**Sustainable Cement based Mix Design Using CatBoost model optimized with Firefly Optimization Algorithm**

Teiboklang Chyne1, \* and Biranchi Panda1

1 Indian Institute of Technology Guwahati

\* Corresponding Author  
[teibokchyne@gmail.com](mailto:teibokchyne@gmail.com%20)

**Abstract.** In structural engineering, concrete compressive strength is the most important performance parameter for designing the conventional concrete, however, challenges lie in optimizing the mix design formulation, which can be time-consuming and material-intensive. This paper focuses on the applicability of CatBoost model constructed using Firefly optimization algorithm for improved forecasting the compressive strength of sustainable concrete. A comprehensive and trustworthy data set is used to devel-op the CatBoost models, which include 785 test results and 15 input features. Shapley Additive Explanations (SHAP) analysis is conducted to interpret the model's predictions and provide insights into the most influential factors affecting the mix design. The results demonstrate that the optimized CatBoost model, with carefully selected hyper parameters including learning rate, depth, and iterations, can reliably predict the performance of sustainable concrete. The SHAP analysis showed that curing time, fiber dosage, and cement dosage significantly influence the compressive strength property. This approach significantly improves the predictive accuracy of the mix designs, reducing both material waste and computational time. Future work will focus on validating these predictions through experiments, potentially extending its application to other critical parameters in the concrete performance.

**Keywords:** Concrete Strength, Machine Learning, CatBoost Algorithm, Firefly Optimization Algorithm, SHAP Analysis.

**1 Introduction**

Concrete plays a vital role in the global construction industry due to its exceptional compressive strength, durability, cost-effectiveness, and adaptability across a wide range of structural applications. In recent years, its importance has also been reinforced by efforts to enhance sustainability using supplementary cementitious materials (SCMs), and low-carbon alternatives, addressing environmental concerns while maintaining performance [1]. Multiple factors shape the qualities of concrete, including the type and content of cement, which influence strength, setting time, and durability. The water-cement ratio is critical for controlling strength, workability, and permeability [2]. Aggregate properties such as size, shape, and gradation affect both the mechanical strength and workability of the mix [3]. The inclusion of chemical or mineral admixtures—such as superplasticizers, retarders, or fly ash—can modify setting time, enhance workability, or improve durability [4]. A uniform mixing process ensures consistent material distribution, while appropriate curing conditions, particularly temperature and moisture, are essential for strength development and long-term durability. Environmental factors like ambient temperature and humidity at the time of mixing and placing also influence workability and set time [5]. Supplementary cementitious materials including silica fume and slag, further enhance properties such as chemical resistance and durability. Lastly, effective placement and compaction techniques eliminate voids within the mix, thereby contributing to the structural integrity and durability of the concrete [6]. Integrating SCMs like silica fume and slag into concrete mixes offers several advantages, including cost reduction, utilization of industrial by-products, and enhanced strength and durability. Crucially, this practice also diminishes dependence on ordinary Portland cement (OPC), thereby reducing the significant CO₂ emissions associated with its production [7].

Given the multifaceted nature of these variables and their interdependencies, accurately predicting the compressive strength of concrete remains a complex task. Consequently, scholars have increasingly adopted ML to predict the properties of cement-based composites. Catboost has proven particularly suitable for this purpose due to its ability to manage nonlinearity, high-dimensional data, missing values, and to generalize effectively. For example, Li et al. [8] demonstrates the utility of combining CatBoost with the Firefly Optimization Algorithm (FOA) to predict sintering characteristics of iron ore powder, reducing reliance on domain expertise. Katlav and Ergen [9] extend the research to ultra-high-performance concrete (UHPC), comparing CatBoost models optimized with various metaheuristics—including Phasor particle swarm optimization (PPSO), dwarf mongoose optimization (DMO), and atom search optimization (ASO). Elshaarawy et al. [10] and Zou et al. [11] reinforce these findings, using SHAP to illuminate CatBoost’s decision pathways and manage high-dimensional inputs in fiber-reinforced concrete applications.

