RESEARCH ON APPLICATION OF RANDOM FOREST MODEL IN PREDICTION OF PILE BEARING CAPACITY BASE ON CPT RESULT

NGHIÊN CỨU ỨNG DỤNG MÔ HÌNH RỪNG NGẪU NHIÊN TRONG DỰ BÁO SỨC CHỊU TẢI CỌC DỰA VÀO KẾT QUẢ XUYÊN CPT

PHẠM TUẤN ANH
Đại học công nghệ Giao thông vận tải
Email: anhpt@utt.edu.vn

Abstract: *The paper presents the results of applying artificial intelligence methods in determining the pile bearing capacity. In this study, an artificial intelligence model namely random forest was developed and applied in pile bearing capacity prediction. The random forest model architecture is optimized by the grid search technique to find the best model. A database of 108 destructive compression results by static pile load method has been synthesized to train and test the model, in which geological data is represented by cone penetration test (CPT) result. In addition, the results of the study are compared with the multi-variable regression model and the traditional formula according to the pile foundation - design standard TCVN 10304:2014, giving the random forest the superiority in determining the load capacity compared to the other two methods. The results of the study show that the random forest with optimum parameters can predict very well the pile load capacity, and has great potential in solving other problems in construction engineering.*

Keywords: *pile bearing capacity, CPT result, artificial intelligence, random forest, multivariable regression, TCVN10304-2014.*

Tóm tắt**:** *Bài báo trình bày kết quả ứng dụng phương pháp trí tuệ nhân tạo trong việc xác định sức chịu tải cọc. Trong nghiên cứu này, một mô hình trí tuệ nhân tạo tên là rừng ngẫu nhiên đã được phát triển và ứng dụng trong việc dự báo sức chịu tải cọc. Kiến trúc mô hình rừng ngẫu nhiên được tối ưu hóa bằng cách khảo sát lần lượt từng tham số để tìm ra mô hình tốt nhất. Một cơ sở dữ liệu gồm 108 kết quả nén tĩnh cọc đã được thu thập để đào tạo và kiểm nghiệm mô hình, trong đó số liệu địa chất được đại diện bằng kết quả xuyên CPT. Kết quả của nghiên cứu được so sánh với mô hình hồi quy đa biến và công thức theo TCVN 10304:2014, cho thấy mô hình rừng ngẫu nghiên mang lại độ chính xác vượt trội trong việc xác định sức chịu tải cọc so với hai phương pháp còn lại. Kết quả của nghiên cứu cho thấy mô hình rừng ngẫu nhiên được tối ưu tốt có khả năng dự báo rất tốt sức chịu tải cọc, đồng thời có tiềm năng lớn trong việc giải quyết các bài toán khác trong lĩnh vực xây dựng.*

Từ khóa: *sức chịu tải cọc, chỉ số CPT, trí tuệ nhân tạo, rừng ngẫu nhiên, TCVN10304-2014.*

1. Introduction

Pile foundation is a type of deep foundation commonly used in the construction industry in general as well as in the field of civil and industrial construction in particular. Practically in the pile design process, the bearing capacity of a single pile plays a decisive role in finding the right pile foundation solution for the project when it affects the determination of the number of piles as well as the size of the foundation cap.

Along with the history of construction, many different methods have been proposed for determining load capacity. There are test methods applied directly in the field such as static load test method [1], dynamic load test method (PDA), static load test method using load cell [2]. The above methods give reliable results, but the disadvantage is that it is time-consuming and uneconomical. To reduce testing costs, many authors have proposed semi-empirical formulas to approximate endurance, using in situ test results (SPT, CPT)[3] [4] [5], etc. These methods give quick results, low cost, high reliability in many cases, however, they are not very general and the calculation results need to be corrected with experimental results. Currently, with the development of the finite element method, many authors have simulated the working of piles and soil and approximated the pile load capacity based on mathematical modeling software such as Plaxis, Ansys, Abacus [6] [7]. However, these methods have the weakness that they have to use many parameters of soil and inductive characters with the output results, the accuracy of the analysis results depends a lot on how these parameters are adjusted for fit.

