Transactions on Machine Intelligence 9(1) (2026) 74–83

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir

Comparative Evaluation of Support Vector Machine and Multiple Linear Regression for Predicting the Compressive Strength of Fly Ash-Containing Concrete

S. R. SalimBahrami [1,*], M. Imani [2]

[1] Assistant Professor, Department of Civil Engineering, Faculty of Civil Engineering and Architecture, Technical and Vocational University (TVU), Tehran, Iran

[2] B.Sc. Student, Department of Civil Engineering, Faculty of Civil Engineering and Architecture, Technical and Vocational University (TVU), Tehran, Iran

ARTICLE INFO

Article History:
Received 14 January 2026
Received in revised form 1 March 2026
Accepted 18 March 2026
Available online 27 March 2026

Keywords:
Fly ash-containing concrete; compressive strength; five-fold cross-validation; Support Vector Machine; Multiple Linear Regression

ABSTRACT

Fly ash-containing concrete has attracted considerable attention as an effective approach to sustainable concrete development due to its potential to reduce cement consumption, reuse an industrial by-product, and mitigate the environmental impacts associated with cement production. However, the complex and nonlinear relationships among mix-design constituents, concrete age, and compressive strength make accurate strength prediction challenging. This study investigates and compares the performance of a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel and Multiple Linear Regression (MLR) for predicting the compressive strength of fly ash-containing concrete. For this purpose, a dataset comprising 471 laboratory observations with seven input variables representing mix-design parameters and concrete age was employed. The data were divided into training and testing sets, and five-fold cross-validation was performed to assess the stability and generalizability of the models. The SVM hyperparameters were optimized using grid search. The results demonstrated that SVM outperformed MLR in modeling the complex relationships governing compressive strength. Specifically, during the training phase, SVM achieved a correlation coefficient of 0.988, compared with 0.814 for MLR. The cross-validation results further confirmed the superior performance and greater stability of the SVM model. These findings highlight the potential of nonlinear machine-learning models for estimating compressive strength and supporting the design of sustainable fly ash-containing concrete.

1. INTRODUCTION

Concrete, as the most widely used engineered material in the construction industry, plays a fundamental role in the development of infrastructure and buildings. However, the extensive use of Portland cement in concrete

* Corresponding Author: rsalimbahrami@tvu.ac.ir
Assistant Professor, Department of Civil Engineering, Faculty of Civil Engineering and Architecture, Technical and Vocational University (TVU), Tehran, Iran

http://dx.doi.org/10.47176/TMI.2026.250984

production is associated with considerable greenhouse gas emissions and the consumption of natural resources. Therefore, the use of supplementary cementitious materials and industrial by-products as partial replacements for cement is regarded as an important strategy for moving the concrete industry toward sustainable development [1–3]. Among these materials, fly ash is one of the most widely used supplementary cementitious materials in concrete because of its pozzolanic properties, favorable particle morphology, and suitability for partially replacing cement. In addition to reducing cement consumption and the environmental impacts associated with its production, the incorporation of fly ash can positively affect workability, strength development at later ages, and certain durability characteristics of concrete [4]. Life-cycle assessment studies have also demonstrated that partially replacing cement with fly ash can reduce the environmental impacts of concrete; however, the magnitude of this reduction depends on the replacement ratio, type of fly ash, concrete characteristics, and transportation conditions [2]. Accordingly, developing methods capable of accurately predicting the mechanical behavior of fly ash-containing concrete can contribute to reducing the number of laboratory tests, accelerating the mix-design process, and facilitating the appropriate selection of raw materials in construction projects.

Compressive strength is one of the most important indicators of concrete performance and a key parameter in the design and quality control of concrete structures. However, compressive strength is influenced by a combination of factors, including the amounts of cement and fly ash, the water-to-cementitious-materials ratio, the quantity and type of aggregates, admixtures, and concrete age. The interactions among these variables are often nonlinear. Consequently, determining compressive strength solely through linear relationships may overlook some of the complexity underlying concrete behavior. In recent years, machine learning has attracted considerable attention as a data-driven approach for identifying relationships between mix-design parameters and the mechanical properties of concrete. These methods can identify patterns embedded in experimental datasets without requiring explicit formulation of complex physical relationships and can subsequently use these patterns to predict new observations [5].

