(E)** V_P for data samples (i.e., 132,855 and 136,315 data samples, respectively). In the boxplot above the histogram, the blue line represents a median value, and the box represents 25 and 75 percentiles of the data.

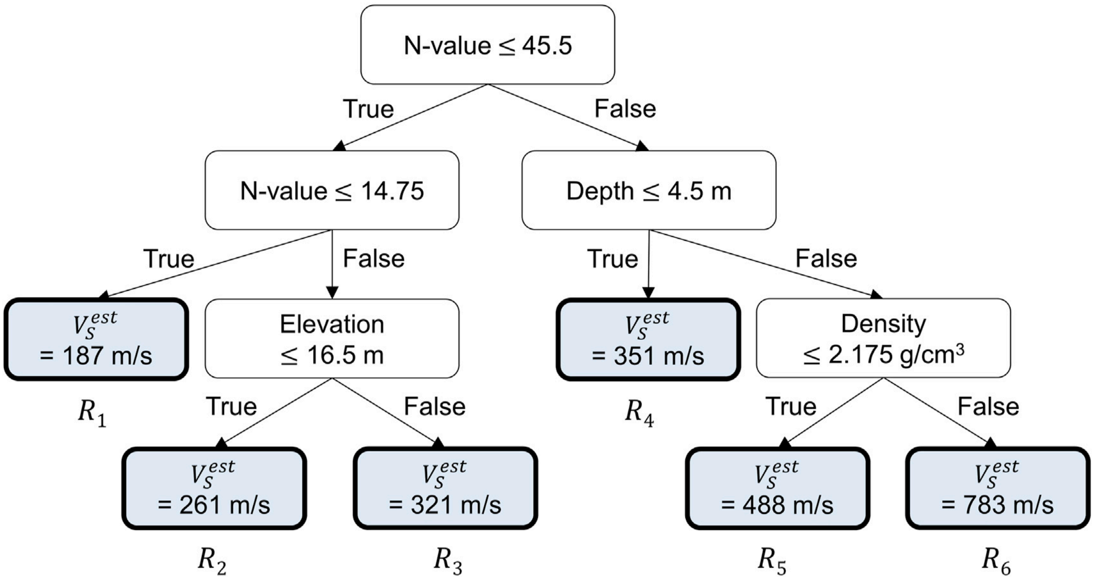

FIGURE 4
Example of the decision tree using the variables of the K-NET dataset.

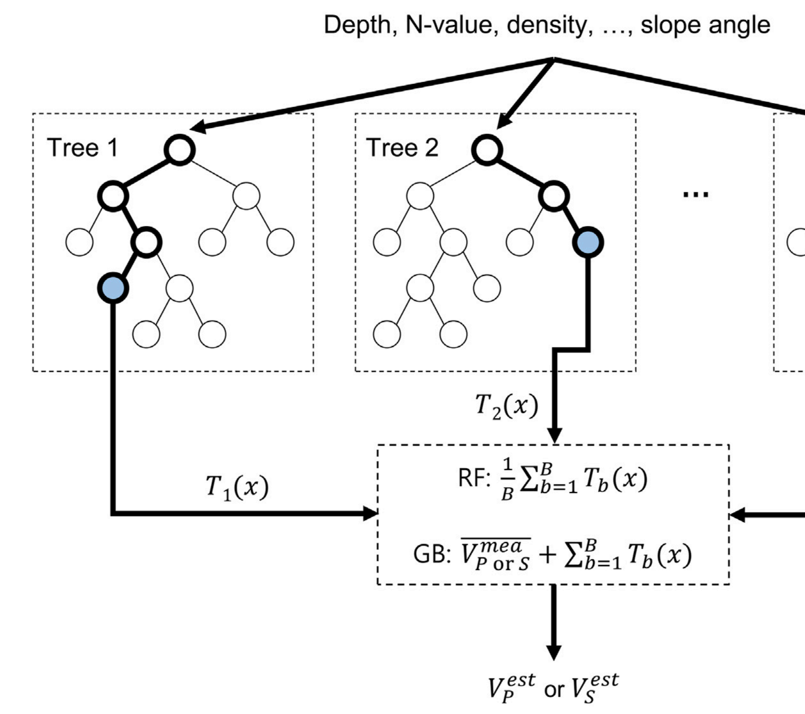
Depth, N-value, density, …, slope angle
Tree 1
Tree 2
…
Tree B
$T_2(x)$
$T_1(x)$
RF: $\frac{1}{B}\sum_{b=1}^{B} T_b(x)$
GB: $\overline{V_{P\text{ or }S}^{mea}} + \sum_{b=1}^{B} T_b(x)$
$T_B(x)$
V_P^{est} or V_S^{est}

FIGURE 5
Architecture of ensemble learning methods (Random Forest and Gradient Boosting).

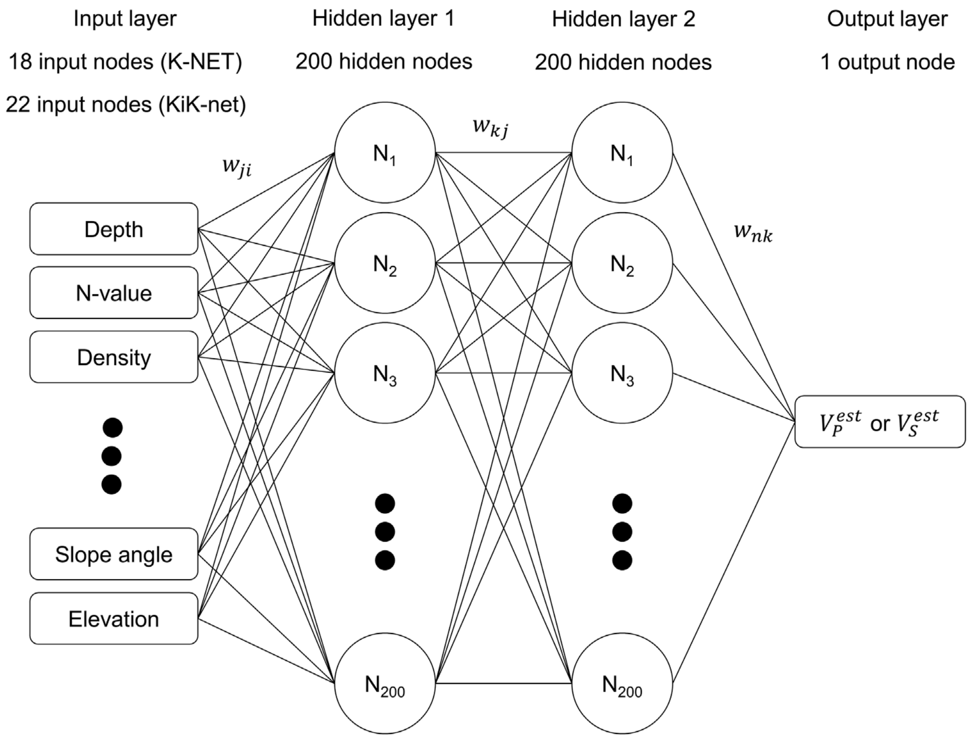

FIGURE 6
Architecture of the ANN-based model consists of an input layer, two hidden layers with 200 nodes (N), and an output layer. The weights between nodes (*w*) are illustrated.

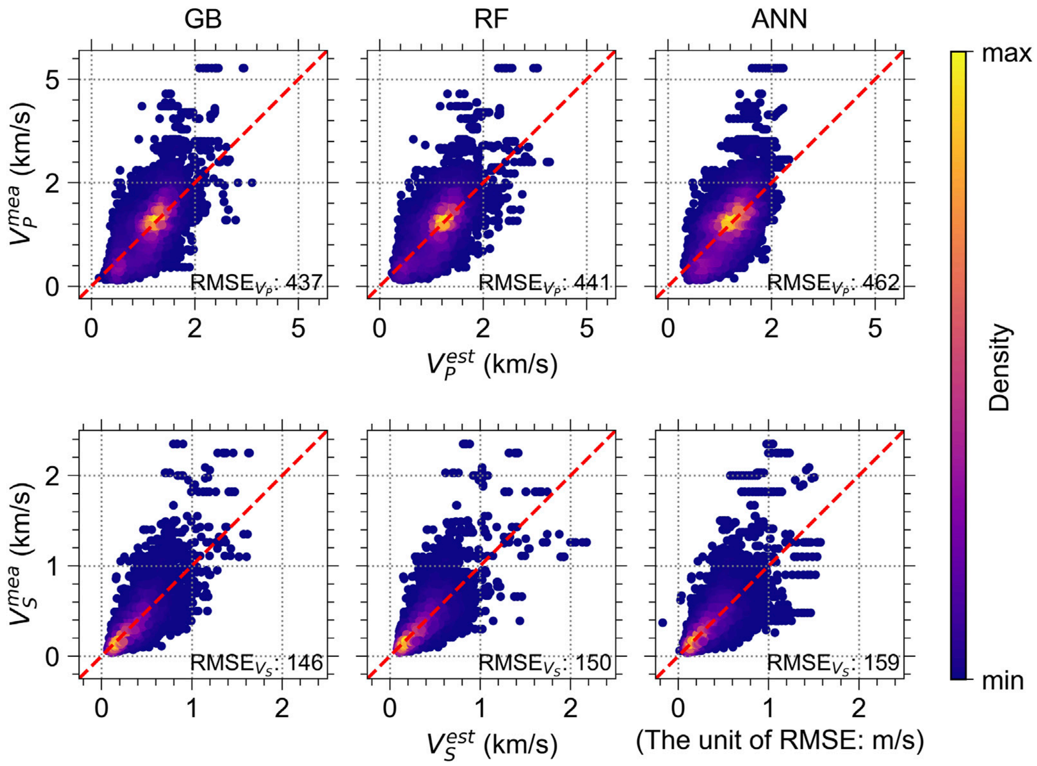

FIGURE 7
Measured velocity values (i.e., V_P^{mea} and V_S^{mea}) versus velocity values estimated by the three ML-based models (i.e., V_P^{est} and V_S^{est}) for the K-NET. The 1:1 lines are depicted as red dashed lines. The color bar on the right side represents the data density. The data were aggregated from five test folds from the five experiments (i.e., five derived ML models).

generalization performance in predicting velocities on the dataset utilized in this study. Furthermore, the hyperparameters used in this study were taken from the suggestions mentioned in the following subsections to present the results of baseline solutions, serving as a fundamental benchmark for assessing their effectiveness in predicting velocities.

