Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Accurate and Generalizable Soil Liquefaction Prediction Model Based on the CatBoost Algorithm

Jiazhi He
University of Jinan
Xianda Feng (✉ feng.xianda@hotmail.com)
University of Jinan https://orcid.org/0000-0002-2737-6046
Mr. Lu
University of Jinan

Research Article

Keywords: Geotechnical engineering, Geological disasters, Soil liquefaction prediction, Machine learning, CatBoost-Optuna model

Posted Date: June 6th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2767016/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Accurate prediction of soil liquefaction is important for preventing geological disasters. Soil liquefaction prediction models based on machine learning algorithms are efficient and accurate; however, the generalizability of some models is weak and they fail to achieve highly precise soil liquefaction predictions in certain areas, which limits the applicability of these models. Thus, a soil liquefaction prediction model was constructed using the CatBoost (CB) algorithm to support categorical features. The model was trained using standard liquefaction datasets from domestic and foreign sources and was optimized with Optuna hyperparameters. Additionally, the model was evaluated using five evaluation metrics and its performance was compared to that of other models that use multi-layer perceptron and support vector machine algorithms. Finally, the prediction capability of the model was verified by a case study. The experimental results demonstrated that the CB-based model generated more accurate soil liquefaction predictions than other comparison models and maintained their performance. Hence, the proposed model accurately predicts soil liquefaction and offers strong generalizability, demonstrating potential to contribute toward the prevention and control of soil liquefaction in engineering projects, and toward ensuring the safety and stability of structures built on or near liquefiable soils.

1. Introduction

Geological disasters can cause significant damage to buildings and bridges. Many of these disasters are caused by lateral displacement, ground settling, soil and water ejection, and foundation instabilities. Each of these causes are the result of soil liquefaction, which is a phenomenon that occurs when saturated soil suddenly loses its bearing capacity and even its strength under the influence of earthquakes and other vibrations, causing the soil to become a fluid. Hence, accurate prediction of soil liquefaction even with limited soil test data is an important topic in the field of geotechnical engineering.

Currently, the standard penetration, static penetration, and shear wave velocity test methods (Rahman and Siddiqua, 2017) are typically used to assess the liquefaction of saturated soil. In studies on the liquefaction prediction based on test data and soil liquefaction principles, machine learning algorithms have exhibited higher accuracies and efficiencies than numerical simulations (Ye et al., 2022) and empirical formulas (Bolton Seed et al., 1985; Zhang et al., 2022).

Researchers typically consider soil liquefaction prediction as a binary classification problem, with an outcome of either liquefaction or non-liquefaction. Various soil liquefaction prediction models are based on the analysis of liquefaction data collected from historical earthquakes, as well as experimental and stratum data, as shown in Table 1.

Table 1
Soil liquefaction prediction models based on machine learning algorithms.

Author	Main features	Algorithm	Data size
(Pan et al., 2008)	*(N1)60*, a_{max}, M_W, *FC%*	logistic regression	200
(Zhang et al.,2013)	a_{max}, σ_v, $\sigma'_{v,}q_{c1N,}CSR$	logistic regression	226
(Xiao et al., 2022)	Mw, *I, N,* d_w, d_s	logistic regression	159
(Chiru-Danzer et al., 2001)	*L,* a_{max}, *D50, FC%, (N1)60*	artificial neural network	443
(Chern et al., 2008)	Mw, σ_v, $\sigma'_{v,}q_{c1N,}a_{max}$	artificial neural network	466
(Fan, 2021)	Mw, *L,* a_{max}, d_s, *D50, FC%*	artificial neural network	485
(Hu et al., 2016)	Mw, *L,* a_{max}, *FC%,* d_s, d_w	Bayesian network	350
(Zhang et al., 2014)	*I, L,* d_w, d_s, *N, D50, Cu, CSR*	Bayesian network	30
(Peng et al., 2020)	Mw, *L, D50, Cu,* d_w, d_s, *N, CSR*	random forest	72
(Liu et al., 2021)	Mw, a_{max}, *Vs1,* d_s	random forest	225
(Mao et al., 2018)	*I,* d_w, *N, D50, Cu,* σ'_v, *CSR*	support vector machine	64
(Wang et al., 2019)	Mw, d_s, *L,* d_w, *N, Time*	support vector machine	40
(Li, 2020)	V_s, d_w, d_s, Mw, *CSR,* a_{max}	support vector machine	154

Note: Mw represents magnitude; *I* represents the intensity; *L* represents the earthquake epicenter distance; *N* represents the number of standard penetration tests; *(N1)60* represents the corrected number of standard penetration tests; a_{max} represents the earthquake peak level acceleration; d_w represents the depth of the groundwater; d_s represents the depth of the soil layer; *FC%* represents the fine content percentage; σ_v represents the vertical total stress; σ'_v represents the effective overburden stress; q_{c1N} represents the static penetration end resistance; V_s represents the corrected shear wave velocity; *CSR* represents the seismic shear stress ratio; *D50* represents the median grain size; C_u represents the coefficient of non-uniformity; *Time* represents the duration of the earthquake.

These models incorporate logistic regression, artificial neural network, Bayesian network, random forest, and support vector machine (SVM) algorithms. The logistic regression model has a simple structure and can efficiently calculate the probability of liquefaction, but it is sensitive to the characteristics of the dataset and consequently has difficulty handling unbalanced data. Artificial neural networks mimic the structure of the brain and can learn from incomplete or inaccurate liquefaction data, but as the complexity of the data increases, excessive fitting can lead to a weakening of the generalization of the model. Bayesian networks can model the hierarchical structure of the factors involved in earthquake liquefaction, resulting in more accurate and robust predictions, but this approach is computationally complex and is not efficient for very large datasets. Random forest models reduce the risk of overfitting and improve the prediction stability by combining multiple decision trees, but they are not effective for datasets with a small number of features. SVM models can handle nonlinear, high-dimensional, and small sample size problems but they are sensitive to parameter requirements and occasionally have difficulty finding appropriate kernel functions.

In contrast, the CatBoost (CB) algorithm is an ensemble machine learning algorithm that avoids overfitting and can robustly process large amounts of data. Additionally, its distributed multi-core operation facilitates the parallel processing of features, which increases the training efficiency. Therefore, in this study, a soil liquefaction model was developed based on the CB algorithm, and its generalization was optimized using multiple feature selection techniques, algorithm comparisons, and performance evaluations. The practical performance of the model was evaluated using specific case study data. The results of this study provide new insights into optimal methods for enhancing the generalization of soil liquefaction prediction models based on machine learning algorithms.

The rest of this paper is organized as follows. In Section 2, the dataset selected for this study is described. Section 3 explains the theory on which the CB algorithm and Optuna hyperparameter optimization are based. In Section 4, the training of the model is detailed. Section 5 discusses the results of the case study. Finally, Section 6 draws the conclusions of this study.