Among machine-learning algorithms, Support Vector Machine (SVM) is a notable option for engineering problems because of its ability to model nonlinear relationships, its effectiveness with medium-sized datasets, and its capability to employ kernel functions. The use of a Radial Basis Function (RBF) kernel enables the mapping of data into a higher-dimensional feature space and facilitates the modeling of nonlinear relationships between input variables and the response variable. In contrast, Multiple Linear Regression (MLR) is a simple, interpretable, and computationally efficient statistical method that represents the relationship between multiple independent variables and a dependent variable within a linear framework. Therefore, MLR can serve as an appropriate baseline model for assessing the added value of nonlinear modeling approaches.

The application of machine learning to predicting the properties of fly ash-containing concrete has expanded considerably in recent years. Azimipour et al. developed linear and nonlinear SVM models to investigate the applicability of this approach for predicting the fresh properties and compressive strength of self-compacting concrete with high-volume fly ash, employing kernel functions and grid search to determine the model parameters [6]. In a study by Song et al., a dataset of fly ash-containing concrete was investigated. Their results indicated that ensemble models, particularly Adaptive Boosting (AdaBoost5) with decision trees and Random Forest, outperformed individual models in predicting compressive strength, while k-fold cross-validation confirmed the suitability of these approaches [7].

Building on this line of research, Li et al. developed several deep-learning and hybrid models for predicting the compressive strength of fly ash-containing concrete using the same seven-variable structure. Their results showed that a hybrid model combining a fully connected neural network and a convolutional neural network achieved a high correlation coefficient. Their analysis of variable importance also indicated that cement content and concrete age had the greatest influence on strength prediction [8]. Furthermore, Kumar7 et al., in 2024, investigated several machine-learning algorithms using 240 observations of high-volume fly ash and silica-fume self-compacting concrete and demonstrated that Random Forest provided the best predictive performance among the models examined [9]. These studies indicate that machine learning has considerable potential for predicting the strength of fly ash-containing concrete. At the same time, however, the performance of machine-learning algorithms depends on the type of concrete, input variables, dataset size and variability, and the manner in which model parameters are tuned.

Despite the numerous studies conducted in this area, a substantial proportion of recent research has focused on complex deep-learning models and sophisticated machine-learning methods. In contrast, a direct comparison between a nonlinear model such as SVM and a linear and interpretable model such as MLR can provide clearer insight into the importance of nonlinear relationships governing the compressive strength of fly ash-containing concrete. Moreover, recent studies have shown that model selection should be based not only on performance using a single training–testing split but also on cross-validation and performance stability [7, 9].

Therefore, the present study aims to evaluate and compare the performance of SVM and MLR in predicting the compressive strength of fly ash-containing concrete using a reliable laboratory dataset and influential input parameters. To this end, the performance of the two models is evaluated based on the correlation coefficient (R), mean squared error (MSE), and mean absolute error (MAE), while five-fold cross-validation is employed to assess model stability and generalizability. The ultimate objective is to determine whether the nonlinear modeling capability of SVM can provide more accurate and reliable predictions of the compressive strength of fly ash-containing concrete than the linear MLR model.

2. PREDICTION MODELS

2.1. Support Vector Machine

Support Vector Machine (SVM) is a powerful machine-learning method for predicting and modeling complex and nonlinear relationships between variables. It is based on the principle of structural risk minimization [10]. Its strong generalization capability, particularly when the number of experimental observations is limited, has made SVM an attractive approach for engineering applications and the prediction of concrete properties. Unlike many neural-network methods, the structure of SVM is based on identifying support vectors and optimizing model parameters. Consequently, it can achieve satisfactory predictive performance even with a relatively limited number of training observations.