3.1 Gradient boosting (GB)

Before we start explaining the GB, we describe the decision tree algorithm, which is the main concept of GB and RF. The decision tree consists of nodes, where a tree is grown on the training dataset. The tree contains three types of nodes: root node, internal node, and leaf node, where the root and internal nodes play a role in splitting the data samples, and the leaf node makes the final decision for the prediction value.

We presented an example tree using the independent variables of the K-NET, as shown in Figure 4 to explain the internal structure. First, the root node splits 15,253 independent data samples into two internal nodes by asking if the N-value ≤ 45.5. If the condition is true, the internal node condition (i.e., N-value ≤ 14.75) works to further divide the allocated data samples into a leaf node (R_1) and another internal node. If the root node condition is false, the data samples are further divided by the internal node (i.e., depth ≤ 4.5 m) into a leaf node (R_4) and another internal node. Following the if-else rules, the model finally creates a tree that consists of a root node, four internal nodes, and six leaf nodes (R_1, R_2, . . ., R_6).

One may wonder how the decision tree model creates the splitting criterion of the node. The model grows a tree by splitting the data samples into two groups by finding the threshold that minimizes the mean of squared errors (MSE), which is calculated as

$$\mathrm{MSE} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n}\left(V_{S_i}^{mea} - V_{S_{R_j}}^{est}\right)^2 \tag{1}$$

where $V_{S_i}^{mea}$ is the measured V_S associated with i^{th} data sample of the K-NET (i.e., $i = 1, 2, \ldots, n$; $n = 15{,}253$), and $V_{S_{R_j}}^{est}$ is the estimated V_S determined by a specific leaf node, R_j, where j is the leaf node index ($j = 1, 2, \ldots, J$).

The estimated V_S (V_S^{est}), which is an output of the trained tree model (T), can be expressed by the following equation:

$$V_S^{est} = T(x_i) = \sum_{j=1}^{J} c_j I\left(x_i \in R_j\right) \tag{2}$$

where x_i represents the independent variables of the i^{th} data sample of the K-NET, j and J are the specific leaf node number and the total number of leaf nodes (i.e., six leaf nodes for the example tree in Figure 4), respectively, c_j is the predicted dependent variable decided by the specific region R_j, and I is an indicator function that takes a value of 0 or 1 (i.e., $I = 1$ if $x \in R_j$ and 0 otherwise).

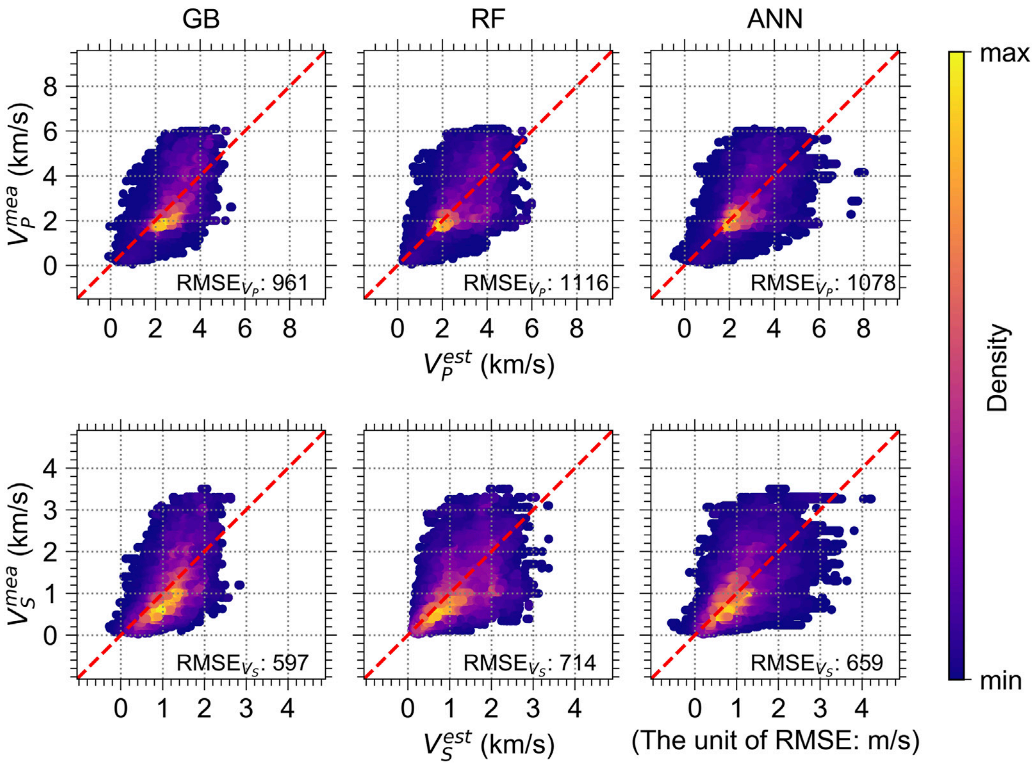

FIGURE 8
Measured velocity values (i.e., V_P^{mea} and V_S^{mea}) versus velocity values estimated by the three ML-based models (i.e., V_P^{est} and V_S^{est}) for the KiK-net. The 1:1 lines are depicted as red dashed lines. The color bar on the right side represents the data density. The data were aggregated from five test folds from the five experiments (i.e., five derived ML models).

However, a single decision tree model is prone to overfitting on a training dataset, resulting in a high variance in new data samples (Geurts et al., 2009; Czajkowski and Kretowski, 2019). The GB algorithm, proposed by Friedman (2001); Friedman (2002), is an ensemble of weak models (i.e., decision trees) and provides robust model performance over the overfitting problem. GB grows many decision trees and connects them in order like links in a chain, where each new tree is grown to modify a mistake made by a previous tree. An example of a GB architecture is presented in Figure 5.

The trees in the GB estimate the residuals between V_S^{mea} and V_S^{est} instead of the V_S itself and are trained to minimize the residuals. The specific steps for training the GB model are as follows. Step 1: The GB has a constant value ($F_0(x)$), which is $\overline{V_S^{mea}}$ from the training dataset. Step 2: From $b = 1$ to B, where B is the last tree index, the GB repeats the following steps (3–5) for successive trees ($T_1, T_2, \ldots, T_B$). Step 3: An individual tree ($T_b(x)$) calculates the residual data for each data sample as:

$$r_{ib} = V_{S_i}^{mea} - F_{b-1}(x_i) \tag{3}$$

where r_{ib} is a residual associated with data sample i in the dataset (i.e., $i = 1, 2, \ldots, 15, 253$) and tree index b, and $F_{b-1}(x_i)$ is the prediction value of the previous GB model ($V_{S_{ib-1}}^{est}$). Step 4: After the residual dataset ($r_{1,b}, r_{2,b}, \ldots, r_{15253,b}$) for $T_b(x)$ is developed, $T_b(x)$ is trained on the dataset, $\{(x_i, r_{ib})\}_{i=1}^{15253}$, instead of $\left\{(x_i, V_{S_i}^{mea})\right\}_{i=1}^{15253}$. The leaf node ($R_{jb}$) is determined during training, where j is the leaf node index of the tree, $T_b(x)$. The mean residual value predicted in R_{jb} is r_{jb}. The r_{jb} is subsequently reduced by a learning rate (v), which is a constant value to reduce the contribution of each tree. Therefore, the tree ($T_b(x)$) can be described as follows:

$$T_b(x) = v\sum_{j=1}^{J_b} r_{jb} I\left(x \in R_{jb}\right) \tag{4}$$

where J_b is the number of leaf nodes in $T_b(x)$. The equation returns the $v^* r_{jb}$ according to independent data (x). Step 5: After a tree is built, it is added to the previous tree. The updated GB model ($F_b(x)$) can be described as follows:

$$F_b(x) = F_{b-1}(x) + T_b(x) = F_0(x) + T_1(x) + \ldots + T_b(x) \tag{5}$$

After the GB finishes developing the last tree ($T_B(x)$), we finally obtain the $F_B(x)$, which is the complete GB model ready to use for predicting V_S (V_S^{est}). In this study, we set the number of trees (b) to 100 and the learning rate (v) to 0.1, as suggested by Pedregosa et al. (2011).

3.2 Random forest (RF)

The RF algorithm, proposed by Breiman (2001), is a bootstrap aggregation (bagging) ensemble algorithm that grows many decision trees using a random subset of the data. Unlike GB, RF trains many weak trees in a parallel manner, where the trees are not affected by each other while being trained. Each tree in the GB predicts the