In recent years, the results of the fourth industrial revolution have been strongly imported into all fields, including the construction sector. Many researchers are looking for ways to apply artificial intelligence solutions to solve various problems in the field of construction in general and pile design in particular. It is an to be some research as Pham et al. (2020) [8], Momeni (2020)[9], Moayedi và Hayati (2019)[10] v.v. However, further research to expand and improve the accuracy of the model is needed. Most of the above publications do not clearly state how to optimize machine-learning models. In addition, random forest is a powerful machine learning method, capable of solving many scientific and engineering problems with fast speed and resistance to overfitting. Specifically, compared with other models such as Artificial Neural Network (ANN), Adaptive Neuro-Fuzzy Inference System (ANFIS), the random forest model has a faster training speed. Along with that is the Boostrap random sampling technique, which helps the model to generalize the research problem and avoid the phenomenon of overfitting with the training data [11], [12].

In this study, the random forest model was used to predict the bearing capacity of piles based on static penetration test (CPT) results. The model hyperparameters are optimized to find the best model by grid search technique. In addition, the research results are also compared with different methods, which are multivariate regression and experiment formula according to TCVN 10304: 2014 to confirm the superiority of the random forest model in determining pile bearing capacity. Finally, importance analysis technique is performed to find out which input variables have the greatest influence on the results of determining the bearing capacity of piles.

2. Development of the Random Forest model

2.1 Random Forest model

The random forest (RF) model is one of the most popular machine-learning methods based on the decision trees model. Forests and Breiman (1999)[13] were the first persons who mention the random forest model, also known as bagging ensemble learning. The typical Decision tree and RF model are illustrated in Figure 1 and Figure 2.

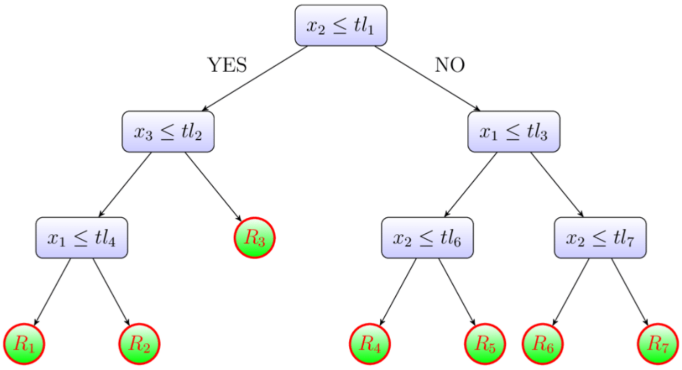

Figure 1. *Visualize the decision tree model for the regression problem*

In the decision tree model, the data is modeled like a tree consisting of branches and leaves. The different instances of the input data (eg x_1, x_2, x_3, etc) are split by branches and the output is at the leaf position (eg R_1, R_2, R_3, etc). More specifically, the architecture of the decision tree model can be considered as a series of if_then_else functions, depending on the input data set, the complexity of the tree as well as the depth of the if the function is calculated and optimized. The individual decision tree model has the major disadvantage of often overfitting the input data. That promotes the development of more advanced models based on decision trees, the random forest model is one of them.

In the RF model, a forest of many decision trees is predefined for training and forecasting. Each tree decision maker is an individual with the exclusive set of forecasting, taking input from a partial data root. The final result of the prediction of the random forest is the average result of the member trees. The interesting point of the random forest model is that the trees are built completely randomly, with the input data of each tree selected according to the bootstrap technique. That will help the model to better generalize the problem and limit the overfitting of individual decision tree models.

The important hyperparameters influences on the model building are: (1) - Number of trees in the forest (n); (2) - Maximum Depth of a tree (D); (3) - Minimum number of samples needed to separate plants (S) and (4) - Minimum number of samples per leaf (L).

