PREDICTION OF BEARING CAPACITY OF PILE USING SUPPORT VECTOR AND CATBOOST REGRESSION

M. M. H. Galib*[1], M. M. Billah[1], S. A. Mofiz[2]

[1] *Undergraduate Student, Department of Civil Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, e-mail: mehedihassangalib@gmail.com*
[2] *Professor, Department of Civil Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, e-mail: samofiz@gmail.com*

***Corresponding Author**

ABSTRACT

The traditional techniques used for the estimation of the bearing capacity of piles are time consuming and costly. Moreover, some of the techniques are empirical, which involve enormous approximations and don't consider all the variables contributing to the strength of the pile. Hence, researchers have grown more interested in several machine learning approaches in order to consider more variables to model practical field situations with great accuracy and overcome the issues of approximation. In this study, cohesion, friction angle, specific weight of soil, pile-soil friction angle, flap number¸ pile area, and pile length have been considered as the contributing factors of pile-bearing capacity. Two of the most widely used machine learning models, Support Vector Regression (SVR) and CatBoost Regression, have been built with different architectures and hyperparameter sets for predicting the bearing capacity of pile. Datasets have been collected from various regions for modelling the heterogeneous nature of the soil and avoiding the overfitting issue to make the model more generalized. The R2 values have been found as 0.87 and 0.95 for concrete and steel piles respectively from SVR and 0.98 and 0.99 for concrete and steel piles respectively from CatBoost Regression. For the further improvement of the prediction accuracy some advance machine learning and deep learning models with more generalized dataset have been suggested.

Keywords: *Bearing capacity, Machine learning, Prediction, Support vector regression, CatBoost regression*

1. INTRODUCTION

The realm of civil engineering is a dynamic interplay between innovation and established practices, ensuring the reliability and safety of the monumental structures that captivate our admiration. In the field of geotechnical engineering, piles are commonly used as a foundation element to support various types of structures such as buildings, bridges, and offshore platforms. The prediction of bearing capacity for piles is a critical aspect of geotechnical engineering and foundation design, as it directly influences the stability and safety of various structures (Niazi et al., 2014). Modern approaches to estimating pile bearing capacity involve numerical, experimental, and analytical methods, providing diverse tools for geotechnical engineering. This integration enhances prediction accuracy, showcasing the progressive nature of contemporary practices in optimizing pile foundation performance. (Shahin, 2010). In the assessment of pile bearing capacity, the Standard Penetration Test (SPT) has been extensively employed (Bouafia & Derbala, 2002). Various hypotheses, rooted in SPT results, have been formulated to predict pile bearing capacity, including empirical equations derived from studies by Meyerhof (1976), Bazaraa and Kurkur (1986), and Shariatmadari et al. (2008). Additionally, a proposed approach involves considering an experimental formula to incorporate the influences of soil type. Specifically, this includes using the SPT value for sandy soil and the untrained shear strength of soil (Cu) for clayey soil (Architectural Institute of Japan, 2004). Nevertheless, the methods mentioned above for estimating pile-bearing capacity have been demonstrated to be both time-consuming and expensive. (Abu-Farsakh et al., 2004). Since the early 1990s, the use of machine learning approaches to predict pile bearing capacity has increased significantly (Debiche et al., 2018). Machine learning, a subset of artificial intelligence (AI) and computer science, revolves around using data and algorithms to replicate human learning processes and enhance accuracy over time. This shift is driven by the advantages of machine learning, such as its capability to handle extensive datasets and navigate highly nonlinear relationships among various parameters. A recent study proposed using Linear regression that it yielded accurate predictions with an average relative error of 6.5% (Liu et al., 2018). Again, a machine learning model based on CatBoost regression to predict the bearing capacity of pile foundations using data from a field test had better performance in predicting the pile capacity compared to other machine learning models. While other recent studies have proposed the use of a random forest algorithm for predicting single pile bearing capacity in various soil types (Wu et al., 2020). Additionally, a K-nearest neighbor algorithm demonstrated accurate predictions for pile-bearing capacity in clayey soils with an average relative error of 9.1% (Jiang et al., 2019). Furthermore, a support vector machine algorithm outperformed conventional empirical methods in predicting pile-bearing capacity (Singh et al., 2019).