Based on the above, this explores the use of Firefly Optimization Algorithm (FOA) along with CatBoost in predicting the compressive strength of concrete, thus facilitating further research on using machine learning to generate new mixture designs.

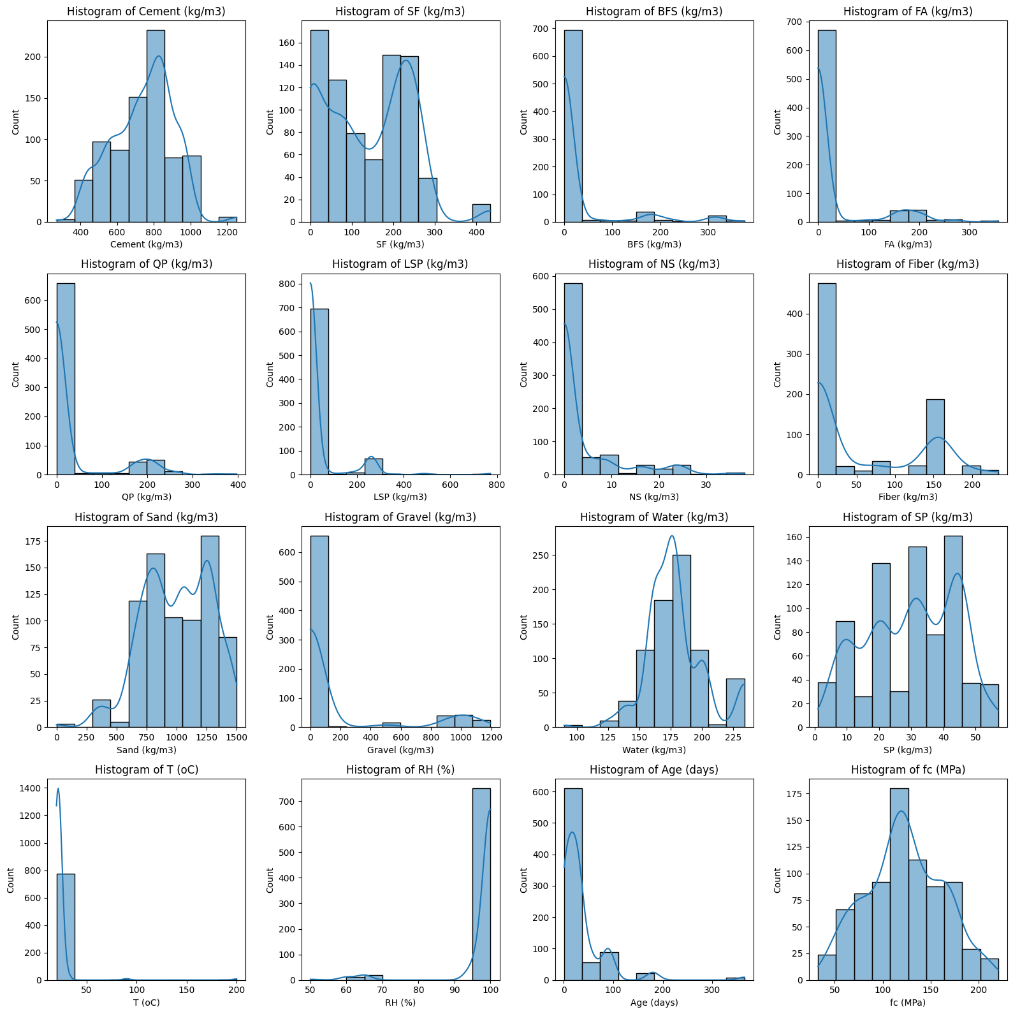
**2 Research Methodology**

**2.1 Data Description**

In this study, the dataset used for analysis was sourced from the work of Katlav and Ergen [9], who improved forecasting of the compressive strength of ultra-high-performance concrete (UHPC) using the CatBoost model optimized with different algorithms. The dataset contains 785 rows of data and 15 features, which were used to train and evaluate the models in this research. The dataset contains several important features relevant to concrete mix design, categorized into two main groups: input features (independent variables) and an output feature (dependent variable). The input features are the mix constituents and environmental parameters that affect the compressive strength of the printed concrete. These features include: Cement (kg/m³), which refers to the amount of cement used in the concrete mix; SF (kg/m³) or silica fume, a fine additive that enhances the concrete’s strength; BFS (kg/m³), blast furnace slag used as a supplementary cementitious material; FA (kg/m³), fly ash, a pozzolanic material used to improve the mix's durability and workability; QP (kg/m³), quarry powder used as a filler material; LSP (kg/m³), limestone powder, another filler material; NS (kg/m³), nano-silica, which improves mechanical properties; Fiber (kg/m³), added to enhance tensile strength and reduce cracking; Sand (kg/m³), fine aggregate; Gravel (kg/m³), coarse aggregate for structural integrity; Water (kg/m³), which affects workability and strength; SP (kg/m³), superplasticizer, a chemical admixture for flowability; T (°C), temperature at the time of mixing and printing, influencing curing and strength development; RH (%), relative humidity during printing, affecting drying and setting; and Age (days), the number of days the concrete has been curing. The output feature is the compressive strength (fc) of the concrete, measured in megapascals (MPa). This feature is used as the dependent variable in the machine learning model, where predictions are made based on the input features.