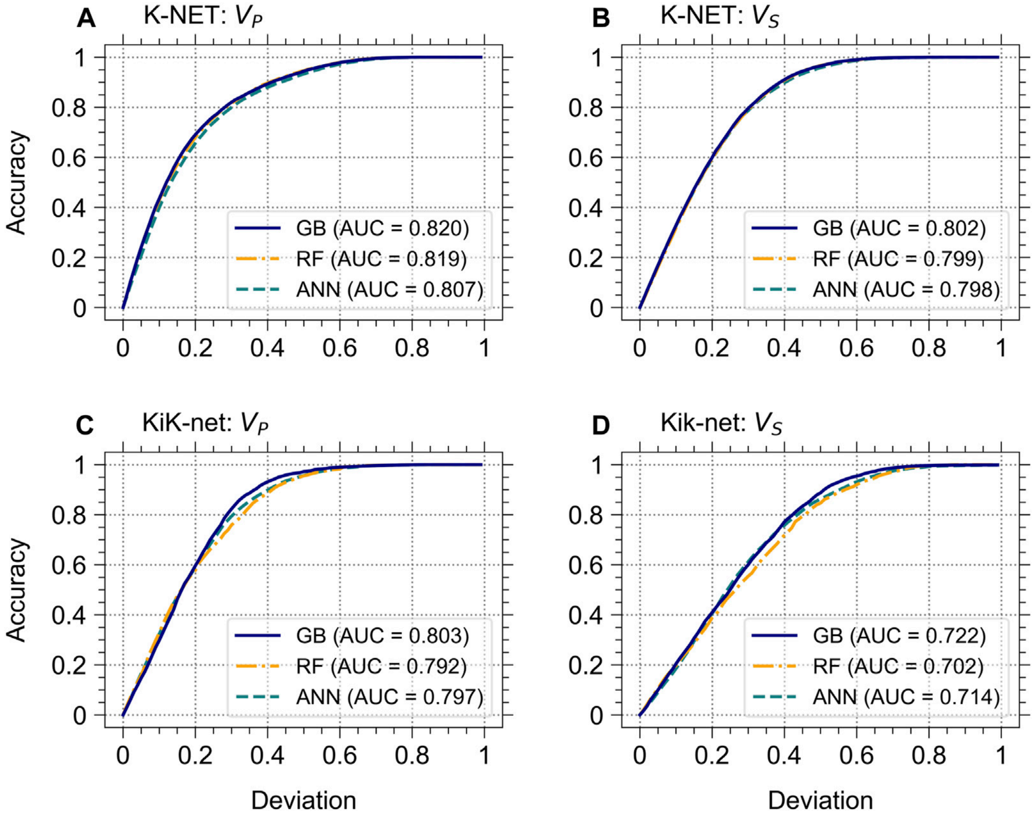

FIGURE 9
REC curves for individual models: **(A)** V_P and **(B)** V_S for K-NET, and **(C)** V_P and **(D)** V_S for KiK-net. Each curve represents the average accuracy across the five test folds from the five experiments (i.e., five derived ML models), with the specified deviation.

residual value, but the tree in the RF directly returns ${V_S}^{est}$. An example of an RF architecture is presented in Figure 5. From $b = 1$ to B, each tree $(T_1, \ldots, T_B)$ is grown on bootstrap samples $(D_1, \ldots, D_B)$, where D_b is the randomly sampled subset data from the training dataset.

While T_b is trained on D_b, the number of variables (p^*) in T_b is randomly chosen from the total number of independent variables (p) to minimize the MSE. The node is further split into two child nodes after choosing the best split points among p^* number of variables.

After RF completes training all trees, it makes a final decision of ${V_S}^{est}$ at a new data sample x by averaging the multiple results of T_b, which can be described as

$$ {V_S}^{est} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{6} $$

In this study, we set the number of trees (b) to 100 and $p^* = p$ (i.e., all variables are considered), as suggested by Pedregosa et al. (2011).

3.3 Artificial neural network (ANN)

The ANN model comprises a collection of nodes grouped in layers, where each node in a layer is connected to the nodes in the next layer. The ANN model includes three types of layers: input layer, hidden layer, and output layer. Figure 6 presents an ANN model containing the two hidden layers used in this study. The number of input variables is nine for K-NET and seven for KiK-net, as described in Table 1. However, we applied the binary encoding method to categorical variables. The total number of variables was increased to 18 for the K-NET and 22 for the KiK-net to train ML models (i.e., GB, RF, and ANN). A detailed explanation of this is provided in the subsequent section. Therefore, the number of input nodes is 18 for K-NET and 22 for KiK-net. We set the number of nodes in the hidden layers to 200, as inspired by Kim et al. (2020). In Figure 6, the values of each node for hidden layer 1 $(h_j^{(1)})$, hidden layer 2 $(h_k^{(2)})$, and output layer (V_s^{est}) can be described as

$$ \begin{aligned} h_j^{(1)} &= f^{(1)}\left(\sum_{i=1}^{18}\left(w_{ji}x_i + b_j\right)\right), \\ h_k^{(2)} &= f^{(2)}\left(\sum_{j=1}^{200}\left(w_{kj}h_j^{(1)} + b_k\right)\right), \\ V_S^{est} &= \sum_{k=1}^{200}\left(w_{nk}h_k^{(2)} + b_n\right) \end{aligned} \tag{7} $$

where x_i is the input value of the i^{th} node in the input layer, w_{ji} is the weight between i^{th} node in the input layer and j^{th} node in the hidden layer 1, w_{kj} is the weight between j^{th} node in hidden layer

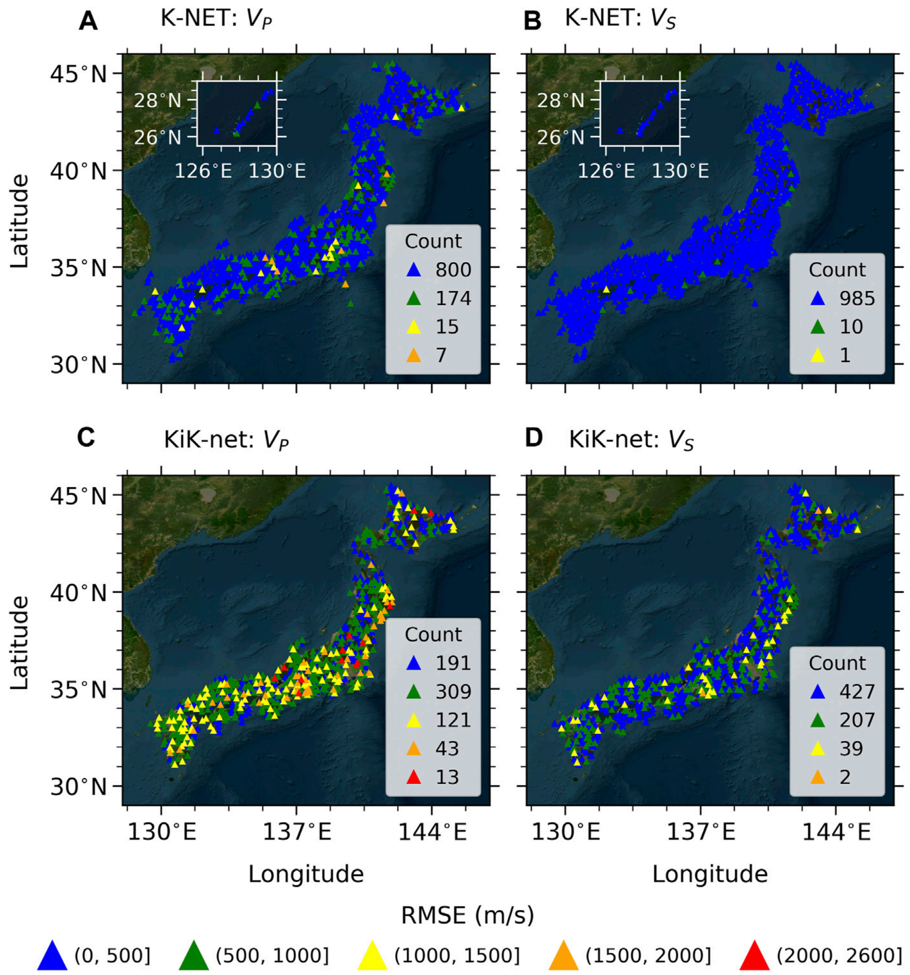

FIGURE 10
Maps for RMSE values of the GB-based model for **(A)** V_P and **(B)** V_S of all of the K-NET sites, and those for **(C)** V_P and **(D)** V_S of all of the KiK-net sites. The data for the sites were aggregated from five test folds of the five experiments (i.e., five derived ML models). The count numbers of the color-coded RMSE ranges are presented inside each of the panels.

1 and k^{th} node in hidden layer 2, and w_{nk} is the weight between k^{th} node in hidden layer 2 and n^{th} node in the output layer (i.e., $n = 1$). b_j, b_k, and b_n are the biases of j^{th} node in hidden layer 1, k^{th} node in hidden layer 2, and n^{th} node in the output layer, respectively.

Each node in an input layer receives an independent variable (e.g., the N-value). At each node, the value (x) is multiplied by the corresponding weight (w), which is summed with other values multiplied by other weights in the same layer. A bias (b) is then added to the sum of all values multiplied by each weight in the layer. The bias provides a better generalization ability to the model by enhancing the fitting flexibility. Then, the activation function (f) is applied to the sum of ($x \times w + b$), where f gives a non-linear property to help the ANN model capture the complex data patterns. The activation function outputs a value that becomes the input value of the node for the next layer.

We applied a rectified linear unit (ReLU) to $f^{(1)}$ and $f^{(2)}$, where the ReLU is a widely used non-linear activation function (Boob et al., 2020). Specifically, the ReLU is defined as $f(\mathrm{X}) = \max(0, \mathrm{X})$, where $f(\mathrm{X}) = 0$ if $\mathrm{X} < 0$ or $f(\mathrm{X}) = \mathrm{X}$ if $\mathrm{X} \geq 0$.

4 Model training strategy

Before training the model, the categorical variables (i.e., geology and soil/rock type) needed to be transformed into numerical variables. We mapped the variables into integers, which were then encoded in a binary format. This method is called binary encoding, which has been popularly utilized in applications (e.g., Jackson and Agrawal, 2019; Yousef et al., 2019). Here is an example using the soil/rock type in the K-NET dataset, which includes 12 unique features (i.e., 12 IDs). First, the length of the encoding vector was determined as $\lceil \log_2(12) \rceil = 4$. Second, each ID is converted into binary format, e.g., 'sandy soil' (ID = 1) to [0, 0, 0, 1], 'fill soil' (ID = 6) to [0, 1, 1, 0], and 'volcanic ash clay' (ID = 12) to [1, 1, 0, 0]. Each bit number in the vector, for example, 1, 1, 0, and 0 in 'volcanic ash clay' (ID = 12) work as independent variables. Using this method, the total number of input variables was increased from 9 to 18 for the K-NET dataset and from 7 to 22 for the KiK-net dataset.