2. Dataset selection

This study utilized over 20 liquefaction datasets collected and categorized by Cetin et al. (2018) for earthquakes that occurred between 1944 and 1995; the data were primarily associated with locations in the United States, Japan, Argentina, China, and the Philippines. Out of the total 208 samples, 113 pertained to liquefaction, and 95 did not. The data contained widely adopted soil liquefaction effect indicators, including the corrected number of standard penetration tests ($(N1)60$), seismic shear stress ratio (CSR), earthquake magnitude, critical depth (m), vertical total stress ($\sigma_{v;}$ (kPa)), and peak level acceleration ($a_{max;}$ (g)), as presented in Table 2.

Table 2
Ranges of widely adopted soil liquefaction impact indicators for previous earthquakes.

Earthquake	Critical depth (m)	σ_v (kPa)	a_{max} (g)	*CSR*	*(N1)60*	Liquefied?	
						No	Yes
1944 Tohnankai M = 8.0	2–4.3	33.75–70.08	0.2–0.2	0.15–0.21	2.2–8.9	0	3
1948 Fukui M = 7.3	2.6–8	43.81–148.04	0.35–0.4	0.27–0.37	6.5–20.6	0	2
1964 Niigata M = 7.5	3.3–11.5	56.5–222.06	0.09–0.18	0.09–0.2	6.6–42	4	6
1968 Tokachioki M = 7.9	2.4–5.5	108.66–108.66	0.2–0.23	0.21–0.25	6.8–38.4	2	3
1971 San Fernando M_w = 6.6	5.4–6.2	45.84–104.64	0.45–0.45	0.28–0.3	3.7–7.9	0	2
1975 Haicheng M_s = 7.3	7–8	94.38–110.25	0.13–0.2	0.13–0.2	7.4–14	0	3
1976 Guatemala M = 7.5	9.1–10.2	129.89–148.79	0.14–0.14	0.12–0.13	4.7–14.3	1	1
1976 Tangshan M_s = 7.8	3.5–5.5	122.4–126.73	0.13–0.5	0.13–0.4	7.9–33.1	2	5
1977 Argentina M = 7.4	2.4–11.7	57.68–106.14	0.2–0.2	0.12–0.18	5.3–13.8	2	3
1978 Miyagiken-Oki M = 6.7	2.4–5.9	45.12–209.76	0.1–0.14	0.09–0.15	2.7–18.7	13	1
1978 Miyagiken-Oki M = 7.4	2.4–7.5	42.88–105.6	0.2–0.32	0.16–0.38	2.7–26	6	14
1979 Imperial Valley M_L = 6.6	1.1–4.3	42.88–146.61	0.13–0.51	0.09–0.41	3.5–46.5	4	4
1980 Mid-Chiba M = 6.1	5.5–14.5	17.75–72.24	0.08–0.08	0.05–0.07	3.7–8.9	2	0
1981 Westmorland M_L = 5.6	1.1–4.3	103.94–274.02	0.16–0.23	0.12–0.23	4–19.9	4	3
1983 Nihonkai-Chubu M = 7.1	3.3–9.3	67.68–86.4	0.12–0.15	0.13–0.15	7.9–16.9	2	1
1983 Nihonkai-Chubu M = 7.7	2.4–10.5	17.75–72.24	0.05–0.28	0.05–0.31	5.2–24	3	12

Earthquake	Critical depth (m)	σ_v (kPa)	a_{max} (g)	*CSR*	*(N1)60*	Liquefied?	
						No	Yes
1987 Elmore Ranch M_w = 6.2	4.3–4.7	49.44–49.44	0.09–0.13	0.08–0.11	6.2–11.3	2	0
1987 Superstition Hills M_w = 6.6	4.3–4.7	56.5–171.65	0.2–0.2	0.18–0.19	6.2–11.3	1	1
1987 Superstition Hills M_w = 6.7	1.1–4.3	39.09–196.11	0.13–0.19	0.11–0.23	3.5–46.5	8	0
1989 Loma Prieta M_w = 7	1.8–8.5	72.24–86.4	0.14–0.46	0.08–0.35	3.5–43.8	7	17
1990 Luzon M_w = 7.6	5–7.3	72.24–86.4	0.25–0.25	0.23–0.23	13.9–25.8	1	1
1993 Kushiro-Oki M_w = 8	3.6–10.8	17.75–76.56	0.4–0.4	0.34–0.4	16.5–29.8	1	2
1994 0rthridge M_w = 6.7	6.3–9	31.5–160.24	0.4–0.69	0.28–0.37	10.8–19	0	3
1995 Hyogoken-Nambu M_L = 7.2	2.5–13	92.01–133.66	0.25–0.7	0.23–0.63	5.6–65.5	30	26

A portion of the dataset was obtained from a single standard penetration test borehole, and the rest was obtained from dense standard penetration test boreholes; thus, the sample information was nonuniform. Thus, borehole data from the same location were assigned to a single historical earthquake case, and these borehole data were combined with the stratum information, reducing the uncertainty of the actual measured number of penetrations (*N1*).

Furthermore, errors in the testing methods resulted in high uncertainties for each factor owing to vast differences in the associated location and collection date of these liquefaction data. Therefore, *N1(60)* was based on a weighted average of *N1* values for the strata, which was corrected for the effective normal stress, hammer energy, equipment rod length, equipment sampler, and drill hole diameter. This value was used as the primary feature for selecting features from the original dataset.

3. Algorithm theory

3.1 CatBoost (CB) algorithm

CB (Prokhorenkova et al., 2019) is an unbiased boosting algorithm that supports categorical features and a high-performance machine learning algorithm that has been evolved from gradient boosting. The principle of gradient boosting, as shown in Fig. 1, involves performing gradient descent on the loss function in the function space, combining weak learners according to the computed loss function value of the model, and iteratively constructing a strong learner. Gradient boosting is the primary method used for solving problems with heterogeneous features, noisy data, and complex dependencies. On this basis, CB preprocesses the categorical features during training and utilizes a rank-boosting strategy to solve the gradient bias and prediction offset problems that are typical in gradient boosting decision trees. In addition, it employs a tree structure model as the base learner. Moreover, it prevents overfitting and improves the generalization and prediction speed of the model by calculating the algorithms for leaf nodes.

The principle primarily responsible for CB's preprocessing of categorical features is ordered target encoding, which randomly orders the dataset and subsequently uses only the objects placed before the current object to calculate the numerical conversion of the categorical feature. If the *i*-th feature of the *k*-th sample is a categorical feature, the conversion formula is expressed as

$$x_k^i = \frac{\sum_{x_j \in D_k} \{x_k^i = x_j^i\} \times y_j + a \times p}{\sum_{x_j \in D_k} \{x_k^i = x_j^i\} + a}$$

1

,

where *D*$_k$ refers to the dataset prior to the *k*-th sample in the random ordering, $\left\{x_k^i = x_j^i\right\} = 1$ when x_k^i and x_j^i belong to the same category and $\left\{x_k^i = x_j^i\right\} = 0$ when they belong to different categories, *p* is the added prior item, and *a* is typically a weighting coefficient greater than 0.