2. METHODOLOGY

2.1 Typical Flow Chart

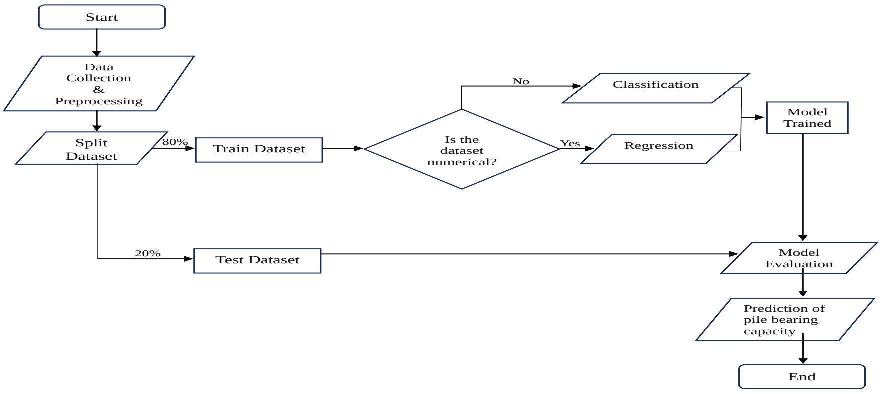

Figure 1: Typical flowchart of a machine learning algorithm

2.2 Data Collection and Preprocessing

The dataset has been collected from an open access scientific journal called Multidisciplinary Digital Publishing Institute (MDPI). Since the dataset contains two types of piles, the whole dataset has been segregated according to the materials of the pile: concrete pile and steel pile.

The dataset has contained 8 variables and 100 observations of steel and concrete piles. The first 7 variables: average cohesion, average friction angle, average soil specific weight, average pile-soil friction angle, flap number, pile area, and pile length have been used as explanatory variables and pile capacity has been used as a response variable. The statistical properties of the dataset have been shown in Table 2.1 and in Table 2.2 for concrete and steel piles respectively.

Table 1: Statistical properties for concrete pile

Variables	**Mean**	**Standard Deviation**	**Correlation with Pile Capacity**
Average Cohesion (kN/m^2)	39.75	47.09	-0.22
Average Friction Angle (°)	22.15	10.50	0.34
Average Soil Specific weight (kN/m^3)	9.65	1.72	0.1
Average Pile-Soil Friction Angle (°)	14.09	1.25	0.18
Flap Number	229.74	370.75	0.34
Pile Area (m^2)	0.17	0.11	0.19
Pile Length (m)	22.96	4.87	0.23
Pile Capacity (kN)	2388.36	677.19	1.00

Table 2: Statistical properties for steel pile

Variables	**Mean**	**Standard Deviation**	**Correlation with Pile Capacity**
Average Cohesion (kN/m^2)	29.94	17.34	0.13
Average Friction Angle (°)	28.32	8.56	0.35
Average Soil Specific weight (kN/m^3)	10.71	2.05	0.21
Average Pile-Soil Friction Angle (°)	13.47	2.12	-0.07
Flap Number	542.53	542.89	0.45
Pile Area (m^2)	0.54	0.50	0.50

Pile Length (m)	24.08	12.71	0.07
Pile Capacity (kN)	3171.25	1538.43	1.00

In the data preprocessing phase, the dataset has been checked to see if it has been fit to be used in machine learning algorithms or if some corrections have been required. Missing values and irrelevant observations have also been checked. Although there were no missing values, some outliers were present in the dataset. These outliers have been removed so that the algorithms would not consider any unusual cases in the models.

2.3 Building up of the model

After preprocessing, each dataset has been split into two parts: the train dataset and the test dataset. 80 percent of the observations have been used as train set for training the models while the remaining 20 percent have been used as a test set for evaluating the prediction accuracy of the models as shown in the flowchart in Figure 1.

