Case Studies in Construction Materials 19 (2023) e02554

Contents lists available at ScienceDirect

Case Studies in Construction Materials

journal homepage: www.elsevier.com/locate/cscm

Sequential backward feature selection for optimizing permanent strain model of unbound aggregates

Samuel Olamide Aregbesola [a], Jongmuk Won [b], Seungjun Kim [c], Yong-Hoon Byun [a,*]

[a] School of Agricultural Civil & Bio-Industrial Engineering, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, South Korea
[b] Department of Civil and Environmental Engineering, University of Ulsan, Daehak-ro 93, Nam-gu, Ulsan 44610, South Korea
[c] School of Civil, Environmental and Architectural Engineering, Korea University, Seoul 02841, South Korea

ARTICLE INFO

Keywords:
Aggregate
Feature selection
Machine learning
Optimization
Permanent strain

ABSTRACT

This study proposes a novel framework for identifying the optimal feature set required to predict the permanent strain of unbound aggregates. An experimental database consisting of 16 input features is preprocessed and the performance of 10 machine learning models is evaluated. The best-performing model is then paired with a sequential backward selection algorithm to determine the optimal feature set for predicting the permanent strain. Finally, the selected features are used to predict the permanent strain, and the performance is compared with those obtained from the principal components analysis. Six features are selected as the optimal feature set. Furthermore, the selected features accurately predict permanent strain with a root mean square error value of 0.014, which is smaller than those obtained from principal components analysis. Thus, the feature selection approach for machine learning models effectively predicts the permanent strain of unbound aggregates using a limited set of input features.

1. Introduction

Machine learning (ML) has received extensive attention and found numerous applications across various fields due to increased availability of data and computational resources. In civil engineering, previous studies have mainly focused on supervised learning, which involves predicting a target variable using carefully selected independent features. Recently, ML algorithms, such as decision trees [1], support vector machines [4,11], random forest [17,41], multivariate adaptive regression splines [33] and k-nearest neighbors [24,31], have been extensively used.

ML models have shown success in various civil engineering problems, including slope stability analysis, tunnel boring machine performance, foundation design, and pavement condition detection [2,21,23,29,37,40,45]. They have been utilized to predict soil properties [5], liquefaction potential [47], and to quantify pavement damage segmentation [15]. Especially for geotechnical engineering, one of the main advantages of these models is the availability of free, open-source packages which can be readily adjusted to suit datasets [48]. Additionally, they have powerful nonlinear mapping abilities and outperform constitutive models in terms of speed, efficiency, and accuracy. However, these models have limitations such as limited interpretability and the need for sufficient data to avoid overfitting or failure to generalize [14,49]. Therefore, further research is necessary to address these limitations and promote the use of ML in geotechnical engineering.

Feature selection is a crucial task that is often overlooked in ML; however, it can considerably affect model performance. It involves

* Corresponding author.
E-mail address: yhbyun@knu.ac.kr (Y.-H. Byun).

https://doi.org/10.1016/j.cscm.2023.e02554
Received 31 August 2023; Accepted 5 October 2023
Available online 5 October 2023

selecting relevant features from a dataset, which can reduce dimensionality and improve model accuracy, training times, and computational cost [8]. However, previous studies in geotechnical engineering have mainly focused on prediction accuracy and feature importance using tree-based algorithms. These studies often fail to eliminate irrelevant parameters during the training process [12,32]. Only a few studies have explored data preprocessing and feature selection methods [13,16,39,46]. For instance, Hu [16], proposed a four-step approach to data cleaning and feature selection for soil liquefaction. A feature selection approach based on global sensitivity analysis and random forest was used to determine optimal feature subsets of tunnel engineering data [46]. In addition, dimensionality reduction techniques, such as principal component analysis (PCA), have been used to process time-based settlement data of road embankments [39]. Nevertheless, there is a need to investigate ML-based feature selection techniques in geotechnical engineering to further improve the model performance and reduce complexity.

ML techniques have found applications in predicting the properties and behavior of aggregates in pavements subjected to dynamic loading [3,9,18,42,43]. Unbound aggregate materials are commonly used as the base and subbase layers in pavement systems. As shown in Fig. 1, these materials exhibit elasto-plastic behavior and experience total deformation when subjected to traffic loads. Total deformation consists of both recoverable and plastic deformations, which are necessary for determining the resilient modulus and permanent strain, respectively [27]. Ikeagwuani [18], utilized multivariate adaptive regression splines, k-nearest neighbors, and support vector machines to determine the resilient modulus of unbound granular materials. Besides, several studies have employed artificial neural networks to predict the resilient modulus of both bound and unbound aggregates [18], while Alnedawi et al. [3] adopted an artificial neural network-based approach to predict the permanent deformation of unbound granular materials. Recently, Won et al. [43], proposed an ML framework for predicting permanent strain using three single-learning and two ensemble algorithms. However, to the best of our knowledge, feature selection studies based on ML models have yet to be applied to predict the permanent strain of unbound aggregates.

This study proposes a novel framework for determining the optimal feature set required to accurately predict the permanent strain of unbound aggregates. To prevent overfitting, a large dataset is used, and the data preprocessing steps are thoroughly documented. Initially, 10 different ML algorithms are applied to predict the permanent strain and the best-performing model is selected based on metrics, such as root mean squared error. A sequential backward selection (SBS) algorithm is then paired with the chosen model to identify the most relevant features for predicting the target variable. Furthermore, the selected feature set is used to predict the permanent strain and compared to the performance obtained from PCA-based dimensionality reduction. Finally, the prediction accuracy of the feature selection-based model is discussed on the basis of the measured experimental data.

2. Database and preprocessing

2.1. Data description

This study utilizes a dataset derived from source and engineered gradations of fifteen crushed aggregates that serve as unbound base courses in flexible pavements [7,35]. The dataset consists of a total of 22,590 data points, obtained from 15 samples at three shear stress ratios. More detailed information on the database can be found in the study by Won et al. [43]. The dataset consists of sixteen input features previously identified as influential factors in determining the permanent deformation of unbound aggregates. These features encompass the number of cycles (N), deviatoric stress (σ_d), six gradation characteristics (C_u, C_c, D_{10}, D_{30}, D_{50}, and D_{60}), and a pair of compaction parameters (optimum moisture content, OMC and maximum dry unit weight, $\gamma_{d(max)}$). Furthermore, strength-related features derived from monotonic triaxial tests (friction angle, ϕ, secant friction angle, ϕ_{sec}, and cohesion, c) and three morphological features (angularity index, AI, surface texture index, STI, and flat and elongated ratio, FER) are included as input features. The output feature for this study is the permanent strain (ε_z). Table 1 provides statistical details regarding the input and output features of the unbound aggregates. All aggregate samples are well-graded, demonstrated by the coefficient of curvature (C_c), which approximately ranges between 1 and 3. As illustrated in Fig. 2, the number of cycles (N) exhibits a uniform distribution, whereas deviatoric stress (σ_d) and permanent strain (ε_z) display a positive skewed distribution. Fig. 3 demonstrates a weak correlation between each input feature and the target variable, ε_z. The internal friction angle (ϕ), maximum dry unit weight ($\gamma_{d(max)}$), and five gradation

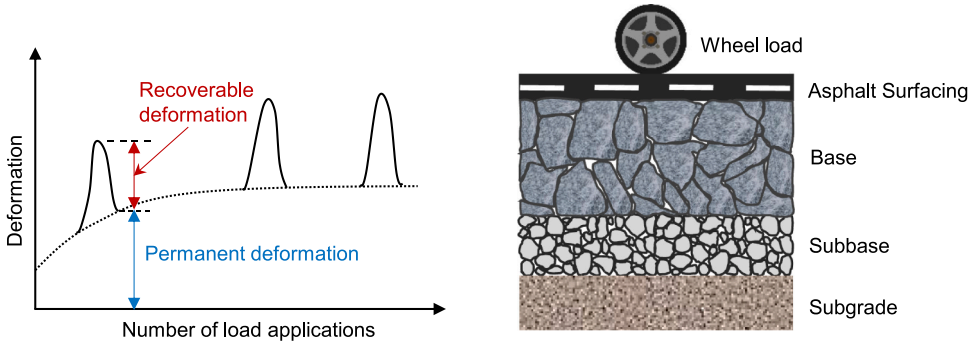

Fig. 1. Typical behavior of unbound aggregate materials under repeated loading.