The amount of information contained in the categorical features has a significant effect on the final performance of the model. CB stores the categorical features used in the model in groups; however, when the number of categorical features becomes extremely large, the final size of the model can increase substantially. Therefore, the storage size of a specific feature depends on the number of values adopted. By splitting the tree model in CB, the final size of the model can be reduced, potential weight of the categorical features in the final model can be estimated, and best split can be chosen. When choosing the split, all scores change according to

$$s^{new} = s^{old} \cdot (1 + \frac{u}{U})^M$$

2

,

where s^{new} is the new score obtained by splitting the categorical feature or combination feature, s^{old} is the old score of the feature split, u is the number of feature values, U is the maximum value of u among all values of the features and combinations, and M is the model size coefficient. All calculated scores are compared, and the split with the best score is selected.

3.2 Optuna hyperparameter optimization

The hyperparameter selection utilizes Optuna, which is an automated hyperparameter optimization framework with a high modularity; it enables the dynamic construction of the search space for hyperparameters. Throughout the model training process, Optuna acts as a pruner, observing intermediate results and halting unpromising trials that optimize the selection of associated tree model parameters.

4. Model training

4.1 Training process

In this study, Python 3.8, including packages such as sklearn, pandas, numpy, and matplotlib, was used as the operating environment for training the developed liquefaction prediction model. The procedure used for building the prediction model is shown in Fig. 2. First, the original data were processed, and the categorical features were converted into numerical features using CB. The impact weights of each feature value were determined by embedding the algorithm, and feature selection was conducted to obtain an appropriate dataset. Subsequently, the dataset was randomly divided into training and test sets using a ratio of 7:3. The CB hyperparameters were optimized using Optuna and cross-validation. Finally, the best parameter combination was determined via training on and the prediction of repeatedly divided datasets, resulting in an earthquake liquefaction prediction model.

4.2 Data processing

The original data had 31 features, out of which 28 were numerical and three were categorical features; the correlations are shown in Fig. 3. The original data were initially imported into the CB algorithm, and the categorical features were converted into numerical features to calculate the impact weights of each feature value, as shown in Fig. 4. Based on these values, redundant features with a high correlation in the data and features with low impact weights were processed.

Based on the criteria obtained from numerous soil liquefaction studies, the dataset used in this study was composed of 12 features: the corrected number of standard penetration tests *(N1)60*, seismic shear stress ratio (*CSR*), earthquake magnitude, critical depth of liquefaction, underground water depth, vertical total stress (σ_v), earthquake peak level acceleration (a_{max}), corrected shear wave velocity (V_s), median grain size (*D50*), fine content percentage (*FC%*), and data class. Feature selection was performed using the correlation analysis and cross-validation comparisons by employing the same model with different feature combinations.

The processed dataset was calculated for feature importance weights using the CB algorithm, as shown in Fig. 3. The corrected number of standard penetration tests (*(N1)60*) is a key factor for model prediction, and the results in the figure conform to the basic principles of soil liquefaction statistics.

4.3 Experimental results

To evaluate the liquefaction prediction model, the following five evaluation metrics were applied: accuracy (the ratio of the number of correctly-classified samples in the test set to the total number of test samples; represents the overall prediction performance); precision (the ratio of the number of correctly-predicted liquefied samples to the number of samples that were predicted to liquefy; indicates the significance of the prediction - if it is poor, the misjudgment of the prediction results may yield excessively conservative design); recall (the ratio of the number of correctly predicted liquefied samples to the number of actual liquefied samples; determines the conservatism of the prediction model as a measure of safety in practical projects); F1-score (the harmonic mean of the precision and recall); and area under the ROC curve (AUC; a value between 0.85 and 0.95 that generally indicates an excellent prediction performance). The ROC curve can clearly show the effect of any threshold on the generalization performance of a learner.

The hyperparameter tuning process employed the Optuna framework for maximizing the AUC of the soil liquefaction prediction model, which optimized its generalization. The trained liquefaction prediction model based on the CB algorithm was compared with the multi-layer perceptron (MLP) and SVM algorithms, which are typical machine learning algorithms, after undergoing similar processing methods.

To verify the stability of the model and avoid random experimental errors, ten experiments, in which the random numbers were adjusted to divide the training and test sets with different ratios, were repeated. Table 3 shows the mean and standard deviation (in parentheses) of the evaluation indices of the test set for the three machine learning algorithms tested in the repeated experiments. The box plot in Fig. 5 presents the distribution of the obtained results; the lines within the boxes represents the medians of the classifiers; the upper and lower edges of the boxes represent the upper and lower quartiles, respectively; the upper and lower margins represent the upper and lower bounds of the data, respectively; and the diamond points represent outliers. The CB algorithm evidently yielded a higher accuracy and stability, with conservative prediction results.

Table 3
Evaluation indices for the prediction models.

Classifier	Accuracy	Precision	Recall	F1-score
CB	0.866 (0.071)	0.869 (0.091)	0.878 (0.121)	0.866 (0.077)
MLP	0.813 (0.098)	0.840 (0.127)	0.762 (0.132)	0.833 (0.093)
SVM	0.823 (0.041)	0.816 (0.100)	0.871 (0.081)	0.835 (0.048)

The ROC curve in Fig. 6 is a line chart, where the x- and y-axis show the liquefaction false-positive rate and probability of correctly predicting liquefaction, respectively. The area under each line, or the AUC of each model, was greater than 0.9, indicating that the predictive models exhibited a favorable generalization.

5. Prediction case study

On September 21, 1999, an earthquake magnitude of 7.6 with a source depth of 7.0 km occurred along the Chelungpu fault in Chi-Chi, Nantou county, Taiwan. Frequent aftershocks occurred throughout the day; the largest earthquake had a magnitude of 6.8, which occurred less than 1 h after the mainshock. The following morning, another aftershock with a magnitude of 6.8 occurred. The earthquake caused widespread damage, which was accompanied by soil liquefaction. Hwang and Yang (2001) conducted an investigation, and collected 232 sample datasets on the geology of the liquefied and non-liquefied sites before and after the earthquake. Subsequently, these data were processed and used to validate subsequent predictions and analyses.

The processing of this case study dataset primarily involved selecting the corresponding feature values after performing the same calculations and filling in the missing values with the minimum value. Subsequently, the processed dataset was analyzed using a pretrained CB-Optuna prediction model, and the results were compared to those of the MLP and SVM prediction models. Table 4 displays the evaluation indices, which were obtained according to the prediction results.

Table 4
Evaluation indices obtained for the prediction models in the Chi-Chi earthquake dataset.