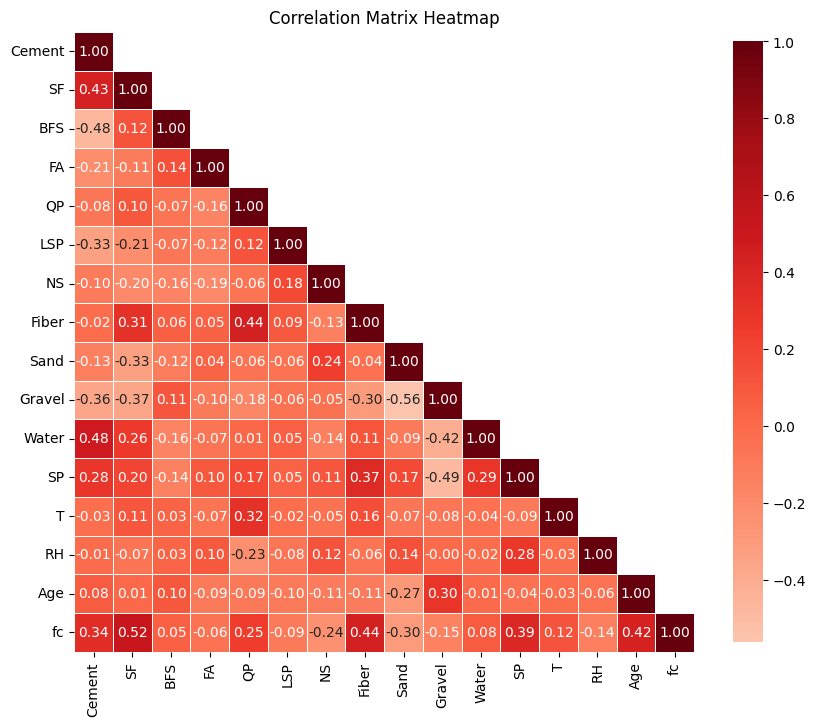
**2.2. Statistics of Variables**

The frequency distribution plots presented in Fig. 1 illustrate the distribution patterns, dominant values, and potential outliers for various concrete mix constituents and related parameters. Cement exhibits a unimodal distribution with a peak at 800 kg/m³ and a mild negative skew, while silica fume (SF) shows a bimodal pattern with peaks at 0 and 200 kg/m³. BFS, FA, QP, and LSP all display extreme right skewness, indicating limited usage and dominant peaks at 0 kg/m³, with occasional higher values. NS, fiber, and gravel also demonstrate low usage, with fiber and gravel exhibiting bimodal and right-skewed distributions, respectively. Sand shows a complex, multimodal distribution, and water follows a slightly right-skewed unimodal pattern centered around 175–180 kg/m³. SP displays a multimodal trend with prominent peaks at specific dosage levels. Temperature and age distributions are highly right-skewed, with temperature peaking at 20–30 °C and age at or below 28 days. Relative humidity is concentrated near 100%, and compressive strength follows a unimodal distribution centered around 125 MPa, with a mild right skew.