Table 1
Range of unbound aggregate properties in the database.

Type	Variable	Min	Max	Mean
Input	N	1	10001	5001
	σ_d [kPa]	27.58	441.95	130.85
	C_u	24	152.42	83.96
	C_c	0.74	3.33	1.82
	D_{10} [mm]	0.07	0.46	0.13
	D_{30} [mm]	0.58	2.99	1.49
	D_{50} [mm]	3	9.64	6.17
	D_{60} [mm]	4.8	14.43	9.6
	OMC [%]	4.2	7.4	5.77
	$\gamma_{d(max)}$ [kN/m^3]	20.63	24.96	22.43
	ϕ [°]	23.6	54.4	42.74
	ϕ_{sec} [°]	42.4	66.9	53.57
	c [kPa]	0	139.4	39.1
	AI	405	558	458.13
	STI	1.69	2.75	2.15
	FER	1.86	2.83	2.4
Target	ε_z [%]	0	5.14	0.6

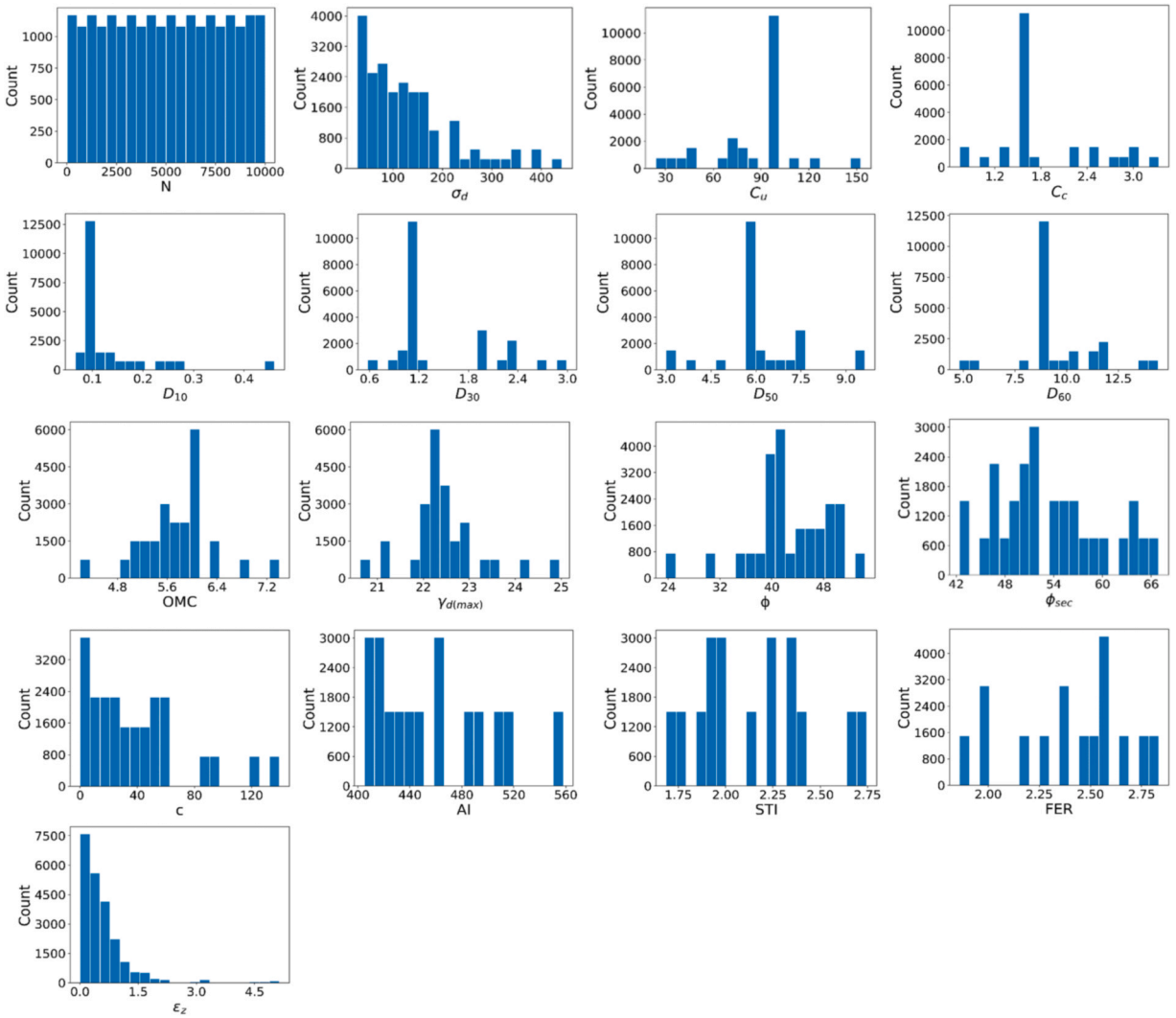

Fig. 2. Data distribution of input features and target ε_z.

properties (C_c, D_{10}, D_{30}, D_{50}, and D_{60}) are slightly negatively correlated with the permanent stain, while other features are slightly positively correlated with the permanent strain.

2.2. Data preprocessing

Data preprocessing plays a crucial role in data analysis as it involves examining and transforming raw data into a format that facilitates efficient analysis. Raw data often have quality and tidiness issues that can be rectified through preprocessing to ensure the accuracy and reliability of analytical results. Furthermore, preprocessing can reveal hidden relationships and patterns within the data. Notably, data preprocessing is an iterative process, and in this study, it comprises outlier detection and data scaling.