Classifier	Accuracy	Precision	Recall	F1-score
CB	0.888	0.836	0.944	0.887
MLP	0.806	0.736	0.907	0.813
SVM	0.759	0.691	0.870	0.770

The results of soil liquefaction using the three models, independent of the training set, are presented in Table 4. Although the results are slightly lower than those used for training the models, as presented in Table 3, employing completed models to directly predict a set of data is more aligned with the actual usage. In addition, only the evaluation indices of the prediction results for the CB model can maintain a high score, whereas the accuracy, precision, and F1-score of the comparison models are slightly decreased, further proving the universality of the proposed CB model.

6. Conclusions

In this study, 208 typical standard penetration test data samples were processed using the CB algorithm, and a prediction model was built with the goal of achieving a strong generalization. Subsequently, its performance was compared with that of two typical machine learning algorithms in terms of five evaluation metrics.

The CB algorithm was selected out of all the ensemble algorithms owing to its overfitting avoidance, high-quality earthquake datasets were collected and organized, the corrected standard was incorporated into the corrected number of standard penetration tests (*(N1)60*) as the main influencing factor, and AUC was used as the Optuna hyperparameter target. These actions all contributed to achieve a strong generalization of the model.

Many factors affect soil liquefaction, and they have complex multidimensional and nonlinear relationships. When predicting cases not used in the training dataset, the CB model proved its strong generalization with accurate and stable predictions. Therefore, the proposed model is widely applicable.

Indeed, the accuracy and universality advantages of the proposed soil liquefaction prediction model based on the CB algorithm effectively help predict and analyze actual engineering projects to prevent geological disasters caused by soil liquefaction and prevent or mitigate uneven settlement of buildings.

The future study will be aimed to improve various aspects of the soil liquefaction prediction model, collect more soil liquefaction feature data, improve the quality of datasets, and integrate multiple models to achieve an advanced liquefaction prediction model.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Funding This work was financially supported by National Natural Science Foundation of China (51809115); Shandong Natural Science Foundation (ZR2019QEE003)

References

1. Bolton Seed H, Tokimatsu K, Harder LF, Chung RM (1985) Influence of SPT procedures in soil liquefaction resistance evaluations. *J. Geotech. Eng.* 111(12):1425-1445. http://doi.org/10.1061/(ASCE)0733-9410(1985)111:12(1425)
2. Cetin KO, Seed RB, Kayen RE, Moss RES, Bilge HT, Ilgac M, Chowdhury K (2018) SPT-based probabilistic and deterministic assessment of seismic soil liquefaction triggering hazard. *Soil Dynamics & Earthquake Engineering* 115:698-709. http://doi.org/10.1016/j.soildyn.2018.09.012
3. Chern S-G, Lee C-Y, Wang C-C (2008) CPT-BASED liquefaction assessment by using fuzzy-neural network. *J. Mar. Sci. Technol.* 16(2):6. http://doi.org/10.51400/2709-6998.2024
4. Chiru-Danzer M, Juang, CH, Christopher RA, Suber J (2001) Estimation of Liquefaction-Induced Horizontal Displacements Using Artificial Neural Networks. *Can. Geotech. J.* 38(1):200–207. http://doi.org/10.1139/t00-087

5. Fan KX (2021) Prediction of earthquake liquefaction displacement based on SGO-RBF neural network. Institute of Seismology, China Earthquake Administration. http://doi.org/10.27055/d.cnki.ggdzy.2021.000016
6. Hu JL, Tang XW, Qiu JN (2016) Prediction of probability of seismic-induced liquefaction based on Bayesian network. *Rock Soil Mech.* 37(6):1745–1752. http://doi.org/10.16285/j.rsm.2016.06.027
7. Hwang JH, Yang CW (2001) Verification of critical cyclic strength curve by Taiwan Chi-Chi earthquake data. *Soil Dynamics & Earthquake Engineering* 21(3):237-257. http://doi.org/10.1016/S0267-7261(01)00002-1
8. Li BY (2020) Study on sand soil seismic liquefaction prediction based on shear wave velocity and support vector machine. *MS Thesis, Jilin University of Architecture & Civil Engineering, Jilin, Jilin Province, China*. http://doi.org/10.27714/d.cnki.gjljs.2020.000108
9. Liu L, Zhang S, Yao X, Gao H, Wang Z, Shen Z (2021) Liquefaction evaluation based on shear wave velocity using random forest. *Adv. Civ. Eng. Mater.* 2021:1-9. http://doi.org/10.1155/2021/3230343
10. Mao ZY, Huang CJ, Lu SC (2018) Seismic liquefaction prediction model based on PSO-SVM. *China Saf. Sci. J.* 28(03):25-30. http://doi.org/10.16265/j.cnki.issn1003-3033.2018.03.005
11. Pan JP, Kong XJ, Zou DG (2008) Evaluation of sand soil liquefaction probability based on logistic regression model. *Rock Soil Mech.* 09:2567-2571. http://doi.org/10.16285/j.rsm.2008.09.050
12. Peng LY, Xie HT, Feng WD (2020) Prediction method for sand soil liquefaction based on random forest algorithm. *Geophys. Geochem. Explor.* 44. http://doi.org/10.11720/wtyht.2020.1501
13. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2019) CatBoost: unbiased boosting with categorical features. arXiv. http://doi.org/10.48550/arXiv.1706.09516
14. Rahman Md Z, Siddiqua S (2017) Evaluation of Liquefaction-Resistance of Soils Using Standard Penetration Test, Cone Penetration Test, and Shear-Wave Velocity Data for Dhaka, Chittagong, and Sylhet Cities in Bangladesh. *Environ. Earth Sci.* 76(5):207. http://doi.org/10.1007/s12665-017-6533-9
15. Wang S, Yu S, Li SK, Yuan Y (2019) Study on sand soil liquefaction prediction method based on RS-PCA-GA-SVM. *Journal of Earthquake Engineering* 41(2):445-453. http://doi.org/10.3969/j.issn.1000-0844.2019.02.445
16. Xiao SH, Cheng XJ, Wang HA, Zhang J (2022) A probabilistic method for discriminating sand soil liquefaction based on standard penetration test. *Journal of Civil, Environmental & Architectural Engineering* 44(5):87-97. http://doi.org/10.11835/j.issn.2096-6717.2021.048
17. Ye B, Song SC, Ni XQ (2022) Discrete element simulation of the impact of sampling methods on the liquefaction mechanics properties of sand soil. *Journal of Tongji University (Natural Science Edition)* 50(7):998-1008. http://doi.org/10.11908/j.issn.0253-374x.21078
18. Zhang ZS, Chen JP, Chen K, Cui C (2014) Bayesian discrimination model for predicting sand soil earthquake liquefaction and its application. *J. Guilin Univ. Technol.* 34(01):63-67. http://doi.org/10.3969/j.issn.1674-9057.2014.01.010
19. Zhang SY, Li ZY, Yuan XM (2022) A new method for liquefaction discrimination based on static probing tests. *Rock Soil Mech.* 06:1-11. http://doi.org/10.16285/j.rsm.2021.1524

20. Zhang J, Zhang LM, Huang HW (2013) Evaluation of Generalized Linear Models for Soil Liquefaction Probability Prediction. *Environ. Earth Sci.* 68(7):1925-1933. http://doi.org/10.1007/s12665-012-1880-z

Tables

Table 1. Soil liquefaction prediction models based on machine learning algorithms.