**Fig. 1.** Frequency Distribution of Variables.

The correlation heatmap (Fig. 2) illustrates the pairwise correlations between features, with a focus on compressive strength. The color gradient ranges from light orange (negative correlation) to brown (positive correlation), where the intensity reflects the correlation strength.



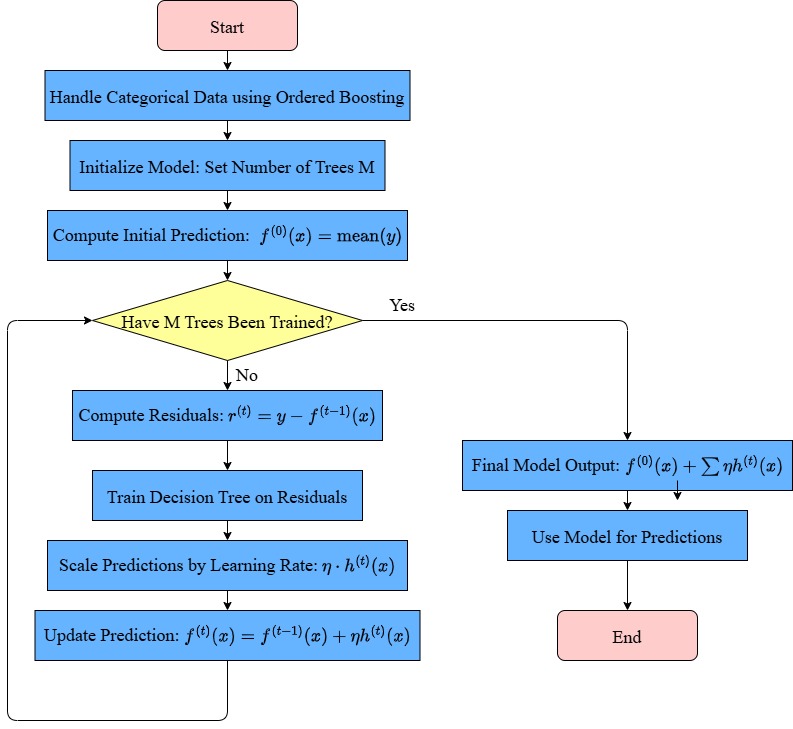
**Fig. 2.** Correlation Matrix Heatmap of Variables.

**2.3 Performance Assessment Metrics**

The performance of the models is evaluated through a series of statistical metrics, which provide an objective measure of the models' ability to predict outcomes accurately. Key metrics employed in this evaluation include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R2).

**2.4 Algorithm structure of CatBoost**

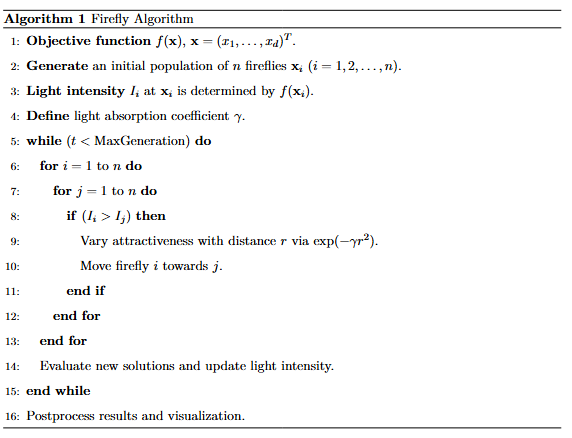
In this section, the basic principles and structure of the CatBoost model are first introduced (Fig. 3). It was developed by Yandex in 2017, is a gradient boosting algorithm designed to handle datasets with categorical features efficiently. CatBoost builds models iteratively by minimizing a loss function using decision trees, with updates guided by pseudo-residuals and a learning rate. The algorithm initializes with a baseline prediction and gradually improves it through multiple iterations. Its mathematical formulation revolves around minimizing a specified loss function using additive tree models.