In this study, the parameters (C) and (ε), as well as the kernel-function parameters, were investigated and tuned to develop an SVM model for predicting the compressive strength of concrete [11]. The Radial Basis Function (RBF) was selected as the model kernel, and its optimal parameters were determined through a model tuning and evaluation procedure. Subsequently, the performance of the optimized model on the training and testing datasets was evaluated using the correlation coefficient (R) and mean squared error (MSE). The optimal SVM parameters and the resulting model evaluation are presented in Table 1.

Table 1. Characteristics of the Optimized Support Vector Machine Model

Data Type	C	ε	σ	Number of Support Vectors
Training	1000	0.008	0.833	48
Testing				

2.2. Multiple Linear Regression

Multiple Linear Regression (MLR) is a commonly used statistical method for determining the relationship between multiple independent variables and a dependent variable [12]. In this approach, the simultaneous effects of the independent variables on the target variable are represented by a linear mathematical relationship, and the model coefficients are estimated based on the available data. Therefore, MLR can be used as an interpretable baseline method for modeling and predicting the compressive strength of concrete. Equation (1) presents the general form of the multiple linear regression equation used in this study.

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_{p-1} + \varepsilon_i i = 1,2,\ldots,n \tag{1}$$

In Equation (1), (Y_i) denotes the predicted value of the dependent variable; (x_1) to (x_{p-1}) represent the independent predictor variables; (β_0) is the value of (Y) when all independent variables ((x_1) to (x_{p-1})) are equal to zero; (β_1) to (β_p) denote the estimated regression coefficients; and (ε) represents the error term of the regression model. Equation (1) can also be expressed in matrix form, as shown in Equation (2).

$$Y_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} X_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \tag{2}$$

The matrix (X) represents the observed values of (p-1) predictor variables for (n) individuals. The vector (Y) contains the observed values of the dependent variable for a sample of size (n). In a regression model, the (β_i) terms represent the model parameters and can be estimated using various methods, such as the least squares method and maximum likelihood estimation. The (ε_i) terms are referred to as error terms and are assumed to follow a normal distribution with a mean of zero and a variance of (σ^2).

Accordingly, Multiple Linear Regression is a commonly used statistical method for examining and predicting the relationship between a dependent variable and multiple independent variables.

In this approach, the simultaneous effects of the independent variables on the target variable are expressed through a mathematical equation, and the model coefficients are estimated using statistical methods. To select appropriate variables and obtain an optimal model, several variable-selection procedures can be employed, including the Enter, Forward, Backward, and Stepwise methods. Each method determines how variables are entered into or removed from the model according to predefined statistical criteria.

3. DATASET AND DESCRIPTION OF VARIABLES

The data used in this study were extracted from laboratory test results for fly ash-containing concrete reported in previous studies published in reputable references [13–15]. The dataset comprised 471 observations, of which 377 observations (80%) were allocated for model training and development, while the remaining 94 observations (20%) were used to evaluate and validate model performance. The input variables included cement, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate, while the output variables were concrete age and compressive strength. The statistical characteristics and ranges of variation of the variables used in the modeling process are presented in Table 2.

To reduce the effects of differences in scale between the input and output variables and to prevent potential bias during model training, all data were subjected to the same preprocessing and normalization procedure before being entered into the algorithms used in this study. Given the substantially different ranges of the mix-design parameters and compressive strength, data scaling places the variables within a specified range and consequently improves the stability of the modeling process. In this study, the values of both the input and output variables were transformed to the range of **0.1 to 0.9**.

