Geomechanics and Engineering, *Vol. 37, No. 6 (2024) 539-554*
https://doi.org/10.12989/gae.2024.37.6.539

Machine learning application to seismic site classification prediction model using Horizontal-to-Vertical Spectral Ratio (HVSR) of strong-ground motions

Francis G. Phi[1a], Bumsu Cho[2b], Jungeun Kim[2c], Hyungik Cho[3d],
Yun Wook Choo*[1], Dookie Kim[1e] and Inhi Kim[4f]

[1]Department of Civil and Environmental Engineering, Kongju National University, 1223-24 Cheonan-daero, Seobuk-gu, Cheonan-si, Chungcheongnam-do, Republic of Korea
[2]Department of Computer Science and Engineering, Kongju National University, 1223-24 Cheonan-daero, Seobuk-gu, Cheonan-si, Chungcheongnam-do, Republic of Korea
[3]Department of Civil Systems Engineering, Andong National University,1375 Gyeongdong-ro, Andong, Gyeongsangbuk-do, 36729, Republic of Korea
[4]Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology, 193 Munji-ro, Yuseong-gu, Daejeon, Republic of Korea

(*Received January 24, 2024, Revised April 13, 2024, Accepted May 2, 2024*)

Abstract. This study explores development of prediction model for seismic site classification through the integration of machine learning techniques with horizontal-to-vertical spectral ratio (HVSR) methodologies. To improve model accuracy, the research employs outlier detection methods and, synthetic minority over-sampling technique (SMOTE) for data balance, and evaluates using seven machine learning models using seismic data from KiK-net. Notably, light gradient boosting method (LGBM), gradient boosting, and decision tree models exhibit improved performance when coupled with SMOTE, while Multiple linear regression (MLR) and Support vector machine (SVM) models show reduced efficacy. Outlier detection techniques significantly enhance accuracy, particularly for LGBM, gradient boosting, and voting boosting. The ensemble of LGBM with the isolation forest and SMOTE achieves the highest accuracy of 0.91, with LGBM and local outlier factor yielding the highest F1-score of 0.79. Consistently outperforming other models, LGBM proves most efficient for seismic site classification when supported by appropriate preprocessing procedures. These findings show the significance of outlier detection and data balancing for precise seismic soil classification prediction, offering insights and highlighting the potential of machine learning in optimizing site classification accuracy.

Keywords: earthquake; machine learning; seismic design; site characterization; site classification prediction

1. Introduction

Earthquakes are a type of natural hazard that arises from the displacement of faults, leading to the discharge of energy that can have catastrophic effects on human society and infrastructures. This calamity is magnified by soil amplification which leads to it being considered in seismic design codes in the form of the site classification system (Adams 1990, Hryciw *et al*. 1991, Nakhaei and Ali Ghannad 2008, Abbas *et al*. 2021, Holzer *et al.* 2005). The effects brought upon by soil amplification are due to soil conditions which enhances the ground motion. This makes determining soil features crucial for assessing and designing structures. (Türköz 2019, Pradhan *et al*. 2021, Lee *et al.* 2018). Hence the usage of average response spectra recorded in various soil types (Seed *et al.* 1976) or extending the empirical scaling to that of pseudo relative velocity spectra (Lee 1987) were empirical studies of the local site conditions and their influence on the amplitudes of the ground motions were researched to determine the values of the soil amplifications. In search for the fitting model (Trifunac 1987), considered eight different linear models in which all of the it agreed with the trend suggested by the amplification formula presented by (Trifunac 1990). By considering these studies, designers can precisely analyze potential amplification effects and integrate suitable solutions into the structural design to mitigate dangers associated with local site effects and consider soil-structure interactions (Fatahi *et al*. 2014).

Seismic site classification system has been developed to categorize different soil types based on the response of the soil layers due to seismic activity. This system enables seismic designers to construct a standard response spectrum for a specific site with various soil strata. Most of national bodies have adopted similar such as National earthquake hazard reduction program (NEHRP 1997), Eurocode 8, ASCE7-22, Japanese highway bridge design and building codes of different nations (BSSC 2003, Verdugo 2019). NEHRP site classification is one of the typical systems which relies on the average of values of the top 30 meters of

*Corresponding author, Professor
E-mail: ywchoo@kongju.ac.kr
[a]MSc. Student
[b]MSc. Student
[c]Professor
[d]Assistant Professor
[e]Professor
[f]Professor

http://www.techno-press.org/?journal=gae&subpage=7

ISSN: 2005-307X (Print), 2092-6219 (Online)

Table 1 NEHRP seismic site classification system

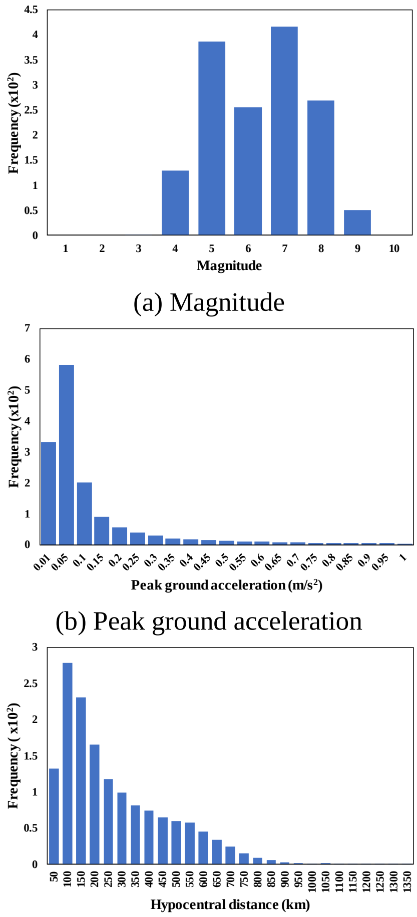
NEHRP Site Class	Model	Shear-Wave Velocity (m/s)
A	Hard Rock	$V_{s30} > 1500$
B	Rock	$760 < V_{s30} < 1500$
C	Very dense soil and soft rock	$360 < V_{s30} < 760$
D	Stiff soil	$180 < V_{s30} < 360$
E	Soft soil	$V_{s30} < 180$

soil for its two parameters: the average shear wave velocity (V_{S30}) or the average standard penetration resistance (N). Vs_{30} is derived from the time it took for the shear wave to propagate through the soil layers unto the ground surface (Dobry *et al.* 2000) it can be also used to evaluate liquefaction potential (Ji *et al.* 2021) due to its capability to define dynamic soil behavior. In this study, the V_{S30} is adopted to classify soil to one of five classes: A, B, C, D, or E (Table 1).

A primary limitation of this classification approach is its dependence on V_{s30}. In instances where a geotechnical report is unavailable due to challenges like difficult accessibility or insufficient funding for invasive techniques like borehole drilling it would make site categorization using NEHRP's method difficult. Thus, nondestructive seismic site categorization methods without soil borings have been investigated. Sites can be categorized by their primary period using average horizontal-to-vertical (H/V) response spectral ratios (Fukushima *et al.* 2007). Non-invasive surface wave interpretations of ambient noise measurements have provided quantitative insights into site amplification (Bard *et al.* 2010). Phung *et al.* (2006) suggested utilizing the pseudo-acceleration response spectrum for the horizontal ground motion component using a 5% damping ratio instead of the usual Fast Fourier transform (FFT) for site assessment. Site categorization uses empirical methods, including peak frequencies and the empirical H/V spectral ratio approach (Ghasemi *et al.* 2009). The horizontal to vertical spectral ratio (HVSR) is a popular seismic site classification method using microtremor or strong ground motion data. Unlike conventional approaches involving sensors or explosive charges, this study utilizes pre-existing earthquake data (Ghasemi *et al.* 2009, Lee *et al.* 2001, Phung *et al.* 2006, Pinzón *et al.* 2019, Wen *et al.* 2010, Zhao *et al.* 2023, 2006). A common empirical formula was proposed by Zhao (2006) utilized to classify sites by using strong-motion attenuation models. Although this method is improved in terms of accuracy, the fluctuating accuracy of different soil types undermines its reliability. Therefore, this research opts for an alternative approach by integrating machine learning techniques into the existing methodologies used in HVSR analysis.

Machine learning has significantly transformed predictive and classification approaches due to its ability to generate forecasts with superior predictive precision compared to traditional statistical methods (Vadyala *et al.* 2022, Zhang *et al.* 2023). Traditional techniques, such as linear regression models and multinomial models hinge on assuming a simplistic straight-line relationship. Usually, traditional regression models are utilized in scenarios with substantial variable correlation (Bangdiwala 2018). While, machine learning techniques create more comprehensive prediction models such as decision trees, support vector machines, and neural networks without the need for high variable correlation (Matsunaga and Fortes 2010, Romero *et al.* 2020). Hence, these studies excel in capturing nonlinear relationships in more complex scenarios. Machine learning techniques have been also broadly utilized in the field of geotechnical engineering some of the examples are: Güllü (2013) utilized strong motion records in developing artificial neural network (ANN) and gene expression programming (GEP) models to forecast V_{S30}. While another example of ANN and GEP is from Al-Swaidani *et al.* (2024), which predicts the strength of problematic clayey soil. Another great example of using neural network was made by Nguyen *et al.* (2022), wherein the where able to use predict the axial load bearing capacity of pile using ANN And also made an evaluation of the residual flexural strength of corroded reinforced concrete by using neural networks (Nguyen *et al.* 2022). On the other hand, Aydın *et al.* (2023) conducted a comparative analysis of various machine-learning techniques for a unified soil classification system (USCS). While, Javadi and Rezania (2009) pioneered the usage of genetic programming (GP) and evolutionary polynomial regression (EPR) for the modeling of the soil behavior as well as data mining in geotechnical engineering. As for Benemaran and Esmaeili-Falak (2023), they used multiple machine learning for predicting the Young's modulus of frozen sand. And, Nguyen-Minh *et al.* (2024) wherein they made an ingenious approach by combining the isogeometric analysis (IGA), limit analysis, and, machine learning by using the multivariate adaptive regression splines model (MARS) to determine the undrained stabilities of specific circular cavities. It is also important to consider pre-processing techniques that may increase the accuracy of final machine learning models in geotechnical fields. In particular, removing of outlier and balancing the data is important for this study as data to be used for this study is inherently imbalanced and has many outliers due to the nature of recording instruments. This consideration can significantly impact the overall data quality and influence the efficacy of constructing a machine-learning model (Maniruzzaman *et al.* 2018). There is currently a lack of research investigating the integration of HVSR data from strong motion records into the dataset for machine learning models.