The final prediction of the model can be made using the following formula (1):

$$y_i = \frac{1}{n}\sum_{j=1}^{n} f_j(x_i) \qquad (1)$$

In which, y_i is the result of predicting the i^{th} sample; n is the number of trees; f_j is the estimator j^{th} in the forest; x_i - the input vector data on the i^{th} sample. How to build decision trees and hyperparameters of random forest can be found in more detail in the literature [13].

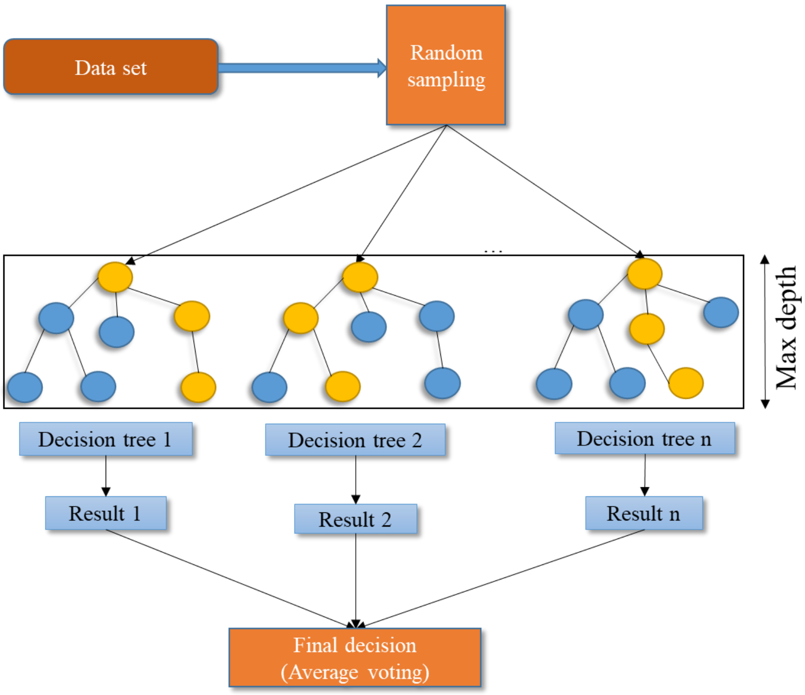

Figure 2. *Random forest model visualization*

2.2 Identification and collection of data

The data used to build and test the model should be collected from various sources to increase the generality of the model. Specifically, a dataset of 108 static pile load tests is compiled and published in the literature of Ghorbani (2018)[14]. This dataset consists of different types of piles, tested with different geological conditions in areas around the world. Therefore, the dataset is highly generalizable and is not localized to a particular locality. All input parameters that can affect the pile load determination are taken according to the input variables in the empirical formula according to Vietnamese national standard TCVN 10304-2014. To be more specific, they are the type of test (T), the type of pile (P), the installation method (denoted as I), end of pile type (EP), the pile tip cross-sectional area (A_t), the shaft area (A_f). The soil properties were shown through parameters obtained from Cone Penetration Test (CPT) results, include the average cone tip resistance along the embedded length of the pile (q_{ca}), the average cone tip resistance over influence zone (q_{ct}), the average sleeve friction along the embedded length of the pile (f_{sa}). The considered output is the ultimate bearing capacity of the pile (denoted as P_u).

The data is divided into two sets: the training set for 80% and the test set for 20% of the total data. Where the training set is used to build the model and testing set is used to evaluate the model. Unlike

Ghorbani's study[14] which initially used only 5 inputs (A_t, A_f, q_{ca}, q_{ct}, f_{sa}), this study will use all 9 input parameters. The statistics of the input data are shown in Table 1.

.