Table 2. Statistical Characteristics and Ranges of the Variables Used in the Modeling Process

Statistic	Compressive Strength (MPa)	Concrete Age (days)	Fine Aggregate (kg/m³)	Coarse Aggregate (kg/m³)	Admixture (kg/m³)	Water (kg/m³)	Fly Ash (kg/m³)	Cement (kg/m³)
Mean	6.27	1	594	801	0	140	0	134.7
Maximum	74.99	365	945	1125	28.2	228	200.1	540
Minimum	31.47	47.84	793.17	1004.04	5.01	181.88	62.58	298.07
Standard Deviation	14.46	65.53	73.85	74.17	5.49	18.01	64.88	100.69

Note: The numerical values have been translated and formatted according to the Persian source provided. However, several values in the original table appear internally inconsistent—for example, the reported minimum and maximum values for compressive strength and fine/coarse aggregate. These should be checked against the original dataset before final publication.

4. EVALUATION OF MODEL PERFORMANCE

To assess the generalizability and accurately compare the performance of the Multiple Linear Regression (MLR) and Support Vector Machine (SVM) models, five-fold cross-validation was employed [10]. For this purpose, the dataset was divided into five approximately equal-sized subsets. In each iteration, four subsets were used to train the model, while the remaining subset was reserved for testing. This procedure was repeated for all five subsets, and the mean values of the evaluation results were subsequently calculated.

The performance of the models was assessed using the correlation coefficient (R), mean squared error (MSE), and mean absolute error (MAE), whose calculation formulas are presented in Equations (3)–(5), respectively. The correlation coefficient indicates the degree of agreement between the predicted and actual values and ranges from −1 to +1. A higher value of R, particularly a value greater than 0.8 and closer to 1, indicates better agreement between the predictions and the experimental data. Accordingly, a model with a higher R and lower MSE and MAE values is considered to have superior predictive performance.

Equation (3) represents the correlation coefficient (R).

$$R = \frac{\sum_{i=1}^{n} t_i o_i - n\bar{t}\bar{o}}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2 \sum_{i=1}^{n}(o_i - \bar{o})^2}} \tag{3}$$

In Equation (3), o_i denotes the model output for the *i-th* observation, represents the actual value of the *i-th* output, $\bar{o}$ is the mean of the model outputs, $\bar{t}$ is the mean of the actual values, and *n* denotes the total number of observations.

The Mean Squared Error (MSE) is defined as the mean of the squared errors, calculated from the difference between the value predicted by the model (or the existing relationship) and the corresponding actual value, as given in Equation (4). The MSE ranges from zero to infinity, and a value closer to zero indicates better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - t_i)^2 \tag{4}$$

The Mean Absolute Error (MAE) represents the average of the absolute errors, calculated as the absolute difference between the value predicted by the model (or the existing relationship) and the corresponding actual value, as given in Equation (5). The MAE ranges from zero to infinity, and a value closer to zero indicates better model performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - t_i| \tag{5}$$

4.1. Comparison and Analysis of the Performance of the Prediction Models

A comparison of the performance of the Multiple Linear Regression (MLR) and Support Vector Machine (SVM) models based on the evaluation criteria obtained from five-fold cross-validation is presented in Figure 2. In addition, Figure 3 illustrates the predicted values versus the experimental values, along with the corresponding performance evaluation metrics, for the two prediction models during the training and testing stages.