Outliers are the data points that deviate significantly from the rest of the sample they belong to, and they can have a negative impact on the results obtained from data analysis. Several methods, such as the z-score, modified z-score, and interquartile range methods, can be used to detect outliers. In this study, the interquartile range (IQR) method is employed to identify outliers, which are defined as the difference between 25 % and 75 % of each feature. The IQR method has been effectively used in previous studies to detect outliers in high-dimensional data [20,25]. Quartiles are markers that divide data into four equal segments. Fig. 4 shows a box plot that uses quartiles to demonstrate the data distribution. The lower and upper parts of the box indicate the 1st and 3rd quartiles (Q_1 and Q_3), corresponding to 25 % and 75 % of the data, respectively. Data points falling outside 1.5 times the IQR below Q_1 or 1.5 times the IQR above Q_3 are considered outliers. Note that, in this study, the outliers were not removed from the dataset. Instead, a unique method to effectively integrate these outliers, robust scaling, was utilized. Scaling is a critical component of ML, particularly when dealing with numerical data that encompasses different units or substantial variations in magnitude. Scaling the features is necessary for such scenarios. A prevalent approach is standard scaling, which transforms features into a standard normal distribution characterized by a mean of 0 and a standard deviation of 1 [19,34]. However, a significant limitation of this method is that outliers can negatively influence the mean or standard deviation. Conversely, this study uses robust scaling, which eliminates the median and scales the data based on the IQR, often resulting in improved performance when a dataset has significant outliers [34]. The equation for robust scaling can be expressed as follows:

$$x_{scaled} = \frac{x - median(x)}{Q_3(x) - Q_1(x)} \tag{1}$$

where x represents the vector of observations for a feature, median(x) denotes the median of the feature, and Q_3(x) and Q_1(x) refer to the third and first quartiles of the feature, respectively.

3. Machine learning techniques

3.1. Multiple machine learning algorithms

In this study, 10 ML regression algorithms were utilized, which included boosting, ensemble, and regularization models. Some of these algorithms have previously been used to predict the behavior of unbound aggregates under various loading conditions [18].

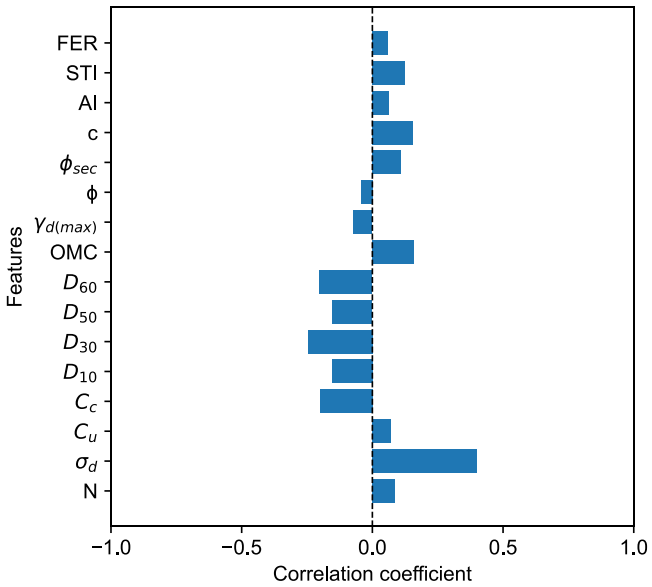

Fig. 3. Correlation of input features with the target ε_z.

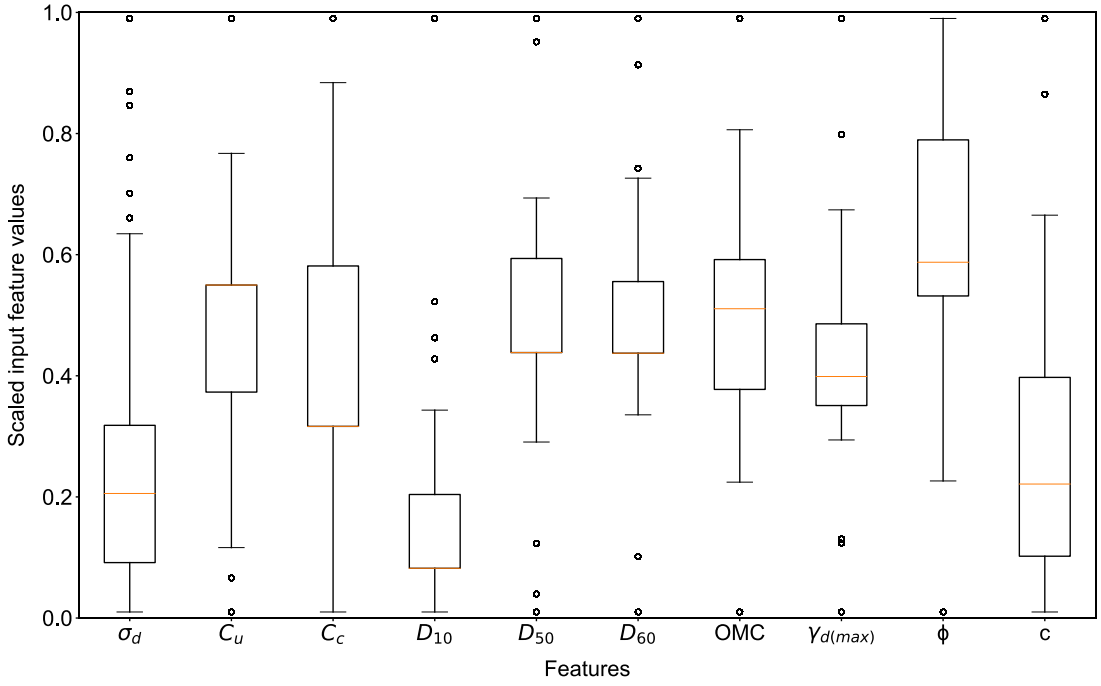

Fig. 4. Box plot of the scaled input feature values. Circles indicate outliers.

Extreme gradient boosting (XGB), light gradient boosting machine, gradient boosting regressor, and AdaBoost regressor employ boosting techniques to enhance model performance by iteratively improving the accuracy of weak learners. Conversely, extra trees and random forest regressors produce predictions by averaging the outputs of multiple decision trees through ensemble techniques. Moreover, the k-neighbors regressor produces predictions by taking into account the output of its k-nearest neighbors. Finally, Ridge and Bayesian ridge regressors employ regularization methods to minimize the complexity of the model and prevent overfitting.

ML algorithms estimate the complexity of a model by minimizing the values of loss functions. In addition, each algorithm contains hyperparameters, which should be set before initiating the learning process. The optimization of these predefined parameters is known as hyperparameter tuning [28]. This study employs a grid search algorithm to identify the best combination of hyperparameters for each algorithm. The grid search algorithm is commonly paired with cross-validation to prevent overfitting while maintaining model performance [30]. The hyperparameters tuned for each ML algorithm used in this study are summarized in Table 2.

3.2. Data splitting and cross-validation

The technique of train-test-validation split is widely used in ML to evaluate model performance on unseen data. The scaled input features are divided into training, testing, and validation subsets. The model is fitted using the training set, enabling the ML algorithm to identify underlying patterns. To optimize the performance of the model, the validation set is used to tune model hyperparameters,

Table 2
Machine learning hyperparameters.

ML algorithms	Hyperparameters
LGBM; XGBoost; Gradient Boosting; Random Forest	Learning rate Max depth Number of estimators
Extra trees regressor	Max depth Number of estimators
Decision Tree	Max depth Minimum sample split
KNN	Number of neighbors Distance metric
AdaBoost	Learning rate
Bayesian Ridge	Alpha (α) Lambda (λ)
Ridge	Alpha (α) fit intercept

such as learning rate and regularization strength. Finally, the test set is used to evaluate the final performance of the model on new data. In this study, an 80 % training/testing splitting ratio was adopted.

Evaluating the performance of ML models typically involves measuring accuracy on unseen data using the test set. However, this approach has limitations because model performance may depend considerably on a few observations within the test set and not all available data is used for training or testing. To overcome these limitations, k-fold cross-validation is employed, where the data is partitioned into k equally sized and nonoverlapping "folds." The model is trained on k−1 folds and the remaining fold is used for testing. This process is repeated k times until each fold has been used as a test set once. The value of k should be chosen carefully because a large value of k may result in high bias, whereas a small value may cause high variance. In this study, a k value of 10 was selected, consistent with recommendations from previous studies [6,26,38].