Table 1. *Statistics of input and output parameter information of the current study*

	T(*)	P(*)	I(*)	EP(*)	A_t	A_f	q_{ca}	f_{sa}	q_{ct}	P_u
Unit	-	-	-	-	(cm^2)	(m^2)	(Mpa)	(kN)	(Mpa)	(kN)
Min	1	1	1	1	20	5.45	0.83	9.39	0.25	60
Mean	-	-	-	-	1736	26.46	5.84	101.89	8.82	1965
Median	2	1	2	1	1230	17.98	5.38	81.91	7.63	1340
Max	2	3	2	2	7854	194.65	24.7	349.64	27.11	10910
SD	-	-	-	-	1674	26.35	4.23	66.29	6.19	1702.2

SD = Standard deviation

T = 1 – Continuous load, 2 – Maintain load; P = 1 – Concrete pile, 2 – Composite pile, 3 – Steel pile; I = 1 – Driven pile, 2 – Bored pile; EP = 1 – closed pile, 2 – Open pile.

2.3 Model validation

In this study, performance indicators including R-squared (R^2), and root mean square error (RMSE) are used to evaluate and compare models, specifically as follows:

$$\text{RMSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(y_i - \overline{y}_i\right)^2} \qquad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{k}\left(y_i - \overline{y}_i\right)^2}{\sum_{i=1}^{k}\left(y_i - \overline{y}\right)^2} \qquad (3)$$

In which, k is the number of samples, yi and $\overline{y}_i$ is the experiment, and predicted result, $\overline{y}$ is the mean value of y_i .

R^2 characterizes the correlation between experimental results and predictions while RMSE characterizes the error between experimental results and predictions. In the ideal case, R^2 reaches 1 while RMSE reaches 0.

3. Result and discussion

3.1 Model optimization results

In this work, the RF model is built based on the Python platform, using the Sklearn library. In addition, the most important hyperparameters of the model are examined in turn to choose the best value within their allowable range. Specifically, the survey scope is given in Table 2.

Table 2. *Range of hyperparameters*

Hyper parameter	Explain	Range
n	Number of trees	2-100
D	Max depth	2-20
S	Min samples to split	2-20
L	Min samples on a leaf	1-20

According to many studies, the maximum number of trees does not need to be too much [13]. Meanwhile, other hyperparameters such as D, S, L determine the complexity of the decision trees. It is important to note that, the more complex the decision tree, the more overfitting the model. Besides, the survey range of other hyperparameters is selected so that when the hyperparameter value changes beyond the survey range, the performance of the model does not change significantly.

That is when the model fits the model excessively with training data and does not predict well for testing the data. To avoid data leakage, the 5 Fold CV technique was used to evaluate the model's performance during the survey. According to this technique, the training set is divided into 5 folds, with 4 folds used for training and the remaining fold used for validation.

3.1.1. Effect of Number of trees on performance of Random Forest models

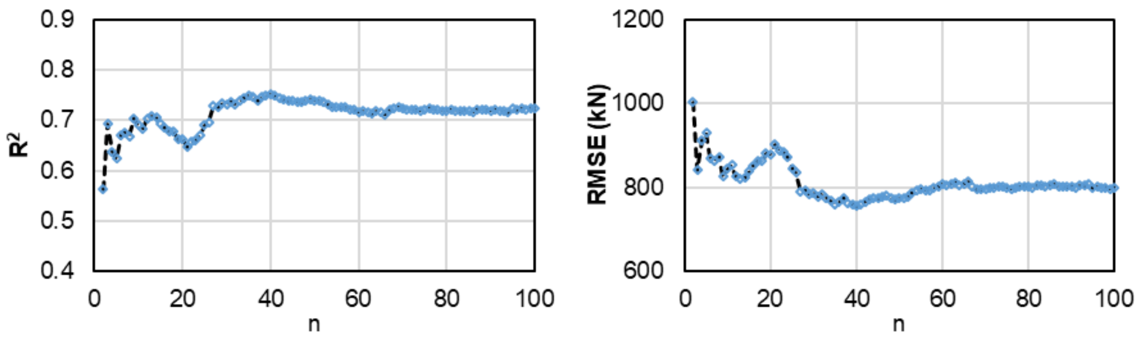

Figure 3. *Result of the model survey according to n*

3.1.2. Effect of Max depth of tree on performance of Random Forest models

The results of the survey on the accuracy of the model when the max depth of tree (D) changes from 2 to 20, n = 40 are shown in Figure 4. It can be seen that the best value of the max depth is 9, then, $R^2 = 0.773$ while RMSE = 722 (kN). Tree depth less than 6 gives very bad prediction results while tree depth greater than 6 does not improve the results much.