As this study has been based on continuous variable, two regression models, such as Support Vector Regression and CatBoost Regression have been used to train the models using the train dataset. After training the model, the accuracy of the models has been evaluated using the test dataset. For evaluating the models, various error metrices such as R^2, Mean Absolute Error, Mean Absolute Percentage Error and Root Mean Square Error have been used.

2.4 Machine Learning Algorithms

2.4.1 Support Vector Regression

The principle involved in Support Vector Regression is to find the best fit line according to the dataset called hyperplane. Both linear and non-linear kernels have been used to find the hyperplane in Support Vector Regression. This hyperplane has been developed in an N-dimensional space, where N is the number of explanatory variables. Then two marginal planes have also been created at a distance equal to ε from the hyperplane in both sides. The equations of these lines have been given as Equation (1), Equation (2) and Equation (3) below.

Equation of hyperplane:

$$y = wx + b \tag{1}$$

Equation of marginal planes:

$$wx + b = +\varepsilon \tag{2}$$

$$wx + b = -\varepsilon \tag{3}$$

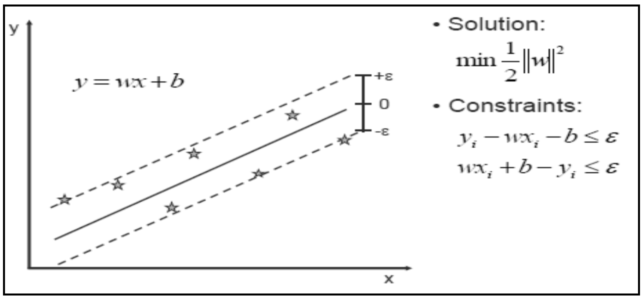

Figure 2: Support vector regression for data points inside the marginal planes.

The principle of Support Vector Regression has been shown in Figure 2. One of the most unique features that makes Support Vector Regression more significant in prediction is that it does not try to minimize the error between the actual observation and the predicted observation like other regression models. Rather it tries to fit the best line within the marginal planes expressed as Equation (2) and Equation (3). In this study, two support vector regression models (SVR-1 and SVR-2) have been built with two hyperparameter sets consisting c, kernel and gamma.

c: The hyperparameter 'c' is inversely proportional to the size of the margin as well as the ε. The trade-off between the insensitive loss and the sensitive loss has been controled by this hyperparameter. The larger value of 'c' indicates smaller ε, while the smaller value of 'c' indicates the larger value of ε. Primarily larger values of 'c' as well as smaller margin helps to improve the regularization and solve the overfitting issue of the model. Furthermore, the insensitive loss will be minimized more with a larger value of 'c'. In this study, two values of c have been used as 0.1 and 1 for SVR-1 and SVR-2 respectively.

Kernel: Kernel is used to find the best suited hyperplane by transforming the data points into the required form. It separates the data points in either linear or non-linear manner. While a linear kernel catches the linear dependencies of the explanatory variables with the response variable, a non-linear kernel considers the non-linear dependencies also. In this study, two kernels such as 'Linear' and 'Radial Basis Function' have been used for SVR-1 and SVR-2 respectively.

Gamma: Another hyperparameter 'Gamma' has been used to control the distance of influence of the individual data points.

The hyperparameters used in SVR-1 and SVR-2 have been listed in Table 2.3.

Table 3: Hyperparameters for Support Vector Regression

Hyperparameters	**SVR-1**	**SVR-2**
c	0.1	1
Kernel	Linear	Radial Basis Function
Gamma	Scale	Auto

2.4.2 CatBoost Regression

CatBoost Regression is a relatively new supervised machine learning algorithm which has been developed in 2017. The principle of this algorithm consists of two main features: the categorical data (the cat) and the gradient boosting (the boost). Although it has been suggested by many researchers that the CatBoost Regression has been also well suited to regression problems. CatBoost Regression incorporates a number of decision trees and improves the accuracy of the prediction through gradient boosting. Gradient boosting is a technique which combines relatively weak models and forms a better model by fitting the decision trees in a sequential manner.

