buildings

MDPI

Article

Comparative Study on Hyperparameter Tuning for Predicting Concrete Compressive Strength

Jeonghyun Kim [1] and Donwoo Lee [2,*]

[1] Faculty of Civil Engineering, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[2] School of Industrial Design & Architectural Engineering, Korea University of Technology & Education, 1600 Chungjeol-ro, Byeongcheon-myeon, Cheonan 31253, Chungcheongnam-do, Republic of Korea
* Correspondence: lov1004ely@koreatech.ac.kr

Abstract

This study assesses the impact of hyperparameter optimization algorithms on the performance of machine learning-based concrete compressive strength prediction models. Three datasets were used to compare the performance of a basic model that had not undergone hyperparameter optimization, with the model incorporating random search, grid search, and Bayesian optimization. A post-hoc analysis using Shapley additive explanations was also conducted. The results demonstrate that the effectiveness of hyperparameter optimization varies depending on the characteristics of the dataset. For Dataset 1, the application of search algorithms appeared to improve prediction accuracy to some extent. However, for Datasets 2 and 3, the performance improvement from the search algorithms was either insignificant or decreased. Despite these contrasting results, the post-analysis showed that the influence of each feature generally aligned with empirical knowledge across all datasets, suggesting that the Shapley additive explanations method alone may have limitations in pinpointing the causes of model overfitting.

Keywords: artificial intelligence; machine learning; strength prediction; hyperparameter; search algorithm

check for updates

Academic Editor: Ray Kai Leung Su

Received: 2 June 2025
Revised: 19 June 2025
Accepted: 20 June 2025
Published: 22 June 2025

Citation: Kim, J.; Lee, D. Comparative Study on Hyperparameter Tuning for Predicting Concrete Compressive Strength. *Buildings* **2025**, *15*, 2173. https://doi.org/10.3390/buildings15132173

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Accurately predicting concrete strength is essential for preventing both overdesign and underdesign. Underdesign occurs when concrete does not meet the required strength. In fact, insufficient concrete strength due to poor prediction or design has been linked to partial or total structural collapse, which poses severe risks to human safety and causes significant economic loss [1], especially when structures are subjected to higher-than-expected loads or external impacts such as earthquakes. Conversely, overdesign involves using more strength than necessary, resulting in excessive material usage, higher costs, and wasted resources. Therefore, precise prediction of concrete strength allows for a balanced design that ensures structural safety while optimizing the use of resources and minimizing costs.

Traditional methods for predicting the compressive strength of concrete have relied on experimental and theoretical models, but achieving accurate predictions remains challenging [2,3]. This is due to the complex nature of concrete, a composite material composed of various ingredients, including cement, aggregates, water, and admixtures. Numerous factors, based on the properties of these materials, influence the strength of concrete. For example, cement properties such as density, strength, particle size, chemical composition,

and usage content affect concrete strength [4,5]. Similarly, the type of rock used for fine and coarse aggregates, as well as factors like maximum particle size, shape, gradation, density, water absorption, and fine content, also play a role [6,7]. Furthermore, mixing methods [8,9] and curing conditions [10,11] also influence the strength of concrete. The applicability of empirical models is typically restricted to specific cases, resulting in increased errors when predicting unseen data. For example, when estimating the strength of concrete incorporating new types of materials, the influence of these new components is often overlooked [12].

Recent advancements in machine learning (ML) have led to the development of concrete strength prediction models that are more accurate and reliable than traditional approaches. In a study by Cai et al. [13] on predicting chloride concentrations on marine concrete surfaces, it was found that the ML model outperformed seven conventional models in terms of prediction accuracy. Several studies have investigated which ML techniques are most effective for predicting concrete strength [2,14]. Furthermore, hybrid ML models combining two or more algorithms or techniques have also been developed [15–17], with ongoing research focused on enhancing prediction accuracy in this field. These advanced ML techniques have been successfully applied not only to traditional Portland cement-based natural aggregate concrete [18,19] but also to various types of concrete, such as those incorporating recycled materials and industrial by-products [20–22], lightweight concrete [23], as well as radiation-shielding concrete [24].

Hyperparameter tuning, which involves parameters like learning rate, tree depth, and batch size, is an important part of optimizing ML model performance [25]. Hyperparameter optimization has been shown to enhance training speed, reduce overfitting, and improve predictive accuracy. In a study by Tang et al. [3], a model with grid search achieved a concrete compressive strength prediction accuracy of 98%, compared to 92% for a model without grid search. Another study [26] demonstrated that applying grid search improved the coefficient of determination (R^2) from 0.922 to 0.961 and reduced the root mean squared error (RMSE) from 3.439 MPa to 2.321 MPa. Similarly, Bayesian optimization for hyperparameter tuning produced improvement in prediction. In research by Joy [27], the R^2 for concrete compressive strength prediction improved from 0.899 to 0.942, while RMSE decreased from 5.10 MPa to 4.28 MPa. Similar results have also been reported in other studies [28,29]. Some research reports that hyperparameter optimization may not significantly improve the accuracy of ML models for complex, high-dimensional problems. This is often because model performance in certain algorithms depends more heavily on factors such as sample size, non-linearity, and feature interactions [30]. A particularly challenging situation in ML arises when high-dimensional data is paired with limited sample size, a phenomenon known as the “curse of dimensionality”. This issue is particularly common in ML applications for concrete science. The properties of concrete are influenced not only by the characteristics of its constituent materials but also by factors such as mixing sequence, mixing time, mixing methods, and external conditions like temperature and humidity. Consequently, datasets related to concrete often exhibit high dimensionality. Furthermore, the key properties of concrete can typically be evaluated after sufficient hydration has occurred (e.g., 28-day compressive strength), which makes the collection of large datasets challenging. A review of 389 ML studies on concrete properties published between 1990 and 2020 shows that over 55% of these studies utilized datasets with fewer than 200 samples [31]. Given these limitations, it is worth questioning whether hyperparameter optimization is consistently effective in ML models for predicting concrete strength. Although the benefits of hyperparameter tuning are well documented, previous studies have predominantly focused on single-method applications. A comparison of multiple optimization techniques, specifically grid search, random search, and Bayesian op-

timization, within the same experimental framework, remains unexplored. This study aims to address this gap by evaluating whether the choice of optimization method significantly influences model performance.

In a previous study by the present authors [32], ML performance was compared by using cement grade, a categorical variable, and cement strength, a continuous variable, as input features. The results indicated that models incorporating cement strength as an input outperformed those using cement grade. However, since the previous study focused on comparing cement grade and cement strength, it did not explore the impact of hyperparameter optimization on model performance, which was recommended for future research. Building on this earlier work, the current study investigates the influence of hyperparameter tuning on predictive performance by applying random search, grid search, and Bayesian optimization to ML models for predicting concrete compressive strength. The research involved training models on three distinct datasets and analyzing performance variations across them. Furthermore, the Shapley additive explanations (SHAP) method was utilized to identify the key features influencing model predictions and to assess their alignment with domain knowledge.

2. Methodology

The research methodology is summarized in Figure 1. Three different databases were constructed using data collected from the literature, which was then split into a training set for ML and a test set for validation. To optimize the performance of the eXtreme Gradient Boosting (XGB) model, three hyperparameter tuning algorithms were applied to the training set. Following hyperparameter optimization, K-Fold cross-validation was conducted to evaluate the generalization performance of models. Five folds were configured, and for each fold, the training set and validation set were separated to train the optimal model. Predictions for both the validation and training sets were evaluated, and performance metrics such as R^2, mean absolute error (MAE), and RMSE were recorded. The performance across folds was averaged to derive the final results. The generalization and predictive performance of the trained XGB model were further evaluated using the test set. To enable explainable artificial intelligence, the SHAP technique was employed to analyze and visualize feature importance, identifying key features influencing the predictions of the model. This approach allowed the study to thoroughly investigate the influence of hyperparameter tuning algorithms on the performance improvement of the XGB-based predictive model. The research was conducted using Python version 3.11.9.