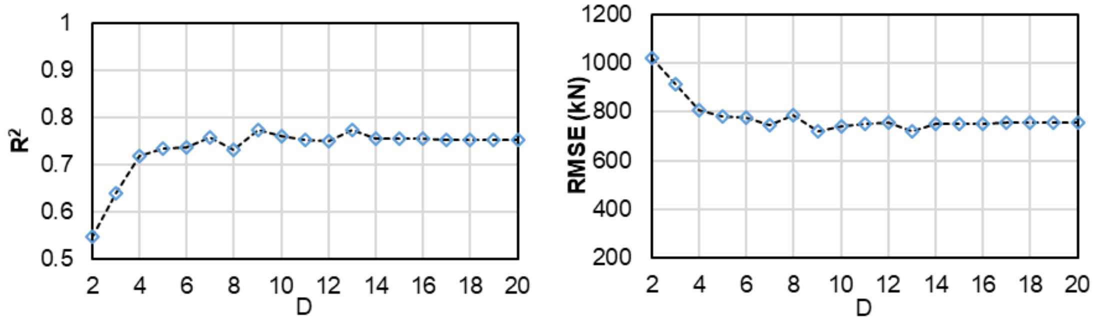

Figure 4. *Result of the model survey according to D*

3.1.3 Effect of Min samples to split of tree on performance of Random Forest models

The results of the survey on the accuracy of the model when min samples to split of tree (S) changes from 2 to 20, n = 40, D =9 are shown in Figure 5. It can be seen that the bigger min samples to split of tree the value, the lower performance of the model, and the best value of S is 2, then, R2 = 0.773 while RMSE = 722 (kN).

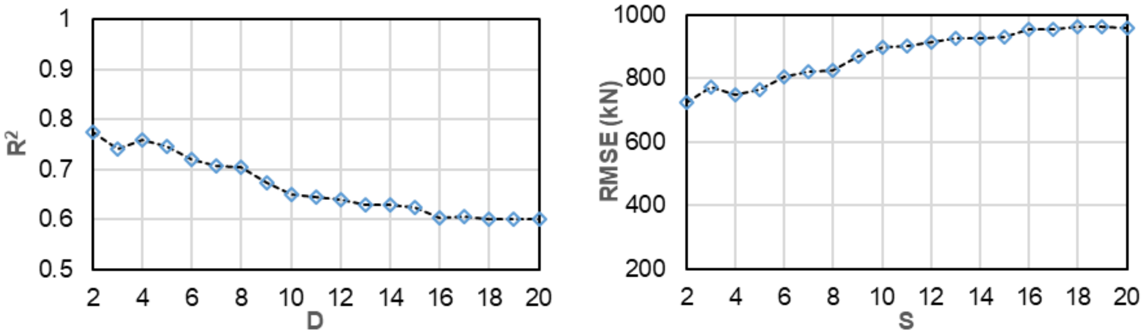

Figure 5. *Result of the model survey according to S*

3.1.4 Effect of Min samples on a leaf of tree on performance of Random Forest models

The results of the survey on the accuracy of the model when Min samples on a leaf of tree (L) changes from 1 to 20, n = 40, D =9, S = 2 are shown in Figure 6. It can be seen that when L is less than 8, the larger L is, the better the prediction result, but conversely, when L is greater than 8, the larger L is, the worse the result. Thus, the best value of L is 8, then, $R^2 = 0.859$ while RMSE = 568 (kN).