Author	Main features	Algorithm	Data size
(Pan et al., 2008)	*(N1)60,* a_{max}, M_W, *FC%*	logistic regression	200
(Zhang et al.,2013)	a_{max}, σ_v, σ'_v, q_{c1N}, *CSR*	logistic regression	226
(Xiao et al., 2022)	Mw, *I, N,* d_w, d_s	logistic regression	159
(Chiru-Danzer et al., 2001)	*L,* a_{max}, *D50, FC%, (N1)60*	artificial neural network	443
(Chern et al., 2008)	Mw, σ_v, σ'_v, q_{c1N}, a_{max}	artificial neural network	466
(Fan, 2021)	Mw, *L,* a_{max}, d_s, *D50, FC%*	artificial neural network	485
(Hu et al., 2016)	Mw, *L,* a_{max}, *FC%,* d_s, d_w	Bayesian network	350
(Zhang et al., 2014)	*I, L,* d_w, d_s, *N, D50, Cu, CSR*	Bayesian network	30
(Peng et al., 2020)	Mw, *L, D50, Cu,* d_w, d_s, *N, CSR*	random forest	72
(Liu et al., 2021)	Mw, a_{max}, *Vs1,* d_s	random forest	225
(Mao et al., 2018)	*I,* d_w, *N, D50, Cu,* σ'_v, *CSR*	support vector machine	64
(Wang et al., 2019)	Mw, d_s, *L,* d_w, *N, Time*	support vector machine	40
(Li, 2020)	V_s, d_w, d_s, Mw, *CSR,* a_{max}	support vector machine	154

Note: Mw represents magnitude; *I* represents the intensity; *L* represents the earthquake epicenter distance; *N* represents the number of standard penetration tests; *(N1)60* represents the corrected number of standard penetration tests; a_{max} represents the earthquake peak level acceleration; d_w represents the depth of the groundwater; d_s represents the depth of the soil layer; *FC%* represents the fine content percentage; σ_v represents the vertical total stress; σ'_v represents the effective overburden stress; q_{c1N} represents the static penetration end resistance; V_s represents the corrected shear wave velocity; *CSR* represents the seismic shear stress ratio; *D50* represents the median grain size; C_u represents the coefficient of non-uniformity; *Time* represents the duration of the earthquake.

Table 2. Ranges of widely adopted soil liquefaction impact indicators for previous earthquakes.

Earthquake	Critical depth (m)	σ_v (kPa)	a_{max} (g)	*CSR*	*(N1)60*	Liquefied?	
						No	Yes
1944 Tohnankai M = 8.0	2–4.3	33.75–70.08	0.2–0.2	0.15–0.21	2.2–8.9	0	3
1948 Fukui M = 7.3	2.6–8	43.81–148.04	0.35–0.4	0.27–0.37	6.5–20.6	0	2
1964 Niigata M = 7.5	3.3–11.5	56.5–222.06	0.09–0.18	0.09–0.2	6.6–42	4	6
1968 Tokachioki M = 7.9	2.4–5.5	108.66–108.66	0.2–0.23	0.21–0.25	6.8–38.4	2	3
1971 San Fernando M_w = 6.6	5.4–6.2	45.84–104.64	0.45–0.45	0.28–0.3	3.7–7.9	0	2
1975 Haicheng M_s = 7.3	7–8	94.38–110.25	0.13–0.2	0.13–0.2	7.4–14	0	3
1976 Guatemala M = 7.5	9.1–10.2	129.89–148.79	0.14–0.14	0.12–0.13	4.7–14.3	1	1
1976 Tangshan M_s = 7.8	3.5–5.5	122.4–126.73	0.13–0.5	0.13–0.4	7.9–33.1	2	5
1977 Argentina M = 7.4	2.4–11.7	57.68–106.14	0.2–0.2	0.12–0.18	5.3–13.8	2	3
1978 Miyagiken-Oki M = 6.7	2.4–5.9	45.12–209.76	0.1–0.14	0.09–0.15	2.7–18.7	13	1
1978 Miyagiken-Oki M = 7.4	2.4–7.5	42.88–105.6	0.2–0.32	0.16–0.38	2.7–26	6	14
1979 Imperial Valley M_L = 6.6	1.1–4.3	42.88–146.61	0.13–0.51	0.09–0.41	3.5–46.5	4	4
1980 Mid-Chiba M = 6.1	5.5–14.5	17.75–72.24	0.08–0.08	0.05–0.07	3.7–8.9	2	0
1981 Westmorland M_L = 5.6	1.1–4.3	103.94–274.02	0.16–0.23	0.12–0.23	4–19.9	4	3
1983 Nihonkai-Chubu M = 7.1	3.3–9.3	67.68–86.4	0.12–0.15	0.13–0.15	7.9–16.9	2	1
1983 Nihonkai-Chubu M = 7.7	2.4–10.5	17.75–72.24	0.05–0.28	0.05–0.31	5.2–24	3	12
1987 Elmore Ranch M_w = 6.2	4.3–4.7	49.44–49.44	0.09–0.13	0.08–0.11	6.2–11.3	2	0
1987 Superstition Hills M_w = 6.6	4.3–4.7	56.5–171.65	0.2–0.2	0.18–0.19	6.2–11.3	1	1

1987 Superstition Hills M_w = 6.7	1.1–4.3	39.09–196.11	0.13–0.19	0.11–0.23	3.5–46.5	8	0
1989 Loma Prieta M_w = 7	1.8–8.5	72.24–86.4	0.14–0.46	0.08–0.35	3.5–43.8	7	17
1990 Luzon M_w = 7.6	5–7.3	72.24–86.4	0.25–0.25	0.23–0.23	13.9–25.8	1	1
1993 Kushiro-Oki M_w = 8	3.6–10.8	17.75–76.56	0.4–0.4	0.34–0.4	16.5–29.8	1	2
1994 0rthridge M_w = 6.7	6.3–9	31.5–160.24	0.4–0.69	0.28–0.37	10.8–19	0	3
1995 Hyogoken-Nambu M_L = 7.2	2.5–13	92.01–133.66	0.25–0.7	0.23–0.63	5.6–65.5	30	26

Table 3. Evaluation indices for the prediction models.

Classifier	Accuracy	Precision	Recall	F1-score
CB	0.866 (0.071)	0.869 (0.091)	0.878 (0.121)	0.866 (0.077)
MLP	0.813 (0.098)	0.840 (0.127)	0.762 (0.132)	0.833 (0.093)
SVM	0.823 (0.041)	0.816 (0.100)	0.871 (0.081)	0.835 (0.048)

Table 4. Evaluation indices obtained for the prediction models in the Chi-Chi earthquake dataset.

