Investigation of the Gaussian Process with Various Kernel Functions for the Prediction of the Compressive Strength of Concrete

Hoang Ha
University of Transport and Communications, Ha Noi, Vietnam
hoangha.utc@gmail.com

Hieu Vu Trong
University of Transport and Technology, Thanh Xuan, Hanoi, Vietnam
vipnoxhd123@gmail.com

Trang Le Huyen
University of Transport and Technology, Thanh Xuan, Hanoi, Vietnam
lehuyentrang0500@gmail.com

Dam Duc Nguyen
University of Transport and Technology, Thanh Xuan, Hanoi, Vietnam
damnd@utt.edu.vn

Indra Prakash
Geological Survey of India, Gandhinagar, Gujarat, India
indra52prakash@gmail.com

Binh Thai Pham
University of Transport and Technology, Thanh Xuan, Hanoi, Vietnam
binhpt@utt.edu.vn (corresponding author)

Received: 28 September 2024 | Revised: 15 November 2024 | Accepted: 27 November 2024

 DOI: https://doi.org/10.48084/etasr.9125

ABSTRACT

The Compressive Strength of Concrete (CSC) is a critical parameter for evaluating the quality of concrete used in various construction projects, including buildings, bridges, and roads. The primary objective of this study is to examine the efficacy of a Gaussian Process (GP) Machine Learning (ML) model employing two kernel functions: Radial Basis Function (RBF) and Polynomial (POL), for predicting the CSC, considering readily quantifiable parameters. Based on these kernel functions, two models were created for this prediction, GP-RBF and GP-POL. The modeling process employed a total of 369 concrete sample data, including compressive strength values and eleven other physico-mechanical properties, collected from the Cua Luc bridge project in Vietnam. This dataset was partitioned into a training set (70%) and a testing set (30%) for model training and validation. Various validation metrics, including R2, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), were used to evaluate and compare the models. The findings of this study demonstrated that both models GP-RBF and GP-POL exhibited strong performance in predicting CSC, with GP-POL demonstrating marginal superiority over GP-RBF. Consequently, it can be concluded that POL is more efficacious than RBF in training the GP model for CSC prediction.

Keywords-machine learning; concrete; compressive strength; Gaussian process; kernel function

I. INTRODUCTION

An accurate prediction of the CSC is of critical importance for ensuring the safety and durability of concrete structures. The compressive strength of concrete, a crucial metric for evaluating its quality, is influenced by a multitude of factors, including the proportions of the mixture, the conditions during curing, and the inherent properties of the materials [1-4]. Consequently, the estimation of this parameter poses a substantial challenge. Conventional prediction methods frequently depend on empirical formulas and require extensive experimental testing, which is not only time-consuming but also resource-intensive. Consequently, there has been a growing interest in leveraging advanced computational techniques to enhance prediction accuracy and efficiency [5]. ML techniques have emerged as a significant advancement in this field, providing robust tools for prediction in various domains [6-8], including the prediction of the properties of construction materials [9, 10]. Among these, GP regression, a non-parametric Bayesian technique, has emerged as a prominent method for modeling complex, non-linear relationships in various fields [11, 12]. The efficacy of GP models is contingent on the selection of an appropriate kernel function, which delineates the covariance structure of the data and exerts a significant influence on the model's capacity to generalize from the training data. The selection of an appropriate kernel function [13], is a critical step in the modeling process, as different kernel functions are capable of capturing distinct types of relationships and patterns in the data [14]. The objective of this study is to address this knowledge gap by investigating the performance of various kernel functions within the GP regression framework for predicting the CSC. The focus is on a comparative analysis of the RBF and POL kernels, with an assessment of their prediction accuracy and robustness through extensive experimentation on concrete strength datasets from real-world contexts. The modeling process involved the utilization of a comprehensive dataset comprising 369 concrete samples, encompassing their compressive strength and an additional 11 physico-mechanical properties, obtained from the Cua Luc bridge project in Vietnam. This dataset was apportioned for the construction of training (70%) and testing (30%) sets, which were employed for the training and validation of the models. Various validation metrics, including R^2, RMSE, and MAE, were employed to assess and compare the models. The findings of this study will contribute to a more comprehensive understanding of how different kernel functions impact the predictive capabilities of GP models. This knowledge will aid in the development of more accurate and efficient predictive models for concrete compressive strength, ultimately enhancing the design and safety of concrete structures.

