TẠP CHÍ KHOA HỌC & CÔNG NGHỆ ĐẠI HỌC DUY TÂN
DTU Journal of Science and Technology 02(45) (2021) 16-22
TRƯỜNG ĐẠI HỌC DUY TÂN
DUYTAN UNIVERSITY

Iteratively reweighted multiple linear regression with applications in civil engineering data modeling

Phân tích hồi quy có trọng số với các ứng dụng vào mô phỏng dữ liệu trong ngành xây dựng

Hoang Nhat Duc[a,b*]
Hoàng Nhật Đức[a,b*]

[a]*Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam*
[a]*Viện Nghiên cứu và Phát triển Công nghệ Cao, Trường Đại học Duy Tân, Đà Nẵng*
[b]*Faculty of Civil Engineering, Duy Tan University, Da Nang, 550000, Vietnam*
[b]*Khoa Xây dựng, Trường Đại học Duy Tân, Đà Nẵng*

(Ngày nhận bài: 27/01/2021, ngày phản biện xong: 22/02/2021, ngày chấp nhận đăng: 15/04/2021)

Abstract

Regression analysis is an important task in civil engineering which depends significantly on knowledge extracted from experimental data. This study develops a computer program and implements the iteratively reweighted least squares (IRLS) used for fitting multiple linear regression models. The capability of the combined model, IRMLR, is demonstrated using two artificial datasets and two real-world applications. The results indicate that the IRMLR model can be a useful tool to assist civil engineers in the task of data modeling.

Keywords: Regression analysis; Iteratively Reweighted Least Squares; Civil engineering; Linear regression.

Tóm tắt

Phân tích hồi quy là một nhiệm vụ quan trọng trong ngành kỹ thuật xây dựng vốn phụ thuộc đáng kể vào kiến thức rút ra từ dữ liệu thực nghiệm. Nghiên cứu này phát triển một chương trình tính toán dựa trên thuật toán bình phương nhỏ nhất có trọng số lặp lại (IRLS) được sử dụng để xây dựng các mô hình hồi quy tuyến tính. Tính ứng dụng của mô hình kết hợp, IRMLR, được minh chứng qua hai bộ dữ liệu mô phỏng và hai ứng dụng thực tế. Do đó, mô hình hồi quy dựa trên IRMLR có thể là một công cụ hữu ích để hỗ trợ các kỹ sư dân dụng trong công việc mô hình hóa dữ liệu.

Từ khóa: Phân tích hồi quy; Bình phương nhỏ nhất có trọng số lặp lại; Kỹ thuật xây dựng; Hồi quy tuyến tính.

1. Introduction

In the field of civil engineering, the task of mining knowledge hidden in experimental data is crucial. Herein, the problem of non-linear function approximation is particularly important for many civil engineering sub-fields such as structural engineering [1-3], hydraulic engineering [4-6], construction material [7-9], building energy [10], etc. Based on collected datasets, analyzers can examine the associations

* *Corresponding Author:* Hoang Nhat Duc; Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam; Faculty of Civil Engineering, Duy Tan University, Da Nang, 550000, Vietnam
Email: hoangnhatduc@duytan.edu.vn

among variables. Knowledge regarding these associations can be very helpful for predicting quantities or variables of interest [11-13].

Regression analysis models provide a sound method for appraising association among variables and for deriving a robust predictive equation/formula used for prediction [14, 15]. The conventional linear regression is a basic model. It is still widely used for examining variable associations as well as constructing simple and transparent predictive equations. Although the prediction accuracy of this approach is inferior to sophisticated nonlinear methods (e.g. artificial neural network [16, 17], piecewise linear regression [3, 18], support vector regression [19-22], etc.), the linear regression often serves as a base model used for result comparison and inspecting the nonlinearity property of datasets.

The traditional ordinary least squares (OLS) approach is widely employed to estimate the parameters of linear regression models. However, this approach is sensitive to outliers and the prediction performance of the models established by the OLS can be poor in the testing phase [11]. To deal with such issues, researchers often resort to the iteratively reweighted least squares (IRLS) as an alternative to the OLS [14]. The IRLS is a robust regression approach to reduce the influence of outliers by using a weight value associated with a data point [23]. Although the IRLS is a well-known method in statistics, its applications in the field of civil engineering are somehow limited. Therefore, this study develops a computer program based on the IRLS and applies this program in modeling several experimental datasets. The computer program is developed using Visual C# .NET and the performance of the IRLS model is compared to that of the model built by the OLS approach.