Classifier	Accuracy	Precision	Recall	F1-score
CB	0.888	0.836	0.944	0.887
MLP	0.806	0.736	0.907	0.813
SVM	0.759	0.691	0.870	0.770

Figures

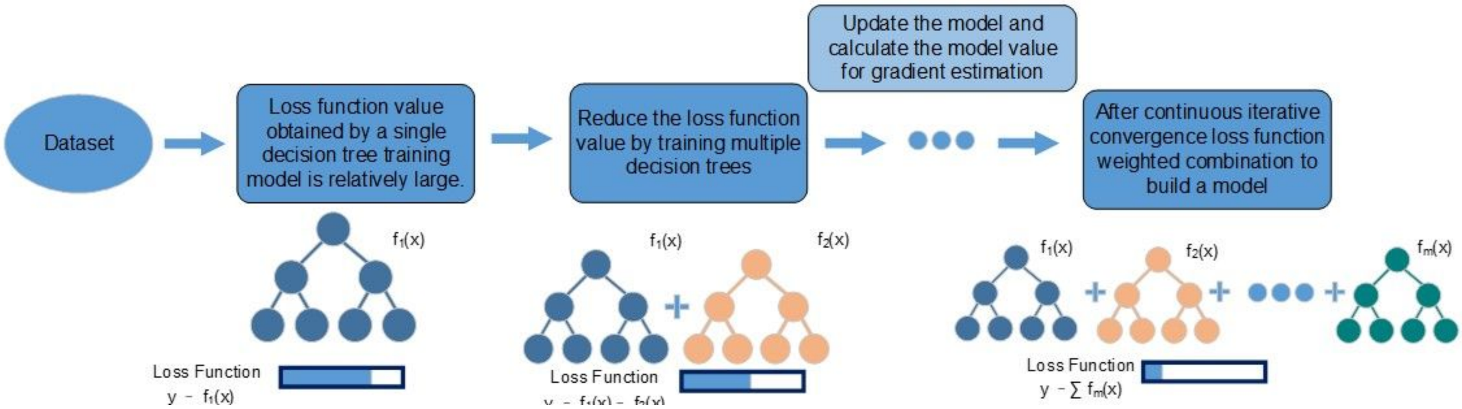
Dataset
Loss function value obtained by a single decision tree training model is relatively large.
Reduce the loss function value by training multiple decision trees
Update the model and calculate the model value for gradient estimation
After continuous iterative convergence loss function weighted combination to build a model
f1(x)
Loss Function
y - f1(x)
f1(x)
f2(x)
Loss Function
y - f1(x) - f2(x)
f1(x)
f2(x)
fm(x)
Loss Function
y - ∑ fm(x)

Figure 1

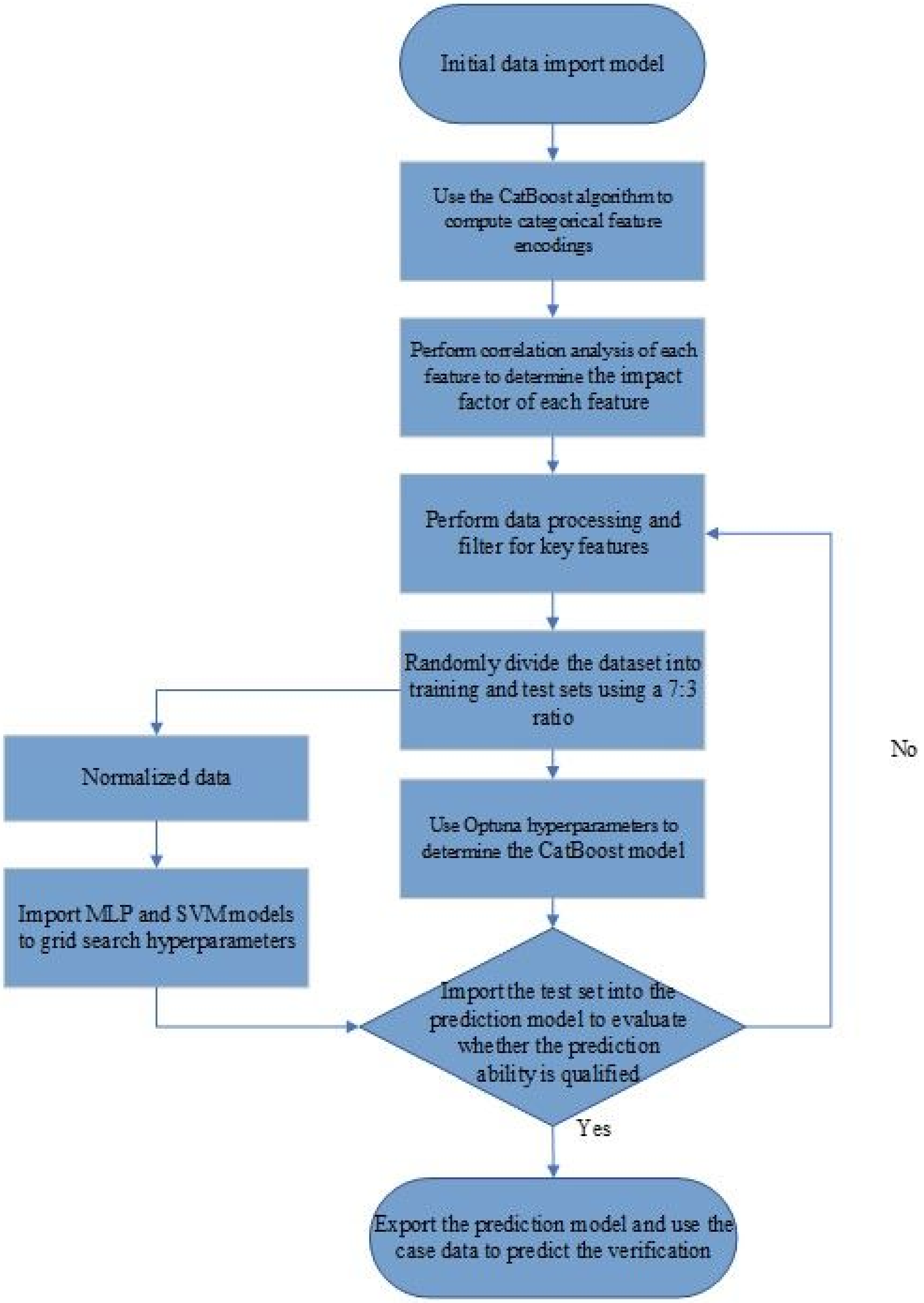
Initial data import model
Use the CatBoost algorithm to compute categorical feature encodings
Perform correlation analysis of each feature to determine the impact factor of each feature
Perform data processing and filter for key features
Randomly divide the dataset into training and test sets using a 7:3 ratio
Normalized data
Import MLP and SVMmodels to grid search hyperparameters
Use Optuna hyperparameters to determine the CatBoost model
Import the test set into the prediction model to evaluate whether the prediction ability is qualified
No
Yes
Export the prediction model and use the case data to predict the verification