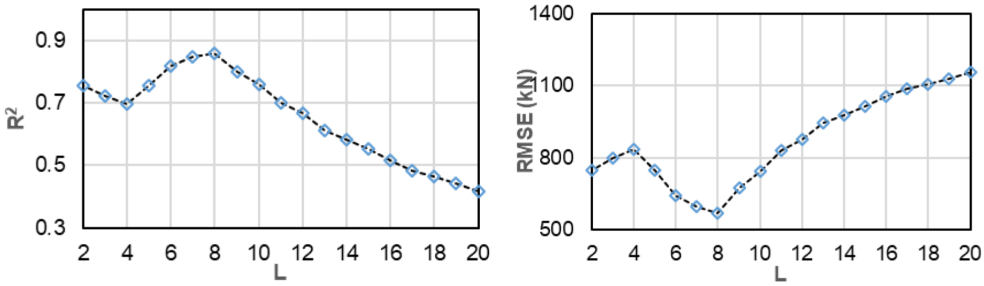

Figure 6. *Result of the model survey according to L*

In general, the best model among the survey models has the number of trees n = 40, D = 9, S = 2, and L = 8. With such valuable parameters, the model is good enough to learn the generality of the data, and at the same time not too complicated to help the model avoid overfitting.

3.2 Compare to different methods

In this section, the results of predicting the capacity of the random forest model with the optimization of the parameters, are compared with the bearing capacity according to the Vietnamese national standard TCVN 10304-2014[15] and the multivariable regression. Result prediction is performed on testing data.

With multivariable regression, the determination system is performed on the Data Analysis tool of EXCEL 2016 software, the multivariable regression weights and bias are determined based on the training set. Multivariable regression weight and bias are showed in table 3.

The general formula of the linear multivariable regression method is as formula (4):

$$P_u = \sum_{i=1}^{9} \beta_i.X_i + \beta_0 \quad (4)$$

In which, β_i is the weight refers to i^th input X_i and β_0 is the bias.

In addition, the formula for calculating the bearing capacity according to the results of the static penetration test according to national standard TCVN 10304 is as formula (5):

$$P_u = k_c.q_{ct}.A_t + \frac{q_{ca}}{\alpha_i} A_f \quad (5)$$

In which, k_c and α_i the coefficient of the cone tip resistance and sleeve friction resistance, see table G2 TCVN 10304: 2014.

Table 3. *Weights and bias value of multivariable regression*

Coefficient	T	P	I	E_P	A_t	A_f	q_{ca}	f_{sa}	q_{ct}	β_0
Value	-2277	-23,5	104,6	181,7	0,3	45,8	60,5	3,7	52	3226,3

The results of the calculation methods can be seen in Figure 7.

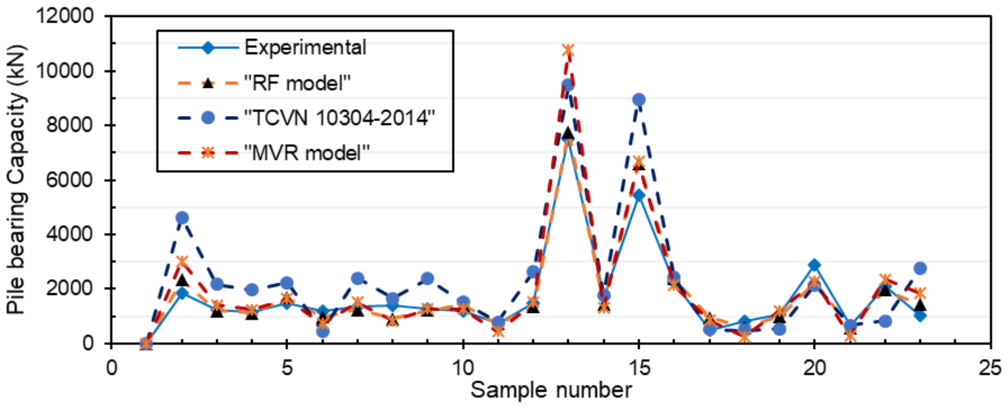

Figure 7. *The comparison of the three methods on the testing set*

Table 4. *Model performance based on three methods*

Criteria	Method		
	RF model	MVR	TCVN 10304-2014
R^2	0.921	0.948	0.82
RMSE (kN)	425	856.32	1287

The results of pile bearing capacity analysis by 3 methods: RF, TCVN 10304, and multivariable regulation showed that the RF model is the best model with R^2 = 0.92 and RMSE = 425 (kN). The multivariate regression model gives quite good results with R^2 = 0.948, however, the RMSE value is very high, reaching 856.32 (kN). It proves that the model predicts the correlation results well, but the root means the squared error is too high due to the difference between the prediction and the experiment error. Finally, the pile bearing capacity determined by the formula in TCVN10304:2014 achieved the lowest accuracy with R^2 = 0.82 and RMSE = 1287 (kN).