2. Iteratively Reweighted Multiple Linear Regression

A general linear regression model can be stated as follows [11]:

$$Y = X\beta + \varepsilon \qquad (1)$$

where Y is the response variable. X is the explanatory variable used to derive predictions of Y.

We define a diagonal matrix W within which entries inside its main diagonal is the weight associated with individual data samples. An entry inside the main diagonal of W is computed as follows:

$$w_i = \frac{1}{\max(\delta, |\Delta_i|)} \qquad (2)$$

where Δ_i denotes residual of the i^{th} data sample and δ = 0.0001 is a small number used for ensuring numerical stability.

When W is available, we can estimate the model parameter β as follows [23]:

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y \qquad (3)$$

The calculation steps of an Iteratively Reweighted Multiple Linear Regression (IRMLR) model can be summarized in **Fig. 1**.

While convergence is not achieved

(i) Initialize W with w_i = 1 and estimate Δ_i using the OLS method

(ii) Update W according to Eq. (2)

(iii) Compute the model parameter β according to Eq. (3)

(iv) Evaluate convergence status

End While

Return the IRMLR model

Fig. 1. The IRMLR construction phase

It is noted that to evaluate the convergence of the model construction phase, this study employs the following criteria:

Criteria 1: $|\Theta|_2 < 0.00001$ (4)

where Θ denotes the difference between 2 consecutive vector β .

Criteria 2: *iter* < *MaxIterNumber* (5)

where *iter* and *MaxIterNumber* denote the current number of iterations and the pre-specified maximum number of iterations.

3. Model Applications

In this section of the article, we develop the aforementioned IRMLR model with Visual C#. NET and apply it to solve 5 data modeling tasks. The C# code used for constructing the model is written by the author. The procedures used for training and implementing an IRMLR model are illustrated in **Fig. 2**. The program has been developed and compiled via a Console App (.NET framework 4.7.2). The program interface is demonstrated in **Fig. 3.**

```
var Train_Result = Train_IRL(Xtr, Ttr, MaxIter, 0.00001);
var Model = Train_Result[0];

var TrackTrainingMse = Train_Result[1];
MyMatrix.WriteMatrixToCsvFile(TrackTrainingMse, SaveResultLoc +
    "TrackTrainingwRmse.csv");

var Ytr = Predict(Model, Xtr);
var Yte = Predict(Model, Xte);
```

Fig. 2. Illustration of the model training and prediction processes

```
C:\Users\Administrator\source\repos\ConsoleVS_Study\ConsoleVS_Study\bin\Debug\ConsoleVS_Study.exe
Iteratively reweighted least squares for regression analysis
hoangnhatduc@duytan.edu.vn
Enter Training Data Location: (eg. D:/csvFiles/App1_TrainData.csv)
D:/csvFiles/App1_TrainData.csv
Enter Testing Data Location: (eg. D:/csvFiles/App1_TestData.csv)
D:/csvFiles/App1_TestData.csv
Enter Location for Saving Result: (eg. D:/csvFiles/Result_IRLS_MLR_App1/)
D:/csvFiles/Result_IRLS_MLR_App1/
Enter Max. Number of Training Iteration : (eg. 10)
10
Training Phase: RMSE-MAE-MAPE-R^2
0.0382425482296473
0.0230174324540203
1.9775802123351
0.917742490623848
Testing Phase: RMSE-MAE-MAPE-R^2
0.0218571224969559
0.0177033383356862
1.4081268662125
0.978136510837361
Execution time = 15538 (ms).
```

Fig. 3 The program interface

3.1. Application 1

In the first application, the model is applied to model an artificial dataset generated via the following rule:

$$Y = 0.5X + 1 + r/50 \quad (6)$$

where r denotes a Gaussian random variable with mean = 0 and standard deviation = 1.