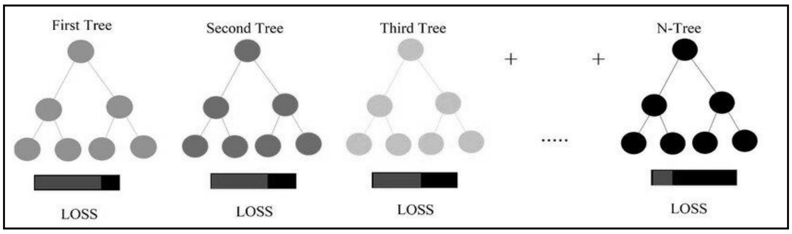

Figure 3: Sequential decision trees in CatBoost Regression

In Figure 2.3, it has been demonstrated that the sequential decision trees of CatBoost Regression have learnt from the previous trees and thus the loss function has been sequentially reduced. One of the distinct features of CatBoost Regression is that the decision trees grow symmetrically, while in Extreme Gradient Boosting, the trees grow depth-wise. Therefore, the nodes in CatBoost Regression remain at the same level and split at the same boundary condition. This precedure allows the CatBoost Regression to overcome the overfitting issue and develop a more generalized prediction model. In comparison to Extreme Gradient Boosting and other models, CatBoost Regression takes shorter time to perform. And it can be applied for both small and large datasets.

2.5 Evaluation of the Model

Once the model has been trained, it has been required to evaluate the model using the test dataset. There are lots of techniques for carrying out performance measurement as well as error metrics to evaluate the models. In this study, coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) have been used as error metrics.

Formula of (R^2):

$$\mathrm{R}^2 = 1 - \frac{ss_{res}}{ss_{tot}} \tag{2.4}$$

In Equation (3.7),

SS_{res} = Residuals sum of squares

SS_{tot} = Total sum of squares

Formula of MAE:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|A_t - P_t\right| \tag{2.5}$$

Formula of MAPE:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \tag{2.6}$$

Formula of RMSE:

$$\mathrm{M} = \sqrt{\sum_{t=1}^{n}\frac{(A_t - P_t)^2}{n}} \tag{2.7}$$

In Equation (2.4), Equation (2.5), Equation (2.6) and Equation (2.7),

M = Error metrics
A_t = Actual value
P_t = Predicted value
n = Number of observations

3. RESULT & DISCUSSION

3.1 Support Vector Regression

The error metrices for steel and concrete piles have been shown in Table 3.1 and Table 3.2 respectively. It has been observed that SVR-2 has performed quite well for both steel and concrete pile as compared to SVR-1. This improvement has been brought through hyperparameter tuning. From further comparison between steel and concrete, it has also been observed that SVR-2 has predicted the bearing capacity of steel pile with more accuracy than that of the concrete pile.

Table 4: Error metrics for Support Vector Regression (steel pile)

Models	Error Metrics			
	R^2	MAE	MAPE (%)	RMSE
SVR-1	0.92	447.19	11.49	534.45
SVR-2	0.95	410.39	10.52	489.68

Table 5: Error metrics for Support Vector Regression (concrete pile)

Models	Error Metrics			
	R^2	MAE	MAPE (%)	RMSE
SVR-1	0.80	688.64	26.71	753
SVR-2	0.87	495.21	18.31	558.23

The R^2 values for steel and concrete pile have been found as 0.95 and 0.87 respectively from SVR-2. The other error metrices such as MAE, MAPE and RMSE have been found as 410.39, 10.52% and 489.68 respectively for steel pile and 495.21, 18.31% and 558.23 respectively for concrete pile from SVR-2. Along with R^2, the other error metrices have also shown that SVR-2 has performed with more accuracy for steel pile as compared to the concrete pile. From SVR-1, The R^2 values for steel and concrete pile have been found as 0.92 and 0.80 respectively. The other error metrices such as MAE, MAPE and RMSE have been found from SVR-1 as 447.19, 11.49% and 534.45 respectively for steel pile and 688.64, 26.71% and 753 respectively for concrete pile. The R^2 of SVR-2 for steel and concrete pile have been shown in Figure 3.1 and Figure 3.2 respectively.