**Fig. 3.** Flowchart of the CatBoost Regression Algorithm.

**2.5 Custom Implementation of Firefly Optimization Algorithm**

The Firefly Optimization Algorithm (FOA) (Fig. 4) is a nature-inspired metaheuristic based on the social behavior of fireflies, where their attraction is determined by the brightness associated with solution quality. In FOA, less fit fireflies move toward more promising solutions, guided by brightness and distance, with brightness evaluated using Mean Squared Error (MSE). The fireflies' movement incorporates both attractiveness and a random component, with distances computed after applying standard scaling and using Euclidean distance. Standard scaling is applied to ensure that each dimension contributes equally to the distance calculation, preventing any single feature from disproportionately influencing the result due to differences in scale. Five distance metrics—Euclidean Distance, Manhattan Distance (L1 norm), Minkowski Distance (generalized form), Cosine Distance, and Chebyshev Distance—were evaluated. A preliminary run of the Firefly Optimization Algorithm over 30 iterations (with all runs initialized using identical populations) indicated that the Euclidean Distance metric yielded the best performance. The algorithm proceeds through initialization, iterative position updates, and termination upon meeting a stopping criterion. Although FOA traditionally moves toward brighter fireflies, in this context, fireflies are directed toward dimmer ones to minimize MSE. The parameters for the FOA were selected based on the guidelines and empirical findings presented in [12, 13]. Its robust exploration-exploitation balance makes FOA suitable for complex optimization tasks across various domains, including engineering and machine learning.



**Fig. 4.** The Firefly Optimization Algorithm.

**2.9 Implementation**

This study used the CatBoost algorithm to predict the compressive strength of concrete, leveraging its ability to handle both numerical and categorical features effectively. The model was initially trained without specifying hyperparameters to establish a baseline. To enhance performance, the FOA was applied to fine-tune seven key hyperparameters, including learning rate, tree depth, and number of iterations. FOA iteratively searched for the optimal parameter set by simulating the behavior of fireflies attracted to dimmer (better-performing) solutions, minimizing prediction error (MSE). The optimized model was compared with the baseline using evaluation metrics. Regression slope analysis on both training and testing sets confirmed improved prediction accuracy after FOA optimization. A 5-fold cross-validation was conducted to ensure the model's robustness and generalization. SHAP analysis was performed to interpret feature contributions, revealing the direction and magnitude of each variable’s influence on compressive strength predictions, and highlighting the most critical mix design and environmental factors.

**3 Results**

**3.1 Baseline Performance of CatBoost Model**

Initially, the CatBoost model was trained without specifying the hyperparameters to establish a baseline for performance. The results for this default model are presented in Table 1. With an R2 score of 0.9594 on the testing dataset, the model could explain approximately 95.94% of the variance in the target variable (compressive strength). The relatively low values of MAE (5.9288) and RMSE (7.9679) indicate that the model’s initial predictions were fairly accurate. However, there was potential to enhance its predictive capability by tuning hyperparameters.