The noise components of the 1th, 10th, and 15th data samples are intentionally inflated to mimic the effect of outlier as follows:

$$Y = 0.5X + 1 + r/10 \quad (7)$$

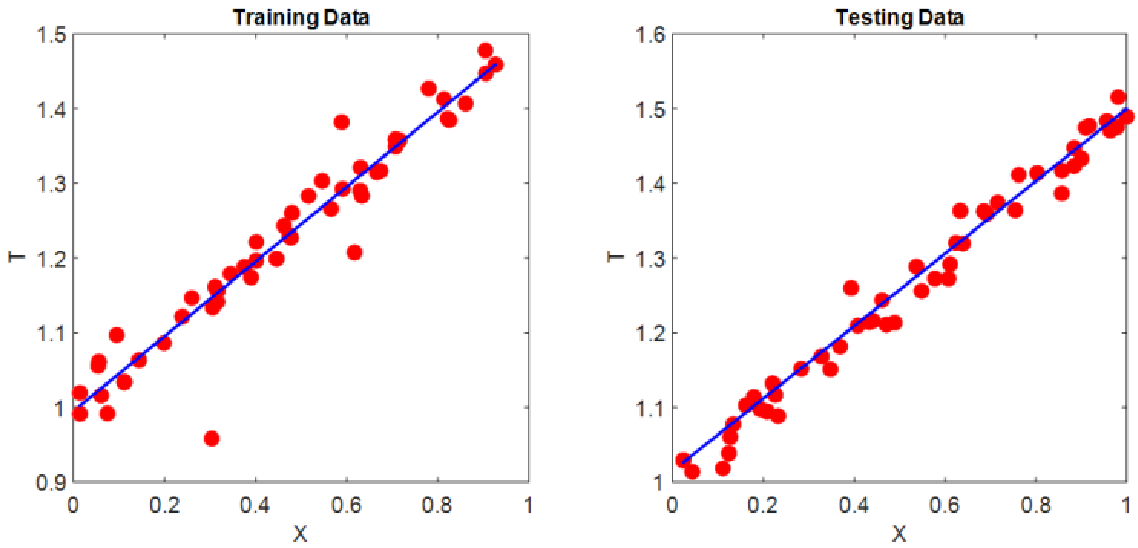

Fig. 4. Prediction results of the 1st application

The IRMLR founds the model parameter as follows: β_0 = 0.500531 and β_1 = 0.99499. The prediction result is graphically shown in **Fig. 4**. It is noted that both of the training and testing datasets include 50 samples. The testing RMSE, MAPE, and R^2 of the IRMLR are 0.021667, 1.391198%, and 0.978351, respectively. These outcome is better than those of the OLS method with RMSE = 0.022105, MAPE = 1.420004%, and R^2 = 0.977568.

3.2 Application 2

In the second application, the model is applied to model another artificial dataset generated via the following rule:

$$Y = 3.5 + 2X_1 - 3X_2 + r/5 \quad (8)$$

where r denotes a Gaussian random variable with mean = 0 and standard deviation = 1.

The noise components of the 1th, 10th, and 15th data samples are multiplied by 1.2 to imitate the effect of outliers on the model training process. The IRMLR founds the model parameter as follows: β_0 = 3.768626, β_1 = 1.75307, and β_2 = -3.26391. The prediction result is graphically shown in **Fig. 5**. The testing RMSE, MAPE, and R^2 of the IRMLR are 0.200155, 7.218727%, and 0.971423, respectively. This performance is better than that of the OLS method with RMSE = 0.279036, MAPE = 10.16683%, and R^2 = 0.954332.