We applied the five-fold cross-validation (CV), which has been widely utilized in model evaluation (Berrar, 2019). This approach

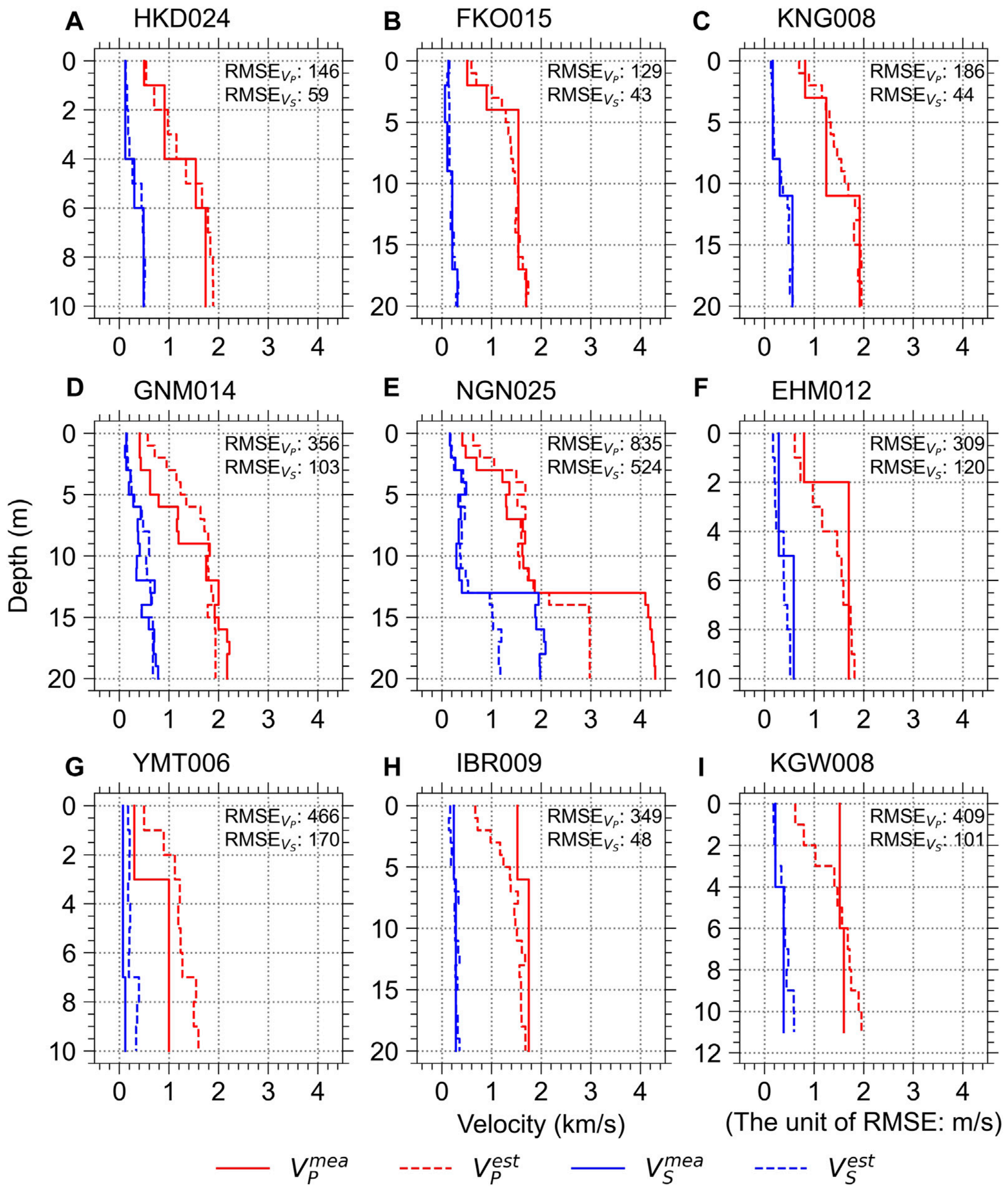

FIGURE 11
Measured wave velocity profiles (V_P^{mea} and V_S^{mea}) versus the velocity profiles predicted by the GB-based model (V_P^{est} and V_S^{est}) at the nine K-NET sites **(A–I)** from five test folds.

assesses the generalization ability of models and prevents overfitting. The five-fold CV divides the entire dataset randomly into five roughly equal folds. Then, the model uses four folds for training and the remaining one fold for testing (i.e., 80% for training and 20% for test dataset). We repeated for five times: i.e., we developed five ML models. The test results from these five experiments were aggregated to evaluate the general performance of the ML algorithm.

This study aims to train ML models using the data for some sites and evaluate the model performance using the data for new sites. Therefore, all data samples were split based on site locations and not on whole data samples. Each fold is allocated 20% of the total sites but may not be divided exactly. With our case as an example, the K-NET sites were divided into training and testing parts as follows: 797:199 (for four experiments), and 796:200 (for one experiment). For the KiK-net dataset, the V_P data were divided into 541:136 (for two experiments) and 542:135 (for three experiments), and V_S data were separated equally for all experiments: 540 for training and 135 for testing.

5 Validation

5.1 Comparison between predictions and measurements

The three ML-based models developed in this study were evaluated for each test fold after training. Figure 7 presents the

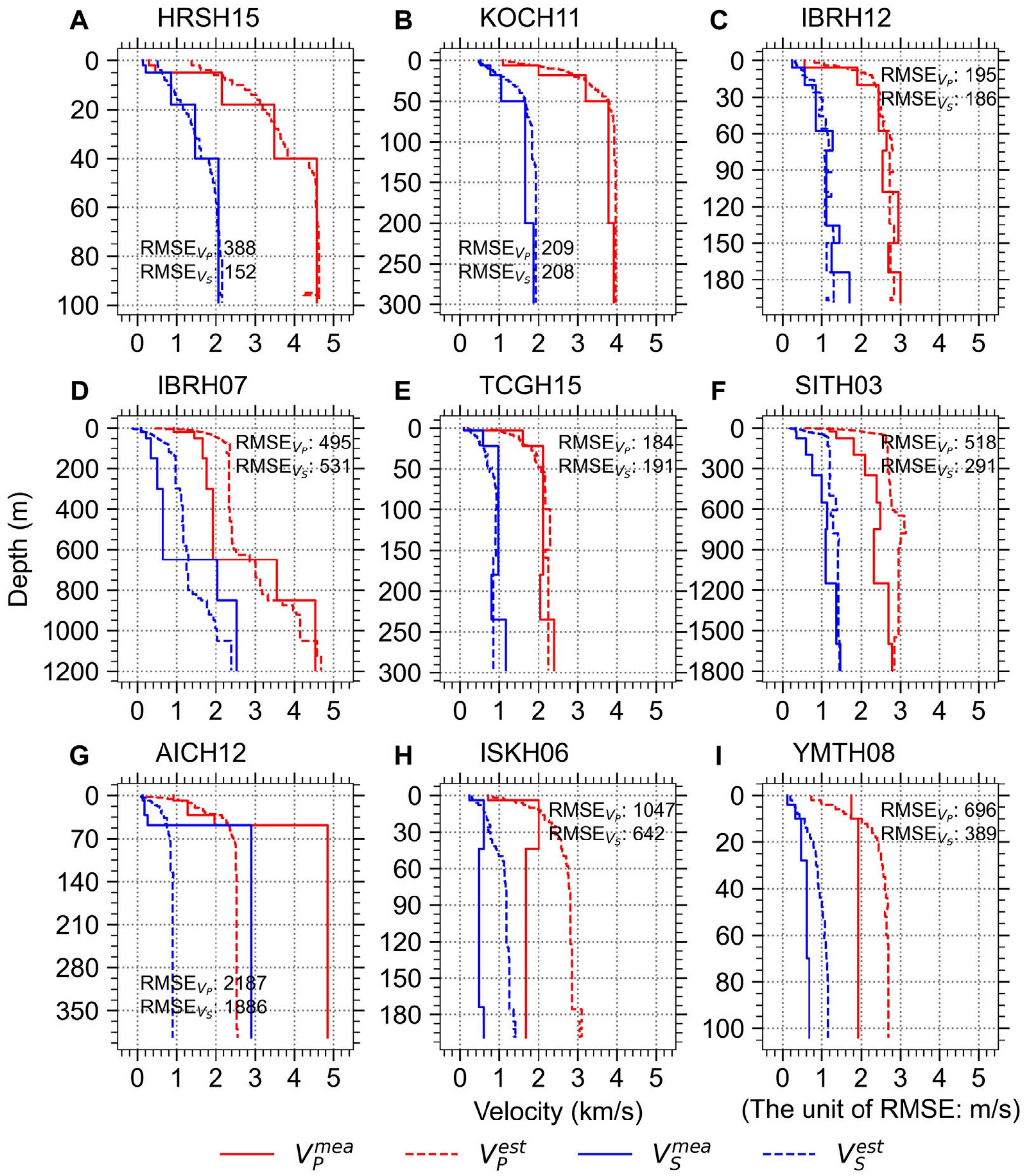

FIGURE 12
Measured wave velocity profiles (V_P^{mea} and V_S^{mea}) versus the velocity profiles predicted by the GB-based model (V_P^{est} and V_S^{est}) at the nine KiK-net sites **(A–I)** from five test folds.

measured wave velocities (i.e., V_P^{mea} and V_S^{mea}) versus the estimated values (V_P^{est}; V_S^{est}) using the three models for the entire K-NET test data samples. Note that these data were aggregated from five test folds from the five experiments (i.e., five derived ML models). As a performance indicator for V_S prediction, we calculated the root mean squared error (RMSE$_{V_S}$) as:

$$\mathrm{RMSE}_{V_S}\ (\mathrm{m/s}) = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(V_{S_i}^{mea} - V_{S_i}^{est}\right)^2} \qquad (8)$$

where $V_{S_i}^{est}$ is the estimated V_S value for the i^{th} data sample, $V_{S_i}^{mea}$ is the corresponding measurement, and n is the data sample size. The RMSE values calculated for each of the five test folds were averaged. The RMSE$_{V_P}$ was computed in the same manner. The RMSE$_{V_S}$ and RMSE$_{V_P}$ are the smallest for the GB-based model and the largest for the ANN-based model.