3.3 Features importance analysis

In this section, the importance of input features was analyzed. Since the RF model randomly selects features to build decision trees, feature importance is determined by the percentage increase in error (% increase in MSE) of the model, when that feature is unused. Features' importance can be measured through the importance index, which is in the range [0,1] and the sum of all feature' indexes is equal to 1. The larger the index, the more important the features. The features importance analysis result is presented in . It can be seen that of all the variables used to build the RF model, the pile tip cross-sectional area (A_t) had the highest importance, with an important score of 0.587. The shaft area (A_f) was the next important input variable when the importance score is 0.319. Thus, the parameters that characterize the pile geometry showed great importance in predicting the pile bearing capacity. The variables that were characteristic of background attributes such as q_{ct}, q_{ca}, f_{sa} achieved 3rd, 4th, 5th ranks in importance, respectively. The type of pile tip (EP) played a less important role while the remaining variables such as T, I, P had almost no influence on the prediction of pile load capacity.

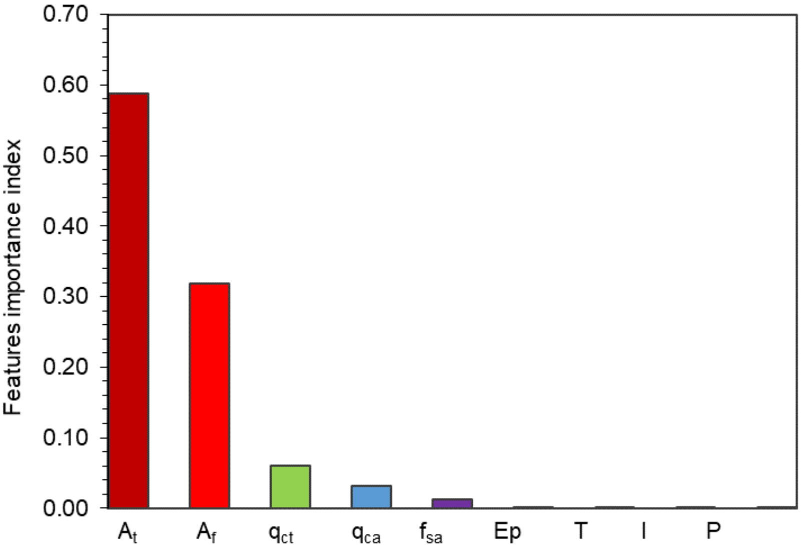

Figure 8. *The feature important analysis result*

4. Conclusions

The present study applied a random forest model, based on artificial intelligence to determine the bearing capacity of piles. The research results show that it is necessary to optimize the parameters so that the model random forest achieves high accuracy when predicting the ultimate bearing capacity of the pile. Specifically, the number of trees around the value 40, the depth of tree greater than 6, the number of samples required to split a node as small as possible, and the number of nodes per leaf should not exceed 8. In addition, the RF model allowed to outperform the two models included for comparison, the multivariable regression model and the formula for determining the load capacity according to the national standards TCVN 10304-2014. The feature importance analysis technique performed on the final RF model showed that the parameters related to the geometrical dimensions of the pile seem to be of greater importance than those related to the soil properties. Based on computing, it is recommended that artificial intelligence models be introduced into the standard foundation. At the same time, continue to calibrate the formulas in the standard to achieve higher accuracy in design practice.