II. MATERIALS AND METHOD

A. Data Used

In the context of regression modeling, two primary variables must be determined: the dependent variable, which corresponds to the output, and the independent variables, which correspond to the input. In the context of this study, the CSC is designated as the dependent variable, while the other physico-mechanical properties of concrete are defined as independent variables. The generation of databases for the modeling of prediction of the CSC prediction involved the usage of a dataset comprising 369 concrete samples collected from the Cua Luc bridge project in Vietnam. A total of 70% of the data was selected for the generation of the training dataset, which was used to train the models, while the remaining 30% was utilized for the generation of the testing dataset, which was used to validate the models. The prediction of the CSC was facilitated by the employment of eleven independent variables, namely the age of concrete, cement content, coarse aggregate 10 mm × 20 mm, coarse aggregate 5 mm × 10 mm, natural sand content, water content, superplasticizer admixture content, silica-fume admixture content, slump, water to cement ratio, and aggregate to cement ratio. The initial data analysis of the parameters used in this study is presented in Table I. The data processing and modeling were conducted deploying the Weka software.

TABLE I. DESCRIPTIVE STATISTICS OF THE STUDY'S INPUT AND OUTPUT

No	Parameters	Unit	Min	Max	Average	STD
1	Age of concrete	(day)	3	28	20	10
2	Cement content	(kg)	230	20	379	81
3	Coarse aggregate 10 mm × 20 mm	(kg)	715	800	755	20
4	Coarse aggregate 5 mm × 10 mm	(kg)	270	370	329	22
5	Natural sand content	(kg)	692	850	781	45
6	Water content	(l)	140	195	160	13
7	Superplasticizer admixture	(l)	2	5	4	1
8	Silica-fume admixture	(kg)	0	26	11	11
9	Slump	(mm)	7	18	14	4
10	Water to cement ratio	-	0	1	0	0
11	Aggregate to cement ratio	-	3	9	5	1
12	CSC	(MPa)	14	73	38	14

B. Gaussian Process

GP is a powerful framework for modeling distributions over functions [15]. It is frequently employed for various regression, classification, and uncertainty quantification tasks. At the core of the Gaussian processes is the notion of treating functions as random variables. Given a set of input-output pairs, training data, a GP defines a distribution over functions, which is consistent with the observed data. This distribution is characterized by a mean function and a covariance function, also known as a kernel function, which specifies how the outputs at different input points are correlated. One of the key strengths of GP is its flexibility in modeling complex data patterns while providing principled uncertainty estimates. The choice of the kernel function plays a crucial role in shaping the characteristics of the inferred functions. There are two main types of kernel functions used in developing the GP model:

$$k_{RBF}(x, x') = exp(-\frac{|x-x'|^2}{2l^2}) \quad (1)$$

where l is the length scale of the kernel, x and x' are two data points, and:

$$k_{Poly}(x, x') = (x^T x' + c)^d \quad (2)$$

where d is the degree of the POL and c is a constant term. GPs offer a versatile and powerful approach to modeling and predicting the behaviour of construction materials. Their

capacity to manage intricate relationships and furnish uncertainty estimates renders them especially beneficial in ensuring the safety, durability, and sustainability of construction projects. By leveraging GPs, engineers and researchers can make more informed decisions, optimize material usage, and enhance the overall quality and performance of construction materials. In addition to the kernel functions employed for training the GP model, other hyperparameters were used, as shown in Table II.

TABLE II. THE PARAMETERS USED FOR THE MODEL DEVELOPMENT

No	Hyper-parameters	GP-POL	GP-RBF
1	Batch size	100	100
2	Debug	False	False
3	Filter Type	Normalize	Standardize
4	Kernel	Poly Kernel	RBF Kernel
5	Noise	1.0	1.0
6	Num Decimal Places	2	2
7	Seed	1	1
8	Cache Size	250,007	250,007
9	Exponent	1.0	1.0
10	Use Lower Order	False	False