3.3. Model performance

A vital component of ML model development is the evaluation of its performance to determine the prediction error. This is typically achieved through the use of performance evaluation metrics, such as mean absolute error (MAE), mean average percentage error (MAPE), and root mean square error (RMSE). These metrics are defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{x_i y_i}{x_i}\right| \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \tag{4}$$

where y_i represents the prediction made by the ML algorithm, x_i corresponds to the measured value, and n represents the total number of data points. The MAE assigns equal weights to all errors, regardless of their magnitude. However, MAPE is often combined with other metrics because it can be sensitive to outliers and extreme values. Finally, RMSE is commonly employed when it is essential to penalize larger errors more than smaller ones.

Table 3 displays the performance of the 10 selected algorithms on the test set. Exceptional performance is shown by the extreme gradient boosting, extra trees, random forest, decision tree regressors, and light gradient boosting machine across all metrics. These five models produce a high coefficient of determination (R^2) of 0.99, indicating that they effectively capture the majority of patterns and relationships within the data. These results further suggest that boosting and ensemble techniques outperform single-learning ML methods in predicting the permanent strain. Consistent with the findings reported by Won et al. [43], the XGB algorithm performs best when all sixteen features are incorporated in the training and testing process. XGB is able to create accurate and robust models by integrating the predictions of multiple individual trees. The model iteratively adds decision trees to a model, with each new tree focused on correcting the errors of the previous trees. Early stopping approach is used to prevent overfitting, enhance generalization, and optimize computational efficiency. The small difference in RMSE between the training and test sets indicates the robustness of the model to new data. Therefore, in this study, the XGB algorithm was selected to determine the optimal feature subset and predict the permanent strains of unbound aggregate materials.

4. Modeling framework based on feature selection

4.1. Sequential backward selection (SBS)

The SBS algorithm is a heuristic search technique that automatically identifies a subset of features most relevant to a specific

Table 3
Performance metrics of various learning algorithms.

Model	MAE	MSE	RMSE	R^2	MAPE
Extreme gradient boosting	0.0043	0.0003	0.0165	0.9993	0.0145
Extra trees regressor	0.0017	0.0004	0.0168	0.9992	0.0077
Random forest regressor	0.0022	0.0005	0.0191	0.9988	0.0085
Decision tree regressor	0.0027	0.0006	0.0207	0.9987	0.0095
Light gradient boosting machine	0.0063	0.0007	0.0242	0.9984	0.0249
K-neighbors regressor	0.0028	0.0009	0.0271	0.9980	0.0125
Gradient boosting regressor	0.0454	0.005	0.0702	0.9886	0.1306
AdaBoost regressor	0.2246	0.0715	0.2671	0.8353	0.9689
Bayesian ridge	0.2698	0.2285	0.4777	0.4767	0.9779
Ridge regression	0.2694	0.2285	0.4777	0.4767	0.9746

MAE = mean absolute error, MSE = mean squared error, RMSE = root mean squared error, R^2 = coefficient of determination, and MAPE = mean average percentage error.

problem, aiding in the development of a more efficient model. The algorithm iteratively eliminates features with the least effect on the model performance, until yielding the desired number of features. The removal of insignificant features promotes model generalization to unseen data and improves computational efficiency. The working principle of the SBS algorithm is illustrated in Fig. 5. The selection process starts with all original features, with a size of 16 ($n = 16$). The first iteration generates all possible feature subsets of size $n-1$, or 15. For each subset, one feature is removed, and the model performance is evaluated using RMSE. The feature that is missing from the subset with the best score is removed. The second iteration starts with a size n of 15 and generates all possible feature subsets of size $n-1$, or 14. After the RMSE evaluation, the feature absent from the subset with the best score is eliminated. This process continues until only one feature remains. Ultimately, the subset with the best evaluation score is selected as the final feature set.

The proposed framework for feature selection-based modeling in estimating the permanent strain of unbound aggregates is displayed in Fig. 6. The framework consists of a database that comprises input and target features, which initially undergo preprocessing, including outlier detection and robust scaling. Subsequently, the scaled dataset is partitioned into training and test sets. Various ML algorithms are utilized to fit the training sets, and ten-fold cross-validation is carried out to enhance the performance of each model. The best-performing ML model is then determined and combined with the SBS algorithm to search for the optimal feature subsets. This combination is referred to as the SBS-based model. Thus, the feature selection-based ML model can predict the permanent strain of unbound aggregates with optimal feature subsets.

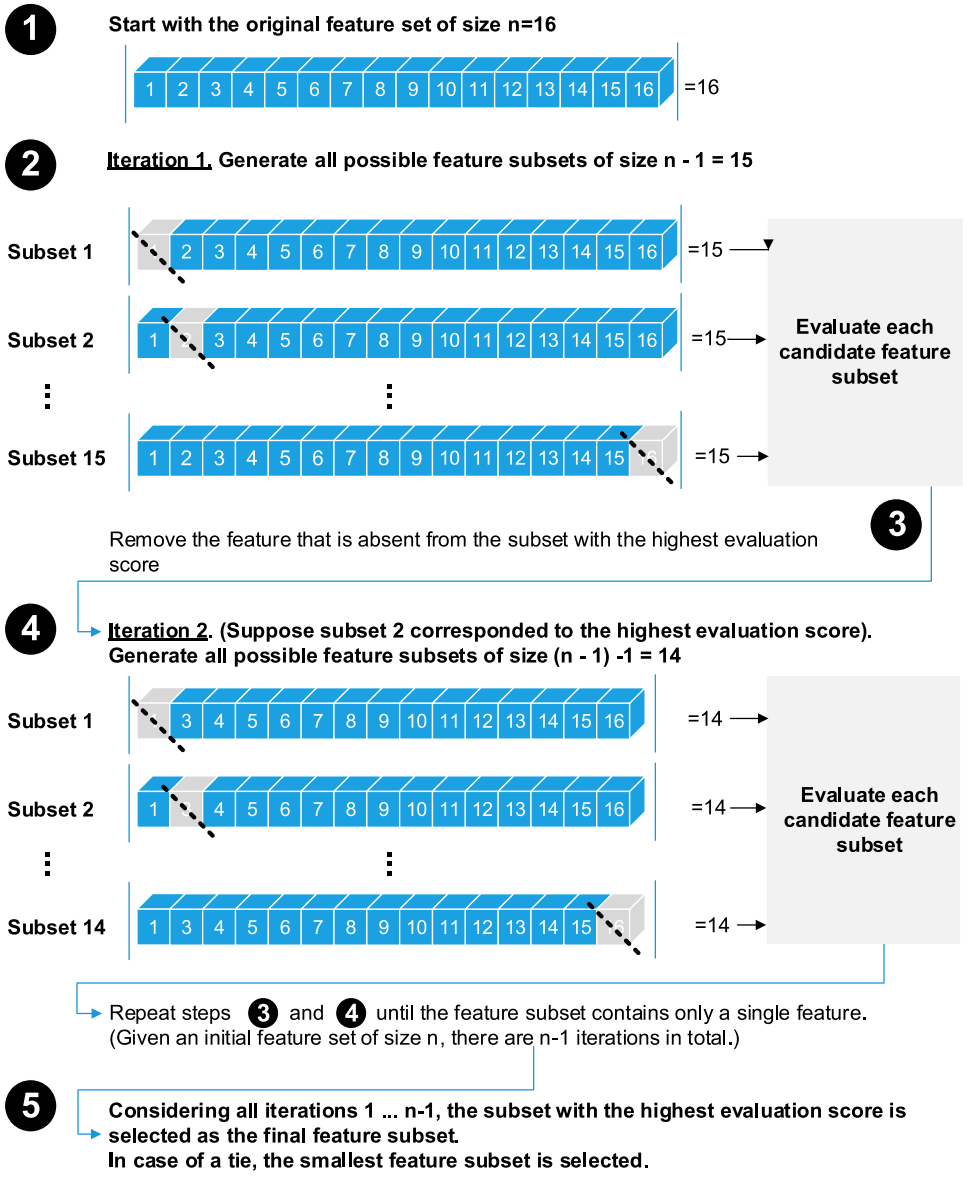

Fig. 5. Working principle of sequential backward selection.