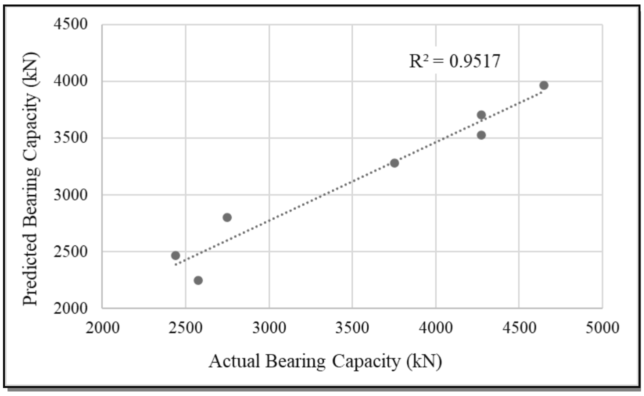

Figure 4: R^2 for steel pile (SVR-2)

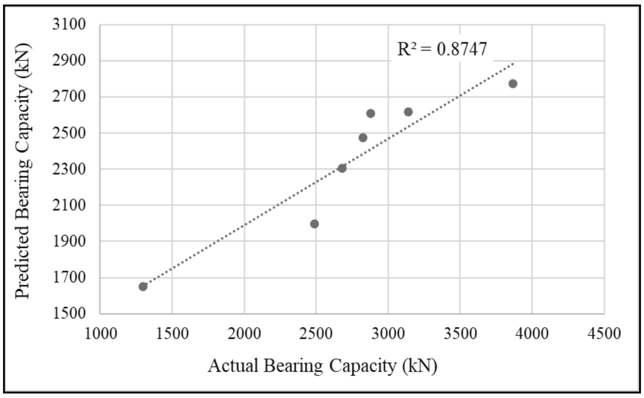

Figure 5: R^2 for concrete pile (SVR-2)

The variations of predicted bearing capacity with respect to actual bearing capacity for steel and concrete pile have been shown in Table 3.3 and Table 3.4 respectively and have been represented through bar chart in Figure 3.3 and Figure 3.4 respectively.

Table 6: Actual vs predicted bearing capacity for steel pile

Actual Observation (kN)	Predicted Observation for SVR-1 (kN)	Predicted Observation for SVR-2 (kN)
4275	3381.73	3526.34
2440	2374.11	2465.46
4650	3873.18	3969.35
3750	3216.55	3279.40
2750	2868.54	2802.24
2575	2302.71	2248.49
4275	3804.94	3706.43

Table 7: Actual vs predicted bearing capacity for concrete pile

Actual Observation (kN)	Predicted Observation for SVR-1 (kN)	Predicted Observation for SVR-2 (kN)
2682.5	2274.98	2303.45
3142	2414.01	2616.61
1295	2002.31	1651.83
2825	2259.71	2476.79
2880	2307.44	2610.37
3867	2481.82	2772.95
2490	2035.40	1996.66