REFERENCES

1. A. Altaee, B. H. Fellenius and E. Evgin (1992), "Axial load transfer for piles in sand. I. Tests on an instrumented precast pile", *Can. Geotech. J., vol. 29, no. 1, pp. 11–20, Feb. doi: 10.1139/t92-002.* https://doi.org/10.1139/t92-002.
2. H. Seo, R. B. Moghaddam and W. D. Lawson (2016), "Assessment of methods for construction of an equivalent top loading curve from O-cell test data", *Soils and Foundations, vol. 56, no. 5, pp. 889–903, Oct. 2016,* doi: 10.1016/j.sandf.2016.08.013.
3. G. G. Meyerhof (1976), "Bearing Capacity and Settlement of Pile Foundations", *Journal of the Geotechnical Engineering Division, vol. 102, no. 3, pp. 197–228,* https://doi.org/10.1061/AJGEB6.0000243.
4. J. H. Schmertmann (1978), "GUIDELINES FOR CONE PENETRATION TEST. (PERFORMANCE AND DESIGN)", *Jul.* https://rosap.ntl.bts.gov/view/dot/958.
5. A. R. Bazaraa and M. M. Kurkur (2020), "N-Values Used to Predict Settlements of Piles in Egypt", *1986, pp. 462–474. Accessed: Mar. 08,* https://cedb.asce.org/CEDBsearch/record.jsp?dockey=0048882.
6. I. Shooshpasha, A. Hasanzadeh and A. Taghavi (2013), "Prediction of the axial bearing capacity of piles by SPT-based and numerical design methods", *International Journal of GEOMATE, vol. 4, no. 2, pp. 560–564,* http://dx.doi.org/10.21660/2013.8.2118.
7. I. Shooshpasha (2013), "Prediction of the Axial Bearing Capacity of Piles by SPT-based and Numerical Design Methods", *geomate, doi: 10.21660/2013.8.2118.*
8. T. A. Pham, H.-B. Ly, V. Q. Tran, L. V. Giap, H.-L. T. Vu and H.-A. T. Duong (2020), "Prediction of Pile Axial Bearing Capacity Using Artificial Neural Network and Random Forest", *Applied Sciences, vol. 10, no. 5, p. 1871, Mar. ,* doi: 10.3390/app10051871.
9. E. Momeni, M. B. Dowlatshahi, F. Omidinasab, H. Maizir and D. J. Armaghani (2020), "Gaussian Process Regression Technique to Estimate the Pile Bearing Capacity", *Arab J Sci Eng, vol. 45, no. 10, pp. 8255–8267, Oct.,* doi: 10.1007/s13369-020-04683-4.
10. H. Moayedi and S. Hayati (2019), "Artificial intelligence design charts for predicting friction capacity of driven pile in clay", *Neural Comput & Applic, vol. 31, no. 11, pp. 7429–7445, Nov.*doi: 10.1007/s00521-018-3555-5.
11. N. Altman and M. Krzywinski (2017), "Ensemble methods: bagging and random forests", *Nature Methods, vol. 14, no. 10, pp. 933–934, Oct.,*doi: 10.1038/nmeth.4438.
12. T. K. Ho (1995), "Random decision forests" in *Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282 vol.1.* doi: 10.1109/ICDAR.1995.598994.
13. L. Breiman (2001), "Random Forests", *Machine Learning,* 45**,** 5–32,. https://doi.org/10.1023/A:1010933404324.
14. B. Ghorbani, E. Sadrossadat, J. Bolouri, P. Rahimzadeh Oskooei (2018), "Numerical ANFIS-Based Formulation for Prediction of the Ultimate Axial Load Bearing Capacity of Piles Through CPT Data", *Geotechnical and Geological Engineering, pp. 1–20, Jan. ,* doi: 10.1007/s10706-018-0445-7.
15. "Pile Foundation - Design Standard TCVN 10304-2014". Vietnamese national standard. *https://tieuchuan.vsqi.gov.vn/tieuchuan/view?sohieu=TCVN+10304%3A2014.*

Ngày nhận bài:4/3/2022.

Ngày nhận bài sửa:01/4/2022.

Ngày chấp nhận đăng:02/4/2022.