This research would benefit the engineering practice by providing a framework for future works with the goal of making a tool that utilizes machine learning and different kinds of waves (strong ground motion, ambient noise., i.e.,) that will be able to serve as an alternative to traditional non-destructive methods of site characterization. Which has a further potential to have great impact in countries that have multiple earthquake occurrence as it would provide a much cost-effective measure for determining site class and would help improve the safety of the area by providing proper basis for the structural design.



(a) Magnitude

(b) Peak ground acceleration

(c) Hypocentral distance

Fig. 1 Distribution of dataset in terms of different input features

This study aims to make a framework by integrating machine learning techniques with the HVSR to build a predictive model and to evaluate performance of various pre-processing methods with up-date machine learning algorithms and by:

(1) utilizing strong ground recordings to obtain different features to create the dataset;

(2) enhancing model accuracy by eliminating outliers by applying an outlier detection method;

(3) applying the data balancing algorithm to even out the inherent imbalance in the dataset; an

(4) evaluating various machine learning models to determine the most effective combination.

2. Dataset and methodology

2.1 Origin of data

The dataset used for this study is publicly available data on the KiK-net website. The KiK-net is one of the well-established strong-motion seismograph networks managed by the national research institute for earth science and disaster prevention (NIED) in Japan. There are 669 KiK-net strong-motion stations scattered throughout Japan, and these strong-motion stations have a minimum 100-m-deep borehole (Aoi *et al.* 2004). Each station has seismographs placed at the bottom of each borehole and on the ground surface. The stations use tri-axial force-balance accelerometers. Earthquake signals are recorded using SMAC-MDK devices up to 2000 gals. Ground motions are stored based on trigger conditions for 6.5 hours of data at 200 Hz. The system's internal clock is automatically calibrated every hour using GPS signals (Aoi *et al.* 2020). Finally, the dataset utilized in this study consists of 15,103 earthquakes recorded during past years from 2000 to 2023 at 625 stations in Kik-net.

Table 2 Class distribution of earthquake records used in this study

NEHRP (1997) Site Class	Number of Records
A	258
B	4312
C	7848
D	2484
E	201

The resulting distribution of earthquake records used in this study is tabulated in Table 2. The output classification is based on the NEHRP general guidelines for site classification (BSSC, 2003) using the given V_{S30} data to determine the soil type of each site. Fig. 1 shows the distribution of the data in terms of the input features: magnitude, peak ground acceleration, hypocentral distance. Based on the Fig. 1, it can be seen that the magnitude of the earthquakes falls on the range of 4 to 9 with 7 having the highest frequency.

While, peak ground acceleration is mostly between the range of 0.01 to 0.45 with 0.05 having the highest frequency. The hypocentral distance of the dataset can be also seen in Fig. 1 which shows that majority of the data is around the range of 50km to 700km with 100 km being prevalent.

2.2 Flow of data processing

The overview of the whole process done in this study is seen in Fig. 2. It will start off with extracting features from the earthquake data gathered from the KiK-net website.

Then, additional processing will be done to address some of the problem of the dataset.

Synthetic minority over-sampling technique (SMOTE) and removing of outliers will applied in the dataset this is to address the issue of imbalance and outliers. After processing the dataset, it would be split into 2 parts one for training and one for testing. The training set would then undergo training using various machine learning models and it will be tested using the testing set which would then result in the different performance metrics.

This study partitions the dataset into training and testing sets with 70 percent and 30 percent of the data. Data splitting is randomized using the Scikit package, which guarantees an unbiased split. Subsequently, the dataset is

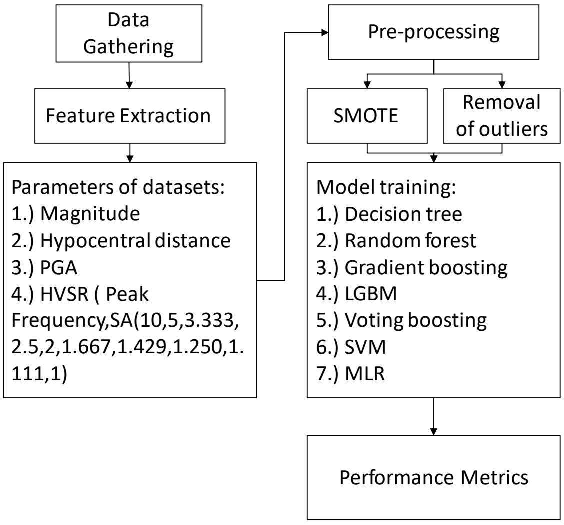


Fig. 2 Flowchart of data process

subjected to three distinct outlier detection techniques: mahalanobis distance, isolation forest, and local outlier factor, examining potential errors that may occur during the recording of earthquakes due to sensors subjected to challenging environmental circumstances (Samara *et al.* 2022). The ensuing datasets are subsequently employed for the seven machine-learning models that are previously discussed.

Imbalanced datasets introduce bias into the outcomes of the machine learning models particularly in multi-class classification as it enhances the accuracy of the classes with more sample size. The proportion ratios of the original dataset to the smallest dataset in this study are 1.28: 21.45: 39.04: 12.36: 1 for classes A, B, C, D, and E, respectively. They result in an uneven dataset and site classes A and E are nearly 20 to 40 times lower than the other classes. In this research, the synthetic minority over-sampling technique (SMOTE) is employed to address the issue of imbalanced class distribution. The SMOTE generates artificial datasets by extrapolating from the minority classes, thereby achieving a balanced dataset with equal samples for each classification. The utilization of synthetic data is limited to the training set and excluded from the testing set, as its purpose is solely to facilitate the model development process rather than its evaluation (Fernández *et al.* 2018, He and Ma 2013, Brownlee 2020).

The models were constructed using a dataset that underwent data balancing, explicitly employing the SMOTE. As mentioned earlier, the balancing technique employed three outlier identification approaches. Subsequently, the accuracy, precision, recall, and F1-score metrics were calculated at Table 3.

2.3 Input features

In this study, the HVSR of each earthquake record is selected as a key input parameter. The HVSR has been adopted as one of the crucial parameters, enabling the characterization of shallow-subsoil properties and assessing side effects. It measures the ratio of Fourier amplitudes between horizontal and vertical ground motion components at the Earth's surface. It is expressed as the following Eq. (1). In which HVSR$_s$ denotes the horizontal to vertical spectral ration at the surface while H$_s$ and V$_s$ are the horizontal and vertical spectral acceleration at the surface.

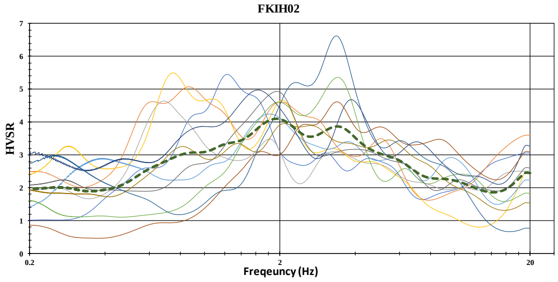


Fig. 3 Example of HVSR plots

$$HVSR_S = \frac{H_S}{V_S} \tag{1}$$

The basic principle of HVSR hinges on two assumptions: (1) the horizontal and vertical waves travel equally on the bedrock, making the horizontal to vertical spectral ratio at the bedrock, HVSR$_b$, equal to 1. (2) The seismic waves travel from the bedrock to the surface with no enlargement in the vertical component which would make the vertical transfer function, TF$_V$ = 1. By combining both of these assumptions and Eq. (1), we can see the relationship between HVSR$_S$ and horizontal transfer function,TF$_H$

$$TF_H = \frac{H_S}{H_B} * \frac{V_B}{V_S} * \frac{H_B}{V_B} = \frac{H_S}{V_S} \tag{2}$$

$$HVSR_S = TF_H \tag{3}$$

HVSR's simplicity in data acquisition and signal processing makes it applicable across various fields, including site effect research, earthquake recordings, and strong-motion data analysis. Its versatility makes it cost-effective and practical for seismic hazard assessment (Gulowaty and Ksieniewicz 2019).

In order to provide a detailed characteristic of HVSR curves, peak frequencies of HVSRs and ten spectral accelerations on the HVSR curves that are evenly distributed at a certain frequency are selected. To clarify, when the term HVSR is used in the study it is pertaining to the horizontal to vertical spectral ration at the surface (HVSR$_s$).

Additional input features include latitude and longitude of earthquake source, magnitude (M$_w$) of earthquake, sampling frequency (Hz) of records, arrival time of earthquake from its source to stations, hypocentral distance of hypocenter to the stations (R$_{hyp}$), and peak ground acceleration (PGA) of records. The majority of these parameters describe characteristics of strong earthquake motion records.

2.4 Outlier detection methods

Outlier detection methods (ODM) are procedures implemented to identify and rectify anomalous data that does not follow the expected pattern (Wang *et al.* 2019). They fulfill a vital function by detecting system malfunctions, fraudulent activities, and errors. Traditional methods are initially arbitrary, but they progress to become more methodical and systematic based on computer science and statistics principles (Hodge and Austin 2004). There are multiple variations of ODM, but in this study, the focus will be on the following: Mahalanobis distance, isolation forest, and local outlier factor.

2.4.1 Mahalanobis distance

Mahalanobis distance is a measure used to quantify the dissimilarity between two groups or populations based on a set of relevant characteristics or measurements. It also allows for considering different scales and correlations between variables, making it a useful metric for comparing groups with multivariate data. Considering the covariance matrix, Σ, this function determines the squared Mahalanobis distance, D^2, for every row in a matrix to the center vector, μ, defined with a singular column vector, x (Vinod 2014).

$$D^2 = (x - \mu)'\Sigma^{-1}(x - \mu) \tag{4}$$

As a metric for identifying outliers, Mahalanobis distance quantifies the impact of data points on regression coefficients. Significant distances are regarded as having greater leverage. Because mean and standard deviation are sensitive to outliers, researchers use more robust metrics to replace the center and covariance in Mahalanobis distance calculations (Rehman *et al*. 2019).

2.4.2 Isolation forest

The Isolation Forest algorithm is a model-based approach used to detect data anomalies in high dimensional datasets. In contrast with traditional methods that generate a profile of normal instances, this algorithm focuses on isolating and identifying anomalies with significant accuracy (Nguyen *et al.* 2023). It takes advantage of the fact that anomalies are the minority and have attribute-values that are very different from normal instances. The proposed methodology constructs an ensemble of isolation trees, whereby each tree isolates instances utilizing random partitioning. Anomalies are classified as instances exhibiting relatively low average path lengths on the trees. Isolation Forest has a linear time complexity, low memory requirement, and can handle large data sets and high-dimensional problems.