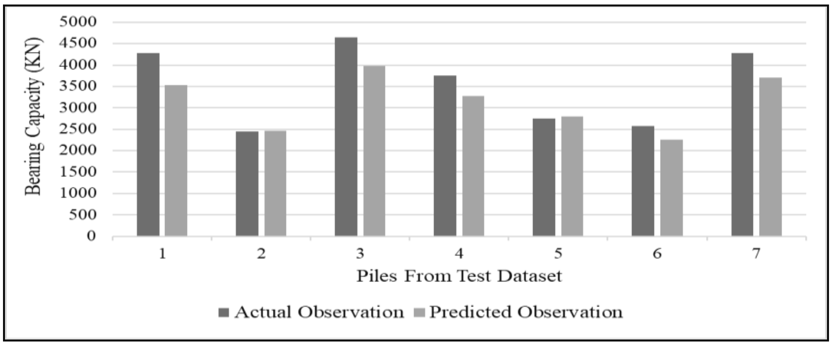

Figure 6: Actual vs predicted bearing capacity for SVR-2 for steel pile

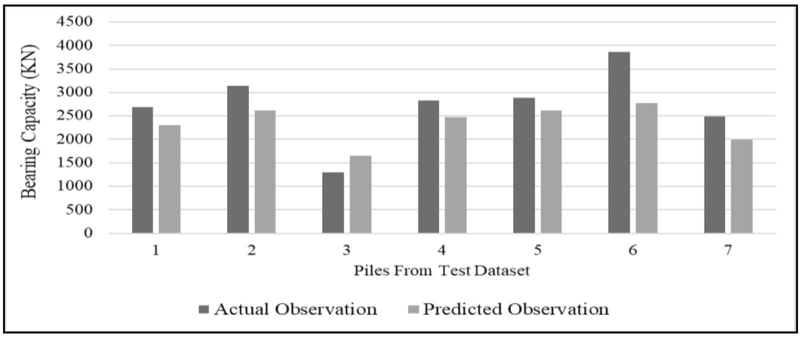

Figure 7: Actual vs predicted bearing capacity for SVR-2 for concrete pile

3.2 CatBoost Regression

Although CatBoost and XGBoost have been based on Decision Tree algorithm, the trees are symmetrical in CatBoost Regression while they grow normal depth-wise in XGBoost Regression. Therefore, it has been found than the CatBoost has performed with quite good prediction accuracy as compared to other models in this study. The MAPE for CatBoost Regression has been found only 6.01% for steel pile and only 9.41% for concrete pile. The error metrices have been shown in Table 3.5. The other error metrices such as MAE, MAPE and RMSE have been found 177.84, 60.1% and 208.07 respectively for steel pile and 231.76, 9.41% and 294.66 respectively for concrete pile. These error matrices have also been found less than the error metrices found from previous models.

Table 8: Error metrics for Extreme Gradient Boosting Regression

Error Metrices	Steel Pile	Concrete Pile
R^2	0.99	0.98
MAE	177.84	231.76
MAPE (%)	6.01	9.41
RMSE	208.07	294.66

The R^2 of CatBoost Regression for steel and concrete pile have been shown in Figure 3.5 and Figure 3.6 respectively.

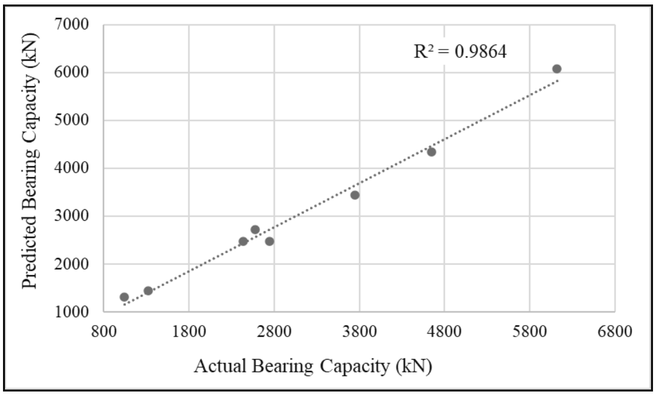

Figure 8: R^2 for CatBoost Regression (steel pile)

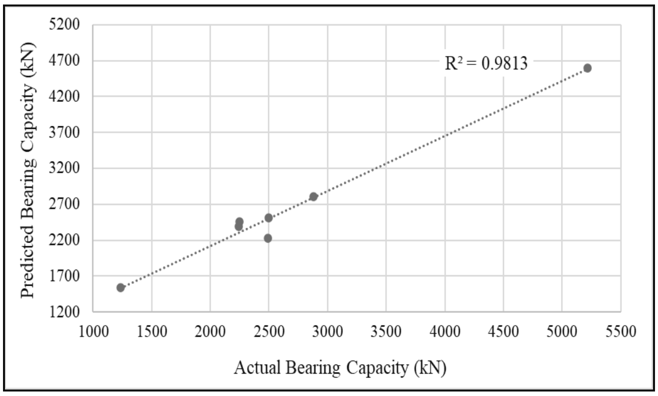

Figure 9: R^2 for CatBoost Regression (concrete pile)

The variations of predicted bearing capacity with respect to actual bearing capacity for steel and concrete pile have been shown in Table 3.6 and Table 3.7 respectively and have been represented through bar chart in Figure 3.7 and Figure 3.8 respectively.