**Table 1**. Comparison of Default CatBoost and FOA-CatBoost Model Results.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Training** | | |  |  | | **Test** | |  |
|  | **MAE** | | **R²** | **RMSE** | | **MAE** | **R²** | | **RMSE** | |
| CatBoost | 4.0771 | | 0.9803 | 5.6401 | | 5.9288 | 0.9594 | | 7.9679 | |
| FOA-CatBoost | 1.7785 | | 0.9928 | 3.402 | | 3.9006 | 0.9786 | | 5.5442 | |

**3.2 Hyperparameter Tuning**

The FOA optimization of the CatBoost model’s hyperparameters led to notable improvements in performance. Key hyperparameters for the FOA-CatBoost model were adjusted as follows (Table 2). These hyperparameters reflect a well-balanced CatBoost model that avoids overfitting while effectively capturing underlying data patterns.

**Table 2**. FOA-CatBoost Hyperparameters.

|  |  |
| --- | --- |
| **Hyperparameter** | **Value** |
| Learning Rate | 0.2286 |
| Depth | 6 |
| L2 Leaf Regularization | 35 |
| Bagging Temperature | 5.7426 |
| Subsample Ratio | 0.6313 |
| Border Count | 146 |
| Iterations | 2117 |

**3.5 Regression Slope Analysis**

The predictive accuracy of the models was evaluated using regression slope analysis, comparing actual and predicted values. As shown in Fig. 5, the CatBoost model before optimization (Fig. 5a) demonstrated regression slopes of 0.984 (training) and 0.959 (validation) with R2 values of 0.98 and 0.96, indicating strong correlations but slight overfitting. After optimization using the Firefly Algorithm, the FOA-CatBoost model (Fig. 5b) achieved improved slopes of 0.992 and 0.978 with R2 values of 0.99 and 0.98 for training and validation, respectively. These results show that FOA-CatBoost aligns more closely with the ideal slope of 1, demonstrating enhanced consistency, reduced overfitting, and improved generalization compared to the unoptimized CatBoost model.

|  |  |
| --- | --- |
|  |  |
| (a) CatBoost model | (b) FOA-CatBoost model |

**Fig. 5.** Regression slope analysis of the models.

**3.6 5-Fold Cross-Validation Results**

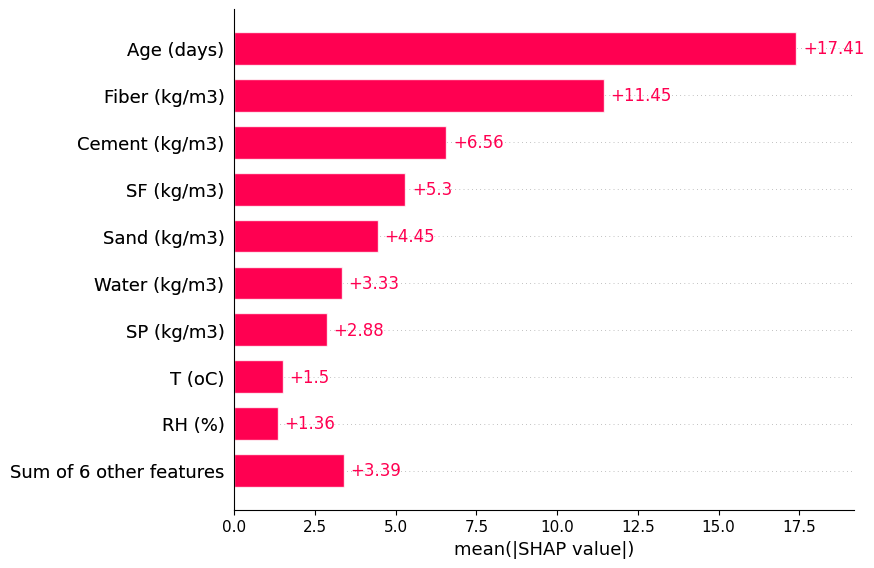
To assess the model's performance and generalizability, 5-fold cross-validation was employed. This technique divides the dataset into five subsets, allowing each to serve once as a validation set while the others form the training set. It helps minimize overfitting and yields reliable performance estimates. The aggregated metrics from all folds were: MAE = 4.5256, R2 score = 0.9715, and RMSE = 6.6891. These results indicate strong predictive performance, with the model explaining approximately 97.15% of the variance in the target variable. The low MAE and RMSE values confirm the model’s accuracy and reliability. Detailed fold-wise results are provided in Table 3.