Figure 8 presents the measured wave velocities (i.e., V_P^{mea} and V_S^{mea}) versus the estimated values (V_P^{est}; V_S^{est}) using the three models for the KiK-net dataset. RMSE$_{V_S}$ and RMSE$_{V_P}$ are the smallest for the GB-based model indicating a stronger alignment along the 1: 1 line, and the largest for the RF-based model. The results for individual experiments are included in Supplementary Appendix I of the Electronic Supplement.

The RMSE depends on the study area and data features including the number of sites and velocities distribution. Many previous studies have utilized varying ranges of V_S to make predictions for different geological regions, resulting in varied

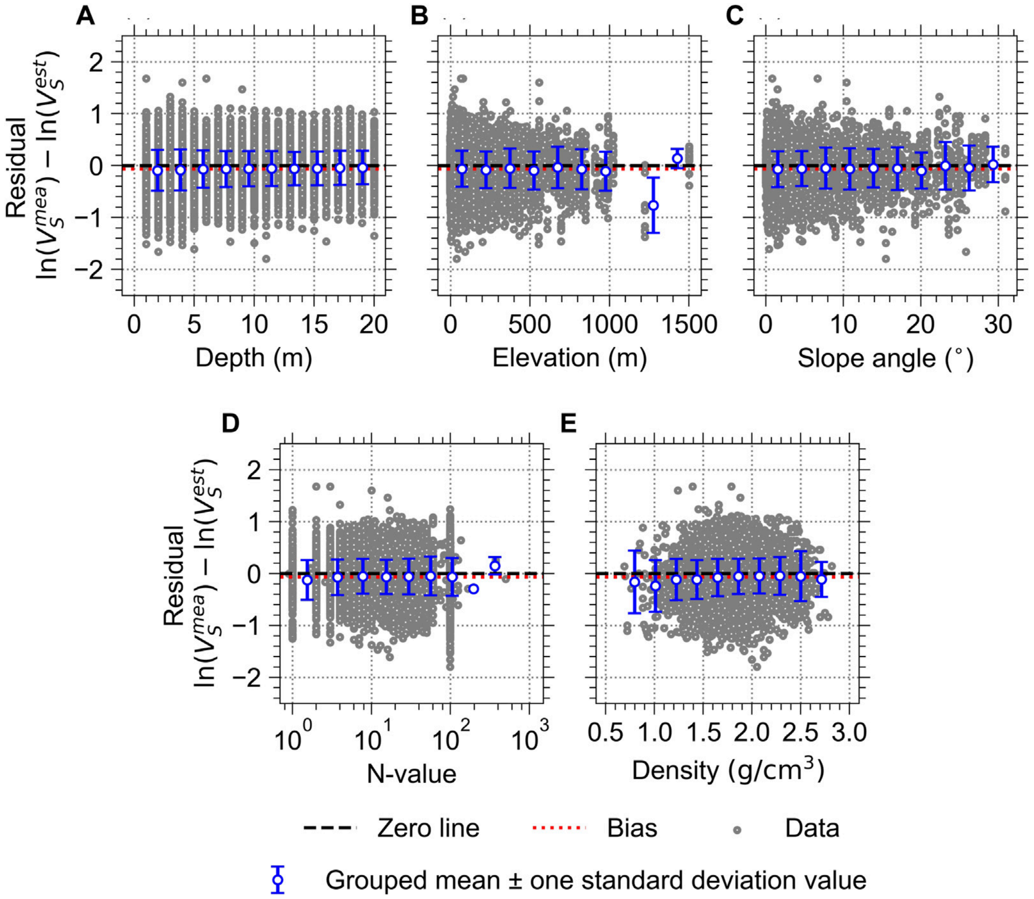

FIGURE 13
Residuals of the V_S estimated by the GB-based model for the K-NET stations with respect to all the continuous variables: **(A)** depth; **(B)** elevation; **(C)** slope angle; **(D)** N-value; and **(E)** density. The data were aggregated from five test folds from the five experiments (i.e., five derived ML models).

RMSEs. For example, Ataee et al. (2019) utilized uncorrected and corrected SPT-N with 88 boreholes to predict V_S. The results using uncorrected SPT-N and V_S under approximately 1,200 m/s presented that the RMSEs of the models ranged from 94.512 to 104.149 m/s. Those using corrected SPT-N and V_S under approximately 600 m/s presented RMSEs ranging from 59.423 to 67.473 m/s. Ghorbani et al. (2012) utilized corrected SPT blow counts, and effective overburden stress to predict V_S. They used 80 boreholes, where the V_S ranges from 66 to 363 m/s. The RMSE of the prediction model is 37.2 m/s. Sun et al. (2013) used tip resistance, sleeve friction, pore pressure, and overburden effective stress to establish the correlation with V_S. They utilized 17 sites, where the measured V_S is under approximately 400 m/s. The RMSEs of the correlation forms are from 30.42 to 38.57 m/s. Furthermore, Dumke and Berndt (2019) used 38 types of variables (e.g., depth below seafloor, surface heat flow, and distance to nearest spreading ridge) to predict V_P. They used 333 sites containing V_P above 4,000 m/s, where the velocity range is not mentioned. The RMSEs vary approximately between 400 and 500 m/s depending on the considered variables. The RMSE presented in this paper may be reasonable, given that the prediction models were made and tested for the larger number of sites distributed throughout Japan, which includes various study areas and a wider range of velocities than other studies. However, discrepancies have been observed, especially for the KiK-net: V_P dataset, implying that more region-specific depth-related variables may be needed to infer the velocity profiles better.

We further investigated the relationship between the measured and estimated velocities by employing the Regression Error Characteristic (REC) curve (Bi and Bennett, 2003). The REC curve depicts the relationship between the specified deviation tolerance on the x-axis, which is the error tolerance, and the y-axis for the proportion of data with prediction deviations smaller than the corresponding deviation. The resulting curve provides an estimation of the cumulative distribution function of the error. Furthermore, the REC curve quantifies the performance of the model by computing the area under the curve (AUC). A higher AUC value indicates better model performance. Figure 9 illustrates the REC curves for each model. The curves were individually computed for the five test folds from the five experiments and were then averaged to make a single curve. The AUC was subsequently calculated based on the single curve, representing the general performance of each model on the dataset. The results for K-NET and KiK-net, both at V_P and V_S, reveal that all AUCs are above 0.702 for the specified deviations ranging from 0 to 1.0. Notably, the GB-based model has the highest AUC across all

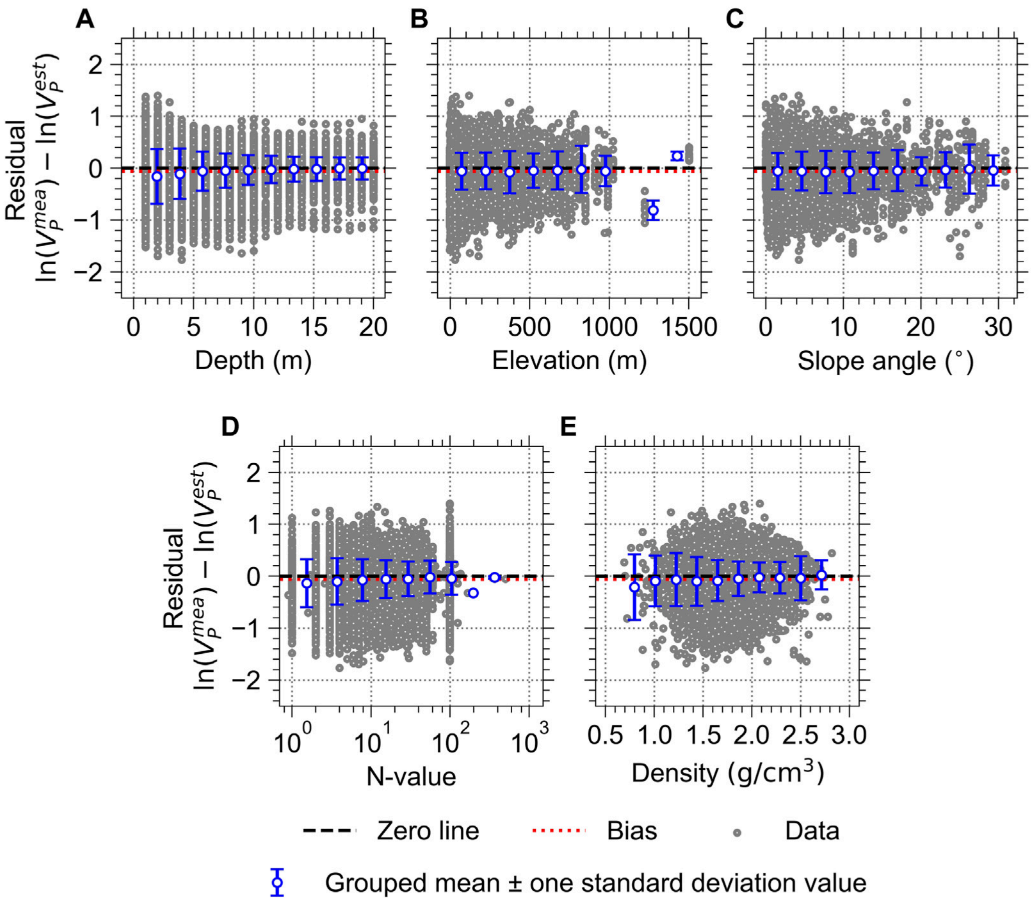

FIGURE 14
Residuals of the V_P estimated by the GB-based model for the K-NET stations with respect to all the continuous variables: **(A)** depth; **(B)** elevation; **(C)** slope angle; **(D)** N-value; and **(E)** density. The data were aggregated from five test folds from the five experiments (i.e., five derived ML models).

cases, indicating its relatively strong predictive performance within the deviation range.