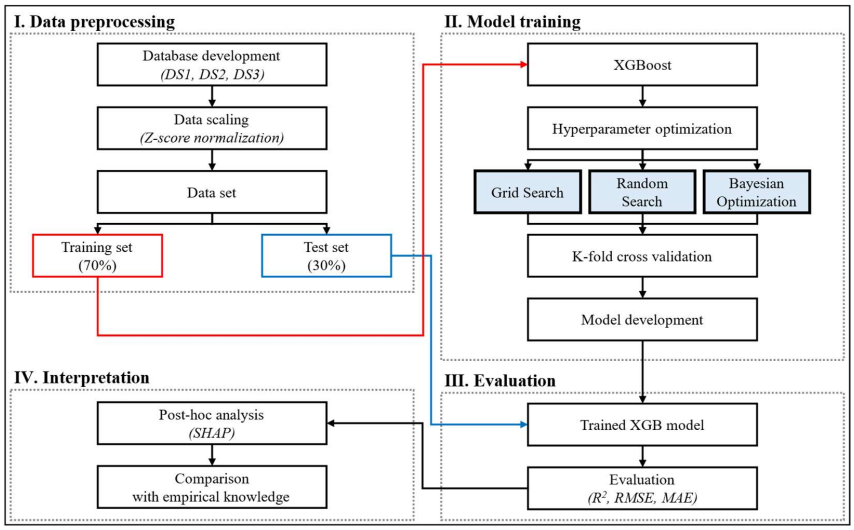

Figure 1. Research flow.

2.1. Data Preprocessing

As mentioned in the Introduction, the present study extends the authors' previous research on developing a concrete compressive strength prediction model that includes cement strength as an input feature [32]. Accordingly, the three datasets used in that earlier work were retained for this study. These datasets vary in input features, sample sizes, and sources, providing a diverse foundation for evaluating and comparing the effectiveness of three hyperparameter optimization techniques under differing data conditions.

- Dataset 1 (DS1): This dataset was sourced from the study by Zheng et al. [33]. To reduce dimensionality, 8 out of the 16 original input variables were selected: cement strength (CCS), water (W), cement (C), slag (S), fly ash (F-ash), coarse aggregate (CA), fine aggregate (FA). The output variable is the 28-day compressive strength of concrete (28CS).
- Dataset 2 (DS2): This dataset was obtained from the study by Zhao et al. [34]. Variables not within the scope of this study (e.g., slump, tensile strength of cement) and data points without recorded compressive strength were excluded from the original dataset. The input variables are CCS, curing age (Age), maximum coarse aggregate size (Dmax), stone powder (SP), fine aggregate fineness modulus (FA-FM), water-cement ratio (w/c), W, and sand-aggregate ratio (S/a). The output variable is the compressive strength of concrete (CS).
- Dataset 3 (DS3): This dataset was compiled by the authors from various studies. The input variables include CCS, coarse aggregate-specific gravity (CA-SG), fine aggregate-specific gravity (FA-SG), C, W, w/c, CA, FA. The output variable is 28CS. To eliminate the size effect of the concrete specimen, compressive strengths for Ø100 $\times$ 200 mm and Ø150 $\times$ 300 mm cylindrical specimens, as well as a 100 mm cube specimen, were converted to equivalent compressive strength for a 150 mm cube specimen [35].

The statistical characteristics of each variable are provided in Table 1, and individual data points can be accessed in the Supplementary Materials.

The Pearson correlation coefficients among the variables in each dataset are shown in Figure 2. As the coefficient approaches 1 or -1, the two factors exhibit strong positive and negative correlations, respectively. Due to the characteristics of concrete mix design based on the raw material quantity needed per cubic meter, variables often show high correlations. For example, in Dataset 1, there is a strong correlation with coefficients of 0.80 between cement content and 28-day compressive strength and 0.85 between coarse aggregate content and fine aggregate content. The former refers to the increase in hydration products as cement content rises, which enhances concrete strength. The latter is due to the nature of the concrete mix design, where an increase in coarse aggregate content leads to a natural reduction in fine aggregate content. In Dataset 3, a significant negative correlation (coefficient of -0.7) was observed between the water-to-cement ratio and 28-day compressive strength; it is a well-established fact in the field that the water-to-cement ratio affects concrete strength. When independent variables exhibit significant correlations, multicollinearity can occur, potentially affecting the performance stability of regression models. However, completely eliminating high correlations among variables is practically challenging, as also confirmed in previous studies [36,37].

Table 1. Descriptive statistics of datasets used in this research.

Dataset	Sample Size	Feature	Unit	Min.	Max.	Mean.	STD.	Skew.	Kurt.	Missing
Dataset 1	610	CCS	MPa	42.9	53.2	47.9	3.8	−0.149	−1.52	0
		W	kg/m^3	143.2	166.4	153.2	6.4	0.231	−1.19	0
		C	kg/m^3	135	335	231.5	43.9	0.066	−0.99	0
		S	kg/m^3	50.0	115.0	81.7	18.0	0.391	−0.97	0
		F-ash	kg/m^3	25.0	80.0	58.4	13.9	0.203	−0.99	0
		FA	kg/m^3	719.1	946.4	870.1	66.4	−1.075	−0.19	0
		CA	kg/m^3	896.4	1004.7	955.2	30.4	0.508	−0.84	0
		28CS	MPa	31.2	73.0	48.7	8.9	0.428	−0.42	0
Dataset 2	388	CCS	MPa	35.5	63.4	48.0	4.7	0.401	0.79	0
		Age	Day	1	388	73	95	1.936	2.74	0
		Dmax	mm	12	80	30.1	13.6	2.727	8.06	10
		SP	%	0	20	8.92	5.45	0.353	−0.75	31
		FA-FM	n.a	2.2	3.5	3.03	0.26	−0.746	0.15	24
		w/c	n.a	0.30	1.01	0.47	0.10	1.027	2.87	0
		W	kg/m^3	104	291	169.9	21.2	−0.662	4.16	0
		s/a	%	26	54	38.0	5.78	0.966	1.05	0
		CS	MPa	4.23	96.3	55.1	19.0	−0.017	−0.71	0
Dataset 3	371	CCS	MPa	32.1	67.5	49.9	8.620	0.280	−0.639	0
		CA-SG	n.a	2.23	2.89	2.60	0.127	−0.854	0.201	28
		FA-SG	n.a	2.24	2.71	2.59	0.116	−1.169	0.392	116
		C	kg/m^3	250	601	390.0	69.8	0.507	0.372	0
		W	kg/m^3	108	320	181.2	28.8	0.891	2.758	0
		w/c	n.a	0.27	0.80	0.48	0.093	0.458	0.132	0
		CA	kg/m^3	680	1366	1098.9	140.6	−0.521	−0.140	0
		FA	kg/m^3	493	1160	717.7	111.6	0.879	1.052	0
		28CS	MPa	10	83.3	40.5	10.5	0.415	0.465	0