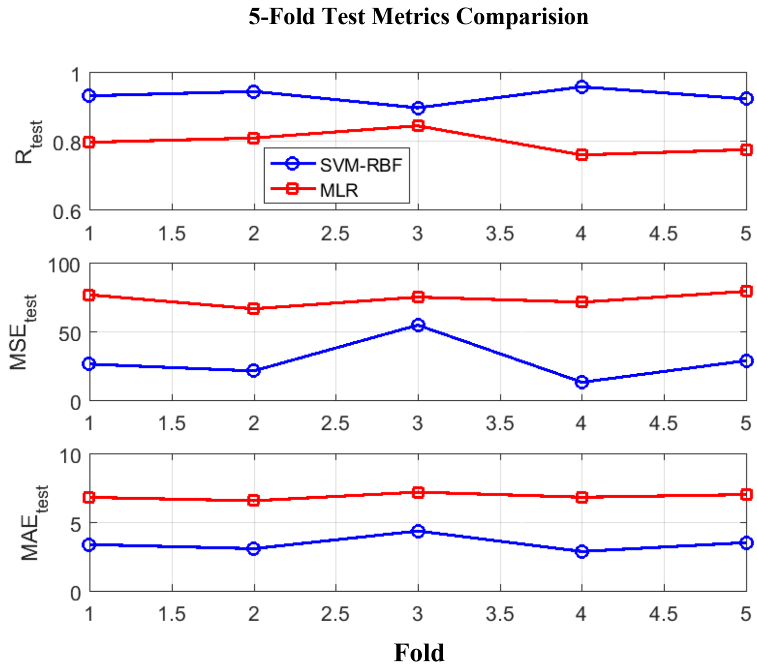

Fig. 2. Comparison of the Performance of Multiple Linear Regression and Support Vector Machine Models in Five-Fold Cross-Validation Based on Performance Evaluation Metrics

The results of five-fold cross-validation indicate that the Support Vector Machine (SVM) model achieved a higher correlation coefficient and lower MSE and MAE values than the Multiple Linear Regression (MLR) model in all five folds. Furthermore, the consistent performance of SVM across the different folds, together with its substantial performance advantage over MLR, demonstrates the greater capability of SVM for predicting the compressive strength of fly ash-containing concrete.

This superior performance can be attributed to the ability of the RBF kernel to identify and model nonlinear and complex relationships between mix-design parameters and compressive strength, whereas the linear structure of MLR has a more limited capacity to represent such relationships.

A comparison of the actual and predicted values shows that both models were able to capture the overall trend of variations in compressive strength. However, the SVM model exhibited less dispersion of the data points around the line of equality and better agreement with the actual values. In the training dataset, SVM achieved a correlation coefficient of approximately 0.988, substantially outperforming MLR, which achieved a value of approximately 0.814. SVM also exhibited slightly lower prediction errors.