Table 9: Actual vs predicted bearing capacity for CatBoost for steel pile

Actual Observation (kN)	**Predicted Observation (kN)**
2440	2482.01
4650	4337.70
3750	3451.93
1321	1440.19
2750	2476.26
6120	6074.35
2575	2728.89
1043	1319.07

Table 10: Actual vs predicted bearing capacity for CatBoost for concrete pile

Actual Observation (kN)	**Predicted Observation (kN)**
2880	2809.53
2240.5	2390.05
2490	2231.80
2250	2462.49
1235	1538.99
2500	2514.45
5215	4601.80

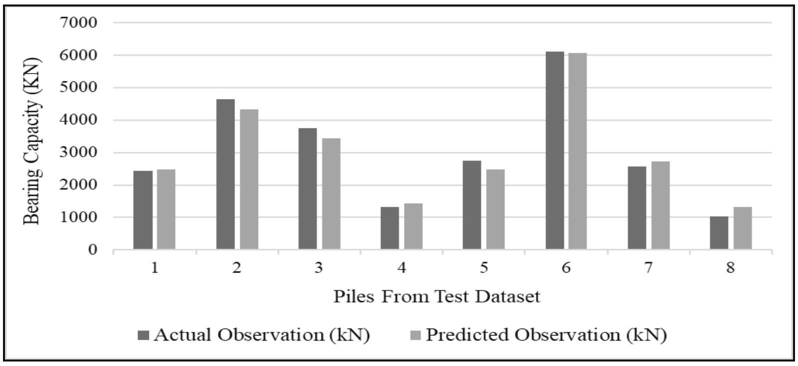

Figure 10: Actual vs predicted bearing capacity for CatBoost for steel pile

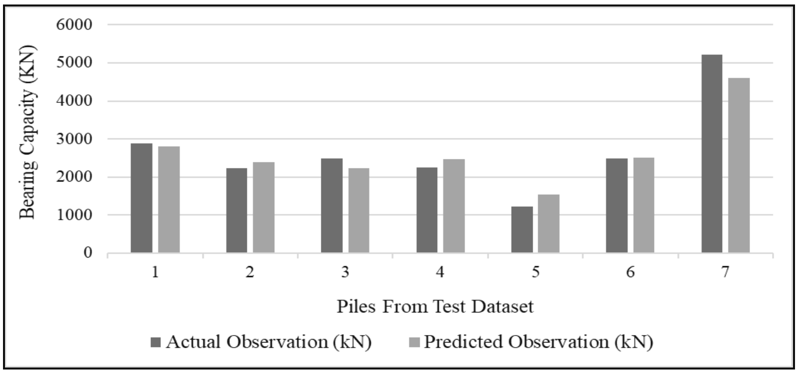

Figure 11: Actual vs predicted bearing capacity for CatBoost for concrete pile

3.3 Comparison between the models

The error metrices for the models used in this study have been shown in Table 3.8 and Table 3.9 for steel and concrete pile respectively.

Table 11: Error metrices for steel pile dataset

Models		**R^2**	**MAE**	**MAPE (%)**	**RMSE**
Support Vector Regression	SVR-1	0.92	447.19	11.49	534.45
	SVR-2	0.95	410.39	10.52	489.68
CatBoost Regression		0.99	177.84	6.01	208.07

Table 12: Error metrices for concrete pile dataset

Models		**R^2**	**MAE**	**MAPE (%)**	**RMSE**
Support Vector Regression	SVR-1	0.80	688.64	26.71	753.00
	SVR-2	0.87	495.21	18.31	558.23
CatBoost Regression		0.98	231.76	9.41	294.66