**Table 3**. 5-Fold Cross-Validation Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Fold** | **MAE** | **R²-score** | **RMSE** |
| 1 | 4.128 | 0.9792 | 5.7107 |
| 2 | 4.2752 | 0.9741 | 6.2558 |
| 3 | 4.9072 | 0.9713 | 7.252 |
| 4 | 5.417 | 0.9543 | 8.6869 |
| 5 | 3.9006 | 0.9786 | 5.5452 |

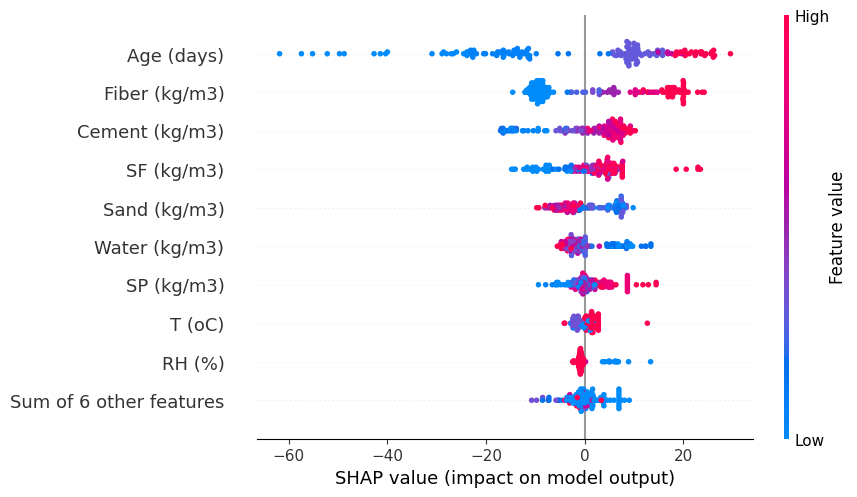
**3.7 SHAP Analysis**

The SHAP analysis was performed on the FOA-CatBoost model due to its superior predictive performance. The mean SHAP plot (Fig. 6) highlights the most influential features affecting the prediction of concrete compressive strength. The analysis reveals that curing age is the most impactful factor, with a mean SHAP value of 17.41, reflecting the natural increase in strength over time due to hydration. Fiber content is the second most influential feature (SHAP value: 11.45), indicating its significant role in enhancing tensile strength and crack resistance. Cement content (6.56) and silica fume (5.3) are also critical, contributing to strength through their binding properties and improved matrix structure. Sand (4.45) and water (3.33) moderately influence strength, with water content being essential due to its impact on the water-cement ratio. Temperature (1.5) and superplasticizer (2.88) also affect strength by influencing curing and workability. Relative humidity (1.36) has a lesser but still relevant impact by affecting the curing environment. The remaining features contribute a cumulative SHAP value of 3.39, indicating a collectively minor but non-negligible influence.



**Fig. 6.** The Mean Absolute SHAP values plot.

A SHAP summary plot (Fig. 7) illustrates each feature's effect on the prediction of concrete compressive strength. The y-axis ranks features by importance, while the x-axis shows SHAP values, which indicate the magnitude and direction of a feature’s influence on the model’s output. Each point represents a data instance, with color indicating feature value—blue for low and red for high. Age is the most influential feature, with higher values positively affecting compressive strength, consistent with the behavior of concrete gaining strength over time due to hydration. Fiber content also shows a strong positive impact, with higher fiber values enhancing strength, likely due to improved tensile properties and crack resistance. Cement content follows a similar trend, positively contributing to strength by forming a denser, more cohesive matrix. Silica fume (SF) further strengthens concrete, with high values improving microstructure and durability. In contrast, sand and water exhibit negative impacts. High values of these features correlate with decreased strength, likely due to increased porosity and weaker bonding. Temperature (T) show a generally positive influence, as higher temperatures can accelerate curing. However, higher humidity (RH) tended to have negative SHAP impact while lower RH tended to have positive SHAP impact.