Figure 2

Flowchart of the prediction model

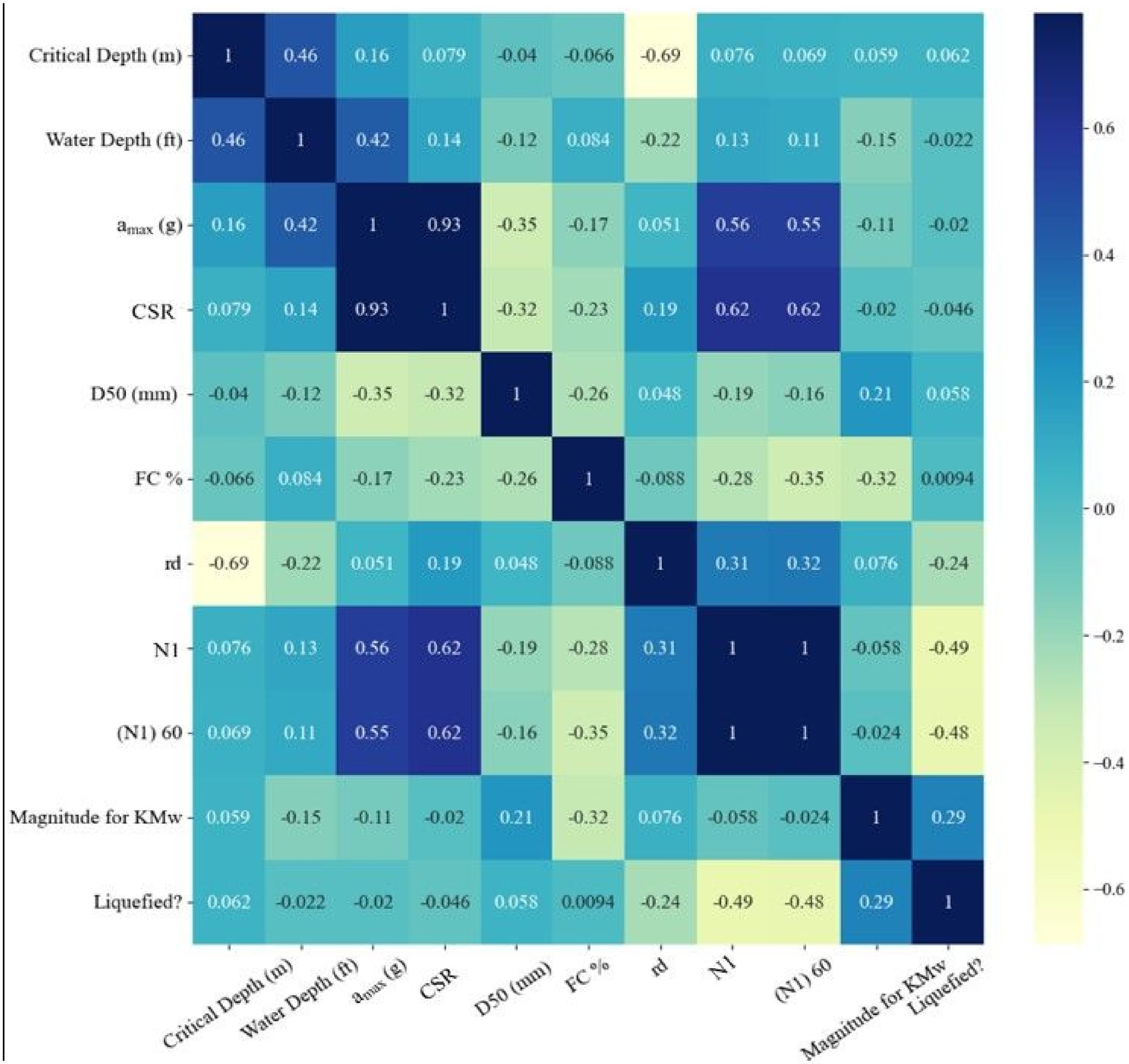

Figure 3

Thermodynamic diagram of the correlation among characteristics.