2.4.3 Local outlier factor

The local outlier factor (LOF) is an outlier detection algorithm that operates on the density of data points. It identifies outliers by assessing the local deviation of a given data point within a dataset, making it particularly effective for datasets with uneven distributions. The algorithm determines outliers by analyzing the density between each data point and its neighboring points. If a data point exhibits lower density than its neighbors, it is more likely to be flagged as an outlier (Cheng *et al*. 2019).

2.5 Data balancing

The resulting number of samples per each classification in this study is summarized in Table 2, presenting a great disparity in the total number of samples for A and E, with only 258 and 201, respectively, whereas C has the most samples with 7848. Itpresents an imbalance with the dataset, leading to a less accurate model. To make the best possible model, data balancing has been introduced in the form of SMOTE; Gulowaty and Ksieniewicz 2019), as shown in Fig. 2. The accuracy of the model is greatly affected by the imbalance in the class distribution. The imbalance in the dataset occurs because of the geographical distribution of Japan, where the majority of the lands with recording stations are classified as soil type C. Oversampling is determined to be more appropriate to balance the class distribution because it works well with moderately imbalanced data (Wongvorachan *et al.* 2023). Aside from that, if under sampling is used, the data would not be sufficient to make an accurate and reliable model.

SMOTE is used in this study to generate synthetic class examples using the pre-existing data in the minority class. By doing so, the classifier would be able to make a larger decision region that uses nearby minority class points. This would result in having a much more balanced dataset with more related minority class samples, which is why SMOTE typically performs better than other data balancing methods (Chawla *et al.* 2002).

2.6 Hyperparameter tuning

Hyperparameter tuning is an essential step in the design process of machine learning models for performance optimization. This process plays an important role in maximizing model performance and improving prediction accuracy by preventing model overfitting and enhancing generalization performance. In this study, we have applied the grid search method among various hyperparameter tuning methods.

The grid search method systematically detects the optimal configuration within a fixed range of hyperparameters using a decision-theoretic approach, providing an intuitive and complete method (Yang and Shami 2020). Table 3 presents the parameters applied to outlier detection and classification models, their ranges, and the optimal parameter values determined through grid search. In the "Range" column, the values in parentheses indicate the start, end, and increment values for the search range of each parameter.

Table 3 Optimal hyperparameter selection via grid search

	Model	Parameter	Range	Optimal
Outlier model	Mahalanobis Distance	scaling	(0.01, 0.1, 0.01)	0.01
	Isolation Forest	max_sample	(0, 200, 5)	140
		contamination	(0.1, 0.5, 0.1)	0.1
	Local outlier factor	n_neighbors	(5, 40, 5)	40
		contamination	(0.01, 0.1, 0.01)	0.01
Classifier model	Decision Tree	criterion	gini, c, log_loss	gini
	Random Forest	n_estimators	(0, 200, 5)	190
	Gradient Boosting	n_estimators	(0, 200, 5)	100
	LGBM	boosting_type	gbdt, dart, rf	gbdt
	SVM	C	(1, 5, 1)	1

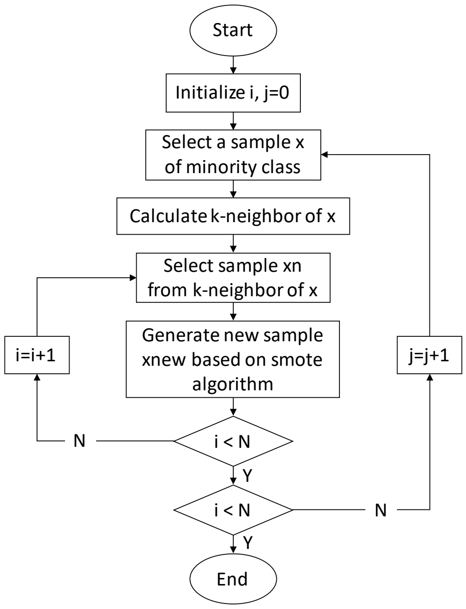

Start
Initialize i, j=0
Select a sample x of minority class
Calculate k-neighbor of x
Select sample xn from k-neighbor of x
Generate new sample xnew based on smote algorithm
i=i+1
j=j+1
i < N
N
Y
i < N
N
Y
End

Fig. 4 Algorithm of SMOTE

2.7 Machine learning

Stuyts and Suryasentna (2023) defined machine learning as a specialized branch of artificial intelligence focuses on automating the learning process from tabular data. It encompasses four distinct sub-disciplines: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, models are created using input-output. The primary objective is to enable the machine to acquire the experience by observing and understanding the relationship between input data and corresponding outcomes.

On the other hand, reinforcement learning involves machine learning through interaction with its environment, primarily employing a trial-and-error approach to maximize rewards until reaching a specified target. Unsupervised learning, in contrast, lacks labeled output, with some algorithmic features emerging organically from the data. Semi-supervised learning is a hybrid approach, combining elements from supervised and unsupervised learning models. Its primary aim is to enhance learning performance by leveraging abundant unlabeled data alongside a limited labeled data set.

Supervised machine learning models are adopted in this study, including decision tree, random forest, gradient boosting, LGBM; voting boosting, Support vector machine (SVM), and Multiple linear regression (MLR).

2.7.1 Decision Tree (DT)

A decision tree is a classification modeling approach that employs a divide-and-conquer strategy for analyzing large databases, illustrated in Fig. 1. It is particularly useful for discovering features and patterns relevant to discrimination and predictive modeling (Myles *et al*. 2004). The tree structure includes nodes (root, internal, and leaf) connected by branches, and key steps involve splitting, stopping, and pruning. This predictive tool aims to create a hierarchy of decision rules, maximizing accuracy while maintaining generalizability to new data (Song and Lu, 2015).

2.7.2 Random Forest (RF)

Random Forests are an ensemble learning method comprising a collection of tree-structured classifiers. Each tree is grown with a randomly sampled vector, independently and with the same distribution across all trees. The method employs a voting mechanism, where the ensemble of trees collectively predicts the most popular class. The generalization error converges as the number of trees increases, and the model's robustness to noise is enhanced through random feature selection. Internal estimates, such as out-of-bag estimates, guide the model's parameters and monitor error, strength, and correlation. Random Forests are applicable to both classification and regression tasks, providing accurate and robust predictions while offering insights into variable importance (Breiman, 2001).

2.7.3 Gradient Boosting (GB)

Gradient boosting is an ensemble machine learning technique that sequentially builds a strong predictive model by adding weak models, often decision trees, to correct errors the existing ensemble makes. The algorithm focuses on minimizing the negative gradient of a chosen loss function, allowing for flexibility in handling various data-driven tasks. The method's adaptability to different loss functions makes it customizable, and its simplicity facilitates implementation and experimentation (Natekin and Knoll 2013).

2.7.4 Light gradient boosting method (LGBM)

Light gradient boosting machine is a gradient boosting decision tree (GBDT) algorithm designed to enhance the efficiency and scalability of GBDT in the face of high-dimensional features and large datasets. To address these challenges it introduces two key techniques, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS selectively excludes data instances with small gradients during training, focusing on those with larger gradients for more accurate information gain estimation. EFB bundles mutually exclusive features to reduce their number, leveraging the sparsity of feature spaces. Through experiments on various datasets, it demonstrates significant acceleration in training speed, up to over 20 times, while maintaining nearly the same level of accuracy compared to conventional GBDT implementations (Ke *et al*. 2017).

2.7.5 Voting Boosting (Random Forest, Gradient Boosting, LGBM)

Hard Voting (HV) model is a type of voting algorithm used as a meta-classifier. In this approach, multiple machine learning classifiers are trained and evaluated in parallel, and each classifier independently predicts a class label. The final prediction of the ensemble is determined by a majority (plurality) vote, where the class label most frequently predicted by the individual classifiers becomes the final output of the ensemble. In the context of the Hard-Voting model, the final class label zˆ is determined by calculating the mode (most frequently occurring value) of the set of predictions from the individual classifiers. HV is described as the simplest case of a voting algorithm with lower error and less overfitting, as it relies on the majority decision of the ensemble to make predictions.

2.7.6 Support vector machine (SVM)

SVM is a robust classifier aiming to establish a hyperplane, or decision boundary, between two classes in a labeled training dataset. The hyperplane maximizes the margin between classes and is positioned to be as far as possible from the closest data points, known as support vectors (Mahesh 2019). Originally proposed for linear classification, SVM also employs the kernel method for handling non-linear problems. Kernel functions map data into higher-dimensional spaces, facilitating linear separation, with the choice of kernel significantly impacting SVM performance. The Radial Basis Function (RBF) kernel is an example, and kernel selection often involves experimentation and cross-validation to identify the most effective kernel for a specific pattern recognition problem (Huang *et al.* 2018).

2.7.7 Multiple linear regression (MLR)

MLR is a statistical technique used for estimating relationships between variables with cause-and-effect associations. Unlike univariate regression, which analyzes the relationship between a dependent variable and one independent variable, MLR deals with regression models having one dependent variable and multiple independent variables (Uyanık and Güler 2013).

2.8 Models from previous literatures

Empirical models were used to properly compare the result of the machine learning with pre-existing models. The first model is by using the Eq. (5) proposed by Kramer, 1996 (KR96)

$$f_{peak} = \frac{\mathrm{Vz}}{4H} \tag{5}$$

Wherein Vz is the shear wave velocity which would become Vs_{30} , average shear wave velocity at 30 m, by considering that H, height, be equal to 30 m. While, $\mathrm{f}_{\mathrm{peak}}$ is the peak frequency found in the average HVSR (Harinarayan and Kumar 2018). The second model would be the one proposed by Ghofrani and Atkinson 2014 (GA14)

$$\log(Vs_{30}) = 2.56 + 0.20\log(f_{peak}) \tag{6}$$

As suggested by Ghofrani and Atkinson, this is only applicable for global region as wells as for $\mathrm{f}_{\mathrm{peak}}$ greater than 1 Hz (Ghofrani and Atkinson 2014). The last model was proposed by Kwak and Seyhan (KS18) which assumes a non-linear relationship between the Vs_{30} and the $\mathrm{f}_{\mathrm{peak}}$ from the average HVSR

$$Vs_{30} = 200\left[\left(\frac{0.85f_{peak}}{1-0.05f_{peak}}\right)^2 - 1\right]^{\frac{1}{6}} \tag{7}$$

Table 4 Confusion Matrix and evaluation parameters

	True	False
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)
Accuracy	(TP + TN) / (TP + FP + TN + FN)	
Precision	TP / (TP + FP)	
Recall	TP / (TP + FN)	
F1-score	2 * ((Precision * Recall) / (Precision + Recall))	

This model is particularly effective with $\mathrm{f}_{\mathrm{peak}}$ in the range of 1-20 Hz but tends to overestimate the Vs_{30} when $\mathrm{f}_{\mathrm{peak}}$ reaches 12 Hz or greater (Kwak and Seyhan).