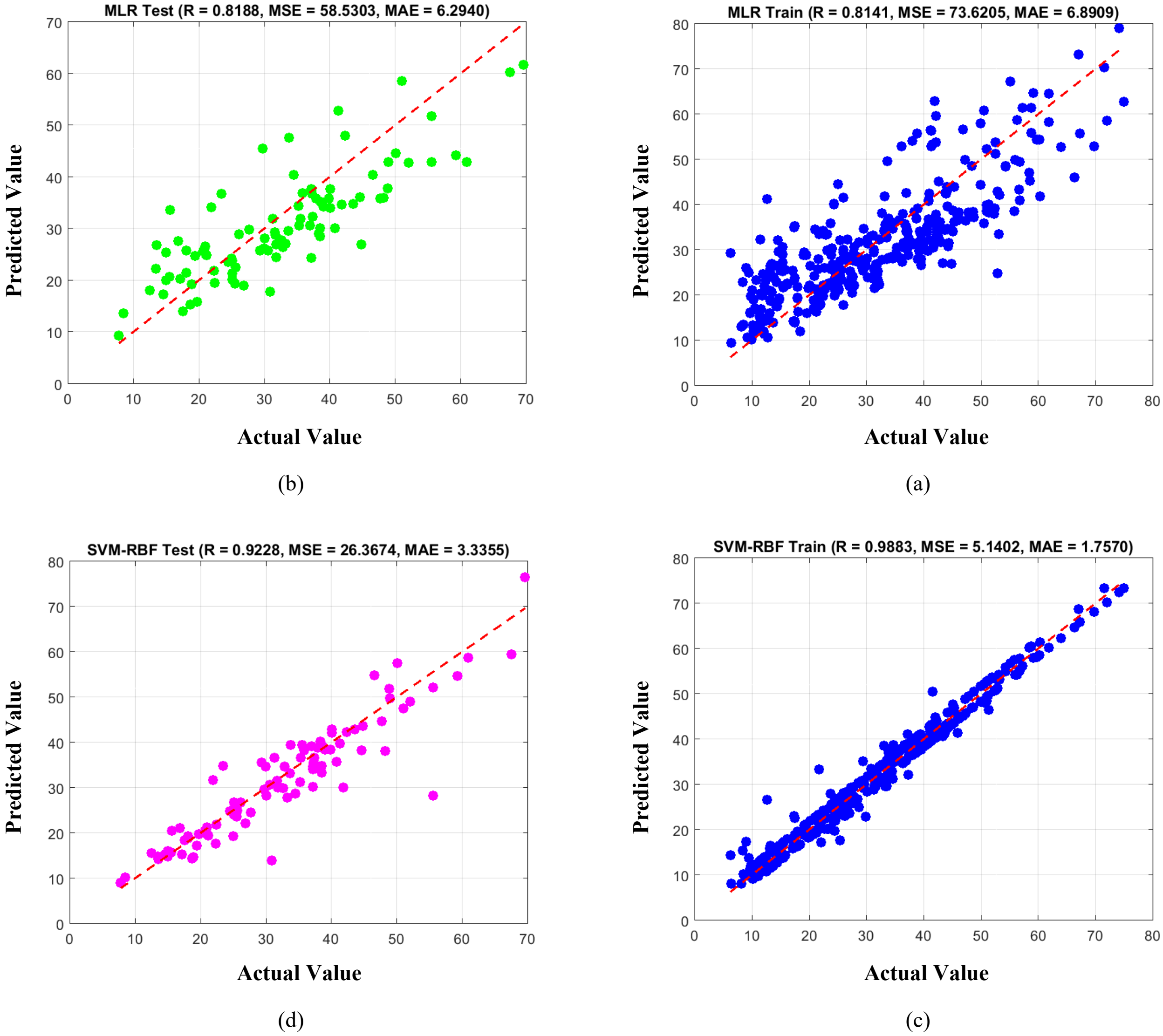

Fig. 3. Plot of Predicted versus Experimental Values with Performance Evaluation Metrics: (a) Multiple Linear Regression at the Training Stage, (b) Multiple Linear Regression at the Testing Stage, (c) Support Vector Machine at the Training Stage, and (d) Support Vector Machine at the Testing Stage

This superiority was also maintained on the test dataset, with the Support Vector Machine achieving an R value of approximately 0.923, compared with approximately 0.818 for Multiple Linear Regression. Moreover, the lower MSE and MAE values obtained by the Support Vector Machine indicate that this method provides better predictive performance on previously unseen data. Overall, the results demonstrate that, owing to its ability to model nonlinear relationships between mix-design constituents and compressive strength, the Support Vector Machine is a more suitable approach for this problem than the linear relationship assumed by Multiple Linear Regression.