4. CONCLUSIONS

The study conducted a comprehensive comparison between CatBoost Regression and Support Vector Regression (SVR) in predicting the bearing capacity of piles, specifically focusing on steel and concrete piles. The results revealed that CatBoost Regression has performed quite well in predicting the bearing capacity of piles as compared to Support Vector Regression in the case of steel piles. It has also been found that machine learning algorithms can model the characteristics of steel piles better as compared to concrete piles. Hence the R2 value has been found 0.99 for the steel pile, while the R2 value of the concrete pile has been found 0.98 in CatBoost Regression. The performance of CatBoost Regression in predicting the bearing capacity of steel piles has also been found satisfactory compared to the concrete pile. The Mean Absolute Percentage Error (MAPE) further underscored the predictive prowess of CatBoost Regression. For steel piles, the MAPE was found to be only 6.01%, indicating a high level of accuracy in predicting the bearing capacity. In comparison, concrete piles exhibited a slightly higher MAPE of 9.41%, still reflecting a commendable predictive performance.

Since the accuracy of prediction obtained from CatBoost Regression and Support Vector Regression is satisfactory for steel and concrete piles respectively, these algorithms can be used in the practical field to predict the bearing capacity of piles. To improve the prediction accuracy, more advanced hyperparameter tuning can be carried out. Furthermore, some other advanced machine learning models along with deep learning models can be used to improve the accuracy in predicting the bearing capacity of a pile.

REFERENCES

Abu-Farsakh, M.Y.; Titi, H.H. (2004). Assessment of Direct Cone Penetration Test Methods for Predicting the Ultimate Capacity of Friction Driven Piles. J. Geotech. Geoenvironmental Eng.130, 935–944.

Architectural Institute of Japan. (2004). Recommendations for Design of Building Foundation; Architectural Institute of Japan: Tokyo, Japan.

Bazaraa, A.R.; Kurkur, M.M. (1986). N-values used to predict settlements of piles in Egypt. In Proceedings of InSitu'86; New York, American Society of Civil Engineers: New York, NY, USA; pp. 462–474.

Bouafia, A.; Derbala, A. (2002). Assessment of SPT-based method of pile bearing capacity–analysis of a database. In Proceedings of the International Workshop on Foundation Design Codes and Soil Investigation in View of International Harmonization and Performance-Based Design, Tokyo, Japan, pp. 369–374.

Debiche, F., & Kettab, R.M., & Benbouras, M.A., & Benbellil, B., & Djerbal, L., & Petrisor, A.-I. (2018). Use of GIS systems to analyze soil compressibility, swelling and bearing capacity under superficial foundations inalgiers region, ALGERIA. Urban. Arhit. Constr. 9, 357–370.

Jiang, J., Li, W., & Wang, J. (2019). Prediction of pile bearing capacity based on K-nearest neighbor algorithm Measurement, 136, 163-173.

Liu, H., Guo, Y., Zhang, X., & Xu, H. (2018). Predicting pile-bearing capacity using linear regression in different soil conditions. Advances in Civil Engineering, 1-8.

Meyerhof, G.G. (1976). Bearing Capacity and Settlement of Pile Foundations. J. Geotech. Eng. Div.102, 197228.

Niazi, F.S.; Mayne, P.W. CPTu-based enhanced UniCone method for pile capacity. *Eng. Geol.* **2016**, *212*, 21–34.

Shahin, M.A. (2010). Intelligent computing for modeling axial capacity of pile foundations. Can. Geotech. J. 47,230–243.

Shariatmadari, N.; Eslami, A.A.; Karim, P.F.M. (2008). Bearing capacity of driven piles in sands from SPT applied to 60 case histories. Iran. J. Sci. Technol. Trans. B Eng. 32, 125–140.

Singh, S., Puri, V., & Kumar, P. (2019). Support vector machine-based prediction of pile bearing capacity. Geotechnical and Geological Engineering, 37(6), 4989-4998.

Wu, X., Hu, H., Fan, X., & Wu, W. (2020). A random forest model for predicting bearing capacity of single piles in different soil types. Engineering Applications of Artificial Intelligence, 94, 103806.