C. Validation Metrics

In order to validate and compare the regression models, three main validation metrics are often used: the determination coefficient (R^2), the MAE, and the RMSE. The following is a description of these metrics: R^2 is a statistical measure used in regression analysis to assess the goodness of fit of a model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variable (s). R^2 values range from 0 to 1, where 0 indicates that the model does not explain any of the variance and 1 indicates that the model explains all the variance in the dependent variable [16]. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where y_i represents the actual values of the dependent variable, $\hat{y}_i$ represents the predicted values of the dependent variable by the regression model, and $\overline{y_i}$ is the mean of the actual values of the dependent variable. MAE is a widely used metric in regression analysis to measure the accuracy of a model in predicting continuous outcomes [17]. MAE quantifies the average magnitude of the errors between the predicted values and actual values, providing a straightforward interpretation of prediction accuracy [18]. A distinguishing feature of MAE is its equitable treatment of all individual differences between the predicted and actual values, a characteristic that lends it resilience as a metric for evaluating model performance. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

where n is the number of data points, y_i represents the actual value of the dependent variable for the i-th observation, and $|y_i - \hat{y}_i|$ represents the predicted value of the dependent variable for the i-th observation. RMSE is a frequently utilized metric for evaluating the accuracy of a regression model. It quantifies the mean absolute deviation between the predicted values and the actual values, thereby offering insights into the model's predictive capability [18]. A notable advantage of RMSE is its sensitivity to outliers, as it places greater emphasis on larger errors [19]. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

where n is the number of observations, y_i represents the actual values of the dependent variable, and $\hat{y}_i$ represents the predicted values of the dependent variable by the regression model.

III. RESULTS AND ANALYSIS

The GP-POL and GP-RBF models were trained and validated using a training and testing dataset, respectively, for the purpose of predicting the CSC. The selection of the optimal hyperparameters for the GP models was guided by the parameters enumerated in Table I, with the objective of attaining the most efficacious performance of the models. The outcomes of the training and validation processes are presented in Figures 1-3. Figure 1 displays the plots of the actual values obtained from the experimental test versus the predicted values attained from the prediction models. It is evident that the actual and predicted values of the GP-RBF model are more proximate than those of the GP-POL model for the training dataset. However, the actual and predicted values of the GP-POL model are more proximate than those of the GP-RBF model. This observation aligns with the error performance metrics depicted in Figure 2. Figure 3 portrays the plots depicting the relationship between the actual and predicted values of the models, with the R^2 values of the models also indicated. It is evident from this analysis that while the R^2 value of the GP-RBF model exceeds that of the GP-POL model for the training dataset, the R2 value of the GP-RBF model is lower than that of the GP-POL model for the testing dataset. Consequently, it can be concluded that the GP-POL model exhibits superior predictive capabilities compared to the GP-RBF model in predicting the CSC.

IV. DISCUSSION AND CONCLUSIONS

The selection of kernel functions has been demonstrated to exert a substantial influence on the efficacy of the Gaussian Process (GP) models [15]. The Radial Basis Function (RBF) kernel is a widely used and effective option for many applications due to its flexibility and smoothness assumptions. However, the RBF kernel can be more effective for data with inherent Polynomial (POL) structures [19]. A recommended approach involves the empirical evaluation of diverse kernels to identify the most suitable one for a given problem. In this study, the GP was trained and validated with two types of kernel functions, RBF and POL, to generate two models: GP-RBF and GP-POL, respectively, for prediction of the Compressive Strength of Concrete (CSC) based on eleven input parameters, such as age of concrete, cement content, coarse aggregate 10 mm × 20 mm, coarse aggregate 5 mm × 10 mm, natural sand content, water content, superplasticizer admixture content, silica-fume admixture content, slump, water to cement ratio, and aggregate to cement ratio. The database, which was collected from the Cua Luc bridge project in

Vietnam, was used for the generation of training and validating datasets for the models. The validation and comparison of the models was conducted utilizing three common validation metrics: R^2, RMSE, and MAE.