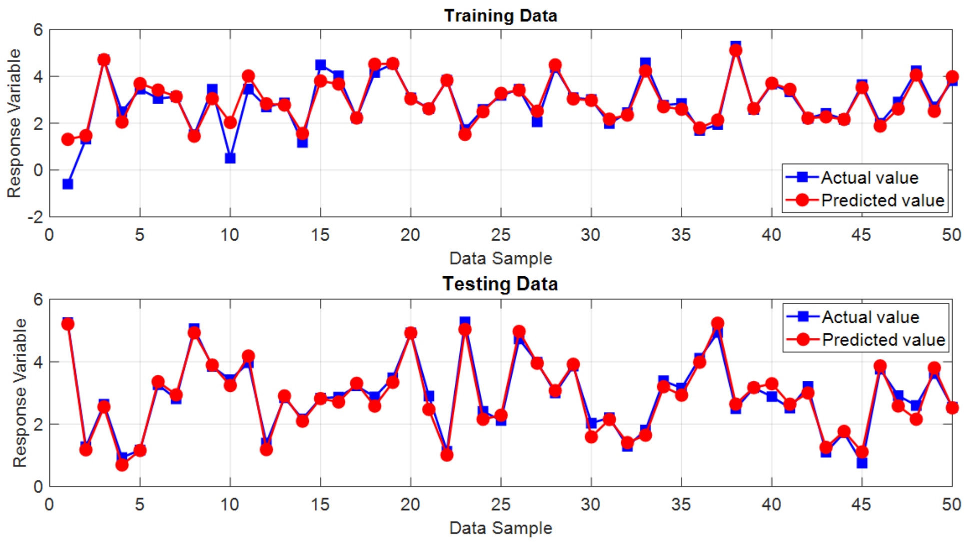

Fig. 5. Prediction results of the 2nd application

3.3. Application 3

In this application, the model is used for predicting the undrained lateral load capacity of pile in clay. The dataset is collected by [24] and compiled in [25]. Diameter of pile, depth of pile embedment, eccentric of load, and undrained shear strength of soil are explanatory variables. This dataset consists of 38 samples. This study employs 35 samples as training data and 4 samples as testing data. The training and prediction results of the IRMLR model are shown in **Fig. 6**. The testing performance of this model (RMSE = 6.146797, MAPE = 11.2392%, and R^2 = 0.811734) is better than that of the OLS based model (RMSE = 9.987739, MAPE = 19.04094%, and R^2 = 0.568654).

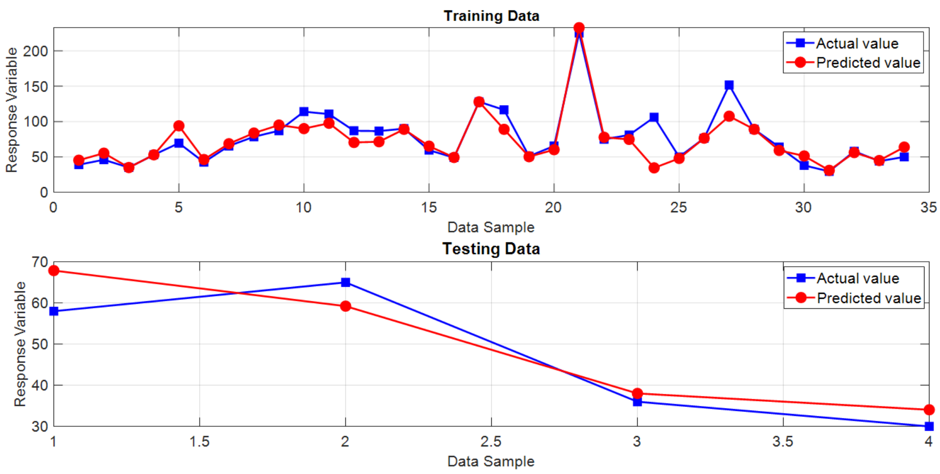

Fig. 6. Prediction results of the 3rd application

3.4 Application 4

In this application, a data set of soil shear strength experiment [26] is used to train and validate the prediction model. The dataset has been collected during the geotechnical investigation phase of the Le Trong Tan Geleximco Project. The explanatory factors are (1) depth of sample (m), (2) sand percentage (%), (3) loam percentage (%), (4) clay percentage (%), (5) moisture content percentage (%), (6) wet density (g/cm3), (7) dry density (g/cm^3), (8) void ratio, (9) liquid limit (%),

(10) plastic limit (%), (11) plastic index (%), and (12) liquidity index. The soil shear strength is the modeled variable. In total, there are 249 samples used for model training and testing (refer to **Fig. 7**). The testing performance of the IRMLR model is as follows: RMSE = 0.035912, MAPE = 8.07538%, and R^2 = 0.780765. The testing performance of the OLS based model is as follows: RMSE = 0.036549, MAPE = 8.178282%, and R^2 = 0.755294.