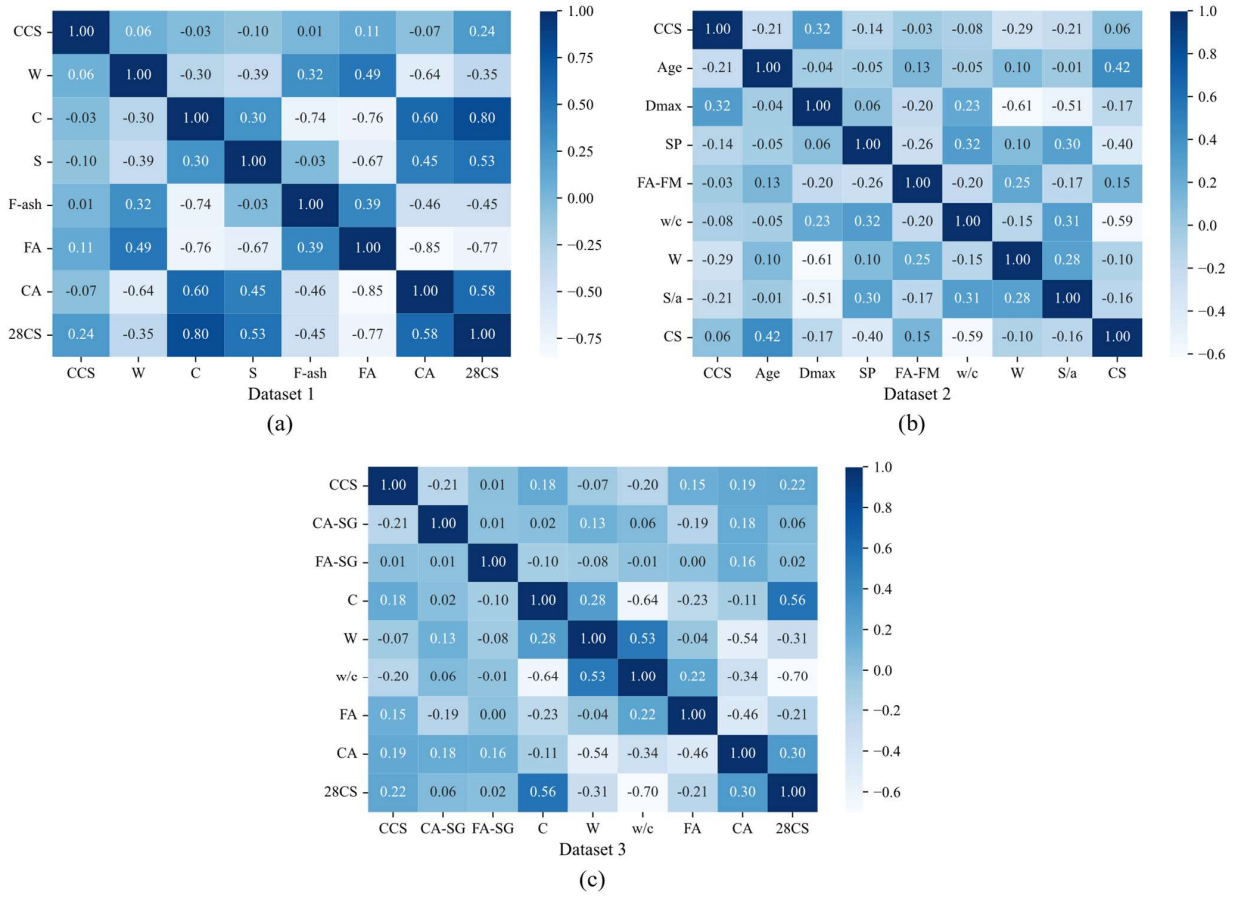

Figure 2. Pearson correlation of datasets: (**a**) dataset 1; (**b**) dataset 2; (**c**) dataset 3.

As shown in Table 1, some data points were missing, and these gaps were addressed using KNNImputer from Sci-kit Learn. KNNImputer replaces missing values by considering the similarity between data points. The KNN imputation is used in concrete property prediction models and has been reported to contribute to improved predictive performance [38,39].

In addition, data scaling was performed to resolve scale differences among input variables. This step prevents variables with larger ranges (e.g., aggregate content) from disproportionately influencing model training compared to variables with smaller ranges (e.g., specific gravity of aggregate). Z-score normalization was applied, adjusting all features to have a mean of 0 and a standard deviation of 1, ensuring equal contribution from all variables during model training.

2.2. Model Training and Evaluation

Selecting the appropriate ML model is a critical step in building a predictive model. To ensure accurate and reliable predictions, it is essential to choose a model that is well-suited to the specific characteristics of the problem. In this study, the XGB model, which has demonstrated excellent predictive accuracy in both the authors' previous research [32] and various other studies [40–42], was adopted. A detailed explanation of the principles and workings of the XGB algorithm can be found in the foundational work by Chen and Guestrin [43]. For model training and validation, the dataset was randomly divided into training and test sets at a 7:3 ratio, allocating 70% of the data to training and 30% to testing. To optimize the model performance, three hyperparameter tuning methods—grid search, random search, and Bayesian optimization—were employed. Each method explored various hyperparameter combinations to identify the best-performing configuration, which was then used for model training. The hyperparameter ranges and the optimal values selected by each tuning method are summarized in Table 2.

Table 2. Hyperparameter settings and selected values by search algorithms.

Algorithms	Parameter	Range	Selected		
			DS1	DS2	DS3
XGB	n_estimators random_state	[100] [42]	n.a	n.a	n.a
Grid search	n_estimator max_depth learning_rate	[50, 100, 150, 200, 250, 300] [1, 3, 5, 7, 9, 11, 13, 15] [0.01, 0.1, 0.2, 0.3, 0.4, 0.5]	250 1 0.3	150 3 0.5	300 3 0.2
Random search	n_estimator max_depth learning_rate	Randint [50, 300] Randint [1, 15] Uniform [0.01, 0.49)	181 2 0.230	283 3 0.119	175 3 0.455
Bayesian optimization	n_estimator max_depth learning_rate	Integer [50, 300] Integer [1, 10] Real [0.01, 0.5] prior = 'log-uniform')	200 2 0.106	271 3 0.141	300 2 0.264

To assess the generalization performance of the model and mitigate overfitting, K-fold cross-validation was adopted. The training dataset was split into five equally sized subsets. The cross-validation was conducted over five iterations. Each iteration of the cross-validation used a different subset as the validation set, while the remaining four subsets served as the training set. After completing the cross-validation, the test set was used to

assess the ability of the trained ML model to generalize to unseen data. The predictive performance of the developed model was evaluated using R^2, RMSE, and MAE (1–3):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - p_i)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|t_i - p_i| \quad (3)$$

where n represents the total number of samples in the dataset; t_i and p_i denote the test and predicted values, respectively; and $\bar{t}$ is the mean of the test values in the dataset.

R^2 evaluates how closely the predicted values match the actual observed values, with a value nearing 1 indicating strong explanatory power. RMSE measures the average prediction error, serving as a key indicator of prediction accuracy. MAE calculates the mean absolute error, highlighting the average magnitude of prediction errors and providing insight into the overall size of the discrepancies.

2.3. SHapley Additive exPlanations

To gain a deeper understanding of the developed ML model, a post-analysis was conducted. Initially, the SHAP method was used to quantitatively assess the impact of key features on the model predictions across different search algorithms. This approach allowed for a clear evaluation of the contribution of each feature and provided valuable insights into how the algorithm interprets data patterns. The results of this analysis were then compared with empirical knowledge to assess how accurately the model reflects real-world phenomena.

3. Results

The quantitative performance metrics of the ML models with and without the applied search algorithms are summarized in Table 3. In the table, NS, RS, GS, and BS represent the basic model with the following settings: no search algorithm applied, the models with random search, grid search, and Bayesian search, respectively.

Table 3. Machine learning performance of model with different search algorithms.

Dataset	Search Method	Test Set			Train Set		
		R^2	RMSE	MAE	R^2	RMSE	MAE
DS1	NS	0.858	3.647	3.228	0.953	1.838	1.345
	RS	0.909	2.918	2.583	0.938	2.128	1.846
	GS	0.911	2.881	2.548	0.915	2.478	2.165
	BS	0.911	2.874	2.556	0.930	2.249	1.973
DS2	NS	0.950	4.022	2.659	1.000	0.401	0.136
	RS	0.960	3.611	2.422	0.997	1.153	0.862
	GS	0.959	3.619	2.398	0.999	0.688	0.478
	BS	0.960	3.600	2.360	0.997	1.071	0.792
DS3	NS	0.800	4.613	3.556	1.000	0.171	0.030
	RS	0.777	4.873	3.527	0.999	0.245	0.130
	GS	0.799	4.619	3.415	0.999	0.375	0.256
	BS	0.774	4.901	3.676	0.992	0.926	0.712

3.1. Prediction Performance

3.1.1. Dataset 1

The scatter plot of the actual compressive strength and the compressive strength predicted by the ML model for Dataset 1 is shown in Figure 3. The baseline model, which did not apply a search algorithm, predicted the compressive strength with an accuracy of 85.8% on test data. In contrast, the models applying random search, grid search, and Bayesian optimization achieved accuracies of 90.9% and 91.1%, respectively, demonstrating that search algorithms can enhance the predictive accuracy of the model.