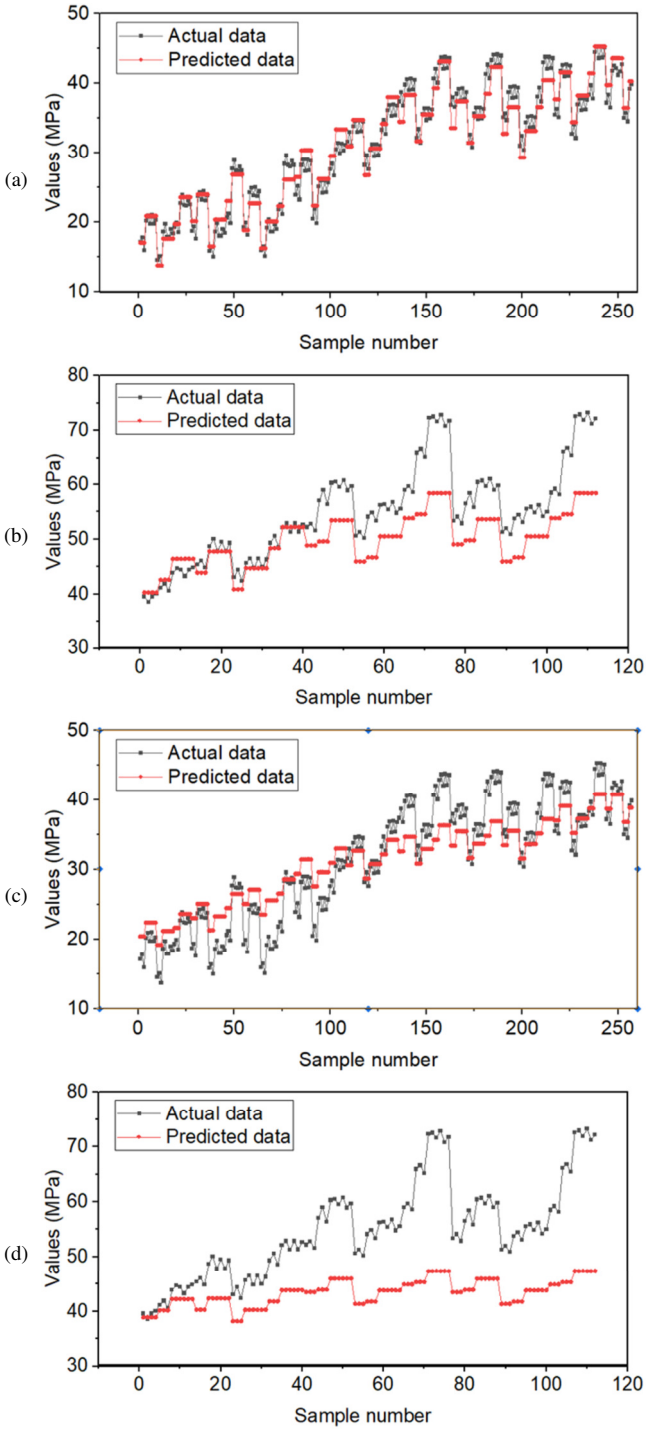


Fig. 1. Predicted versus measured values of the compressive strength using the applied models: (a) training GP-POL, (b) testing GP-POL, (c) training GP-RBF, (d) testing GP-RBF.

The findings of this study demonstrated that both the GP-RBF and GP-POL models exhibited commendable performance in predicting the CSC.