2.9 Metrics

Performance metrics are essential for assessing algorithm or model performance with specific datasets, particularly in classification problems. In this study, the confusion matrix was used to evaluate model's performance to calculate accuracy, precision, recall, and F1-score (Grandini *et al.* 2020, Aydin *et al.* 2023).

Evaluation parameters are calculated with confusion matrix as shown in Table 4 to offer a comprehensive overview of a model's performance under various conditions and scenarios.

3. Results & discussion

The overall accuracy of the models are tabulated in Tables 6 and 7, ranging between 0.61 and 0.77 when the original dataset was utilized, with the decision tree model shows the highest level of accuracy. The overall precision values for the original dataset range from 0.51 to 0.87, whereas overall recall and F1-score values range from 0.23 to 0.61. Upon applying outlier detection methods to the datasets, it was observed that all three methods yielded comparable levels of accuracy, ranging from 0.62 to 0.91. A discernible distinction becomes evident when alternative measurements are investigated, showing that the local outlier detection algorithm achieved the highest recall and F1-score. Meanwhile, the isolation forest algorithm had the highest precision value. In the context of the study, it was observed that the application of SMOTE resulted in a decrease in the lower boundary of accuracy for the original dataset to 0.36.

In contrast, the maximum accuracy increased to 0.85. A similar impact occurred when SMOTE was used with the outlier detection method, resulting in an accuracy range of 0.36 to 0.84 for all variations. The minimum accuracy is calculated for MLP and SVM, ranging 0.36 to 0.38 when SMOTE was used. On the other hand, the LGBM model increased accuracy, reaching a range of 0.84 to 0.85. Regarding the remaining metrics, the majority of the upper bounds remained relatively consistent. However, the lower bounds showed a decrease in value.

The overall performance of the classical HVSR methods were shown in Table 5 which is compared to the machine

Table 5 Comparison of overall performance metrics between machine learning models and classical HVSR models

		MLP	SVM	LGBM	GB	VB	DT	RF	KR96	GA14	KS18
Original Dataset	Accuracy	0.62	0.61	0.69	0.66	0.68	0.77	0.76	0.33	0.28	0.44
	Precision	0.68	0.84	0.51	0.51	0.57	0.61	0.87	0.30	0.19	0.15
	Recall	0.28	0.23	0.38	0.34	0.36	0.60	0.43	0.39	0.23	0.22
	F1-score	0.29	0.21	0.41	0.37	0.40	0.61	0.48	0.27	0.14	0.18

Table 6 Overall Performance Metrics for Outlier Detection Methods

		MLP	SVM	LGBM	GB	VB	DT	RF
Original Dataset	Accuracy	0.62	0.61	0.69	0.66	0.68	0.77	0.76
	Precision	0.68	0.84	0.51	0.51	0.57	0.61	0.87
	Recall	0.28	0.23	0.38	0.34	0.36	0.60	0.43
	F1-score	0.29	0.21	0.41	0.37	0.40	0.61	0.48
Local Outlier Factor	Accuracy	0.62	0.62	0.91	0.82	0.85	0.75	0.77
	Precision	0.64	0.84	0.84	0.69	0.73	0.53	0.88
	Recall	0.27	0.23	0.75	0.57	0.58	0.54	0.44
	F1-score	0.28	0.22	0.79	0.81	0.84	0.74	0.49
Isolation Forest	Accuracy	0.62	0.62	0.91	0.82	0.85	0.79	0.77
	Precision	0.64	0.83	0.88	0.78	0.93	0.68	0.88
	Recall	0.27	0.23	0.71	0.60	0.61	0.65	0.43
	F1-score	0.28	0.22	0.77	0.66	0.69	0.66	0.48
Mahalanobis Distance	Accuracy	0.62	0.61	0.90	0.82	0.84	0.76	0.76
	Precision	0.63	0.84	0.80	0.69	0.71	0.54	0.85
	Recall	0.27	0.23	0.70	0.56	0.57	0.55	0.42
	F1-score	0.28	0.21	0.74	0.81	0.82	0.74	0.47

learning models. We have used 3 models from previous literatures in estimating the Vs_{30} to get the site classification. The accuracy, precision, recall, and, F1-score were all below 50 percent, ranging from 0.14 to 0.44, which is low in comparison with the machine learning models. In particular, GA14 performed the least out of the three empirical models. Some of the can be attributed to the nature of the model which is not a good indicator for deep soil site as well as for f_{peak} less than 1 Hz (Ahn *et al.* 2021). This is because of the sample size present for each of the recording station. The current sample size is not sufficient for the empirical models because using peak frequencies require larger sample size per station and its accuracy may decrease significantly depending on the number of samples (Zhao *et al.* 2006).

3.1 The effect of eliminating outliers

The impacts of removing outliers using three distinct ODMs of Mahalanobis distance, isolation forest, and local outlier factor, were investigated in this study. The effect of removing the outlier can also be observed in Table 6, which presents the outcomes of each model. According to the findings presented in the tables, the decision tree algorithm exhibited superior performance when applied to the original dataset. However, removing outliers did not provide have any notable impact on the algorithm's performance. This outcome is similar to MLP, SVM, and RF, as deleting the outliers had little effect on these models. In the case of LGBM, GB, and VB, removing outliers resulted in a notable increase in their respective performance metrics. Consequently, LGBM emerged as the most proficient model among the abovementioned models when outliers were removed.

In order to improve understanding of the effects associated with removing outliers, graphical representations were generated to illustrate the comparative analysis of accuracy, precision, and F1-score across various site classifications before and after implementing ODM in Fig. 5. Meanwhile, Figs. 6 and 7 present a comparative analysis for precision and F1-score. The diagonal lines in figures, represents where the results remain unchanged irrespective of ODMs. If a point is above this line, it indicates a positive change in the outcome after removing outliers. The size of the symbols represents the data size for the different site class.

Conversely, if a point is positioned below the line, it suggests an adverse change in the outcome. Based on the data presented in Figs. 5 to 7, a significant proportion of the data points are above the diagonal line. In contrast, only a minority of points are positioned below this line. This result suggests an improvement in the performance of the models in terms of their accuracy, precision, and F1-score for each specific site class.

3.2 The Impact of data balancing

Considering the dataset's inherent imbalanced class

Table 7 Overall Performance Metrics for SMOTE

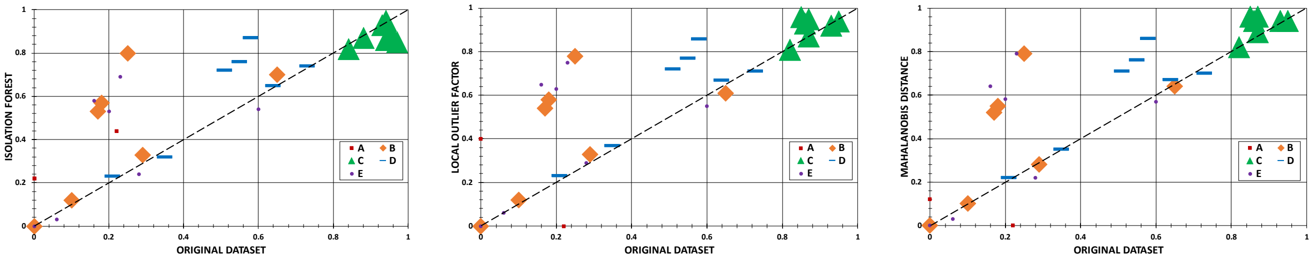
		MLP	SVM	LGBM	GB	VB	DT	RF
Original Dataset	Accuracy	0.36	0.38	0.85	0.70	0.81	0.66	0.76
	Precision	0.30	0.31	0.77	0.50	0.64	0.46	0.60
	Recall	0.51	0.52	0.78	0.72	0.76	0.55	0.60
	F1-score	0.28	0.29	0.77	0.55	0.67	0.48	0.60
Local Outlier Factor	Accuracy	0.37	0.38	0.84	0.69	0.79	0.69	0.76
	Precision	0.31	0.31	0.64	0.50	0.62	0.51	0.61
	Recall	0.51	0.51	0.75	0.72	0.72	0.58	0.61
	F1-score	0.28	0.29	0.71	0.55	0.65	0.53	0.61
Isolation Forest	Accuracy	0.36	0.37	0.84	0.69	0.79	0.67	0.74
	Precision	0.30	0.30	0.83	0.51	0.63	0.51	0.63
	Recall	0.57	0.54	0.76	0.74	0.69	0.57	0.61
	F1-score	0.27	0.28	0.87	0.56	0.65	0.53	0.61
Mahalanobis Distance	Accuracy	0.36	0.36	0.84	0.69	0.80	0.66	0.75
	Precision	0.30	0.30	0.83	0.49	0.59	0.50	0.56
	Recall	0.44	0.47	0.67	0.62	0.63	0.57	0.55
	F1-score	0.27	0.28	0.65	0.53	0.81	0.53	0.76

Table 8 Feature Importance Analysis Based on Random Forest

Original		**Mahalanobis Distance**		**Isolation Forest**		**Local Outlier Factor**	
Feature	**MDI**	**Feature**	**MDI**	**Feature**	**MDI**	**Feature**	**MDI**
Station Code	0.115	Station Code	0.118	Station Code	0.118	Station Code	0.121
Peak Frequency	0.081	Peak Frequency	0.086	Peak Frequency	0.089	Peak Frequency	0.093
SR10	0.068	SR10	0.074	SR10	0.069	SR1.667	0.068
SR1.667	0.063	Rhyp	0.063	SR1.667	0.068	SR3.333	0.068
SR3.333	0.06	SR1.1667	0.061	Rhyp	0.06	SR10	0.063
Rhyp	0.06	SR3.333	0.06	SR3.333	0.06	SR5	0.055
SR1.429	0.059	SR5	0.055	SR5	0.058	Rhyp	0.055
SR5	0.054	SR1.429	0.054	SR1.429	0.054	SR1.429	0.054
SR2	0.053	SR2.5	0.052	SR2	0.053	SR2	0.052
SR1.111	0.05	SR2	0.05	SR1	0.046	SR1	0.045
SR1	0.049	SR1	0.044	SR1.111	0.044	SR2.5	0.045
SR2.5	0.046	SR1.111	0.042	Lat	0.042	SR1.111	0.043
Lat	0.041	SR1.25	0.041	SR2.5	0.041	SR1.25	0.04
Long	0.041	Long	0.04	SR1.25	0.04	Lat	0.04
SR1.25	0.04	Mag	0.039	Long	0.039	Long	0.04
Mag	0.035	Lat	0.038	Mag	0.036	Arrival Time	0.039
Arrival Time	0.035	Arrival Time	0.033	Arrival Time	0.035	Mag	0.033
PGA	0.031	PGA	0.03	PGA	0.032	PGA	0.03
Sampling Freq	0.019	Sampling Freq	0.019	Sampling Freq	0.015	Sampling Freq	0.017