Figure 10 presents maps of RMSEs for the GB-based model for all the sites considered in this study, which were aggregated from five test folds from the five experiments (i.e., five derived ML models). Overall, the models for V_P and V_S of the K-NET sites (Figures 10A, B) indicate that almost 80% of the sites have RMSE values within the range of (0, 500] for V_P, and almost 99% are within the same range for V_S. In contrast, for KiK-net (Figures 10C, D), approximately 46% of the sites have RMSE values within the (500, 1,000] range for V_P, and about 63% are within the (0, 500] range for V_S. It can be noticed that the RMSE values larger than 1,000 m/s for the estimated V_P values at the K-NET sites are concentrated in the area around 139 °E and 35.5 °N (Figure 10A). Furthermore, the RMSE values greater than 1,500 m/s and 1,000 m/s for the estimated V_P and V_S values, respectively, at the KiK-net sites are mainly clustered in the region around 137 °E and 35 °N (Figures 10C, D). The RMSEs for the KiK-net sites show a certain pattern along the east coast (from 140 to 142 °E and from 36 to 40 °N) (see Figure 10D). These observations imply that there could be factors that can affect the V_S and V_P values, other than those considered in this study.

Figure 11 shows examples of the wave velocity profiles predicted by the GB-based model compared with the measured profiles at the nine K-NET sites. The eight and one sample profiles were randomly selected from the V_S RMSE bands of (0, 500] m/s and (500, 1,000] m/s, respectively, from the entire test folds from five experiments. The wave velocities predicted for the HKD024, FKO015, and KNG008 sites (Figures 11A–C, respectively) are in good agreement with the measured profiles when compared to the other illustration, producing RMSE values ≤186 m/s. In contrast, there are some discrepancies between the measured and predicted profiles at certain depth ranges for the rest of the sites. At GNM014, V_P is overestimated at depths of up to 9 m (Figure 11D). At NGN025, the wave velocities are underestimated at depths greater than 13 m (Figure 11E). At EHM012, V_P is underestimated at depths from 2 to 6 m (Figure 11F). There are also some discrepancies in the V_P profiles at the YMT006, IBR009, and KGW008 sites (Figures 11G–I, respectively). In detail, the V_P is consistently overestimated across the entire depth range at the YMT006 site (Figure 11G). At the IBR009 site, V_P is underestimated up to a depth of 5 m (Figure 11H). Similarly, the KGW008 site demonstrates underestimation up to 3 m and overestimation at depths beyond 10 m (Figure 11I).

Figure 12 presents examples of the wave velocity profiles estimated by the GB-based model compared with the measured profiles at the nine KiK-net sites. The six, two, and one sample

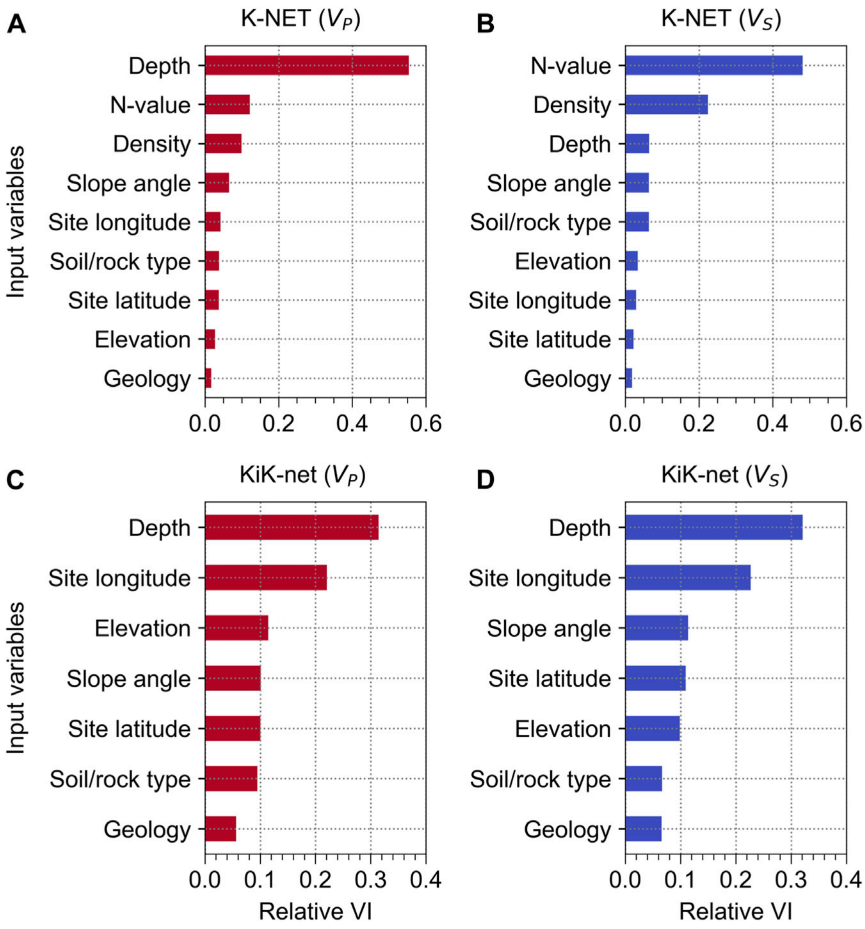

FIGURE 15
Relative variable importance (VI) for the GB-based models for **(A)** K-NET (V_P), **(B)** K-NET (V_S), **(C)** KiK-net (V_P), and **(D)** KiK-net (V_S). The variables are presented in an order of descending relative VI.

profiles were randomly selected from the V_S RMSE bands of (0, 500], (500, 1,000], and (1,500, 2000], respectively, from the entire test folds from five experiments. The estimated wave velocities for the HRSH15, KOCH11, IBRH12, and TCGH15 sites (Figures 12A–C, E, respectively) comparatively match well with the measured profiles, producing RMSE values ≤388 m/s. Some discrepancies are observed at specific depth ranges for the other sites. At IBRH07, the wave velocities are overestimated at depths from 51 m to 650 m and underestimated at depths from 651 m to 1,050 m (Figure 12D); however, the estimated wave velocities show relatively close agreement beyond 1,050 m. At SITH03, the velocities are overestimated almost throughout the depth (Figure 12F). At AICH12, the velocities are underestimated at depths greater than 49 m (Figure 12G). Overestimations are observed at the ISKH06 site for depths greater than 44 m (Figure 12H). For the YMTH08 site, velocities are overestimated for depths exceeding 18 m, while the V_P is underestimated for depths up to 6 m (Figure 12I).

Some discrepancies were observed in relation to a certain depth and profile patterns, as shown in Figure 11 and Figure 12. It is possible that the model could not well predict velocities for sites that have unusual profile patterns or for those that have not been frequently used when training or are not included in the training dataset. The model was trained to reduce the overall error for the entire sites used in training, so it might not well generalize the unseen patterns. As seen in the samples in Figure 11 and Figure 12, the model was trained to predict slower velocities near the ground surface and faster velocities at greater depths. Furthermore, model predicts velocities gradually increasing, and does not predict well the abrupt velocity changes (e.g., Figure 11E; Figure 12G). There are velocity reversals at depths greater than 44 m at ISKH06 (Figure 12H), and the V_P values are unusually faster near the ground surface at YMTH08 (Figure 12I). It turned out that the model was not able to capture these profiles.

For the systematic evaluation of discrepancies between measured and estimated wave velocities, we computed the residuals for all the continuous variables as

$$\begin{aligned}\mathrm{Res}^{V_S} &= \ln\left(V_S^{mea}\right) - \ln\left(V_S^{est}\right)\\ \mathrm{Res}^{V_P} &= \ln\left(V_P^{mea}\right) - \ln\left(V_P^{est}\right)\end{aligned} \qquad (9)$$

where Res^{V_S} and Res^{V_P} are the residuals for V_S and V_P, respectively. We also calculated the standard deviations of the residuals and biases (i.e., mean values of the residuals) for all trained models.

Figure 13 shows the Res^{V_S} of the GB-based model for the K-NET stations, which were aggregated from five test folds from the five experiments. The Res^{V_S} ranges from −1.798 to 1.677 and is aligned along zeros with respect to all continuous variables with a small bias (i.e., −0.066). Figure 14 shows the Res^{V_P} of the GB-based model for the K-NET stations, which were aggregated from five test folds from the five experiments. The Res^{V_P} ranges from −1.765 to 1.396 and is also aligned along with zeros with a bias value of −0.059. The standard deviation values for Res^{V_S} are 0.354, 0.358, and 0.374 for the GB-, RF-, and ANN-based models, respectively. The standard deviation values for Res^{V_P} are 0.360, 0.361, and 0.383 for the GB-, RF-, and ANN-based models, respectively. The residuals do not show any trend with the considered variables, indicating that all the variables have certain contributions to the model, or that some of the variables do not have influence on wave velocities. Moreover, the results seem reasonable when compared to those of Kwak et al. (2015), who presented the range of standard deviation of residuals for V_S prediction models made using K-NET for each soil/rock type, which was from 0.245 to 0.462. Because the Res^{V_S} and Res^{V_P} of the GB-based model for KiK-net stations show the same aspects, we describe the results in Supplementary Appendix II of the Electronic Supplement.

5.2 Variable importance

We examined the contribution levels of the independent variables to the prediction accuracy of the best model, the GB-based model. The method is called variable importance (VI), which is computed as the sum of the decrease in error when a variable splits a tree node (e.g., a node split by an N-value ≤ 14.75). The variable importance for variable x (i.e., VI(x)) is calculated as follows:

$$\text{VI}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t\,\in\,T_b, v(s_t)=x} p(t)\Delta i(t) \qquad (10)$$

where b is the tree index (b = 1 to B), t is the node in a specific tree model (T_b), and $v(s_t)$ is the variable used for splitting the node in which s_t is the splitting criterion (s) at note t. $p(t)$ is the proportion ($\frac{N_t}{N}$) of data samples reaching t, where N is the number of total training data samples, and N_t is the number of data samples at node t. $\Delta i(t)$ is the impurity reduction at node t, which can be expressed as follows:

$$\Delta i(t) = i(t) - \frac{N_{t_l}}{N_t} i(t_l) - \frac{N_{t_r}}{N_t} i(t_r) \qquad (11)$$

where $i(t)$ is the MSE at node t, $i(t_l)$ and $i(t_r)$ are the MSEs at the left child node (t_l) and right child node (t_r), respectively, split from node t. N_{t_l} and N_{t_r} are the numbers of data samples at t_l and t_r, respectively.