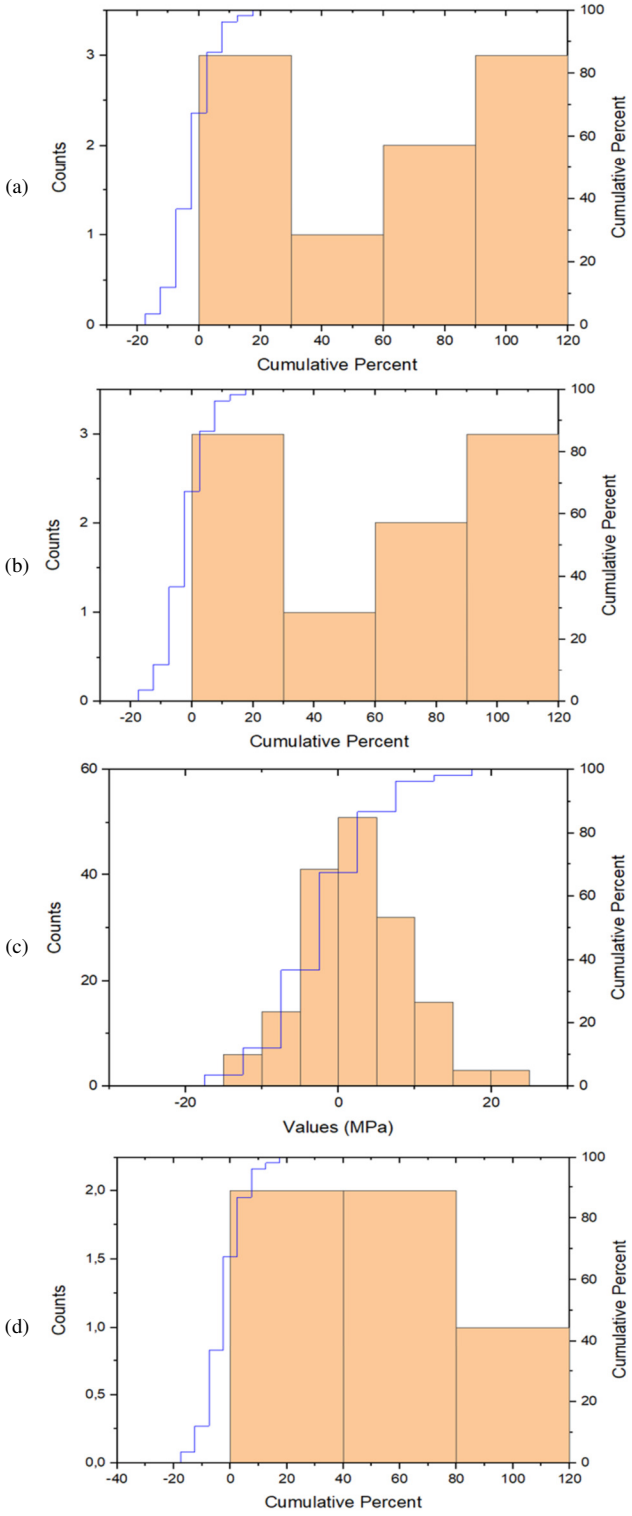
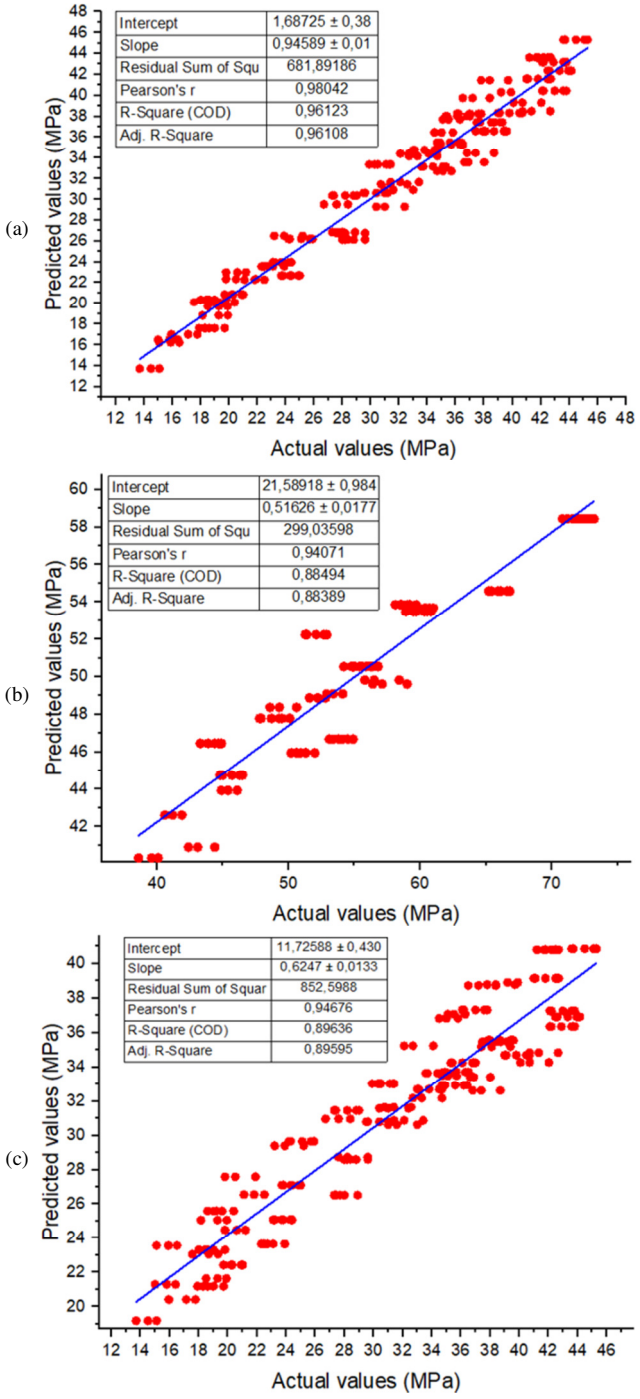


Fig. 2. Error analysis of the applied models: (a) training GP-POL, (b) testing GP-POL, (c) training GP-RBF, (d) testing GP-RBF.