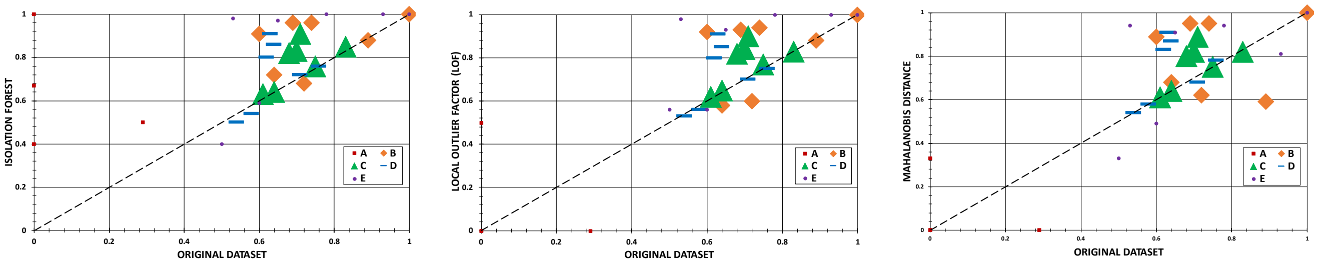
distributions in this study, the effect of a data balancing technique, SMOTE, was examined. The initial hypothesis suggested that the process of balancing the dataset would lead to an enhancement in the overall accuracy of all models, as compared to the scenario where SMOTE was not employed. However, when comparing the outcomes depicted in Tables 6 and 7, before and after utilizing SMOTE, it is evident that this expectation still needs to be met. It is apparent that a comparable trend with the effects of removing outliers emerged wherein the performance of MLP, SVM, and DT models showed a decrease in their performance when SMOTE was used. This is in contrast to LGBM, GB and VB, which showed increased accuracy. On the other hand, Random Forest exhibited the same accuracy whether data balancing was utilized or not.

In order to analyze the impacts of data balancing, the outcomes of the distinct site classes for each model are compared in Figs. 8 to 11. There is a noticeable rise in the precision of the site classifications, namely for classes A, B, and E. In contrast, the application of SMOTE resulted in a decrease in accuracy for site class C. Regarding accuracy, most of the findings are situated in the upper part of the



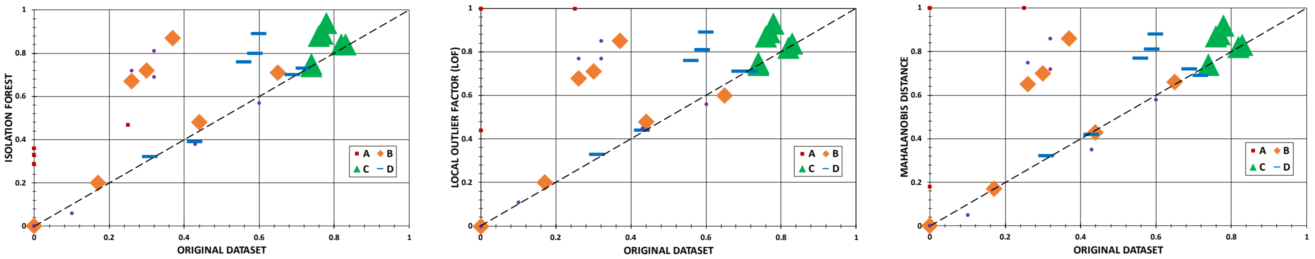
(a) Isolation forest (b) Local outlier factor (c) Mahalanobis distance

Fig. 5 Comparison of before and after removing the outliers (Accuracy) (The size of the figures indicates the sample size)



(a) Isolation forest (b) Local outlier factor (c) Mahalanobis distance

Fig. 6 Comparison of before and after removing the outliers (Precision) (The size of the figures indicates the sample size)



(a) Isolation forest (b) Local outlier factor (c) Mahalanobis distance

Fig. 7 Comparison of before and after removing the outliers (F1-score) (The size of the figures indicates the sample size)

graph and in proximity to the diagonal line, indicating a lack of substantial variation in their outcomes. Although there is a rise in the F1-score, its magnitude is less substantial than accuracy.

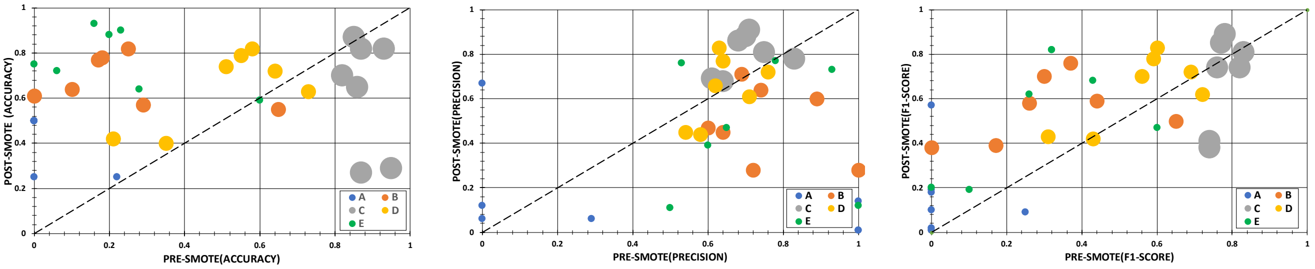
The overall metrics indicate a decline in performance for specific models when SMOTE was employed, it was observed that other site classes experienced a significant improvement in accuracy and F1-score, particularly the minority classes. This outcome aligns with the primary objective of implementing data balancing techniques: mitigating bias towards classes with a high sample count.

The SMOTE algorithm is employed in conjunction with removing outliers, as depicted in Figs. 8 to 11. In contrast to the data presented in Table 7, Figs. 8 to 11 indicate a comparatively lesser degree of improvement across the various site classes, except for classes A and E. Regarding, the precision and F1-score, the outcome remained comparatively consistent. Although this is the case, there remains a compelling case for combining SMOTE with outlier detection methods, specifically the combinations of LGBM plus Isolation and Forest plus SMOTE, which yields the highest overall F1-score and maintains a near-perfect accuracy ranking in this study.

3.3 Feature importance analysis

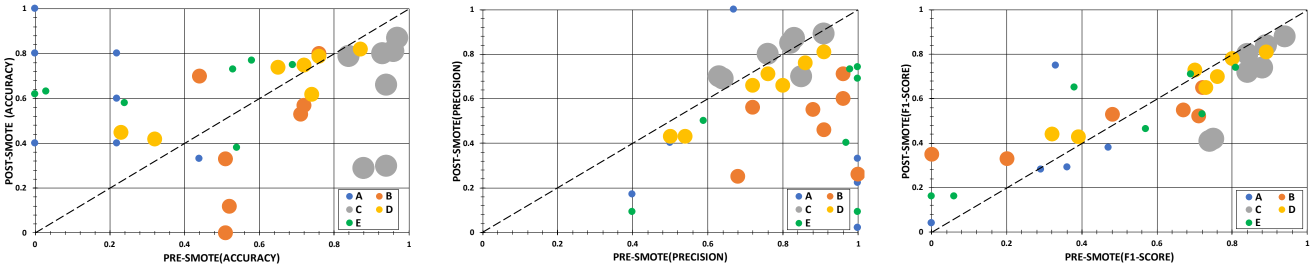
Feature importance analysis is made to determine which feature contributes the most in different machine learning modes. By doing so, it would help in having a greater understanding on how the features affect the models (Nguyen-Minh *et al.* 2023). By doing so, it can also improve the reliability and performance while also decreasing the complexity of the models by removing features that were shown to be the least significant. Table 8 presents a list of feature importance rankings based on the Mean Decrease in Impurity (MDI) values using the Random Forest model. The rankings compare the features of the original dataset, which has not undergone outlier processing, with the features by outlier detection model (Archer and Kimes 2008). The analysis is based on the top and bottom five features according to their frequency of occurrence across all scenarios. The top five features, which include 'Station Code' and 'Peak Frequency,' consistently show the highest importance across all scenarios, indicating their crucial role in the model's predictive performance.

While, 'SR10' and 'SR1.667,' despite some variation in importance rankings depending on the scenario,



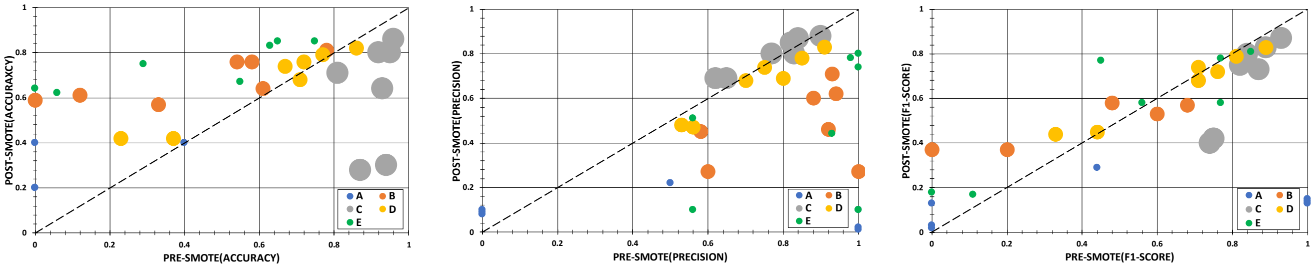
(a) Accuracy (b) Precision (c) F1-score

Fig. 8 Comparison of the pre-SMOTE and post-SMOTE using the original dataset (The size of the figures indicates the sample size)