**Fig. 7.** The SHAP feature impact plot.

**4 Conclusion**

The optimized CatBoost model, configured with the hyperparameters outlined in Table 2, demonstrated considerable improvements in the testing phase as illustrated in Table 1. The decrease in MAE and RMSE, along with the increase in R2 score, highlights the optimized model’s ability to provide more precise predictions with reduced variance when compared to the default configuration. Fine-tuning critical parameters such as learning rate, depth, and L2 regularization enabled the model to more accurately capture the underlying data patterns while avoiding overfitting in the testing phase. The 5-fold cross-validation results highlight the model’s consistency and reliability across different subsets of data. The slight variations in performance metrics across the folds are typical and indicate the model’s robustness in generalizing to unseen data. The overall high R2 score suggests that the model is well-tuned and capable of providing accurate predictions, making it a valuable tool for the intended application. Additionally, the SHAP analysis provided insights into feature importance, revealing the primary factors influencing the model’s predictions.

**References**

|  |  |
| --- | --- |
| 1. | Scrivener, K.L., John, V.M., Gartner, E.M.: Eco-efficient cements: Potential economically viable solutions for a low-CO₂ cement-based materials industry. Cem Concr Res. 114, 2–26 (2018). https://doi.org/10.1016/j.cemconres.2018.03.015. |
| 2. | Ghasemipour, V., others: Experimental Evaluation of the Effect of Water-Cement Ratio on Compressive, Abrasion Strength, Hydraulic Conductivity Coefficient and Porosity of Nano-Silica Concretes. Journal of Applied Engineering Sciences. 8, 17–24 (2018). |
| 3. | Akinnolu, A., others: Effect of Aggregate Shapes on the Properties of Concrete. LAUTECH Journal of Civil and Environmental Studies. 5, 1–8 (2020). |
| 4. | Ramachandran, V.S.: Effect of chemical admixtures on workability and strength of concrete. Cem Concr Res. 25, 353–362 (1995). |
| 5. | Technology, C.E.: The Critical Role of Temperature in Concrete: From Mixing to Curing. Civil Engineering Technology. (2023). |
| 6. | Ghosh, S., Chattopadhyay, B.: Compaction of concrete: A critical review. Journal of Civil Engineering Research. 8, 1–9 (2018). |
| 7. | Supplementary Cementitious Materials Improving Sustainability of Concrete. NBM&CW. (2024). |
| 8. | Li, Y., Duan, Y., Zhou, Y., Yang, J., Li, F., Yang, A.: Research on prediction model of iron ore powder sintering foundation characteristics based on FOA-Catboost algorithm. Alexandria Engineering Journal. 86, 603–615 (2024). |
| 9. | Katlav, M., Ergen, F.: Improved forecasting of the compressive strength of ultra-high-performance concrete (UHPC) via the CatBoost model optimized with different algorithms. Structural Concrete. (2024). |
| 10. | Elshaarawy, M.K., Alsaadawi, M.M., Hamed, A.K.: Machine learning and interactive GUI for concrete compressive strength prediction. Sci Rep. 14, 16694 (2024). |
| 11. | Zou, D.L., Wu, L.L., Hao, Y.F., Xu, L., Chen, J.J.: Composition-strength relationship study of ultrahigh performance fiber reinforced concrete (UHPFRC) using an interpretable data-driven approach. Constr Build Mater. 392, 131973 (2023). |
| 12. | Yang, X.-S.: Firefly algorithms for multimodal optimization. In: International Symposium on Stochastic Algorithms. pp. 169–178. Springer (2009). |
| 13. | Fister, I., Yang, X.-S., Fister Jr, I., Fister, D., Brest, J.: A comprehensive review of firefly algorithms. Swarm Evol Comput. 13, 34–46 (2013). |