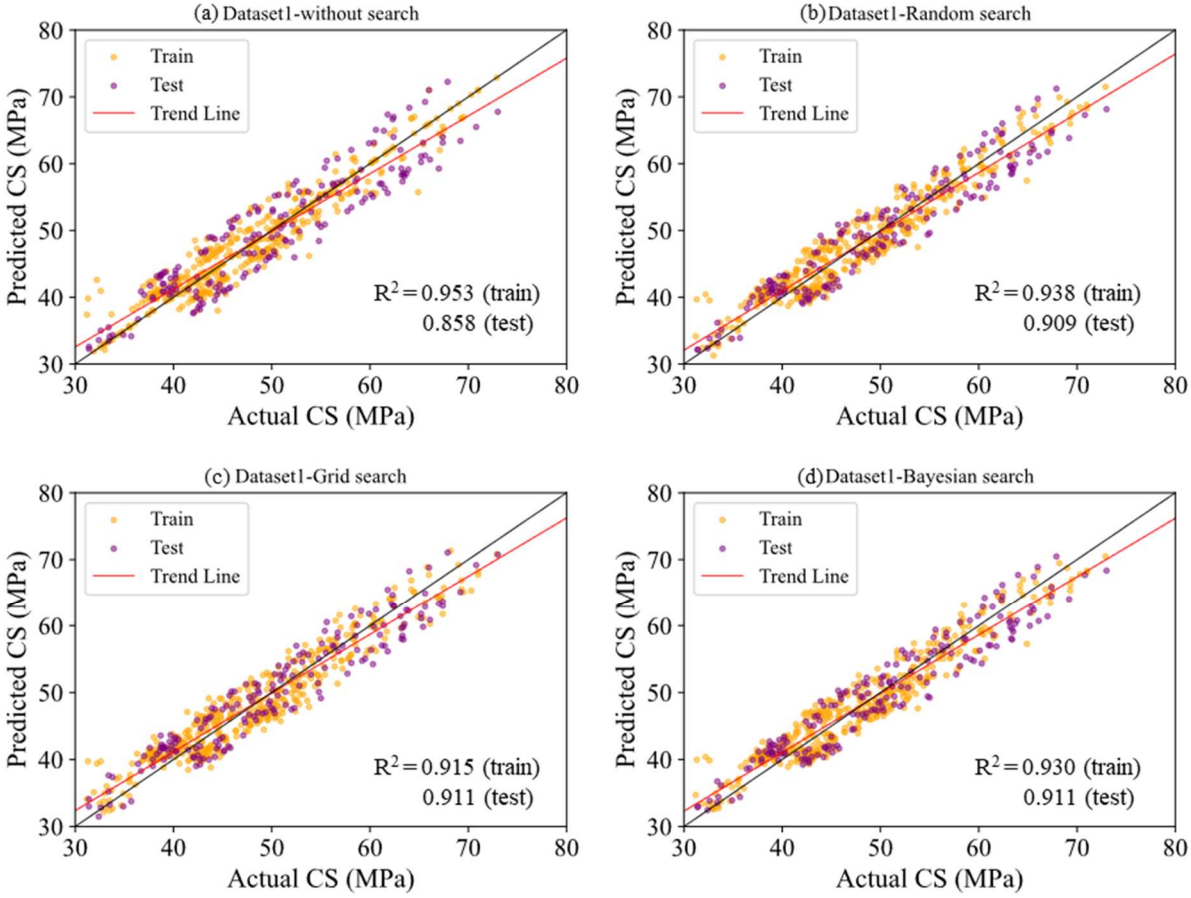

Figure 3. Scatter plots of predicted versus actual CS values of Dataset 1-based ML model: (**a**) without hyperparameter tuning; (**b**) with random search; (**c**) grid search; (**d**) Bayesian optimization.

In developing an ML model, overfitting is a critical issue that must be continuously monitored [44,45]. Overfitting occurs when a model becomes overly optimized for the training data, resulting in degraded performance on new, unseen data. Specifically, if there is a significant difference between the performance metrics of the training set and the test set, it suggests that the model may be overfitting to the training data. Thus, comparing these metrics provides meaningful insights into the generalization ability of the ML model. Figure 4 presents the performance metrics for Dataset 1. The analysis showed that, for the basic model, the differences in R^2, RMSE, and MAE between the training and test sets were 0.096, 1.81 MPa, and 1.88 MPa, respectively, which were notably larger than those observed in models with search techniques. Among the search techniques, the model using grid search exhibited the smallest differences in performance metrics, with R^2, RMSE, and MAE differences of 0.004, 0.4 MPa, and 0.38 MPa, respectively. The random search model demonstrated performance metric differences of 0.029, 0.79 MPa, and 0.74 MPa, while the Bayesian optimization model showed differences of 0.019, 0.63 MPa, and 0.58 MPa. These

findings suggest that the application of search algorithms enhances model performance and reduces error, indicating that such models are more likely to generalize effectively when exposed to new, unseen data.

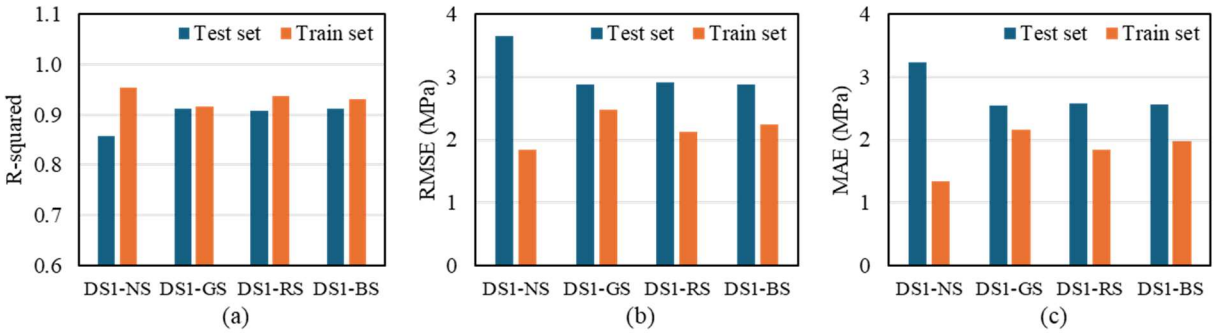

Figure 4. Performance metrics of the training and test sets of Dataset 1-based ML model: (**a**) coefficient of determination; (**b**) root mean squared error; (**c**) mean absolute error.

3.1.2. Dataset 2

Figure 5 presents a scatter plot showing the relationship between the predicted and actual compressive strength for Dataset 2. The basic model predicted compressive strength with an accuracy of 95.0% on the test data. Unlike Dataset 1, where the application of search algorithms improved prediction accuracy by approximately 4–5%, no such performance improvement was observed for Dataset 2. The models with search algorithms achieved R^2 values that were 0.9% to 1.0% higher than that of the basic model. However, given the inherent heterogeneity of concrete materials and the variability associated with ML models, such marginal improvements may not represent statistically significant differences. Furthermore, since the baseline model already achieved a high level of accuracy on Dataset 2, the practical impact of hyperparameter tuning appears to be limited.