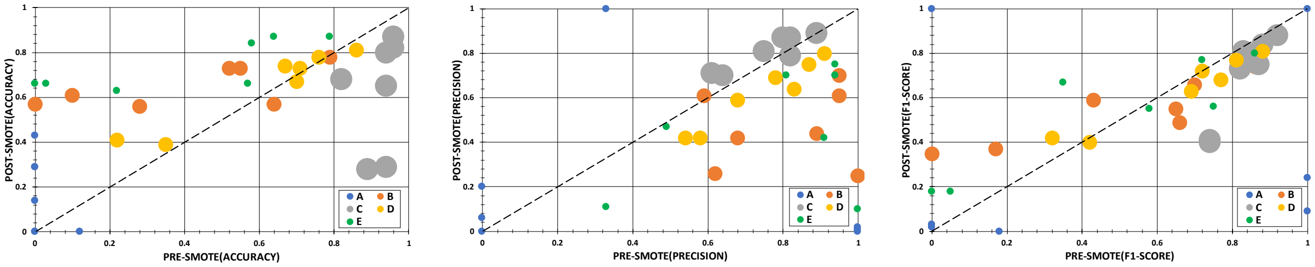
(a) Accuracy (b) Precision (c) F1-score

Fig. 9 Comparison of the pre-SMOTE and post-SMOTE in conjunction with Isolation Forest (The size of the figures indicates the sample size)



(a) Accuracy (b) Precision (c) F1-score

Fig. 10 Comparison of the pre-SMOTE and post-SMOTE in conjunction with Local Outlier Factor (The size of the figures indicates the sample size)



(a) Accuracy (b) Precision (c) F1-score

Fig. 11 Comparison of the pre-SMOTE and post-SMOTE in conjunction with Local Outlier Factor (The size of the figures indicates the sample size)

consistently emerge as significant variables. 'SR3.333' and 'Rhyp' exhibit changes in importance across different models such as Mahalanobis Distance, Isolation Forest, the Original, and the Local Outlier Factor.

For the bottom five features, 'Sampling Freq', 'PGA', and 'Arrival Time' consistently display the lowest importance, suggesting that they have little influence on the model's decision-making process. 'Mag' shows variability in importance rankings across different scenarios but consistently appears, indicating a lower impact. 'SR1.25', 'Lat', and 'Long' are not consistently significant but appear frequently, hinting at their limited influence. This suggests that the impact of these variables on model performance can vary depending on the given conditions and scenarios.

3.4 Optimal combinations

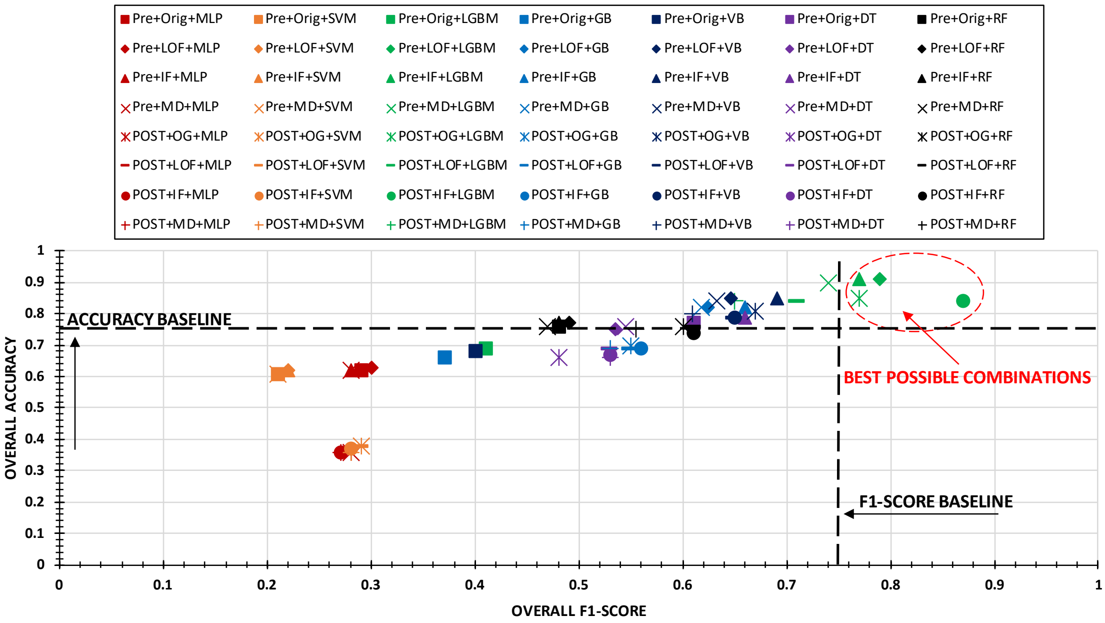


Fig. 12 Overall Accuracy vs Overall F1-score

The results indicate that LGBM has the highest overall accuracy of 0.84 to 0.91 when outlier detection methods and SMOTE are used. This is excellent, given that there are five classes in the study. At first, LGBM achieved an accuracy of 0.69 without ODM or SMOTE. However, upon applying ODM or SMOTE, its overall classification increased from third to first, and it now possessed the highest accuracy among all models and combinations. Subsequently, voting boosting maintains an accuracy range of 0.68 to 0.85. Although the remaining models exhibit a decent accuracy ranging from 0.51 to 0.82, The sole anomalies seen were in the outcomes of Support Vector Machines (SVM) and Multiple Linear Regression (MLR), where the overall accuracy fell below 50 percent (about 0.34 to 0.40) when SMOTE was employed. In summary, the accuracy of LGBM exhibited the most substantial improvement when outliers were eliminated, both before and after the use of SMOTE. This phenomenon can be explained in greater detail by referring to Fig. 12.

Fig. 12 depicts a graphical representation that compares the overall accuracy and F1-score. The 64 data points represent a unique combination of ODM, SMOTE, and machine learning models. A value of 0.75 for both accuracy and F1-score can be set as a baseline for dataset with mode input features and real-world data (Guo *et al.* 2022). This baseline would show the best possible combination of pre-processing and machine learning model. Based on the data presented in Fig. 12, it is evident that the majority of accuracy values for the various combinations fall within the range of 0.6 to 0.9.

In contrast, the F1-score has a broader distribution from 0.2 to 0.9. Several combinations have demonstrated an accuracy exceeding 0.75, although they failed to attain an F1-score of 0.75. The only combinations that exhibited accuracy and F1-score values exceeding 0.75 were those using LGBM as the machine learning model. Specifically, the combination of LGBM with Isolation Forest and SMOTE and LGBM with LOF yielded the highest F1-score and accuracy, respectively. This finding demonstrates that the LGBM model exhibited superior performance compared to other models when either ODMs or SMOTE techniques were employed. Based on the results of this research, it is recommended to use outlier detection methods particularly LOF or IF which was shown to have significant impact in both the accuracy and the F1-score. On the other hand, it is recommended to use SMOTE for a fairly imbalanced dataset as it would help in having a balanced accuracy across all classes.

4. Conclusions

This study investigated the utilization of seven machine learning models to classify the locations of 625 different stations based on a dataset of 15,103 earthquakes from 2000 up to January 2023 reported at KiK-net from Japan. In order to enhance the outcomes of the analysis, the following procedures were implemented in this research: outlier detection approach and data balancing. Three outlier detection approaches, namely Mahalanobis Distance, Isolation Forest, and Local Outlier Factor, are used to eliminate any outliers. Data balancing is used in this paper due to the dataset's intrinsic imbalance by SMOTE. In summary, the following conclusions were drawn from this study:

- Outlier detection methods are proven to increase the accuracy of LGBM, gradient boosting, and voting boosting.
- The SMOTE is also proven to increase the accuracy of the site classes with lower count.
- In this study, it is concluded that LGBM is the best performer out of all of the machine learning models as it outperformed the other model in both accuracy and F1-score.

• Peak frequency is shown to have significant impact based on the MDI. This further proves its applicability as a parameter for site characterization.
• The results of this research suggest that the findings and the machine learning model can serve as an initial framework for in the researcher's goal of developing a tool that can classify soils by using pre-existing earthquake recordings that would be beneficial in engineering practices as it would serve as another means of determining soil classification using non-destructive methods and pre-existing sensors.
• The results of this research prove that the process is highly accurate. Thus, the process shown in this research can be used with other or local pre-existing strong motion records, which gives engineers a preliminary soil characterization.

Based on the experiences of the researchers during this research, here are some recommendations for future works: Larger dataset would be beneficial in making improving the machine learning model. While, other outlier detection method, data balancing method, and feature importance analysis method can be done to get more insights. And, exploration of new features like geological topography which has several empirical models to predict *Vs₃₀* are great directions for this research.

Lastly, the final goal of this research is to develop a practical tool that is able assess the site classification and predict Vs profile by using weak ground motion as well as ambient mechanical noise recording while only relying on a single accelerometer which continuously gather data in a cheap and non-destructive way.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government and Ministry of Science and ICT (MSIT) (No.2021R1A4A1031509 and No.2021R1A2C2009985).

References

A Design Spectrum Model Featuring Resonant-Like Soil-Amplification (2013).

Abbas, M., Elbaz, K., Shen, S.L. and Chen, J. (2021), "Earthquake effects on civil engineering structures and perspective mitigation solution a review", *Arabian J. Geosci.*, **14**(14), 1350. https://doi.org/10.1007/s12517-021-07664-5.

Adams, R.D. (1990), "Earthquake occurrence and effects: Injury", **21**(1), 17-20. https://doi.org/10.1016/0020-1383(90)90146-L.

Ahn, J.K., Kwak, D.Y. and Kim, H.S. (2021), "Estimating VS30 at Korean Peninsular seismic observatory stations using HVSR of event records", *Soil Dyn. Earthq. Eng.*, **146**(106650). https://doi.org/10.1016/j.soildyn.2021.106650.

Aljanabi, K.R. and AL-Azzawi, O.M. (2021), "Neural network application in forecasting maximum wall deflection in homogenous clay", *Int. J. Geo-Eng.*, **12**(1), 29. https://doi.org/10.1186/s40703-021-00158-z.

Aoi, S., Asano, Y., Kunugi, T., Kimura, T., Uehira, K., Takahashi, N., Ueda, H., Shiomi, K., Matsumoto, T. and Fujiwara, H. (2020), "MOWLAS: NIED observation network for earthquake, tsunami and volcano", *Earth. Planet. Sp.*, **72**(1), 126. https://doi.org/10.1186/s40623-020-01250-x.

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

Al-Swaidani, A.M., Meziab, A., Khwies, W.T., Al-Bali, M. and Lala, T. (2024), "Building MLR, ANN and FL models to predict the strength of problematic clayey soil stabilized with a combination of nano lime and nano pozzolan of natural sources for pavement construction", *Int. J. Geo-Eng.*, **15**(1), 2. https://doi.org/10.1186/s40703-023-00201-1.