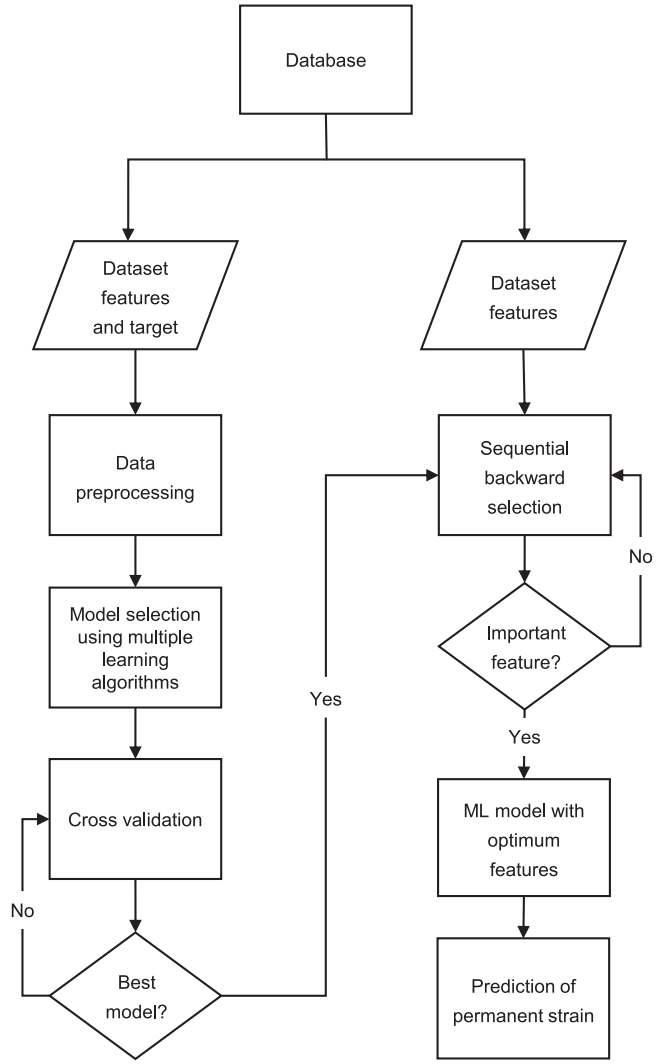

Fig. 6. Flowchart of the feature selection-based modeling for estimating permanent strain of unbound aggregates.

Table 4
Performance of the ML model using selected features through sequential backward selection.

Number of features	Feature	RMSE
16	N, σ_d, C_u, C_c, D_{10}, D_{30}, D_{50}, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, ϕ_{sec}, c, AI, STI, FER	0.0151
15	N, σ_d, C_u, C_c, D_{10}, D_{30}, D_{50}, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, ϕ_{sec}, c, AI, FER	0.0150
14	N, σ_d, C_u, C_c, D_{10}, D_{50}, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, ϕ_{sec}, c, AI, FER	0.0149
13	N, σ_d, C_u, C_c, D_{10}, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, ϕ_{sec}, c, AI, FER	0.0147
12	N, σ_d, C_u, C_c, D_{10}, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, ϕ_{sec}, AI, FER	0.0151
11	N, σ_d, C_u, C_c, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, ϕ_{sec}, AI, FER	0.0149
10	N, σ_d, C_u, C_c, D_{60}, OMC, $\gamma_{d(max)}$, ϕ, AI, FER	0.0149
9	N, σ_d, C_u, C_c, OMC, $\gamma_{d(max)}$, ϕ, AI, FER	0.0141
8	N, σ_d, C_c, OMC, $\gamma_{d(max)}$, ϕ, AI, FER	0.0147
7	N, σ_d, C_c, OMC, $\gamma_{d(max)}$, ϕ, FER	0.0144
6	N, σ_d, OMC, $\gamma_{d(max)}$, ϕ, FER	0.0138
5	N, σ_d, OMC, $\gamma_{d(max)}$, ϕ	0.0153
4	N, σ_d, OMC, $\gamma_{d(max)}$	0.0153
3	N, σ_d, OMC	0.0184
2	σ_d, OMC	0.1144
1	σ_d	0.1814

4.2. Sequential backward selection (SBS)-based model

The proposed SBS-based model employs RMSE as the evaluation metric to search for the optimal feature subset. The feature subsets at each iteration along with their corresponding evaluation scores are presented in Table 4. Fluctuations in RMSE values as the size of features changes are shown in Fig. 7. When the number of components decreases from 16 to 10, there is no remarkable change in the RMSE values. The RMSE varies when the number of components ranges from ten to six, and subsequently, the RMSE considerably increases as the number of features reduces from five to one. The feature subset with a size of six provides the best evaluation with a score of 0.0138. This feature subset includes σ_d, *OMC*, *N*, $\gamma_{d(max)}$, ϕ, and *FER*. The results of the SBS-based model indicate that deviatoric stress is the most crucial feature required for predicting permanent strain. This outcome is consistent with previous studies, which also demonstrated that higher deviatoric stress values are associated with an increase in permanent strain [7,36]. Additionally, Xiao et al. [44], reported that the number of load cycles, *N*, is a primary factor affecting the permanent strain of unbound aggregates, and compaction properties are also highly relevant. Moreover, it has been suggested that shear strength parameters, such as ϕ, considerably influence the permanent strain [10]. Lower ϕ values often indicate less resistance to permanent strain, resulting in substantial rutting. Therefore, using only the selected features, the long-term permanent strain of unbound aggregate materials can be predicted accurately.

4.3. Principal component analysis (PCA)-based model

Dealing with large datasets that have a high number of features can be challenging, making data exploration and visualization a difficult task. To overcome this challenge, PCA is often used as a method to significantly reduce dimensionality while preserving most of the information in the dataset. The main objective of PCA is to transform a large set of features into a smaller one while maintaining as much variance or statistical information as possible [22]. Consequently, a set of principal components is obtained that can be used to reduce the dimensionality of the dataset [46]. Different approaches can be used to determine the required number of principal components. Some studies have chosen the first principal components that cumulatively account for 70 % of the total variance [22]. Alternatively, the residual sum of squares or RMSE can be used to determine the number of principal components [19].