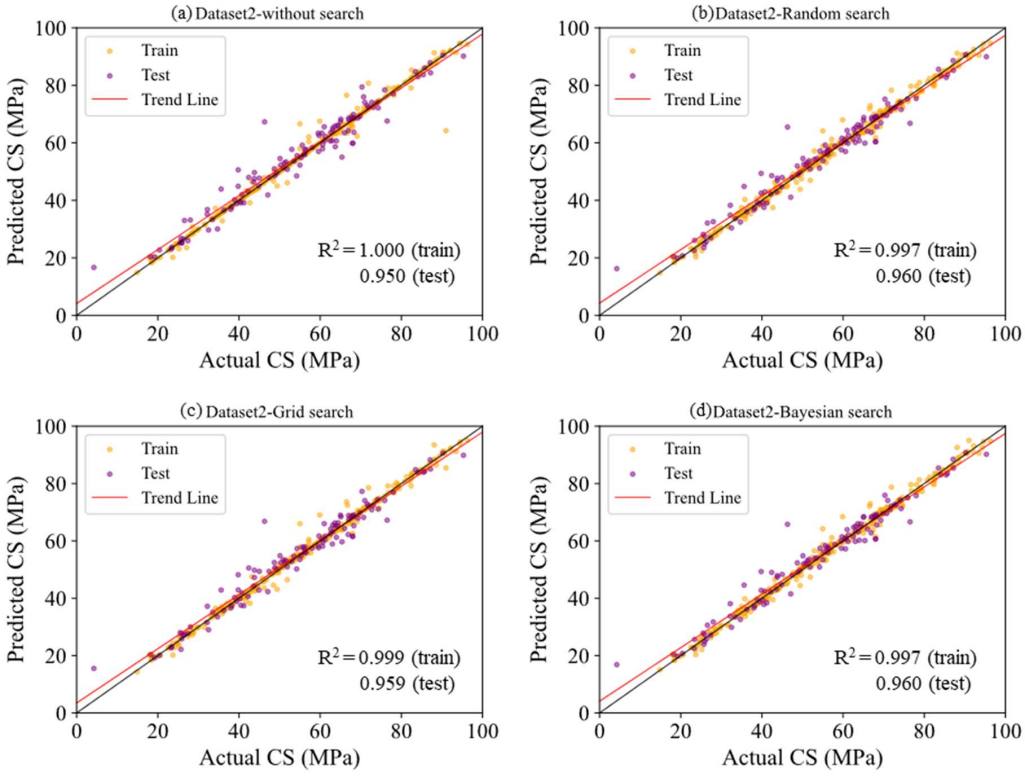

Figure 5. Scatter plots of predicted versus actual CS values of Dataset 2-based ML model: (**a**) without hyperparameter tuning; (**b**) with random search; (**c**) grid search; (**d**) Bayesian optimization.

Figure 6 presents the performance metrics for the model using Dataset 2. Regardless of the application of search algorithms, the difference in R^2 between the training and test sets ranged from 0.037 to 0.050. For RMSE and MAE, the differences between the training and test sets were observed to be between 2.46–3.62 MPa and 1.56–2.52 MPa, respectively. Similar performance metric differences between the training and test sets were reported in previous studies utilizing the XGB model to predict concrete compressive strength. In the study by Li and Song [46], the R^2, RMSE, and MAE differences between the training and test sets were 0.067, 3.189 MPa, and 2.621 MPa, respectively, while Nguyen et al. [41] reported differences of 0.098, 5.690 MPa, and 4.380 MPa for the same metrics. For Dataset 2, the hyperparameter tuning had little effect on the R^2 but contributed to a slight reduction in prediction errors. Specifically, the models with search algorithms exhibited RMSE values 10.0–10.5% lower and MAE values 8.9–11.1% lower than the baseline model.

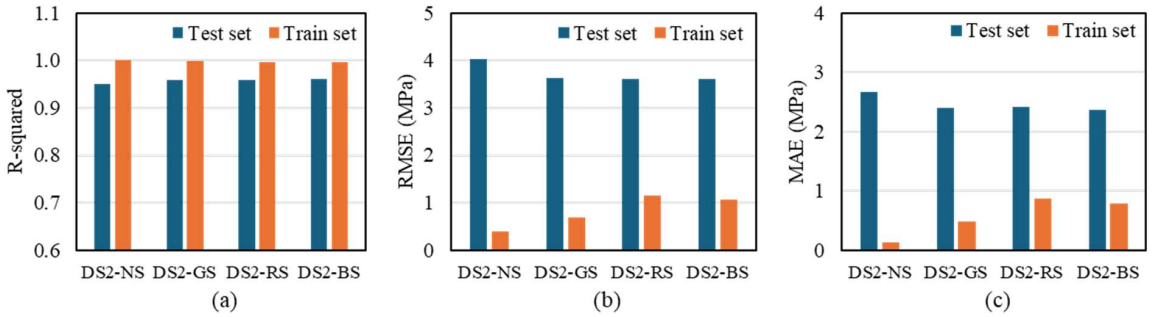

Figure 6. Performance metrics of the training and test sets of Dataset 2-based ML model: (**a**) coefficient of determination; (**b**) root mean squared error; (**c**) mean absolute error.

3.1.3. Dataset 3

Figure 7 presents the scatter plot of actual versus predicted values for Dataset 3. The prediction accuracy of the models ranged from 77.4% to 80.0%. The basic model achieved the highest accuracy, while the models with search algorithms showed slightly lower accuracy, with a decrease ranging from 0.1% to 2.6%.

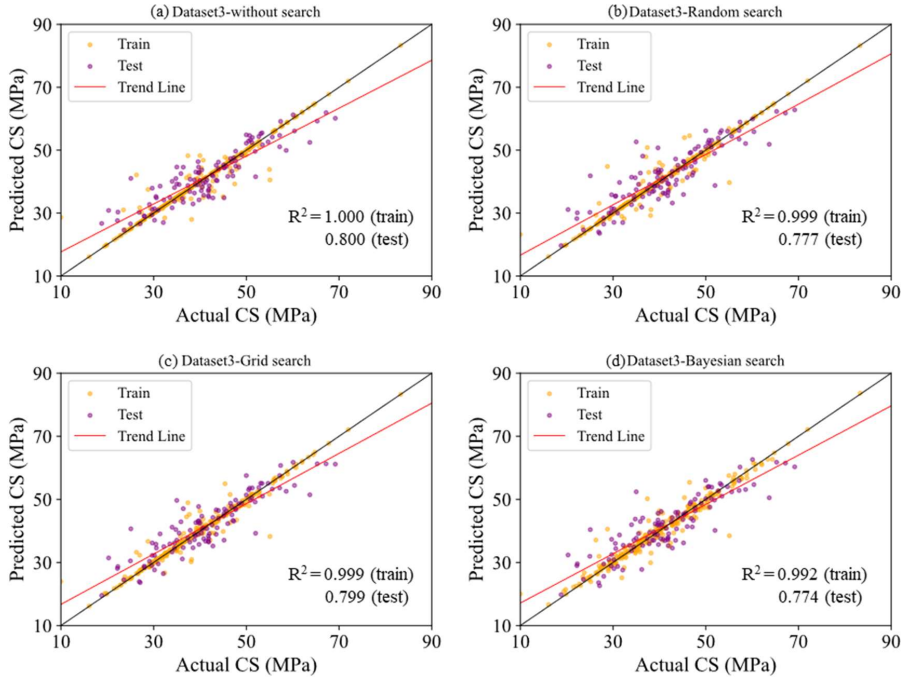

Figure 7. Scatter plots of predicted versus actual CS values of Dataset 3-based ML model: (**a**) without hyperparameter tuning; (**b**) with random search; (**c**) grid search; (**d**) Bayesian optimization.

As shown in Figure 8, the models based on Dataset 3 exhibit significant differences between the performance metrics of the training and test sets. This suggests that the models may have overfitted to the training data, which could lead to poor generalization and inaccurate predictions of new data. Moreover, contrary to the trends observed in Datasets 1 and 2, the application of search algorithms in Dataset 3 does not appear to have contributed to improving the model performance.

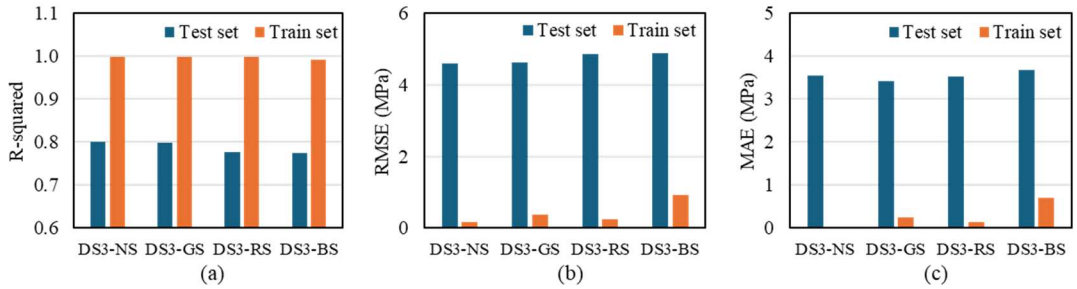

Figure 8. Performance metrics of the training and test sets of Dataset 3-based ML model: (**a**) coefficient of determination; (**b**) root mean squared error; (**c**) mean absolute error.