Figures 15A, B show the relative VIs for the K-NET independent variables for the GB-based models. The relative VI was computed by VI for each variable divided by the total VI for all variables. The VI was calculated on each test fold, and the VIs for all the five test folds were averaged. The VIs computed for binary codes were summed for the categorical variables. Three depth-dependent variables (i.e., depth, N-value, and density) have the highest VIs for both V_P and V_S models. The depth is ranked at the top for the V_P model, whereas the N-value is ranked at the top for the V_S model. Figures 15C, D present the relative VIs for the KiK-net dataset. The depth turned out to be the most critical variable for both V_P and V_S models. The effect of the site location is more significant for the KiK-net model than for the K-NET model. The slope angle and elevation have a certain influence on the models, whereas the soil/rock type and geology have the least influence. Although the influence of the geology turned out to be insignificant, the performance of the GB-based model was enhanced by including it. The RMSEs of the model were reduced from 615 m/s to 597 m/s for V_S, and from 979 m/s to 961 m/s for V_P, implying that it is also related to wave velocities at a deeper depth.

The confining pressure increases with depth, leading to an increase in the density, N-value, and wave velocities. Therefore, the depth and associated variables were determined to be most strongly correlated, as revealed by VI. The slope and elevation are related with shear stiffnesses, which eventually affect wave velocities. The site coordinates are relatively high VI, implying that they may be associated with site conditions that were not captured by other variables. The geology has the lowest VI, as it is for the ground surface. However, we included it in the model because of its certain effect in enhancing the predictive performance.

6 Conclusion

This paper presented three ML-based models (i.e., GB-, RF-, and ANN-based models) predicting V_P and V_S in Japan. We used borehole databases from the two seismograph networks, K-NET and KiK-net. We considered various factors such as depth, N-value, density, slope angle, elevation, geology, soil/rock type, and site coordinates. The number of trees was designated as 100 to train the RF- and GB-based models. We developed an ANN-based model with four layers, where each hidden layer included 200 nodes.

The models were trained and evaluated on the datasets using the five-fold cross-validation. The average RMSEs across all test folds showed that the GB-based model provided the best estimation among the other models for both K-NET and KiK-net sites. The RMSEs of the GB-based model for V_S and V_P of the K-NET sites were 146 and 437 m/s, respectively, and those of the KiK-net sites were 597 and 961 m/s, respectively, while those of the other models ranging from 150 to 462 m/s for K-NET and from 659 to 1,116 m/s for KiK-net. Furthermore, the REC curve indicated that the GB-based model revealed relatively high performance within the deviation range. We also validated the GB-based model by checking the residuals between the measured and estimated wave velocities with respect to various variables. The variable importance of the model for K-NET indicated that depth, N-value, and density were the essential variables in predicting the V_P and V_S of the K-NET sites. Note that we used the unnormalized N-values for the K-NET sites, which might lower the prediction capability. For KiK-net sites, depth was the most influential variable. The site longitude also had a high relative variable importance value, indicating the roles of factors other than those considered in this study. The geology has the smallest VI values, as shown in Figure 15. However, it turned out that including the geology can improve the model performance, decreasing the RMSE values. In addition, we consider that including latitude and longitude is necessary because these improved prediction performances of the models, capturing the effects that were not captured by other variables.

This paper proposed a model for predicting wave velocities based on various factors, which can be used for site exploration in various fields, including rock engineering and petroleum engineering. The key findings of this study highlight that common machine learning algorithms can reasonably predict the wave velocity profiles across the region of Japan as an example. The results from cross-validation present the general performances of models on the dataset and site-specific performances specifically for the GB-based model. The study reveals the importance of input variables contributing to predicting accuracy. Moreover, it suggests that considering more region-specific variables including site coordinates can assist the models in interpreting complicated relationships.

As for the limitations of this study, the models are limited by their reliance on borehole databases exclusively obtained from specific seismograph networks in Japan. This approach may present a bias towards the conditions within these networks. Consequently, predictive performance could be constrained when extending the applications to regions with different geological attributes. In this context, ensuring the consistency of the environmental and experimental conditions, and the employed measured data is crucial to guarantee the validity of results beyond the area considered in this study. Additionally, even though the various variables were included, an incomplete representation remains for specific regions. This implies the presence of intricate geological properties that necessitate analysis to understand their influence on the prediction of wave velocities in a particular area. Furthermore, the study reveals that incorporating site coordinates can influence predictive performance. Nevertheless, the specific contributions of these variables to predictive performance concerning geological characteristics remain subject to consideration. While this study confirmed that the most commonly used machine-learning techniques could be successfully applied for predicting wave velocities, exploring more advanced techniques and investigating additional factors in the future will enhance the prediction performance.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kyoshin.bosai.go.jp/, Strong-motion Seismograph Networks (K-NET, KiK-net).

Author contributions

JK: Conceptualization, Data curation, Formal Analysis, Methodology, Validation, Visualization, Writing–original draft, Writing–review and editing. J-DK: Conceptualization, Data curation, Formal Analysis, Writing–original draft. BK: Conceptualization, Formal Analysis, Funding acquisition, Project administration, Supervision, Writing–original draft, Writing–review and editing.

Funding

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2022-Tech-03) and the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure, and Transport (Grant 21CATAP-C164148-01). The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication. The opinions, findings, and conclusions or recommendations expressed in this article are solely those of the authors and do not represent those of the funders.

Acknowledgments

We are indebted to the National Research Institute for Earth Science and Disaster Resilience (NIED), Japan, for making the resources of K-NET and KiK-net seismographs publicly available.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2023.1267386/full#supplementary-material

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). *Tensorflow: large-scale machine learning on heterogeneous distributed systems*. arXiv preprint arXiv:1603.04467.

Akin, M. K., Kramer, S. L., and Topal, T. (2011). Empirical correlations of shear wave velocity (Vs) and penetration resistance (SPT-N) for different soils in an earthquake-prone area (Erbaa-Turkey). *Eng. Geol.* 119 (1-2), 1–17. doi:10.1016/j.enggeo.2011.01.007

Ameen, M. S., Smart, B. G., Somerville, J. M., Hammilton, S., and Naji, N. A. (2009). Predicting rock mechanical properties of carbonates from wireline logs (A case study: arab-D reservoir, ghawar field, Saudi arabia). *Mar. Petroleum Geol.* 26 (4), 430–444. doi:10.1016/j.marpetgeo.2009.01.017

Andrus, R. D., Piratheepan, P., Ellis, B. S., Zhang, J., and Juang, C. H. (2004). Comparing liquefaction evaluation methods using penetration-V_S relationships. *Soil Dyn. Earthq. Eng.* 24 (9-10), 713–721. doi:10.1016/j.soildyn.2004.06.001

Anemangely, M., Ramezanzadeh, A., Amiri, H., and Hoseinpour, S.-A. (2019). Machine learning technique for the prediction of shear wave velocity using petrophysical logs. *J. Petroleum Sci. Eng.* 174, 306–327. doi:10.1016/j.petrol.2018.11.032

Ataee, O., Moghaddas, N. H., and Lashkaripour, G. R. (2019). Estimating shear wave velocity of soil using standard penetration test (SPT) blow counts in Mashhad city. *J. Earth Syst. Sci.* 128 (3), 1–25. doi:10.1007/s12040-019-1077-x

Bajaj, K., and Anbazhagan, P. (2019). Seismic site classification and correlation between V_S and SPT-N for deep soil sites in Indo-Gangetic Basin. *J. Appl. Geophys.* 163, 55–72. doi:10.1016/j.jappgeo.2019.02.011

Berrar, D. (2019). Cross-validation. *Encycl. Bioinforma. Comput. Biol.* 1, 542–545. doi:10.1016/B978-0-12-809633-8.20349-X

Bi, J., and Bennett, K. P. (2003). "Regression error characteristic curves," in Proceedings of the 20th international conference on machine learning (ICML-03)), Washington, DC USA, August 21 - 24, 2003, 43–50.

Boob, D., Dey, S. S., and Lan, G. (2020). Complexity of training relu neural network. *Discrete Optim.* 2020, 100620. doi:10.1016/j.disopt.2020.100620

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/A:1010933404324

Chang, C., Zoback, M. D., and Khaksar, A. (2006). Empirical relations between rock strength and physical properties in sedimentary rocks. *J. Petroleum Sci. Eng.* 51 (3-4), 223–237. doi:10.1016/j.petrol.2006.01.003

Czajkowski, M., and Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. *Expert Syst. Appl.* 137, 392–404. doi:10.1016/j.eswa.2019.07.019

Deng, C., Pan, H., Fang, S., Konaté, A. A., and Qin, R. (2017). Support vector machine as an alternative method for lithology classification of crystalline rocks. *J. Geophys. Eng.* 14 (2), 341–349. doi:10.1088/1742-2140/aa5b5b

Ding, P., Wang, D., Di, G., and Li, X. (2019). Investigation of the effects of fracture orientation and saturation on the Vp/Vs ratio and their implications. *Rock Mech. Rock Eng.* 52 (9), 3293–3304. doi:10.1007/s00603-019-01770-3

Dumke, I., and Berndt, C. (2019). Prediction of seismic P-wave velocity using machine learning. *Solid earth.* 10 (6), 1989–2000. doi:10.5194/se-10-1989-2019

Eberli, G. P., Baechle, G. T., Anselmetti, F. S., and Incze, M. L. (2003). Factors controlling elastic properties in carbonate sediments and rocks. *Lead. Edge* 22 (7), 654–660. doi:10.1190/1.1599691

Fiorentino, G., Quaranta, G., Mylonakis, G., Lavorato, D., Pagliaroli, A., Carlucci, G., et al. (2019). Seismic reassessment of the leaning tower of pisa: dynamic monitoring, site response, and SSI. *Earthq. Spectra* 35 (2), 703–736. doi:10.1193/021518EQS037M

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statistics* 29, 1189–1232. doi:10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Statistics Data Analysis* 38 (4), 367–378. doi:10.1016/S0167-9473(01)00065-2

Geological Survey of Japan (2015). *Seamless digital geological map of Japan 1: 200,000.* Japan: National Institute of Advanced Industrial Science and Technology.