In this study, PCA was combined with XGB to determine the optimal number of principal components. PCA was applied to the 16 input variables used in the XGB model, and its performance was compared to that of the SBS-based model. Table 5 presents the variance explained by the 16 principal components, which represent the input features of the dataset. The first principal component corresponds to the direction in the data where the most significant variations in observations occur. As shown in Fig. 8, it was found that the individual contribution to variance decreases with an increasing number of components.

Table 6 displays the performance of the PCA-based model evaluated using RMSE across 16 iterations. Notably, using only two principal components, RMSE values as low as 0.029 can be obtained, indicating the efficacy of this dimensionality reduction method. However, it should be noted that the PCA does not provide feature subsets. Fig. 9 shows that a principal component of seven with an RMSE of 0.019 provides the best combination, which is still lower than that of the SBS-based model and uses only six different features.

4.4. Prediction accuracy

The SBS- and PCA-based models are used here to evaluate the prediction accuracy of permanent strain by modifying the test dataset. The SBS-based model was modified to have six features, while the PCA-based model transformed the original dataset into seven principal components. Fig. 10 demonstrates the performance of both models after predicting the permanent strain of the test sets. Both models provided excellent predictions; however, the SBS-based model outperformed the PCA-based model.

In Fig. 11, the permanent strain predicted by the SBS-based model is compared to the measured strain of the same aggregate material at two gradations under three different deviatoric stresses. For all specimens, the SBS-based model showed R^2 values higher than 0.924, regardless of the gradation and applied stress level. Especially for the low permanent strains, a slight improvement in performance was found compared to those obtained by Won et al. [43]. This could be attributed to the difference in selected hyperparameters [30]. Note that in the SBS-based model the optimal feature set and its size were determined according to the lowest RMSE values. On the other hand, in the previous study, a fixed number of features was used for each scenario [43]. Therefore, the SBS approach was found to be an effective method for optimizing feature subsets in ML models to predict the permanent strain of unbound aggregates.

5. Summary and conclusions

This paper proposed a machine learning framework for selecting optimal features to predict the permanent strain of unbound aggregates in pavements. The laboratory characterization of 15 crushed aggregates was compiled into a large database with 16 input features and one target variable. Preprocessing steps, including outlier detection and robust scaling, were undertaken to ensure the integrity of the dataset. The dataset was then split into training and test sets. Ten ML models, including ensemble, boosting, and single-learning algorithms, were evaluated using well-defined metrics. The XGB model outperformed the others and was paired with the SBS algorithm to identify optimal features for predicting the permanent strain of unbound aggregates. The SBS-based feature selection model effectively selected six crucial features. PCA was also conducted to reduce the dimensionality of the dataset, and the performance of the PCA-based model was compared to that of the SBS-based model.

The result showed that ensemble and boosting algorithms perform better than single-learning algorithms in predicting the long-

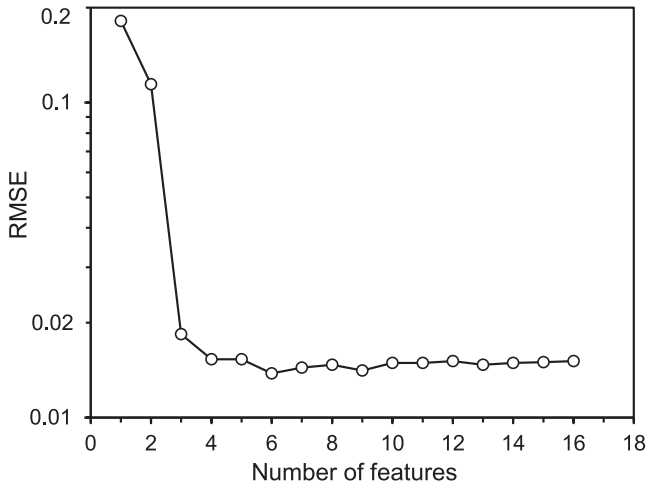

Fig. 7. RMSE values of SBS-based model with the number of features.

Table 5
Contribution rate and cumulative contribution of sixteen principal components.

Principal component	Contribution	Cumulative contribution
1	0.2744	0.2744
2	0.1716	0.446
3	0.1253	0.5713
4	0.1087	0.68
5	0.0796	0.7596
6	0.0625	0.8221
7	0.0622	0.8844
8	0.0369	0.9213
9	0.0284	0.9497
10	0.0265	0.9762
11	0.012	0.9881
12	0.0046	0.9928
13	0.0031	0.9959
14	0.0028	0.9987
15	0.0012	0.9998
16	0.0002	1

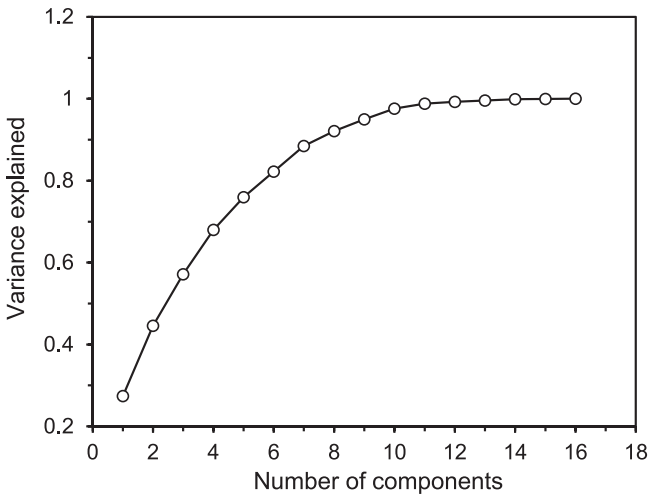

Fig. 8. Variance explained by the principal components.

Table 6
Performance metrics of principal component analysis.

Principal component	RMSE
1	0.384
2	0.029
3	0.024
4	0.025
5	0.024
6	0.022
7	0.019
8	0.020
9	0.021
10	0.021
11	0.022
12	0.022
13	0.022
14	0.022
15	0.023
16	0.022

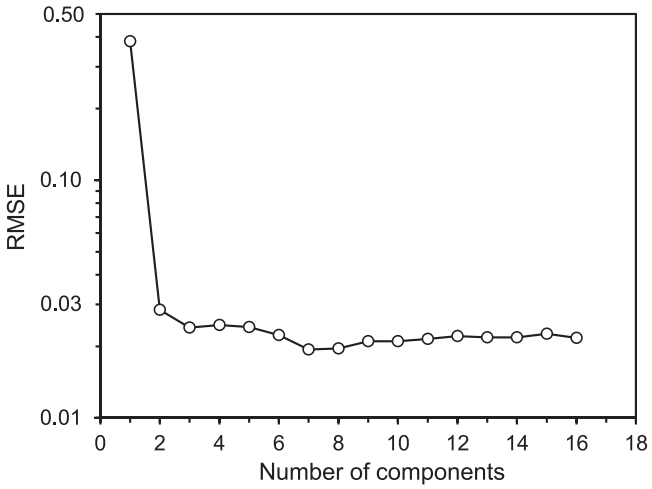

Fig. 9. RMSE values of PCA-based model with the number of principal components.

term permanent strain of unbound aggregates. Furthermore, the SBS-based model with six features outperformed the model with all 16 features, demonstrating the efficacy of feature selection in enhancing predictive capabilities while reducing complexity. Notably, the feature selection model exhibits generalization ability on test datasets, with R^2 values greater than 0.99. This suggests the reliability of the model in predicting permanent strain under diverse conditions. In addition, the SBS-based model provides clearer interpretation and superior prediction performance compared to dimension reduction techniques, such as PCA. The findings of this study have significant implications for geotechnical engineering and pavement design. By utilizing a limited number of crucial features, the proposed feature selection model offers a practical and efficient approach for accurately predicting permanent strain. This has the potential to streamline the prediction process and optimize pavement design for enhanced durability and performance.