3.2. Post-Hoc Analysis

Post-analysis was carried out to investigate how each input feature influences the output feature under different search algorithms and to assess whether these features align with empirical knowledge. SHAP summary plots for Datasets 1 and 2 are presented in Figure 9. Regardless of the application of search algorithms, the SHAP results exhibited similar patterns within each dataset. For Dataset 1, the most influential features in the model were cement content and cement strength, followed by slag and fine aggregate contents. In Dataset 2, the w/c ratio, curing time, and the amount of water were found to be the primary factors affecting compressive strength. Additionally, the key input features generally exhibited behavior that aligned with established empirical knowledge. Specifically, the increase in SHAP values associated with a decrease in the w/c ratio (depicted by blue dots in the SHAP plot), the increase in curing days, cement content, cement strength, and the increase in the specific gravity of aggregates (represented by red points in the SHAP plot) corresponded with experimental knowledge.

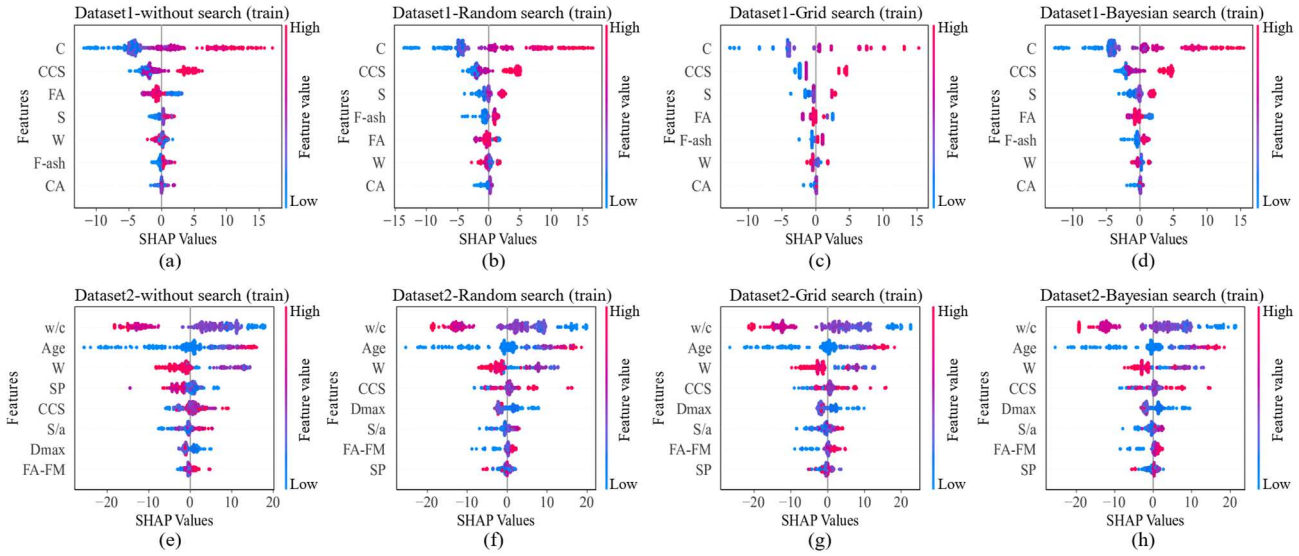

Figure 9. SHAP summary plot for the ML model based on Datasets 1 (**top**) and 2 (**bottom**): (**a**,**e**) without a search algorithm; (**b**,**f**) with random search; (**c**,**g**) grid search; (**d**,**h**) Bayesian optimization.

For Dataset 3, which exhibited the largest discrepancies in performance metrics between the training and test sets, SHAP summary plots for both sets are shown in Figure 10. As expected, in all models, the w/c ratio had the most significant influence on compressive strength, followed by features such as the specific gravity of fine aggregates, cement strength, and cement content, though these secondary influences varied slightly across models. Despite the considerable differences in performance metrics between the training and test sets, the SHAP summary plots showed similar feature behaviors in both sets, which were consistent with empirical knowledge. This suggests that feature importance analysis techniques like SHAP may not fully explain the causes of performance degradation or generalization issues. Thus, the incorporation of additional interpretability techniques or model evaluation methods may be necessary for a more comprehensive analysis of overfitting and generalization problems.

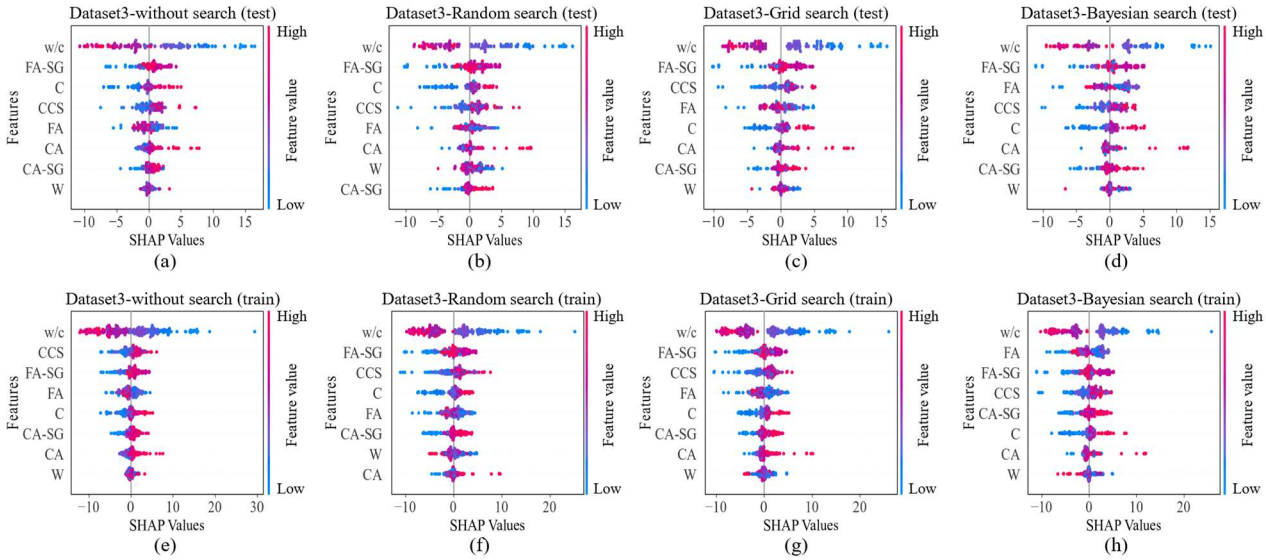

Figure 10. SHAP summary plot for the ML model based on Dataset 3: (**a**–**d**) correspond to the test set, and (**e**–**h**) to the training set. From left, model without a search algorithm (**a**,**e**); with random search (**b**,**f**); grid search (**c**,**g**); Bayesian optimization (**d**,**h**).

3.3. Comparison with Other Studies

To further examine whether the effects of hyperparameter tuning are specific to a single model, additional experiments were conducted using Random Forest (RF) and Artificial Neural Network (ANN) models. Table 4 shows the results, and similar overall trends were observed across models. In Dataset 1, tuning produced performance gains in some cases, particularly for the ANN model. In Dataset 2, moderate improvements were observed, although the impact of tuning varied depending on the model and the optimization method applied. For Dataset 3, hyperparameter tuning increased the R^2 from 0.588 to between 0.688 and 0.742. However, given the likelihood of overfitting, it remains uncertain whether these improvements reflect true gains in generalization. Further investigation is needed to assess the robustness and practical significance of these results.

According to the results of this study, the effectiveness of hyperparameter tuning search methods varied depending on the characteristics of the dataset. However, the performance differences among the applied search methods were generally negligible. These findings are consistent with previous research, which suggests that it is challenging to determine whether any specific search method consistently outperforms others. Specifically, in studies predicting the compressive strength of concrete, the shear strength of reinforced

concrete beams, and the seismic demand of bridges using ML approaches, the differences among grid search, random search, and Bayesian optimization were insignificant.

Table 4. Comparison of model performances from this study with other research.