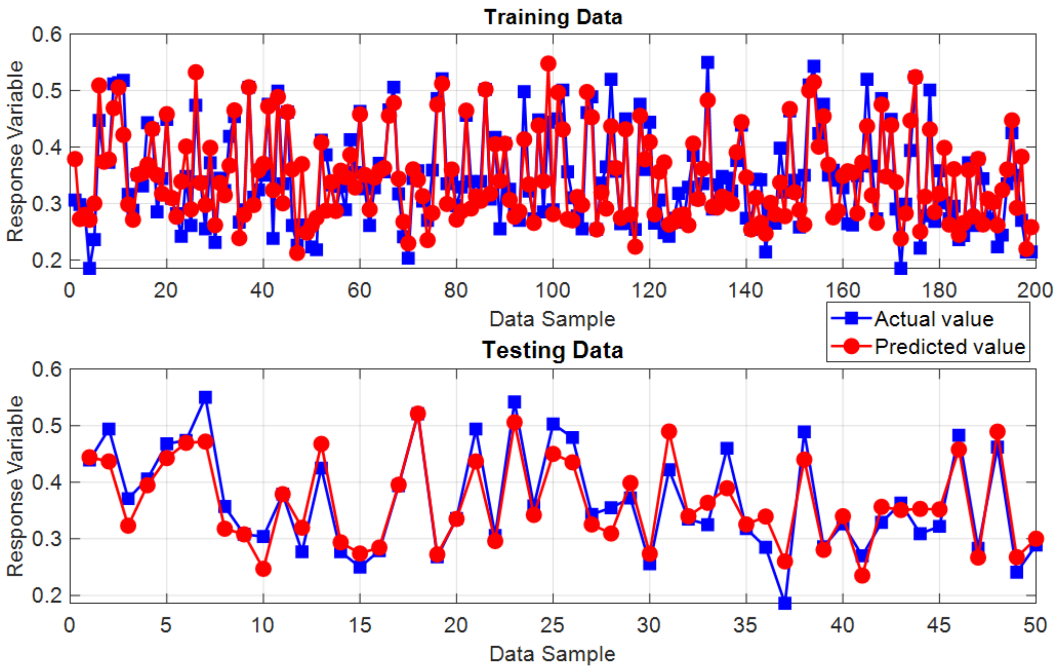

Fig. 7. Prediction results of the 5th application

4. Concluding remarks

Regression analysis is a crucial task in civil engineering which depends significantly on knowledge extracted from experimental data. Compared to the OLS based model, the IRMLR shows better capability in dealing with noisy data. Using experimental results with two artificial datasets and two real-world applications, the advantages of the IRMLR is clearly demonstrated. To facilitate the implementation of the model, this study has developed a computer program that is based on the IRMLR model. The program has been constructed with Visual C# .NET and can be openly downloaded. Future extensions of the current work may include the applying the IRMLR program to model other datasets in the field of civil engineering and investigation of other sophisticated approaches used for robust regression.

Supplementary materials

The compiled program and the experimental datasets can be accessed via: https://github.com/NDHoangDTU/IRWLS_MLR

References

[1] H. D. Nguyen, Q. Zhang, E. Choi, and W. Duan, "An improved deflection model for FRP RC beams using an artificial intelligence-based approach," *Engineering Structures*, vol. 219, p. 110793, 2020/09/15/ 2020.

[2] D. Prayogo, M.-Y. Cheng, Y.-W. Wu, and D.-H. Tran, "Combining machine learning models via adaptive ensemble weighting for prediction of shear capacity of reinforced-concrete deep beams," *Engineering with Computers*, April 30 2019.

[3] N.-D. Hoang, "Estimating Punching Shear Capacity of Steel Fibre Reinforced Concrete Slabs Using Sequential Piecewise Multiple Linear Regression

and Artificial Neural Network," *Measurement,* vol. 137, pp. 58-70, 2019/01/18/ 2019.

[4] K.-W. Liao, N.-D. Hoang, and F.-S. Chien, "A Multi-Hazard Safety Evaluation Framework for a Submerged Bridge using Machine Learning Model," ed, 2019.

[5] N.-D. Hoang, K.-W. Liao, and X.-L. Tran, "Estimation of scour depth at bridges with complex pier foundations using support vector regression integrated with feature selection," *Journal of Civil Structural Health Monitoring,* June 02 2018.

[6] J.-S. Chou, W. K. Chong, and D.-K. Bui, "Nature-Inspired Metaheuristic Regression System: Programming and Implementation for Civil Engineering Applications," *Journal of Computing in Civil Engineering,* vol. 30, p. 04016007, 2016.