Archer, K.J. and Kimes, R.V. (2008), "Empirical characterization of random forest variable importance measures", *Comput. Stat. Data Anal.*, **52**(4), 2249-2260. https://doi.org/10.1016/j.csda.2007.08.015.

Aydın, Y., Işıkdağ, Ü., Bekdaş, G., Nigdeli, S.M. and Geem, Z.W. (2023), "Use of machine learning techniques in soil classification", *Sustainability*, **15**(3), 2374. https://doi.org/10.3390/su15032374.

Baise, L.G., Kaklamanos, J., Berry, B.M. and Thompson, E.M. (2016), "Soil amplification with a strong impedance contrast", *Boston, Massachusetts Engineering Geology*, **202**, 1-13, https://doi.org/10.1016/j.enggeo.2015.12.016.

Bangdiwala, S.I. (2018), "Regression: simple linear", *Int. J. Injury Control Saf. Promotion*, **25**(1), 113-115. https://doi.org/10.1080/17457300.2018.1426702.

Bard, P.Y., Cadet, H., Endrun, B., Hobiger, M., Renalier, F., Theodulidis, N., Ohrnberger, M., Fäh, D., Sabetta, F., Teves-Costa, P., Duval, A.M., Cornou, C., Guillier, B., Wathelet, M., Savvaidis, A., Köhler, A., Burjanek, J., Poggi, V., Gassner-Stamm, G., Havenith, H.B., Hailemikael, S., Almeida, J., Rodrigues, I. Veludo, I. and Kristekova, M. (2010), "From non-invasive site characterization to site amplification: Recent advances in the use of ambient vibration measurements", *Earthquake Engineering in Europe*, (Eds., Garevski, M. and Ansal, A.), Geotechnical, Geological, and Earthquake Engineering, Springer Netherlands, Dordrecht, 105-123.

Benemaran, R.S. and Esmaeili-Falak, M. (2023), "Predicting the Young's modulus of frozen sand using machine learning approaches: State-of-the-art review", *Geomech. Eng.*, **34**(5), 507-527. https://doi.org/10.12989/gae.2023.34.5.507.

Biau, G. and Scornet, E. (2016). "A random forest guided tour: TEST, **25**(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7.

Bisong, E. (2019a), "Matplotlib and Seaborn", Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, (Ed., Bisong, E.), Apress, Berkeley, CA.

Bisong, E. (2019b), "NumPy." Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners, (Ed., Bisong, E.), Apress, Berkeley, CA.

Breiman, L. (2001), "Random forests: Machine learning", **45**(1), 5-32. https://doi.org/10.1023/A:1010933404324.

Brownlee, J. (2020a), *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python.*

Brownlee, J. (2020b), *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning.*

BSSC (2003), *Nehrp Recommended Provisions for Seismic Regulations for New Buildings and Other Structures (Fema 450), 2003rd Ed.*, NEHRP.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002), "SMOTE: Synthetic minority over-sampling technique", *J. Artif. Intel. Res.*, **16**, 321-357. https://doi.org/10.1613/jair.953.

Cheng, Z., Zou, C. and Dong, J. (2019), "Outlier detection using isolation forest and local outlier factor", *Proceedings of the conference on research in adaptive and convergent systems, RACS '19, Association for computing machinery,* New York, NY, USA.

Dobry, R., Borcherdt, R.D., Crouse, C.B., Idriss, I.M., Joyner, W.B., Martin, G.R., Power, M.S., Rinne, E.E. and Seed, R.B. (2000), "New site coefficients and site classification system used in recent building seismic code provisions", *Earthq. Spectra*, **16**(1), 41-67. https://doi.org/10.1193/1.1586082.

Fatahi, B., Tabatabaiefar, S.H.R. and Samali, B. (2014), "Soil-structure interaction vs Site effect for seismic design of tall buildings on soft soil", *Geomech. Eng.*, **6**(3), 293-320. https://doi.org/10.12989/gae.2014.6.3.293.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., and Herrera, F. (2018), Leaning from *Learning from imbalanced data sets*, Springer International Publishing, Cham.

Fukushima, Y., Bonilla, L.F., Scotti, O. and Douglas, J. (2007), "Site classification using horizontal-to-vertical response spectral ratios and its impact when deriving empirical ground-motion prediction equations", *J. Earthq. Eng.*, **11**(5), 712-724. https://doi.org/10.1080/13632460701457116.

Gallipoli, M.R. and Mucciarelli, M. (2009), "Comparison of site classification from vs30, vs10, and hvsr in Italy", *Bull. Seismol. Soc. Am.*, **99**(1), 340-351. https://doi.org/10.1785/0120080083.

Ghasemi, H., Zare, M., Fukushima, Y. and Sinaeian, F. (2009a), "Applying empirical methods in site classification, using response spectral ratio (H/V): A case study on Iranian strong motion network (ISMN)", *Soil Dyn. Earthq. Eng.*, **29**(1), 121-132. https://doi.org/10.1016/j.soildyn.2008.01.007.

Ghasemi, H., Zare, M., Fukushima, Y. and Sinaeian, F. (2009b). "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.*, **29**(1), 121-132. https://doi.org/10.1016/j.soildyn.2008.01.007.

Grandini, M., Bagli, E. and Visani, G. (2020), "Metrics for multi-class classification: an overview", https://doi.org/10.48550/arXiv.2008.05756.

Ghofrani, H. and Atkinson, G.M. (2014), "Site condition evaluation using horizontal-to-vertical response spectral ratios of earthquakes in the NGA-West 2 and Japanese databases", *Soil Dyn. Earthq. Eng.*, **67**, 30-43. https://doi.org/10.1016/j.soildyn.2014.08.015.

Güllü, H. (2013). "On the prediction of shear wave velocity at local site of strong ground motion stations: an application using artificial intelligence", *Bull. Earthq. Eng.*, **11**(4), 969-997. https://doi.org/10.1007/s10518-013-9425-8.

Gulowaty, B. and Ksieniewicz, P. (2019), "SMOTE algorithm variations in balancing data streams", Intelligent Data Engineering and Automated Learning – IDEAL 2019, (Eds., Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A.J., Menezes, R. and Allmendinger, R.), Lecture Notes in Computer Science, Springer International Publishing, Cham.

Guo, H., Zhuang, X., Chen, J. and Zhu, H. (2022), "Predicting earthquake-induced soil liquefaction based on machine learning classifiers: A comparative multi-dataset study", *Int. J. Comput. Method.*, **19**(8), 2142004. https://doi.org/10.1142/S0219876221420044.

Harinarayan, N.H. and Kumar, A. (2018), "Determination of NEHRP site class of seismic recording stations in the Northwest Himalayas and its adjoining area using HVSR method", *Pure Appl. Geophys.*, **175**(1), 89-107. https://doi.org/10.1007/s00024-017-1696-6.

Harinarayan, N.H. and Kumar, A. (2017), "Site classification of the strong motion stations of uttarakhand, India, based on the model horizontal to vertical spectral ratio", 141-149. https://doi.org/10.1061/9780784480489.015.

Hays, W.W. (1993), "The National Earthquake Hazards Reduction Program (NEHRP): Postearthquake investigations ; A Report of the Interagency Coordinating Committee's Subcommittee on Postearthquake Investigations, U.S. Government Printing Office".

He, H. and Ma, Y. (2013), "Imbalanced learning: Foundations", Algorithms, and Applications, John Wiley & Sons.

Hodge, V. and Austin, J. (2004), "A survey of outlier detection methodologies", *Artif. Intell. Rev.*, **22**(2), 85-126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

Hollender, F., Cornou, C., Dechamp, A., Oghalaei, K., Renalier, F., Maufroy, E., Burnouf, C., Thomassin, S., Wathelet, M., Bard, P.-Y., Boutin, V., Desbordes, C., Douste-Bacqué, I., Foundotos, L., Guyonnet-Benaize, C., Perron, V., Régnier, J., Roullé, A., Langlais, M. and Sicilia, D. (2018), "Characterization of site conditions (soil class, VS30, velocity profiles) for 33 stations from the French permanent accelerometric network (RAP) using surface-wave methods", *Bull. Earthq. Eng.*, **16**(6), 2337-2365, https://doi.org/10.1007/s10518-017-0135-5.

Holzer, T.L., Padovani, A.C., Bennett, M.J., Noce, T.E. and Tinsley, J.C. (2005), "Mapping NEHRP V S30 site classes", *Earthq. Spectra*, **21**(2), 353-370. https://doi.org/10.1193/1.1895726.

Hryciw, R., Shewbridge, S., Rollins, K., McHood, M. and Homolka, M. (1991), "Soil amplification at Treasure Island during the Loma Prieta earthquake", *Proceedings of the International Conferences on Recent Advances in Geotechnical Earthquake Engineering and Soil Dynamics*.

Huang, S., Cai, N., Pacheco, P.P., Narrandes, S., Wang, Y. and Xu, W. (2018), "Applications of Support Vector Machine (SVM) learning in cancer genomics", *Cancer Genomics & Proteomics*, **15**(1), 41-51.

Javadi, A. and Rezania, M. (2009), "Applications of artificial intelligence and data mining techniques in soil modeling", *Geomech. Eng.*, **1**(1), 53-74. https://doi.org/10.12989/gae.2009.1.1.053.

Ji, Y., Kim, B. and Kim, K. (2021), "Evaluation of liquefaction potentials based on shear wave velocities in Pohang City, South Korea", *Int. J. Geo-Eng.*, **12**(1), 3. https://doi.org/10.1186/s40703-020-00132-1.

Ji, K., Ren, Y. and Wen, R. (2017), "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.*, **147**, 79-94. https://doi.org/10.1016/j.jseaes.2017.07.032.

Ji, K., Zhu, C., Yaghmaei-Sabegh, S., Lu, J., Ren, Y. and Wen, R. (2023), "Site classification using deep-learning-based image recognition techniques", *Earthq. Eng. Struct. D.*, **52**(8), 2323-2338. https://doi.org/10.1002/eqe.3801.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017), "LightGBM: A highly efficient gradient boosting decision tree", Advances in Neural Information Processing Systems, Curran Associates, Inc.

Kramer, S.L. (1996), *Geotechnical earthquake engineering*, Engineering, **6**, 653.

Kwak, D.Y. and Seyhan, E. (2018), "Development of peak frequency-site condition correlation models using H/V spectral ratio", *Geotech. Earthq. Eng.Soil Dynam.*, https://doi.org/10.1061/9780784481462.033.