However, it was observed that the GP-POL model demonstrated a slight edge in its predictive capabilities over the GP-RBF model. Therefore, it can be concluded that the performance of GP is significantly affected by the selection of kernel functions. In this study, the POL kernel function was found to be more effective than the RBF kernel function in training the GP model for predicting the CSC.



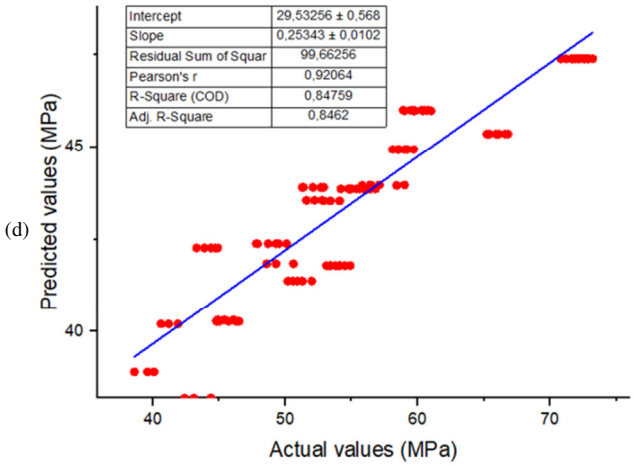


Fig. 3. R analysis using the applied models: (a) training GP-POL, (b) testing GP-POL, (c) training GP-RBF, (d) testing GP-RBF.

In the context of predicting the CSC, the usage of a POL kernel within a GP model can offer numerous advantages over an RBF kernel. For instance, the POL kernel is well-suited for capturing nonlinear relationships that are characteristic of concrete data. Furthermore, the POL kernel introduces model complexity more gradually than the RBF kernel, thereby reducing the risk of overfitting, particularly in the presence of noisy or limited data. The flexibility of the POL kernel to accommodate both linear and nonlinear effects, depending on the selected degree, renders it a versatile tool for a range of relationship complexities in the data. The findings of this study demonstrate that the GP-POL model is a promising tool for the rapid and accurate prediction of the CSC. However, to further enhance the model's performance, it is recommended to conduct a sensitivity analysis to evaluate the importance of the input variables. This analysis would facilitate the identification and removal of superfluous parameters, thereby enhancing the efficiency and precision of the model.

REFERENCES

[1] S. Ahmed, Z. Al-Dawood, F. Abed, M. A. Mannan, and M. Al-Samarai, "Impact of using different materials, curing regimes, and mixing procedures on compressive strength of reactive powder concrete - A review," *Journal of Building Engineering*, vol. 44, Dec. 2021, Art. no. 103238, https://doi.org/10.1016/j.jobe.2021.103238.

[2] C.-C. Vu, O. Plé, J. Weiss, and D. Amitrano, "Revisiting the concept of characteristic compressive strength of concrete," *Construction and Building Materials*, vol. 263, Dec. 2020, Art. no. 120126, https://doi.org/10.1016/j.conbuildmat.2020.120126.

[3] X. H. Vu, T. C. Vo, and V. T. Phan, "Study of the Compressive Strength of Concrete with Partial Replacement of Recycled Coarse Aggregates," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7191–7194, Jun. 2021, https://doi.org/10.48084/etasr.4162.

[4] S. A. Chandio, B. A. Memon, M. Oad, F. A. Chandio, and M. U. Memon, "Effect of Fly Ash on the Compressive Strength of Green Concrete," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5728–5731, Jun. 2020, https://doi.org/10.48084/etasr.3499.