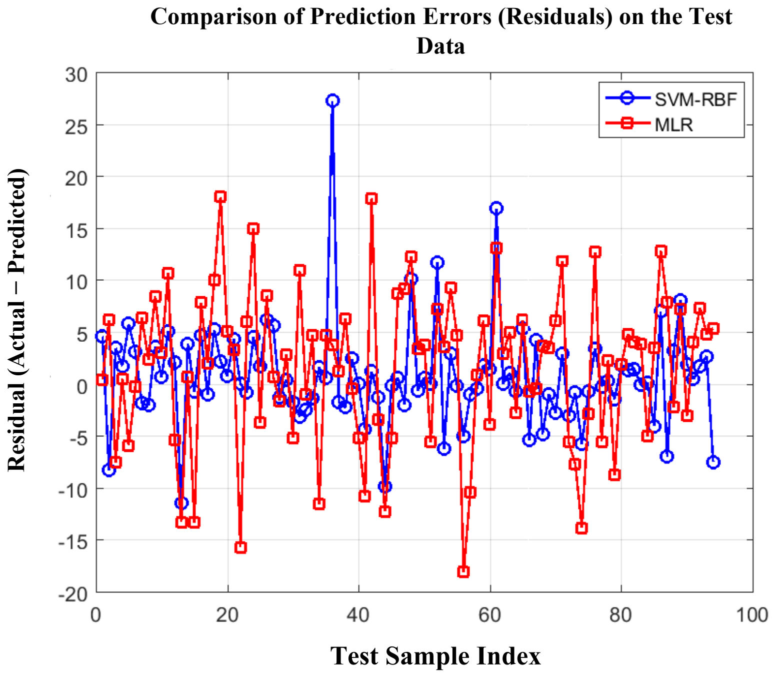

Fig. 4. Comparison of the Distribution of Residual Errors of the Support Vector Machine and Multiple Linear Regression Models on the Test Dataset

Figure 4 presents a comparison of the distribution of residual errors produced by the Support Vector Machine (SVM) and Multiple Linear Regression (MLR) models on the test dataset. Examination of the residual-error plot indicates that the errors of both models are distributed around zero, with no clear or systematic pattern in their variations. This finding suggests that, overall, neither model exhibits substantial systematic bias in predicting the compressive strength of concrete. However, the range of error dispersion for the SVM model is narrower than that of MLR for a considerable proportion of the observations, indicating relatively more stable predictions by SVM.

Although both models encounter relatively large positive or negative errors for some observations, SVM demonstrates a more consistent pattern in controlling prediction errors compared with MLR. This behavior can be attributed to the ability of the RBF kernel to account for nonlinear relationships between mix-design parameters and compressive strength, whereas MLR relies on a linear structure between the input and output variables. Overall, the residual analysis further supports the results obtained from the performance metrics and reinforces the suitability of SVM for this prediction problem.

5. CONCLUSION

This study investigated and compared the capabilities of two approaches, namely Support Vector Machine with a Radial Basis Function kernel (SVM-RBF) and Multiple Linear Regression (MLR), for predicting the compressive strength of fly ash-containing concrete. For this purpose, 471 laboratory observations were used, with cement, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate considered as input variables and compressive strength as the output variable. The results demonstrated that both approaches were capable of capturing the overall trend in compressive-strength variations; however, SVM-RBF exhibited superior predictive performance compared with MLR.

This superiority was observed in both the training and testing datasets as well as in five-fold cross-validation. Specifically, SVM achieved a correlation coefficient of approximately 0.922 on the test dataset, whereas the corresponding value for MLR was approximately 0.818. The MSE and MAE values were also lower for SVM.

Analysis of the plots comparing actual and predicted values, together with the residual analysis, demonstrated that the SVM predictions exhibited better agreement with the experimental data and more controlled error dispersion than those of MLR. The results of five-fold cross-validation further showed that the superiority of SVM was not limited to a particular data split and that its better predictive performance was consistently maintained across different folds.

This behavior can be attributed to the ability of the RBF kernel to identify nonlinear relationships between mix-design constituents and compressive strength, whereas MLR relies on the assumption of a linear relationship between the independent variables and the response. Therefore, SVM can be considered a more suitable approach for modeling and predicting the compressive strength of fly ash-containing concrete within the range of the data investigated in this study.

Nevertheless, the difference in the performance of the two models on the test dataset is not particularly large. Therefore, the selection of SVM should not be based solely on a single performance metric. Instead, cross-validation and an independent test dataset should be considered simultaneously when assessing the generalizability of the model. Overall, the findings of this study demonstrate that machine-learning methods, particularly SVM, can serve as complementary tools to conventional concrete mix-analysis and mix-design approaches, helping reduce the need for extensive laboratory testing and supporting decision-making in the selection of appropriate mix proportions for fly ash-containing concrete.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Akbulut, Z. F., Yavuz, D., Tawfik, T. A., Smarzewski, P., & Guler, S. (2024). Enhancing concrete performance through sustainable utilization of class-C and class-F fly ash: A comprehensive review. *Sustainability, 16*(12), 4905. https://doi.org/10.3390/su16124905