Geurts, P., Irrthum, A., and Wehenkel, L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. *Mol. Biosyst.* 5 (12), 1593–1605. doi:10.1039/B907946G

Ghorbani, A., Jafarian, Y., and Maghsoudi, M. S. (2012). Estimating shear wave velocity of soil deposits using polynomial neural networks: application to liquefaction. *Comput. Geosciences* 44, 86–94. doi:10.1016/j.cageo.2012.03.002

Harmon, J., Hashash, Y. M., Stewart, J. P., Rathje, E. M., Campbell, K. W., Silva, W. J., et al. (2019). Site amplification functions for central and eastern north America–Part II: modular simulation-based models. *Earthq. Spectra* 35 (2), 815–847. doi:10.1193/091117EQS179M

Hasancebi, N., and Ulusay, R. (2007). Empirical correlations between shear wave velocity and penetration resistance for ground shaking assessments. *Bull. Eng. Geol. Environ.* 66 (2), 203–213. doi:10.1007/s10064-006-0063-0

Heath, D. C., Wald, D. J., Worden, C. B., Thompson, E. M., and Smoczyk, G. M. (2020). A global hybrid V_{S30} map with a topographic slope–based default and regional map insets. *Earthq. Spectra* 36 (3), 1570–1584. doi:10.1177/8755293020911137

Jackson, E., and Agrawal, R. (2019). "Performance Evaluation of different feature Encoding schemes on cybersecurity logs," in 2019 SoutheastCon., 11-14 April 2019.

Jamshidi, A., Zamanian, H., and Sahamieh, R. Z. (2018). The effect of density and porosity on the correlation between uniaxial compressive strength and P-wave velocity. *Rock Mech. Rock Eng.* 51 (4), 1279–1286. doi:10.1007/s00603-017-1379-8

Jena, R., Pradhan, B., Almazroui, M., Assiri, M., and Park, H.-J. (2023). Earthquake-induced liquefaction hazard mapping at national-scale in Australia using deep learning techniques. *Geosci. Front.* 14 (1), 101460.doi:10.1016/j.gsf.2022.101460

Jun, M.-J. (2021). A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the seoul metropolitan area. *Int. J. Geogr. Inf. Sci.* 35 (11), 2149–2167. doi:10.1080/13658816.2021.1887490

Karthikeyan, J., and Samui, P. (2014). Application of statistical learning algorithms for prediction of liquefaction susceptibility of soil based on shear wave velocity. *Geomatics, Nat. Hazards Risk* 5 (1), 7–25. doi:10.1080/19475705.2012.757252

Kim, B. (2019). Mapping of ground motion amplifications for the fraser river delta in greater vancouver, Canada. *Earthq. Eng. Eng. Vib.* 18 (4), 703–717. doi:10.1007/s11803-019-0531-8

Kim, S., Hwang, Y., Seo, H., and Kim, B. (2020). Ground motion amplification models for Japan using machine learning techniques. *Soil Dyn. Earthq. Eng.* 132, 106095. doi:10.1016/j.soildyn.2020.106095

Kottke, A. R., Hashash, Y., Stewart, J. P., Moss, C. J., Nikolaou, S., Rathje, E. M., et al. (2012). "Development of geologic site classes for seismic site amplification for central and eastern North America," in 15th World Conf. on Earthquake Engineering, Lisbon, Portugal, September 24 to September 28, 2012.

Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. *Eur. J. Operational Res.* 259 (2), 689–702. doi:10.1016/j.ejor.2016.10.031

Kwak, D. Y., Brandenberg, S. J., Mikami, A., and Stewart, J. P. (2015). Prediction equations for estimating shear-wave velocity from combined geotechnical and geomorphic indexes based on Japanese data set. *Bull. Seismol. Soc. Am.* 105 (4), 1919–1930. doi:10.1785/0120140326

Kwok, O. L. A., Stewart, J. P., Kwak, D. Y., and Sun, P.-L. (2018). Taiwan-specific model for V_{S30} prediction considering between-proxy correlations. *Earthq. Spectra* 34 (4), 1973–1993. doi:10.1193/061217EQS113M

National Research Institute for Earth Science and Disaster Resilience (2019). NIED K-NET, KiK-net, national research Institute for Earth science and disaster resilience. *Natl. Res. Inst. Earth Sci. Disaster Resil.* 2019. doi:10.17598/NIED.0004

Ohta, Y., and Goto, N. (1978). Empirical shear wave velocity equations in terms of characteristic soil indexes. *Earthq. Eng. Struct. Dyn.* 6 (2), 167–187. doi:10.1002/eqe.4290060205

Panza, E., Agosta, F., Rustichelli, A., Vinciguerra, S., Ougier-Simonin, A., Dobbs, M., et al. (2019). Meso-to-microscale fracture porosity in tight limestones, results of an integrated field and laboratory study. *Mar. Petroleum Geol.* 103, 581–595. doi:10.1016/j.marpetgeo.2019.01.043

Pappalardo, G. (2015). Correlation between P-wave velocity and physical–mechanical properties of intensely jointed dolostones, Peloritani mounts, NE Sicily. *Rock Mech. Rock Eng.* 48 (4), 1711–1721. doi:10.1007/s00603-014-0607-8

Parker, G. A., Harmon, J. A., Stewart, J. P., Hashash, Y. M., Kottke, A. R., Rathje, E. M., et al. (2017). Proxy-based V_{S30} estimation in central and eastern North America. *Bull. Seismol. Soc. Am.* 107 (1), 117–131. doi:10.1785/0120160101

Paul, S., Ali, M., and Chatterjee, R. (2018). Prediction of compressional wave velocity using regression and neural network modeling and estimation of stress orientation in Bokaro Coalfield, India. *Pure Appl. Geophys.* 175 (1), 375–388. doi:10.1007/s00024-017-1672-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pickett, G. R. (1963). Acoustic character logs and their applications in formation evaluation. *J. Petroleum Technol.* 15 (6), 659–667. doi:10.2118/452-PA

Rahimi, S., Wood, C. M., and Wotherspoon, L. M. (2020). Influence of soil aging on SPT-Vs correlation and seismic site classification. *Eng. Geol.* 272, 105653. doi:10.1016/j.enggeo.2020.105653

Rahman, T., and Sarkar, K. (2021). Lithological control on the estimation of uniaxial compressive strength by the P-wave velocity using supervised and unsupervised learning. *Rock Mech. Rock Eng.* 54, 3175–3191. doi:10.1007/s00603-021-02445-8

Roy, D. G., Singh, T., Kodikara, J., and Das, R. (2017). Effect of water saturation on the fracture and mechanical properties of sedimentary rocks. *Rock Mech. Rock Eng.* 50 (10), 2585–2600. doi:10.1007/s00603-017-1253-8

Samui, P., Kim, D., and Sitharam, T. (2011). Support vector machine for evaluating seismic-liquefaction potential using shear wave velocity. *J. Appl. Geophys.* 73 (1), 8–15. doi:10.1016/j.jappgeo.2010.10.005

Seo, H., Kim, J., and Kim, B. (2022). Machine-learning-based surface ground-motion prediction models for South Korea with low-to-moderate seismicity. *Bull. Seismol. Soc. Am.* 112 (3), 1549–1564. doi:10.1785/0120210244

Si, W., Di, B., Wei, J., and Li, Q. (2016). Experimental study of water saturation effect on acoustic velocity of sandstones. *J. Nat. Gas Sci. Eng.* 33, 37–43. doi:10.1016/j.jngse.2016.05.002

Sil, A., and Haloi, J. (2017). Empirical correlations with standard penetration test (SPT)-N for estimating shear wave velocity applicable to any region. *Int. J. Geosynth. Ground Eng.* 3 (3), 1–13. doi:10.1007/s40891-017-0099-1

Singh, S., and Kanli, A. I. (2016). Estimating shear wave velocities in oil fields: a neural network approach. *Geosciences J.* 20 (2), 221–228. doi:10.1007/s12303-015-0036-z

Sousa, L. M., del Río, L. M. S., Calleja, L., de Argandona, V. G. R., and Rey, A. R. (2005). Influence of microfractures and porosity on the physico-mechanical properties and weathering of ornamental granites. *Eng. Geol.* 77 (1-2), 153–168. doi:10.1016/j.enggeo.2004.10.001

Sun, C.-G., Cho, C.-S., Son, M., and Shin, J. S. (2013). Correlations between shear wave velocity and *in-situ* penetration test results for Korean soil deposits. *Pure Appl. Geophys.* 170 (3), 271–281. doi:10.1007/s00024-012-0516-2

Tsai, C.-C., Kishida, T., and Kuo, C.-H. (2019). Unified correlation between SPT–N and shear wave velocity for a wide range of soil types considering strain-dependent behavior. *Soil Dyn. Earthq. Eng.* 126, 105783. doi:10.1016/j.soildyn.2019.105783

Wang, P., and Peng, S. (2019). On a new method of estimating shear wave velocity from conventional well logs. *J. Petroleum Sci. Eng.* 180, 105–123. doi:10.1016/j.petrol.2019.05.033

Xiao, S., Zhang, J., Ye, J., and Zheng, J. (2021). Establishing region-specific N–Vs relationships through hierarchical Bayesian modeling. *Eng. Geol.* 287, 106105. doi:10.1016/j.enggeo.2021.106105

Yasar, E., and Erdogan, Y. (2004). Correlating sound velocity with the density, compressive strength and Young's modulus of carbonate rocks. *Int. J. Rock Mech. Min. Sci.* 41 (5), 871–875. doi:10.1016/j.ijrmms.2004.01.012

Yousef, W. A., Ibrahime, O. M., Madbouly, T. M., and Mahmoud, M. A. (2019). *Learning meters of Arabic and English poems with recurrent neural networks: A step forward for language understanding and synthesis*. arXiv preprint arXiv: 1905.05700.

Zhang, Y., Zhong, H.-R., Wu, Z.-Y., Zhou, H., and Ma, Q.-Y. (2020). Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs. *J. Petroleum Sci. Eng.* 192, 107234. doi:10.1016/j.petrol.2020.107234