Reference	R^2				RMSE			
	Basic	**GS**	**RS**	**BS**	**Basic**	**GS**	**RS**	**BS**
DS1 (this study)—XGB	0.858	0.911	0.909	0.911	3.647	2.881	2.918	2.874
DS1 (this study)—RF	0.879	0.892	0.893	0.894	3.354	3.172	3.159	3.152
DS1 (this study)—ANN	0.794	0.913	0.914	0.911	4.384	2.847	2.831	2.877
DS2 (this study)—XGB	0.950	0.959	0.960	0.960	4.022	3.619	3.611	3.600
DS2 (this study)—RF	0.934	0.936	0.932	0.938	4.628	4.543	4.685	4.490
DS2 (this study)—ANN	0.872	0.869	0.943	0.919	6.420	6.498	4.297	5.119
DS3 (this study)—XGB	0.800	0.799	0.777	0.774	4.613	4.619	4.873	4.901
DS3 (this study)—RF	0.763	0.760	0.763	0.760	5.024	5.052	5.017	5.055
DS3 (this study)—ANN	0.588	0.696	0.688	0.742	6.620	5.687	5.765	5.234
Zhang et al. [47]	-	0.795	0.795	0.807	-	-	-	-
Truong et al. [48]	0.896	0.927	0.927	0.936	79.995	67.124	66.845	62.904
Lei et al. [49]	-	0.885	0.893	0.918	-	-	-	-

These findings may be attributed to the limited number of data examples, which could have made it difficult to discern significant differences between the methods. Thus, future research should address these limitations by increasing the amount and diversity of data to enhance the robustness of the findings. Additionally, further investigations are needed to comprehensively analyze performance differences between search methods under a broader range of conditions. Such efforts could provide clearer insights into the impact of search method selection in hyperparameter tuning and contribute to identifying optimal strategies for diverse applications.

4. Summary and Remarks

This study investigates the effect of search algorithms on the performance of ML models for predicting concrete strength. Three distinct search algorithms were applied to three different datasets, showing that their effects varied depending on the datasets, with the following key findings:

1. In Dataset 1, the application of the search algorithm resulted in a significant improvement in the prediction accuracy of the ML model. The R^2 value increased while both RMSE and MAE decreased, indicating an overall enhancement in performance. In addition, models incorporating the search algorithm demonstrated a reduced performance gap between the training and test sets, suggesting a lower risk of overfitting.
2. In Dataset 2, although there was no significant change in the R^2 metric, improvements in RMSE and MAE indicated a slight enhancement in the prediction performance. This suggests that the search algorithm contributed to reducing prediction errors.
3. In Dataset 3, the application of the search algorithm resulted in neither significant improvements in prediction accuracy nor enhancements in overall performance. ML models based on this dataset exhibited a potential risk of overfitting and demonstrated a tendency toward lower prediction stability.
4. SHAP provided valuable insights into feature importance and produced results that were generally consistent with empirical knowledge. However, it fell short in explaining the performance degradation observed in certain datasets. This limitation may stem from the SHAP approach of evaluating features independently, which does not

capture complex feature interactions or multicollinearity. In addition, the potential presence of noise data and high-dimensional variables could obscure interpretability, particularly when working with small datasets. These challenges suggest that SHAP alone may not fully uncover the underlying causes of poor generalization or instability. To advance explainable AI in this context, future research should investigate the factors limiting the reliability of SHAP and explore complementary methods to improve model interpretability.

This study demonstrates that the effectiveness of optimization methods in ML models can vary depending on the characteristics of the dataset. Specifically, hyperparameter optimization does not always guarantee positive outcomes, and performance improvements may be limited or even negative in some cases. Given the rapid advancements in AI technology and the continuous development of various regression models using ML, this study emphasizes the critical need for careful consideration of hyperparameter optimization as a core element in model design.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings15132173/s1.

Author Contributions: Conceptualization, J.K. and D.L.; methodology, J.K. and D.L.; software, J.K. and D.L.; formal analysis, J.K.; investigation, J.K.; writing—original draft preparation, J.K.; writing—review and editing, D.L.; visualization, J.K.; funding acquisition, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2023-00244008).

Data Availability Statement: Data from this study can be accessed upon request by contacting the first author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nascimbene, R. Investigation of Seismic Damage to Existing Buildings by Using Remotely Observed Images. *Eng. Fail. Anal.* **2024**, *161*, 108282. [CrossRef]
2. Nguyen, H.; Vu, T.; Vo, T.P.; Thai, H.T. Efficient Machine Learning Models for Prediction of Concrete Strengths. *Constr. Build. Mater.* **2021**, *266*, 120950. [CrossRef]
3. Tang, F.; Wu, Y.; Zhou, Y. Hybridizing Grid Search and Support Vector Regression to Predict the Compressive Strength of Fly Ash Concrete. *Adv. Civ. Eng.* **2022**, *2022*, 3601914. [CrossRef]
4. Kargari, A.; Eskandari-Naddaf, H.; Kazemi, R. Effect of Cement Strength Class on the Generalization of Abrams' Law. *Struct. Concr.* **2019**, *20*, 493–505. [CrossRef]
5. Mehdipour, I.; Khayat, K.H. Effect of Particle-Size Distribution and Specific Surface Area of Different Binder Systems on Packing Density and Flow Characteristics of Cement Paste. *Cem. Concr. Compos.* **2017**, *78*, 120–131. [CrossRef]
6. Jung, H.; Kim, J.; Yang, H.; Kim, N. Evaluation of Residual Properties and Recovery of Fire-Damaged Concrete with Repeatedly Recycled Fine Aggregates. In *Construction Materials and Their Properties for Fire Resistance and Insulation*; Elsevier: Amsterdam, The Netherlands, 2025; pp. 165–178.
7. Kim, N.; Kim, J. Effect of Maximum Aggregate Size and Powder Content on the Properties of Self-Compacting Recycled Aggregate Concrete. *Period. Polytech. Civ. Eng.* **2023**, *67*, 1038–1047. [CrossRef]
8. Tam, V.W.Y.; Gao, X.F.; Tam, C.M. Microstructural Analysis of Recycled Aggregate Concrete Produced from Two-Stage Mixing Approach. *Cem. Concr. Res.* **2005**, *35*, 1195–1203. [CrossRef]
9. Sičáková, A.; Kim, J. Some Aspects of the Suitability of Three-Stage Mixing for Ready-Mixed Concrete. *J. Croat. Assoc. Civ. Eng.* **2024**, *76*, 709–718. [CrossRef]
10. López Gayarre, F.; López-Colina Pérez, C.; Serrano López, M.A.; Domingo Cabo, A. The Effect of Curing Conditions on the Compressive Strength of Recycled Aggregate Concrete. *Constr. Build. Mater.* **2014**, *53*, 260–266. [CrossRef]
11. Sidhu, A.S.; Siddique, R. Review on Effect of Curing Methods on High Strength Concrete. *Constr. Build. Mater.* **2024**, *438*, 136858. [CrossRef]