[2] Orozco, C., Babel, S., Tangtermsirikul, S., & Sugiyama, T. (2024). Comparison of environmental impacts of fly ash and slag as cement replacement materials for mass concrete and the impact of transportation. *Sustainable Materials and Technologies, 39*, e00796. https://doi.org/10.1016/j.susmat.2023.e00796

[3] Orozco, C., Tangtermsirikul, S., Sugiyama, T., & Babel, S. (2023). Examining the endpoint impacts, challenges, and opportunities of fly ash utilization for sustainable concrete construction. *Scientific Reports, 13*(1), 18254. https://doi.org/10.1038/s41598-023-45632-z

[4] Sharma, V., Dash, S., & Gupta, P. (2024). Comprehensive review of fly ash: Environmental impact and applications. *Environmental Quality Management, 34*(2), e22338. https://doi.org/10.1002/tqem.22338

[5] Rathnayaka, M., Karunasinghe, D., Gunasekara, C., Wijesundara, K., Lokuge, W., & Law, D. W. (2024). Machine learning approaches to predict compressive strength of fly ash-based geopolymer concrete: A comprehensive review. *Construction and Building Materials, 419*, 135519. https://doi.org/10.1016/j.conbuildmat.2024.135519

[6] Azimi-Pour, M., Eskandari-Naddaf, H., & Pakzad, A. (2020). Linear and non-linear SVM prediction for fresh properties and compressive strength of high volume fly ash self-compacting concrete. *Construction and Building Materials, 230*, 117021. https://doi.org/10.1016/j.conbuildmat.2019.117021

[7] Song, Y., et al. (2021). Prediction of compressive strength of fly-ash-based concrete using ensemble and non-ensemble supervised machine-learning approaches. *Applied Sciences, 12*(1), 361. https://doi.org/10.3390/app12010361

[8] Li, H., Chung, H., Li, Z., & Li, W. (2024). Compressive strength prediction of fly ash-based concrete using single and hybrid machine learning models. *Buildings, 14*(10), 3299. https://doi.org/10.3390/buildings14103299

[9] Kumar, S., Kumar, R., Rai, B., & Samui, P. (2024). Prediction of compressive strength of high-volume fly ash self-compacting concrete with silica fume using machine learning techniques. *Construction and Building Materials, 438*, 136933. https://doi.org/10.1016/j.conbuildmat.2024.136933

[10] Dou, B., et al. (2023). Machine learning methods for small data challenges in molecular science. *Chemical Reviews, 123*(13), 8736. https://doi.org/10.1021/acs.chemrev.3c00189

[11] Umutoni, L., & Samadi, V. (2024). Application of machine learning approaches in supporting irrigation decision making: A review. *Agricultural Water Management, 294*, 108710. https://doi.org/10.1016/j.agwat.2024.108710

[12] Aalami, M. T., Hejabi, N., Nourani, V., & Saghebian, S. (2019). Investigation of artificial intelligence approaches capability in predicting the wastewater treatment plant performance (case study: Tabriz wastewater treatment plant). *Amirkabir Journal of Civil Engineering*. https://doi.org/10.22060/ceej.2019.16757.6334

[13] Pham, V.-N. (2025). Practical formulation for estimating the compressive strength of self-compacting fly ash concrete using gene-expression programming. *Journal of Science and Technology in Civil Engineering (JSTCE-HUCE), 19*(3), 98–109. https://doi.org/10.31814/stce.huce2025-19(3)-08

[14] Pham, V.-N., & Nguyen, V.-Q. (2024). Effects of alkaline activators and other factors on the properties of geopolymer concrete using industrial wastes based on GEP-based models. *European Journal of Environmental and Civil Engineering, 28*(16), 3770–3792. https://doi.org/10.1080/19648189.2024.2357677

[15] Pham, V. N. (2025). *Fly ash concrete* (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/c8g7smdgzs.1