Future research could explore the application of the proposed framework to a broader range of aggregate types and consider external factors that may influence permanent strain, such as environmental conditions and traffic loading patterns. Additionally, investigating the transferability of the model to different geographic locations and field-scale pavements would further validate its applicability and potential impact.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

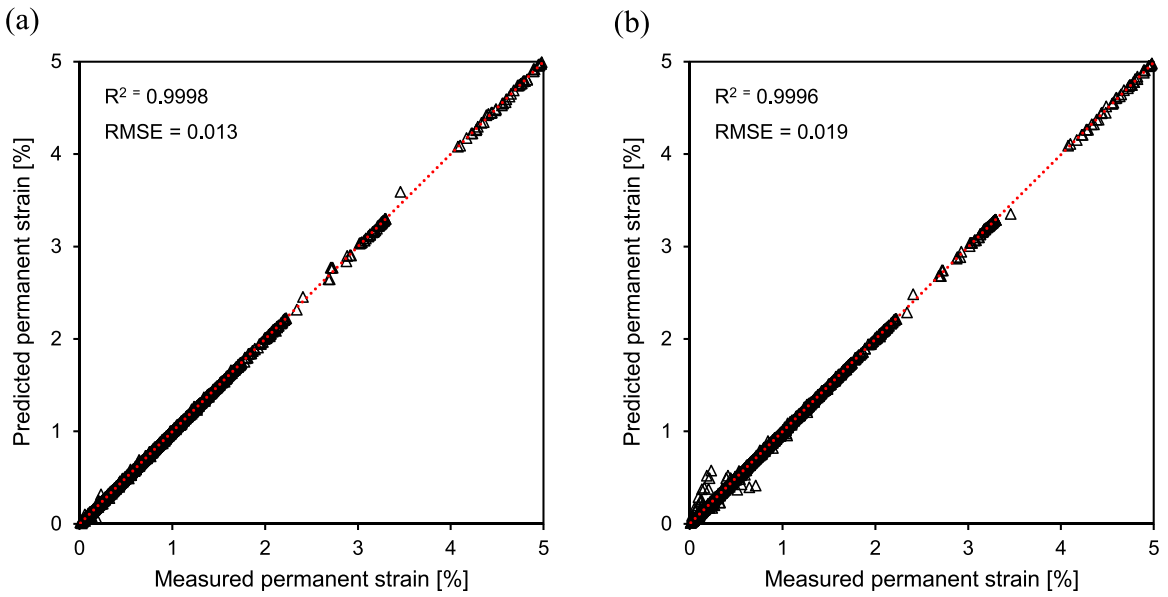

Fig. 10. Predicted and measured permanent strains for the test dataset (4518 data points) using two different ML models: (a) Sequential Backward Selection (SBS); (b) Principal Component Analysis (PCA).

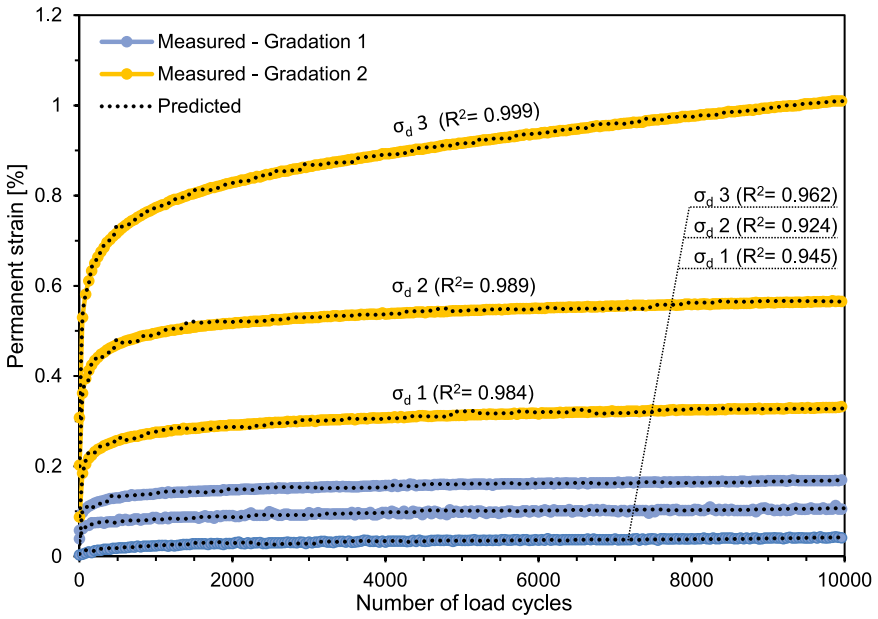

Fig. 11. Measured and modeled permanent strains of a sample under three different deviatoric stress conditions.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2020R1C1C1008925; NRF-2021R1A5A1032433).

References

[1] N.M. Ali, A.I.B. Farouk, S.I. Haruna, H. Alanazi, M. Adamu, Y.E. Ibrahim, Feature selection approach for failure mode detection of reinforced concrete bridge columns, Case Stud. Constr. Mater. 17 (2022), e01383.
[2] K.R. Aljanabi, O.M. AL-Azzawi, Neural network application in forecasting maximum wall deflection in homogenous clay, Int. J. Geo-Eng. 12 (1) (2021) 1–18.