[5] O. R. Abuodeh, J. A. Abdalla, and R. A. Hawileh, "Assessment of compressive strength of Ultra-high Performance Concrete using deep machine learning techniques," *Applied Soft Computing*, vol. 95, Oct. 2020, Art. no. 106552, https://doi.org/10.1016/j.asoc.2020.106552.

[6] I. Prakash, D. D. Nguyen, N. T. Tuan, T. V. Phong, and L. V. Hiep, "Landslide Susceptibility Zoning: Integrating Multiple Intelligent

Models with SHAP Analysis," *Journal of Science and Transport Technology*, vol. 4, no. 1, pp. 23–41, Mar. 2024, https://doi.org/10.58845/jstt.utt.2024.en.4.1.23-41.

[7] M. V. Le, I. Prakash, and D. D. Nguyen, "Predicting Load-Deflection of Composite Concrete Bridges Using Machine Learning Models," *Journal of Science and Transport Technology*, vol. 3, no. 4, pp. 43–51, Dec. 2023, https://doi.org/10.58845/jstt.utt.2023.en.3.4.43-51.

[8] Y. Reich, "Machine Learning Techniques for Civil Engineering Problems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 12, no. 4, pp. 295–310, 1997, https://doi.org/10.1111/0885-9507.00065.

[9] R. K. Tipu, Suman, and V. Batra, "Enhancing prediction accuracy of workability and compressive strength of high-performance concrete through extended dataset and improved machine learning models," *Asian Journal of Civil Engineering*, vol. 25, no. 1, pp. 197–218, Jan. 2024, https://doi.org/10.1007/s42107-023-00768-1.

[10] S. Ghani, N. Kumar, M. Gupta, and S. Saharan, "Machine learning approaches for real-time prediction of compressive strength in self-compacting concrete," *Asian Journal of Civil Engineering*, vol. 25, no. 3, pp. 2743–2760, Apr. 2024, https://doi.org/10.1007/s42107-023-00942-5.

[11] S. S. Ghosh, S. Dey, N. Bhogapurapu, S. Homayouni, A. Bhattacharya, and H. McNairn, "Gaussian Process Regression Model for Crop Biophysical Parameter Retrieval from Multi-Polarized C-Band SAR Data," *Remote Sensing*, vol. 14, no. 4, Jan. 2022, Art. no. 934, https://doi.org/10.3390/rs14040934.

[12] C. Zhang, H. Wei, X. Zhao, T. Liu, and K. Zhang, "A Gaussian process regression based hybrid approach for short-term wind speed prediction," *Energy Conversion and Management*, vol. 126, pp. 1084–1092, Oct. 2016, https://doi.org/10.1016/j.enconman.2016.08.086.

[13] T. Zhou and Y. Peng, "Kernel principal component analysis-based Gaussian process regression modelling for high-dimensional reliability analysis," *Computers & Structures*, vol. 241, Dec. 2020, Art. no. 106358, https://doi.org/10.1016/j.compstruc.2020.106358.

[14] Y. Pan, X. Zeng, H. Xu, Y. Sun, D. Wang, and J. Wu, "Evaluation of Gaussian process regression kernel functions for improving groundwater prediction," *Journal of Hydrology*, vol. 603, Dec. 2021, Art. no. 126960, https://doi.org/10.1016/j.jhydrol.2021.126960.

[15] S. Hwang, B. L'Huillier, R. E. Keeley, M. J. Jee, and A. Shafieloo, "How to use GP: effects of the mean function and hyperparameter selection on Gaussian process regression," *Journal of Cosmology and Astroparticle Physics*, vol. 2023, no. 02, Feb. 2023, Art. no. 014, https://doi.org/10.1088/1475-7516/2023/02/014.

[16] S. Menard, "Coefficients of Determination for Multiple Logistic Regression Analysis," *The American Statistician*, vol. 54, no. 1, pp. 17–24, 2000, https://doi.org/10.2307/2685605.

[17] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, Jul. 2016, https://doi.org/10.1016/j.ijforecast.2015.12.003.

[18] F. M. Talaat, A. Aljadani, B. Alharthi, M. A. Farsi, M. Badawy, and M. Elhosseini, "A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing," *Mathematics*, vol. 11, no. 18, Jan. 2023, Art. no. 3930, https://doi.org/10.3390/math11183930.

[19] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ. Computer Science*, vol. 7, 2021, Art. no. e623, https://doi.org/10.7717/peerj-cs.623.