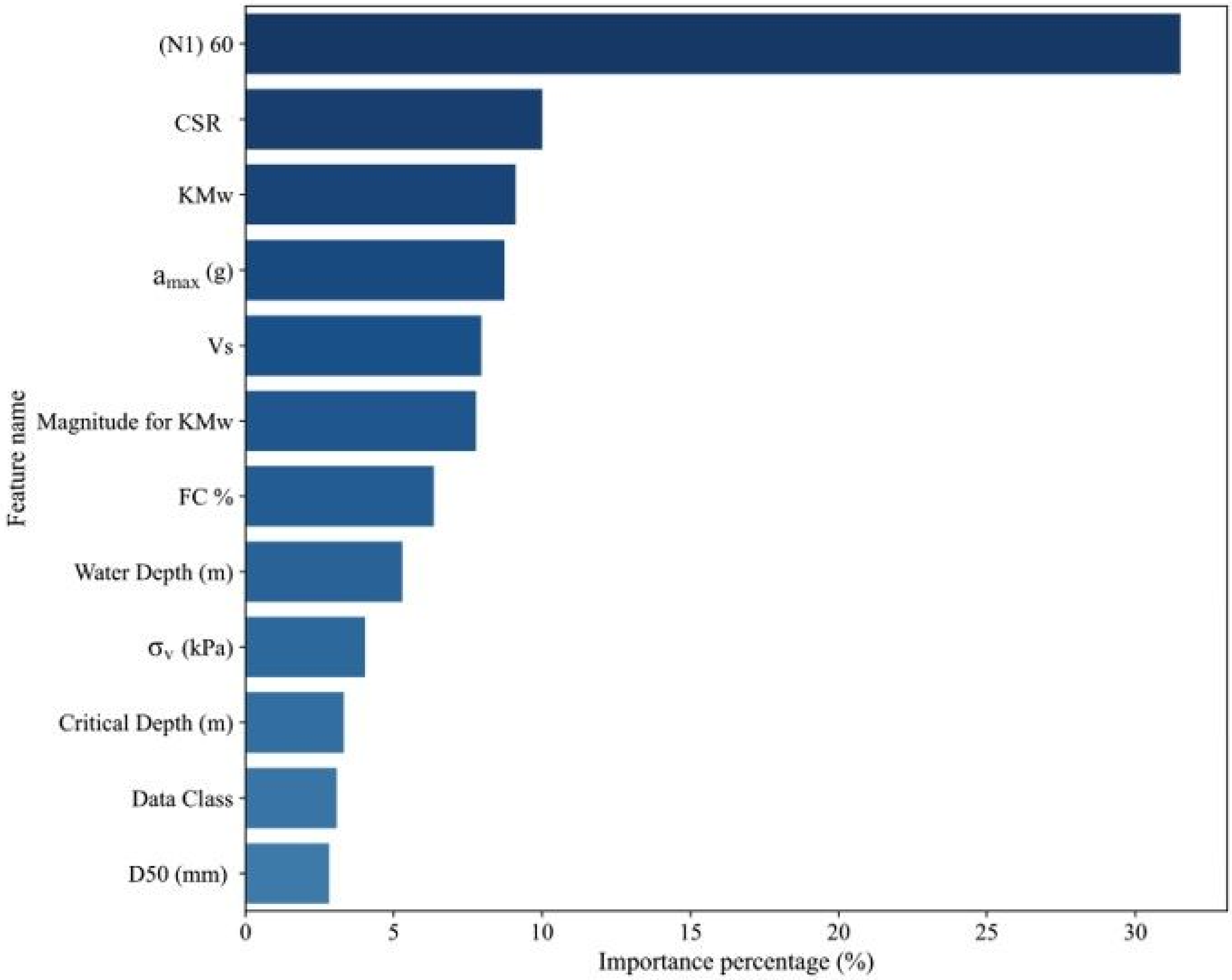

Figure 4

Characteristic importance weights.

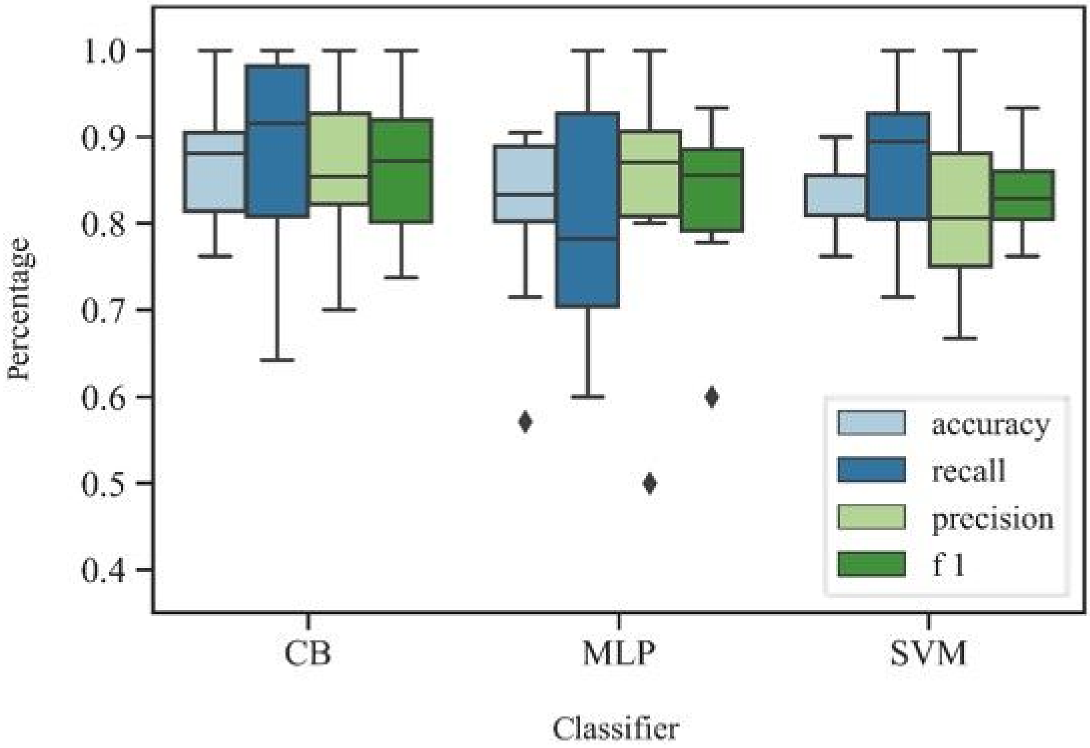

Figure 5

Box plot of the distribution of the evaluation indices for the prediction models. The diamonds indicate the outliers.

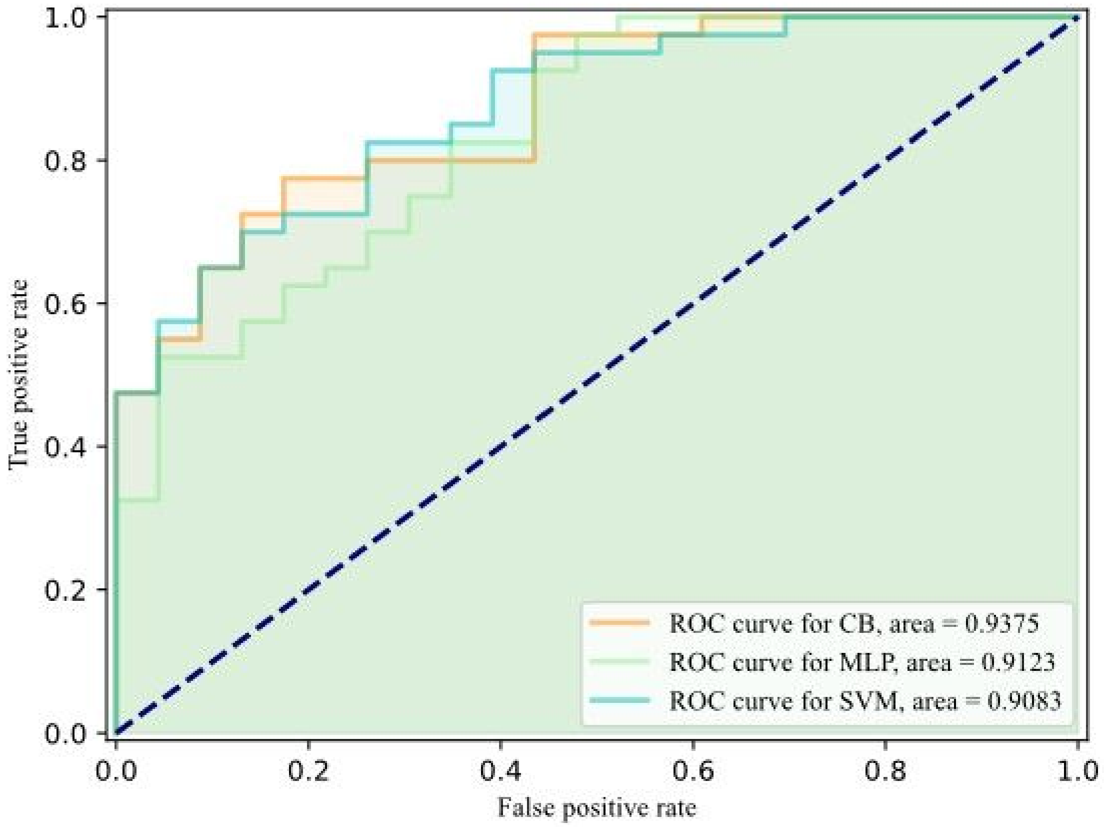

Figure 6

ROC curves for the prediction models.