Lee, C.T., Cheng, C.T., Liao, C.W. and Tsai, Y.B. (2001), "Site classification of Taiwan free-field strong-motion stations", *Bull. Seismol. Soc. Am.*, **91**(5), 1283-1297. https://doi.org/10.1785/0120000736.

Lee, J.H., Kim, J.H. and Jae, K.K. (2018), "Amplification characteristics of domestic and overseas intraplate earthquake ground motions in Korean soil and standard horizontal design spectrum for soil sites", *J. Earthq. Eng. Soc. Korea*, **22**(7), 391-399. https://doi.org/10.5000/EESK.2018.22.7.391.

Lin, S., Gucunski, N., Shams, S. and Wang, Y. (2021), "Seismic site classification from surface wave data to Vs,30 without inversion", *J. Geotech. Geoenviron. Eng.*, **147**(6), 04021029.

https://doi.org/10.1061/(ASCE)GT.1943-5606.0002526.

Mahesh, B. (2019), Machine Learning Algorithms -A Review.

Maniruzzaman, Md., Rahman, Md.J., Al-MehediHasan, Md., Suri, H.S., Abedin, Md.M., El-Baz, A. and Suri, J.S. (2018), "Accurate diabetes risk stratification using machine learning, role of missing value and outliers", *J. Medical Syst.*, **42**(5), 92. https://doi.org/10.1007/s10916-018-0940-7.

Mathur, U., Kumar, N., Pandey, T.N. and Choudhary, A. (2017), "Classification and identification of soil", *Int. J. Adv. Res. Innov. Ideas in Education*, **3**(3), 780-785.

Matsunaga, A. and Fortes, J.A.B. (2010), "On the use of machine learning to predict the time and resources consumed by applications", *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, IEEE, Melbourne, Australia.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D. (2004), "An introduction to decision tree modeling", *J. Chemometrics*, **18**(6), 275-285. https://doi.org/10.1002/cem.873.

Nakhaei, M. and Ali Ghannad, M. (2008), "The effect of soil–structure interaction on damage index of buildings", *Eng. Struct.*, **30**(6), 1491-1499. https://doi.org/10.1016/j.engstruct.2007.04.009.

Natekin, A. and Knoll, A. (2013). "Gradient boosting machines, a tutorial", *Frontiers in Neurorobotics*, **7**.

National Research Institute for Earth Science and Disaster Resilience (2019), NIED K-NET, KiK-net, National Research Institute for Earth Science and Disaster Resilience, https:/doi.org/10.17598/NIED.0004.

Nguyen, T., Ly, D.K., Huynh, T.Q. and Nguyen, T.T. (2023). "Soft computing for determining base resistance of super-long piles in soft soil: A coupled SPBO-XGBoost approach", *Comput. Geotech.*, **162**, 105707. https://doi.org/10.1016/j.compgeo.2023.105707.

Nguyen, T., Ly, K.D., Nguyen-Thoi, T., Nguyen, B.P. and Doan, N.P. (2022a), "Prediction of axial load bearing capacity of PHC nodular pile using Bayesian regularization artificial neural network", *Soils Found.*, **62**(5), 101203. https://doi.org/10.1016/j.sandf.2022.101203.

Nguyen, T., Truong, T.T., Nguyen-Thoi, T., Van Hong Bui, L. and Nguyen, T.H. (2022b), "Evaluation of residual flexural strength of corroded reinforced concrete beams using convolutional long short-term memory neural networks", *Structures*, **46**, 899-912. https://doi.org /10.1016/j.istruc.2022.10.103.

Nguyen-Minh, T., Bui-Ngoc, T., Shiau, J., Nguyen, T. and Nguyen-Thoi, T. (2023), "Coupling isogeometric analysis with deep learning for stability evaluation of rectangular tunnels", *Tunn. Undergr. Sp. Tech.*, **140**, 105330. https://doi.org10.1016/j.tust.2023.105330.

Nguyen-Minh, T., Bui-Ngoc, T., Shiau, J., Nguyen, T. and Nguyen-Thoi, T. (2024), "Undrained sinkhole stability of circular cavity: a comprehensive approach based on isogeometric analysis coupled with machine learning", *Acta Geotechnica*, https://doi.org 10.1007/s11440-024-02266-3.

Phung, V., Atkinson, G.M. and Lau, D.T. (2006), "Methodology for site classification estimation using strong ground motion data from the Chi-Chi, Taiwan, earthquake", *Earthq. Spectra*, **22**(2), 511-531. https://doi.org/10.1193/1.2198873.

Pinzón, L.A., Pujades, L.G., Macau, A., Carreño, E. and Alcalde, J.M. (2019), "Seismic site classification from the horizontal-to-vertical response spectral ratios: Use of the Spanish strong-motion database", *Geosciences*, **9**(7), 294. https://doi.org/10.3390/geosciences9070294.

Pradhan, M.K., Chakraborty, S., Ready, G.R. and Srinivas, K. (2021), "Experimental study of soil amplification and soil-pile-structure interaction performing shake table test", *Proceedings of the Indian Geotechnical Conference 2019*, (Eds., Patel, S., Solanki, C.H., Reddy, K.R. and Shukla, S.K.), Lecture Notes in Civil Engineering, Springer, Singapore.

Rehman, N. ur, Khan, B. and Naveed, K. (2019), "Data-driven multivariate signal denoising using mahalanobis distance", *IEEE Signal Pr. Lett.*, **26**(9), 1408-1412. https://doi.org/10.1109/LSP.2019.2932715.

Romero, M.P., Chang, Y.M., Brunton, L.A., Parry, J., Prosser, A., Upton, P., Rees, E., Tearne, O., Arnold, M., Stevens, K. and Drewe, J.A. (2020), "Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making", *Preventive Veterinary Medicine*, **175**, 104860. https://doi.org/10.1016/j.prevetmed.2019.104860.

Samara, M.A., Bennis, I., Abouaissa, A. and Lorenz, P. (2022), "A survey of outlier detection techniques in IoT: Review and classification", *J. Sensor Actuat. Networks*, **11**(1), 4. https://doi.org/10.3390/jsan11010004.

Song, Y. and Lu, Y. (2015), "Decision tree methods: applications for classification and prediction", *Shanghai Archives of Psychiatry*, **27**(2), 130-135. https://doi.org/10.11919/j.issn.1002-0829.215044.

Stančin, I. and Jović, A. (2019), "An overview and comparison of free Python libraries for data mining and big data analysis", Opatija, Croatia.

Stepanek, H. (2020), "Thinking in Pandas: How to use the python Data Analysis Library the Right Way", Apress, Berkeley, CA.

Stuyts, B. and Suryasentna, S. (2023), "Applications of data science in offshore geotechnical engineering: State of practice and future perspectives", *The Society of Underwater Technology*, London, UK, 1972-1993.

The pandas development team. (2023), "Pandas-dev/pandas: Pandas", Zenodo. https://doi.org/10.5281/zenodo.8364959.

Trifunac, M.D. (1990). "How to model amplification of strong earthquake motions by local soil and geologic site conditions", *Earthq. Eng. Struct. D.*, **19**(6), 833-846. https://doi.org/10.1002/eqe.4290190605.

Türköz, M. (2019), "The effect of soil type and different in-situ test results on soil amplification analysis", *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, **10**(3), 1187-1196. https://doi.org/10.24012/dumf.589196.

Uyanık, G.K. and Güler, N. (2013), "A Study on multiple linear regression analysis", *Procedia - Social and Behavioral Sciences*, **106**, 234-240. https://doi.org/10.1016/j.sbspro.2013.12.027.

Vadyala, S.R., Betgeri, S.N., Matthews, J.C. and Matthews, E. (2022), "A review of physics-based machine learning in civil engineering", *Results in Eng.*, **13**, 100316. https://doi.org/10.1016/j.rineng.2021.100316.

Verdugo, R. (2019), "Seismic site classification: Soil dynamics and earthquake engineering", **124**, 317-329. https://doi.org/10.1016/j.soildyn.2018.04.045.

Vinod, H.D. (2014), "Matrix algebra topics in statistics and economics using R", Handbook of Statistics, (Eds., Rao, M.B. and Rao, C.R.), Computational Statistics with 143-176.

Wang, H., Bah, M.J. and Hammad, M. (2019), "Progress in outlier detection techniques", *A Survey: IEEE Access*, **7**, 107964-108000. https://doi.org/10.1109/ACCESS.2019.2932769.

Wen, R., Ren, Y., Zhou, Z. and Shi, D. (2010), "Preliminary site classification of free-field strong motion stations based on Wenchuan earthquake records", *Earthq. Sci.*, **23**(1), 101-110. https://doi.org/10.1007/s11589-009-0048-8.

Wongvorachan, T., He, S. and Bulut, O. (2023), "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining", *Information*, **14**(1), 54. https://doi.org/10.3390/info14010054.

Xu, R. and Wang, L. (2021), "The horizontal-to-vertical spectral ratio and its applications", *EURASIP J. Adv. Signal Pr.*, **2021**(1), 75. https://doi.org/10.1186/s13634-021-00765-z.

Yang, L. and Shami, A. (2020), "On hyperparameter optimization of machine learning algorithms", *Theor. Pract. Neurocomput.*, **415**(295-316). https://doi.org/10.1016/j.neucom.2020.07.061.

Zhang, W., Gu, X., Hong, L., Han, L. and Wang, L. (2023), "Comprehensive review of machine learning in geotechnical reliability analysis, Algorithms, applications and further challenges", *Appl. Soft Comput.*, **136**, 110066. https://doi.org/10.1016/j.asoc.2023.110066.

Zhao, J.X., Irikura, K., Zhang, J., Fukushima, Y., Somerville, P.G., Asano, A., Ohno, Y., Oouchi, T., Takahashi, T. and Ogawa, H. (2006a), "An empirical site-classification method for strong-motion stations in Japan using h/v response spectral ratio", *Bull. Seismol. Soc. Am.*, **96**(3), 914-925. https://doi.org/10.1785/0120050124.

Zhao, J., Irikura, K., Zhang, J., Fukushima, Y., Somerville, P., Asano, A., Saiki, T., Okada, H. and Takahashi, T. (2023), Site classification for strong-motion stations in Japan using H/V response spectral ratio.

Zhao, J.X., Zhang, J., Asano, A., Ohno, Y., Oouchi, T., Takahashi, T., Ogawa, H., Irikura, K., Thio, H.K., Somerville, P.G., Fukushima, Y. and Fukushima, Y. (2006b), "Attenuation relations of strong ground motion in Japan using site classification based on pPredominant period", *Bull. Seismol. Soc. Am.*, **96**(3), 898-913. https://doi.org/10.1785/0120050122.

IC