[7] E. M. Golafshani and A. Behnood, "Application of soft computing methods for predicting the elastic modulus of recycled aggregate concrete," *Journal of Cleaner Production,* vol. 176, pp. 1163-1176, 2018/03/01/ 2018.

[8] A. Goetzke-Pala, A. Hoła, and Ł. Sadowski, "A non-destructive method of the evaluation of the moisture in saline brick walls using artificial neural networks," *Archives of Civil and Mechanical Engineering,* vol. 18, pp. 1729-1742, 2018/09/01/ 2018.

[9] D.-K. Bui, T. Nguyen, J.-S. Chou, H. Nguyen-Xuan, and T. D. Ngo, "A modified firefly algorithm-artificial neural network expert system for predicting compressive and tensile strength of high-performance concrete," *Construction and Building Materials,* vol. 180, pp. 320-333, 2018/08/20/ 2018.

[10] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-Term Residential Load Forecasting Based on Resident Behaviour Learning," *IEEE Transactions on Power Systems,* vol. 33, pp. 1087-1088, 2018.

[11] F. A. Graybill and H. K. Iyer, *Regression Analysis: Concepts and Applications*: Duxbury Pr, 1994.

[12] H. Agrawal and A. K. Mishra, "Modified scaled distance regression analysis approach for prediction of blast-induced ground vibration in multi-hole blasting," *Journal of Rock Mechanics and Geotechnical Engineering,* vol. 11, pp. 202-207, 2019/02/01/ 2019.

[13] K. O. Achieng, "Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural networks vs support vector regression models," *Computers & Geosciences,* vol. 133, p. 104320, 2019/12/01/ 2019.

[14] W. Mendenhall and T. T. Sincich *A Second Course in Statistics: Regression Analysis (7th Edition)*: Pearson, 2011.

[15] S. Weisberg, *Applied Linear Regression, Third Edition*: John Wiley & Sons, Printed in the United States of America, 2005.

[16] M. Mishra, A. Agarwal, and D. Maity, "Neural-network-based approach to predict the deflection of plain, steel-reinforced, and bamboo-reinforced concrete beams from experimental data," *SN Applied Sciences,* vol. 1, p. 584, 2019/05/17 2019.

[17] T.-H. Tran and N.-D. Hoang, "Predicting Colonization Growth of Algae on Mortar Surface with Artificial Neural Network," *Journal of Computing in Civil Engineering,* vol. 30, p. 04016030, 2016.

[18] N. D. Hoang and C. H. Le, "Sequential Piecewise Linear Regression software program for nonlinear regression analysis in structural engineering," *DTU Journal of Science and Technology,* vol. 05, pp. 03-09, 2019.

[19] T.-D. Nguyen, T.-H. Tran, H. Nguyen, and H. Nhat-Duc, "A success history-based adaptive differential evolution optimized support vector regression for estimating plastic viscosity of fresh concrete," *Engineering with Computers,* December 18 2019.

[20] T.-H. Tran and N.-D. Hoang, "Estimation of algal colonization growth on mortar surface using a hybridization of machine learning and metaheuristic optimization," *Sādhanā,* vol. 42, pp. 929-939, June 01 2017.

[21] S. Heddam and O. Kisi, "Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree," *Journal of Hydrology,* vol. 559, pp. 499-509, 2018/04/01/ 2018.

[22] T.-D. Nguyen, T.-H. Tran, and N.-D. Hoang, "Prediction of interface yield stress and plastic viscosity of fresh concrete using a hybrid machine learning approach," *Advanced Engineering Informatics,* vol. 44, p. 101057, 2020/04/01/ 2020.

[23] C. S. Burrus, "Iterative Reweighted Least Squares," *OpenStax-CNX module: m45285* 2012.

[24] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological Modelling,* vol. 178, pp. 389-397, 2004/11/01/ 2004.

[25] S. K. Das and P. K. Basudhar, "Undrained lateral load capacity of piles in clay using artificial neural network," *Computers and Geotechnics,* vol. 33, pp. 454-459, 2006/12/01/ 2006.

[26] M.-T. Cao, N.-D. Hoang, V. H. Nhu, and D. T. Bui, "An advanced meta-learner based on artificial electric field algorithm optimized stacking ensemble techniques for enhancing prediction accuracy of soil shear strength," *Engineering with Computers,* 2020/11/02 2020.