12. Ben Chaabene, W.; Flah, M.; Nehdi, M.L. Machine Learning Prediction of Mechanical Properties of Concrete: Critical Review. *Constr. Build. Mater.* **2020**, *260*, 119889. [CrossRef]
13. Cai, R.; Han, T.; Liao, W.; Huang, J.; Li, D.; Kumar, A.; Ma, H. Prediction of Surface Chloride Concentration of Marine Concrete Using Ensemble Machine Learning. *Cem. Concr. Res.* **2020**, *136*, 106164. [CrossRef]
14. Chou, J.S.; Tsai, C.F.; Pham, A.D.; Lu, Y.H. Machine Learning in Concrete Strength Simulations: Multi-Nation Data Analytics. *Constr. Build. Mater.* **2014**, *73*, 771–780. [CrossRef]
15. Cook, R.; Lapeyre, J.; Ma, H.; Kumar, A. Prediction of Compressive Strength of Concrete: Critical Comparison of Performance of a Hybrid Machine Learning Model with Standalone Models. *J. Mater. Civ. Eng.* **2019**, *31*, 04019255. [CrossRef]
16. Pham, A.-D.; Ngo, N.-T.; Nguyen, Q.-T.; Truong, N.-S. Hybrid Machine Learning for Predicting Strength of Sustainable Concrete. *Soft Comput.* **2020**, *24*, 14965–14980. [CrossRef]
17. Lu, X.; Yvonnet, J.; Papadopoulos, L.; Kalogeris, I.; Papadopoulos, V. A Stochastic FE2 Data-Driven Method for Nonlinear Multiscale Modeling. *Materials* **2021**, *14*, 2875. [CrossRef]
18. Sobhani, J.; Najimi, M.; Pourkhorshidi, A.R.; Parhizkar, T. Prediction of the Compressive Strength of No-Slump Concrete: A Comparative Study of Regression, Neural Network and ANFIS Models. *Constr. Build. Mater.* **2010**, *24*, 709–718. [CrossRef]
19. Lee, S.-C. Prediction of Concrete Strength Using Artificial Neural Networks. *Eng. Struct.* **2003**, *25*, 849–857. [CrossRef]
20. Asteris, P.G.; Skentou, A.D.; Bardhan, A.; Samui, P.; Pilakoutas, K. Predicting Concrete Compressive Strength Using Hybrid Ensembling of Surrogate Machine Learning Models. *Cem. Concr. Res.* **2021**, *145*, 106449. [CrossRef]
21. Yang, J.; Zeng, B.; Ni, Z.; Fan, Y.; Hang, Z.; Wang, Y.; Feng, C.; Yang, J. Comparison of Traditional and Automated Machine Learning Approaches in Predicting the Compressive Strength of Graphene Oxide/Cement Composites. *Constr. Build. Mater.* **2023**, *394*, 132179. [CrossRef]
22. Kim, J. Challenges with Hard-to-Learn Data in Developing Machine Learning Models for Predicting the Strength of Multi-Recycled Aggregate Concrete. *Appl. Soft Comput.* **2025**, *175*, 113110. [CrossRef]
23. Abd, A.M.; Abd, S.M. Modelling the Strength of Lightweight Foamed Concrete Using Support Vector Machine (SVM). *Case Stud. Constr. Mater.* **2017**, *6*, 8–15. [CrossRef]
24. Juncai, X.; Qingwen, R.; Zhenzhong, S. Prediction of the Strength of Concrete Radiation Shielding Based on LS-SVM. *Ann. Nucl. Energy* **2015**, *85*, 296–300. [CrossRef]
25. Mirmozaffari, M.; Yazdani, M.; Boskabadi, A.; Ahady Dolatsara, H.; Kabirifar, K.; Amiri Golilarz, N. A Novel Machine Learning Approach Combined with Optimization Models for Eco-Efficiency Evaluation. *Appl. Sci.* **2020**, *10*, 5210. [CrossRef]
26. Alhakeem, Z.M.; Jebur, Y.M.; Henedy, S.N.; Imran, H.; Bernardo, L.F.A.; Hussein, H.M. Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques. *Materials* **2022**, *15*, 7432. [CrossRef]
27. Joy, R.A. Fine Tuning the Prediction of the Compressive Strength of Concrete: A Bayesian Optimization Based Approach. In Proceedings of the 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, 25–27 August 2021; IEEE: Amsterdam, The Netherlands, 2021; pp. 1–6.
28. Ahmed, A.; Song, W.; Zhang, Y.; Haque, M.A.; Liu, X. Hybrid BO-XGBoost and BO-RF Models for the Strength Prediction of Self-Compacting Mortars with Parametric Analysis. *Materials* **2023**, *16*, 4366. [CrossRef]
29. Zhang, S.; Chen, W.; Xu, J.; Xie, T. Use of Interpretable Machine Learning Approaches for Quantificationally Understanding the Performance of Steel Fiber-Reinforced Recycled Aggregate Concrete: From the Perspective of Compressive Strength and Splitting Tensile Strength. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109170. [CrossRef]
30. Du, X.; Xu, H.; Zhu, F. Understanding the Effect of Hyperparameter Optimization on Machine Learning Models for Structure Design Problems. *Comput. Aided Des.* **2021**, *135*, 103013. [CrossRef]
31. Li, Z.; Yoon, J.; Zhang, R.; Rajabipour, F.; Srubar III, W.V.; Dabo, I.; Radlińska, A. Machine Learning in Concrete Science: Applications, Challenges, and Best Practices. *npj Comput. Mater.* **2022**, *8*, 127. [CrossRef]
32. Kim, J.; Lee, D.; Ubysz, A. Comparative Analysis of Cement Grade and Cement Strength as Input Features for Machine Learning-Based Concrete Strength Prediction. *Case Stud. Constr. Mater.* **2024**, *21*, e03557. [CrossRef]
33. Zheng, W.; Shui, Z.; Xu, Z.; Gao, X.; Zhang, S. Multi-Objective Optimization of Concrete Mix Design Based on Machine Learning. *J. Build. Eng.* **2023**, *76*, 107396. [CrossRef]
34. Zhao, S.; Hu, F.; Ding, X.; Zhao, M.; Li, C.; Pei, S. Dataset of Tensile Strength Development of Concrete with Manufactured Sand. *Data Brief.* **2017**, *11*, 469–472. [CrossRef]
35. Yi, S.T.; Yang, E.I.; Choi, J.C. Effect of Specimen Sizes, Specimen Shapes, and Placement Directions on Compressive Strength of Concrete. *Nucl. Eng. Des.* **2006**, *236*, 115–127. [CrossRef]
36. Quan Tran, V.; Quoc Dang, V.; Si Ho, L. Evaluating Compressive Strength of Concrete Made with Recycled Concrete Aggregates Using Machine Learning Approach. *Constr. Build. Mater.* **2022**, *323*, 126578. [CrossRef]
37. Kamath, M.V.; Prashanth, S.; Kumar, M.; Tantri, A. Machine-Learning-Algorithm to Predict the High-Performance Concrete Compressive Strength Using Multiple Data. *J. Eng. Des. Technol.* **2024**, *22*, 532–560. [CrossRef]

38. Lyngdoh, G.A.; Zaki, M.; Krishnan, N.M.A.; Das, S. Prediction of Concrete Strengths Enabled by Missing Data Imputation and Interpretable Machine Learning. *Cem. Concr. Compos.* **2022**, *128*, 104414. [CrossRef]
39. Dong, Y.; Tang, J.; Xu, X.; Li, W.; Feng, X.; Lu, C.; Hu, Z.; Liu, J. A New Method to Evaluate Features Importance in Machine-Learning Based Prediction of Concrete Compressive Strength. *J. Build. Eng.* **2025**, *102*, 111874. [CrossRef]
40. Nguyen-Sy, T.; Wakim, J.; To, Q.-D.; Vu, M.-N.; Nguyen, T.-D.; Nguyen, T.-T. Predicting the Compressive Strength of Concrete from Its Compositions and Age Using the Extreme Gradient Boosting Method. *Constr. Build. Mater.* **2020**, *260*, 119757. [CrossRef]
41. Nguyen, N.-H.; Abellán-García, J.; Lee, S.; Garcia-Castano, E.; Vo, T.P. Efficient Estimating Compressive Strength of Ultra-High Performance Concrete Using XGBoost Model. *J. Build. Eng.* **2022**, *52*, 104302. [CrossRef]
42. Duan, J.; Asteris, P.G.; Nguyen, H.; Bui, X.-N.; Moayedi, H. A Novel Artificial Intelligence Technique to Predict Compressive Strength of Recycled Aggregate Concrete Using ICA-XGBoost Model. *Eng. Comput.* **2021**, *37*, 3329–3346. [CrossRef]
43. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
44. Vasicek, D. Artificial Intelligence and Machine Learning: Practical Aspects of Overfitting and Regularization. *Inf. Serv. Use* **2020**, *39*, 281–289. [CrossRef]
45. Ying, X. An Overview of Overfitting and Its Solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]
46. Li, Q.-F.; Song, Z.-M. High-Performance Concrete Strength Prediction Based on Ensemble Learning. *Constr. Build. Mater.* **2022**, *324*, 126694. [CrossRef]
47. Zhang, X.; Dai, C.; Li, W.; Chen, Y. Prediction of Compressive Strength of Recycled Aggregate Concrete Using Machine Learning and Bayesian Optimization Methods. *Front. Earth Sci.* **2023**, *11*, 1112105. [CrossRef]
48. Truong, G.T.; Choi, K.-K.; Nguyen, T.-H.; Kim, C.-S. Prediction of Shear Strength of RC Deep Beams Using XGBoost Regression with Bayesian Optimization. *Eur. J. Environ. Civ. Eng.* **2023**, *27*, 4046–4066. [CrossRef]
49. Lei, X.; Feng, R.; Dong, Y.; Zhai, C. Bayesian-Optimized Interpretable Surrogate Model for Seismic Demand Prediction of Urban Highway Bridges. *Eng. Struct.* **2024**, *301*, 117307. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.