[3] A. Alnedawi, R. Al-Ameri, K.P. Nepal, Neural network-based model for prediction of permanent deformation of unbound granular materials, J. Rock. Mech. Geotech. Eng. 11 (6) (2019) 1231–1242.
[4] M.N. Amin, W. Ahmad, K. Khan, S. Nazar, A.M.A. Arab, A.F. Deifalla, Evaluating the relevance of eggshell and glass powder for cement-based materials using machine learning and SHapley Additive exPlanations (SHAP) analysis, Case Stud. Constr. Mater. 19 (2023), e02278.
[5] M. Ashfaq, M. Iqbal, M.A. Khan, F.E. Jalal, M. Alzara, M. Hamad, A.M. Yosri, GEP tree-based computational AI approach to evaluate unconfined compression strength characteristics of Fly ash treated alkali contaminated soils, Case Stud. Constr. Mater. 17 (2022), e01446.
[6] D. Berrar, Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics 1–3, 2019, pp. 542–545.
[7] Y.-H. Byun, B. Feng, I.I.A. Qamhia, E. Tutumluer, Aggregate properties affecting shear strength and permanent deformation characteristics of unbound–base course materials, J. Mater. Civ. Eng. 32 (1) (2020).
[8] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: a new perspective, Neurocomputing 300 (2018) 70–79.
[9] S. Chen, C. Chen, T. Ma, C. Han, H. Luo, S. Wang, Y. Yang, Rapid extraction of pavement aggregate gradation based on point clouds using deep learning networks, Autom. Constr. 154 (2023), 105023.
[10] L.C. Chow, D. Mishra, E. Tutumluer, Framework for development of an improved unbound aggregate base rutting model for mechanistic-empirical pavement design, Transp. Res. Rec. 2401 (2014) 11–21.
[11] H.A. Dahish, M.S. Alfawzan, B.A. Tayeh, M.A. Abusogi, M. Bakri, Effect of inclusion of natural pozzolan and silica fume in cement - based mortars on the compressive strength utilizing artificial neural networks and support vector machine, Case Stud. Constr. Mater. 18 (2023), e02153.
[12] S. Demir, E.K. Sahin, An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost, Neural Comput. Appl. 35 (4) (2022) 3173–3190.
[13] S. Demir, E.K. Şahin, Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing, Environ. Earth Sci. 81 (18) (2022) 1–17.
[14] S. Demir, E.K. Sahin, Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost, Acta Geotech. (2023).
[15] J. Dong, N. Wang, H. Fang, R. Wu, C. Zheng, D. Ma, H. Hu, Automatic damage segmentation in pavement videos by fusing similar feature extraction siamese network (SFE-SNet) and pavement damage segmentation capsule network (PDS-CapsNet), Autom. Constr. 143 (2022), 104537.
[16] J. Hu, Data cleaning and feature selection for gravelly soil liquefaction, Soil Dyn. Earthq. Eng. 145 (2021), 106711.
[17] T. Hu, H. Zhang, J. Zhou, Machine learning-based model for recognizing the failure modes of FRP-strengthened RC beams in flexure, Case Stud. Constr. Mater. (2023) *18*.
[18] C.C. Ikeagwuani, Determination of Unbound Granular Material Resilient Modulus with MARS, PLSR, KNN and SVM, Int. J. Pavement Res. Technol. 15 (4) (2022) 803–820.
[19] James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. *An Introduction to Statistical Learning*.
[20] J. Jeong, E. Park, W.S. Han, K. Kim, S. Choung, I.M. Chung, Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends, J. Hydrol. 548 (2017) 135–144.
[21] N. Jibanchand, K.R. Devi, Application of ensemble learning in predicting shallow foundation settlement in cohesionless soil, Int. J. Geotech. Eng. (2023).
[22] I.T. Jollife, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2065) (2016).
[23] A. Kheirati, A. Golroo, Machine learning for developing a pavement condition index, Autom. Constr. 139 (2022), 104296.
[24] A. Kodsy, B. Ozturk, M. Iskander, Forecasting of pile plugging using machine learning, Acta Geotech. 18 (7) (2023) 3697–3714.
[25] O. Koren, M. Koren, O. Peretz, A procedure for anomaly detection and analysis, Eng. Appl. Artif. Intell. 117 (2023), 105503.
[26] Košir, A., Odić, A., Tkalčič, M., 2013. How to improve the statistical power of the 10-fold cross validation scheme in recommender systems. *ACM International Conference Proceeding Series*, 3–6.
[27] F. Lekarp, U. Isacsson, A. Dawson, State of the art. II: permanent strain response of unbound aggregates, J. Transp. Eng. 126 (1) (2000) 76–83.
[28] W. Li, X. Xiaoxue, L. Fu, Z. Yu, Application of improved grid search algorithm on SVM for classification of tumor gene, Int. J. Multimed. Ubiquitous Eng. 9 (11) (2014) 181–188.
[29] J. Li, T. Liu, X. Wang, J. Yu, Automated asphalt pavement damage rate detection based on optimized GA-CNN, Autom. Constr. 136 (2022), 104180.
[30] Liashchynskyi, P.B., Liashchynskyi, P., 2019. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. *ArXiv*.
[31] H.S. Marie, K. Abu el-hassan, E.M. Almetwally, M. A. El-Mandouh, Joint shear strength prediction of beam-column connections using machine learning via experimental results, Case Stud. Constr. Mater. 17 (2022), e01463.
[32] N. Micheletti, L. Foresti, S. Robert, M. Leuenberger, A. Pedrazzini, M. Jaboyedoff, M. Kanevski, Machine learning feature selection methods for landslide susceptibility mapping, Math. Geosci. 46 (1) (2014) 33–57.
[33] A.H. Naser, A.H. Badr, S.N. Henedy, K.A. Ostrowski, H. Imran, Application of multivariate adaptive regression splines (MARS) approach in prediction of compressive strength of eco-friendly concrete, Case Stud. Constr. Mater. 17 (2022), e01262.
[34] F. Pedregosa, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, G. Varoquaux, A. Gramfort, B. Thirion, V. Dubourg, A. Passos, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830.
[35] I.I.A. Qamhia, M. Moaveni, Y.H. Byun, B. Feng, E. Tutumluer, Implementation framework of the UIUC aggregate base rutting model, Int. J. Pavement Eng. 22 (10) (2021) 1305–1317.
[36] Rahman, M.S., Erlingsson, S., Ahmed, A., 2022. Modelling the permanent deformation of unbound granular materials in pavements.
[37] R. Ray, S.S. Choudhary, L.B. Roy, M.R. Kaloop, P. Samui, P.U. Kurup, J. Ahn, J.W. Hu, Reliability analysis of reinforced soil slope stability using GA-ANFIS, RFC, and GMDH soft computing techniques, Case Stud. Constr. Mater. 18 (2023), e01898.
[38] J.D. Rodríguez, A. Pérez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 569–575.
[39] F. Siddiqui, P. Sargent, G. Montague, The use of PCA and signal processing techniques for processing time-based construction settlement data of road embankments, Adv. Eng. Inform. 46 (2020), 101181.
[40] E. Soranzo, C. Guardiani, Y. Chen, Y. Wang, W. Wu, Convolutional neural networks prediction of the factor of safety of random layered slopes by the strength reduction method, Acta Geotech. (2022).
[41] M. Wang, G. Zhao, W. Liang, N. Wang, A comparative study on the development of hybrid SSA-RF and PSO-RF models for predicting the uniaxial compressive strength of rocks, Case Stud. Constr. Mater. (2023) *18*.
[42] Z. Weng, G. Ablat, D. Wu, C. Liu, F. Li, Y. Du, J. Cao, Rapid pavement aggregate gradation estimation based on 3D data using a multi-feature fusion network, Autom. Constr. 134 (2022), 104050.
[43] J. Won, E. Tutumluer, Y.-H. Byun, Predicting permanent strain accumulation of unbound aggregates using machine learning algorithms, Transp. Geotech. (2023), 101060.
[44] Y. Xiao, E. Tutumluer, D. Mishra, Performance evaluations of unbound aggregate permanent deformation models for various aggregate physical properties, Transp. Res. Rec. 2525 (1) (2015) 20–30.
[45] H. Yu, M. Mooney, Characterizing the as-encountered ground condition with tunnel boring machine data using semi-supervised learning, Comput. Geotech. 154 (2023), 105159.
[46] P. Zhang, A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Appl. Soft Comput. J. (2019) *85*.

[47] Z. Zhao, W. Duan, G. Cai, M. Wu, S. Liu, CPT-based fully probabilistic seismic liquefaction potential assessment to reduce uncertainty: integrating XGBoost algorithm with Bayesian theorem, Comput. Geotech. 149 (2022), 104868.
[48] W. Zhang, X. Gu, L. Hong, L. Han, L. Wang, Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges, Appl. Soft Comput. 136 (2023), 110066.
[49] W. Zhang, Y. Zhang, X. Gu, C. Wu, L. Han, Deep Learning and Applications. Application of Soft Computing, Machine Learning, Deep Learning and Optimizations in Geoengineering and Geoscience, Springer Singapore, 2022, pp. 41–45.