buildings

MDPI

Article

Interpretable Prediction of Geopolymer Concrete Compressive Strength Using DBO–CatBoost and SHAP Analysis

Nima Saeedi [1], Zahra Mohammadipour Novin [2], Amirreza Shirini [3], Sina Samadi Gharehveran [1], Siamak Pedrammehr [4] and Mohammad Fotouhi [5,*]

[1] Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz 5166616471, Iran; nimasaeidi69@gmail.com (N.S.); sinasamadi.gh@gmail.com (S.S.G.)
[2] Faculty of Mechanical Engineering, University of Tabriz, Tabriz 5166616471, Iran; zahrampnavin84@gmail.com
[3] Faculty of Civil Engineering, Sahand University of Technology, Tabriz 5164736931, Iran; amirshirini81620@gmail.com
[4] Faculty of Design, Tabriz Islamic Art University, Tabriz 5164736931, Iran; s.pedrammehr@tabriziau.ac.ir
[5] Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, The Netherlands
* Correspondence: m.fotouhi-1@tudelft.nl

Abstract

The construction sector faces a critical need to minimize its carbon footprint, which is currently stimulating the development of geopolymer concrete using recycled coarse aggregates as an eco-friendly material compared with Portland cement. Accurate prediction of the compressive strength of this eco-efficient concrete is complex, however, as a result of the complex, non-linear interactions between many of the mix-design and curing parameters. Although modern scientific literature and engineering practices have increasingly adopted machine learning (ML) for concrete strength prediction, a significant scientific gap remains. Most existing studies rely on "black-box" models that lack sufficient interpretability and frequently overlook the severe risk of data leakage during validation, limiting their practical engineering application. To address this gap, this study proposes a robust, data-leakage-aware framework driven by a rigorous nested GroupKFold cross-validation strategy. By grouping concrete samples by their unique Mix_ID, this approach ensures genuine generalization to entirely unseen mixtures. Within this reliable validation scheme, the CatBoost algorithm is utilized for compressive-strength prediction, with the Dung Beetle Optimizer (DBO) serving as an effective tool for hyperparameter tuning. The evaluation results across multiple random seeds show that the DBO–CatBoost model significantly outperforms the default CatBoost, rigorously tuned baseline models (Support Vector Regression and Random Forest), and a comparative metaheuristic benchmark (PSO–CatBoost). It achieves the most stable distribution of errors and excellent predictive accuracy (Test $R^2 = 0.9995 \pm 0.0002$, RMSE = 0.3828 ± 0.0909). In addition, the model predictions were demystified using the methods of SHapley Additive exPlanations (SHAP) and partial dependence plots (PDPs). The interpretability analysis revealed strong statistical associations, showing that Curing Time and Coarse Aggregate are the most prominent predictive features and the strongest pairwise interaction between each other; the NaOH molar concentration is the most important second-level influence on optimization of strength. Overall, the framework provides a robust data-driven screening tool that can assist in preliminary mix-design evaluation. By reducing the reliance on extensive empirical "trial and error" approaches, this predictive model supports more efficient material usage and facilitates preliminary optimization of low-carbon concrete formulations. Theoretically, this study advances the fundamental science of geopolymer materials by explicitly quantifying the

Academic Editors: Ruby Mejía de Gutiérrez and Rafael A. Robayo-Salazar

Received: 16 July 2026
Revised: 12 August 2026
Accepted: 15 August 2026
Published: 21 August 2026

complex, non-linear interactions between alkaline activators, curing conditions, and recycled aggregates. This provides a robust data-driven theoretical foundation for designing and optimizing next-generation eco-friendly concrete products and structures.

Keywords: geopolymer concrete; recycled coarse aggregate; alkaline activator; compressive strength; eco-friendly construction materials; machine learning; SHAP analysis

1. Introduction

The disproportionate contribution of ordinary Portland cement (OPC) to anthropogenic CO_2 emissions (approximately 7–8% of global total emissions) has been a driver of a concerted move toward alternative binder systems. Among these, geopolymer concrete (GPC) is one of the most significant, produced by the alkali activation of aluminosilicate precursors (such as fly ash and ground granulated blast-furnace slag (GGBFS)) with sodium hydroxide (NaOH) and sodium silicate (Na_2SiO_3) solutions, rather than Portland clinker [1,2]. Hence, the integration of recycled coarse aggregate from construction and demolition waste (CDW) into geopolymer systems is a promising route toward circular low-carbon construction materials [1,3].

However, there are several mix-design and curing variables that interact with one another in a highly non-linear manner. For instance, the effect of curing time on compressive strength is strongly influenced by the coarse aggregate replacement level, and the benefit of an elevated NaOH concentration is not independent of curing temperature. Because of these complex interactions, empirical mix-design equations and single-variable regression models are structurally incapable of capturing the true behavior of the material. Furthermore, determining this behavior experimentally for every candidate mix is costly and slow, as specimens must be cast, cured, and tested at several ages before a single strength value is available [1–3].

To reduce the experimental workload, data-driven machine learning (ML) prediction has emerged as a rapidly developing research trend in concrete technology. Ensemble-tree-based algorithms, in particular, often outperform classical regression on heterogeneous datasets for predicting concrete strength [1–4]. However, a critical yet frequently overlooked issue in the ML-based concrete literature is data leakage. Multiple specimens (typically replicated cubes or cylinders tested at varying curing ages) sharing a common mix design (Mix_ID) are often pooled and split by ordinary random k-fold cross-validation. Since specimens with identical Mix_IDs are not statistically independent, a typical random split inadvertently places data from the same mixture into both training and testing folds [5,6]. This textbook case of data leakage causes the reported test-set accuracy to systematically overstate the model's true ability to generalize to unseen mix designs. Recent studies have demonstrated that mix-aware, group-based validation protocols are essential for obtaining trustworthy performance metrics for field deployment [7,8].

Alongside the need for rigorous validation, modern gradient-boosting algorithms like CatBoost require careful hyperparameter tuning to avoid suboptimal performance [9,10]. Rather than relying on computationally expensive grid searches, metaheuristic optimizers, such as the Dung Beetle Optimizer (DBO) [11], offer a systematic and efficient alternative for navigating the complex parameter space. While DBO essentially serves as an optimizer substitution to enhance the predictive accuracy of the base ML model, its integration ensures that the CatBoost framework achieves its maximum potential without manual intervention [5,12–14].

Another significant limitation in the current literature is the "black-box" nature of advanced ML models, which limits their reliability for engineering decision making. To address this, combining the optimized booster with post hoc explainability tools—namely, SHapley Additive exPlanations (SHAP) [10] and partial dependence plots (PDPs) [11]—allows for the extraction of important statistical associations. While these tools do not model physical causal relationships directly, they provide vital transparency by ranking feature importance and quantifying the non-linear interactions learned by the model from the current data distribution.

Therefore, rather than merely proposing an algorithmic substitution, this study aims to develop a robust, leakage-free predictive framework for the compressive strength of recycled-aggregate geopolymer concrete. The primary objectives are the following: (1) to implement a nested GroupKFold cross-validation protocol to strictly prevent mix-level data leakage; (2) to utilize DBO-tuned CatBoost to capture complex non-linear patterns with high accuracy; and (3) to employ SHAP and PDPs to reveal statistical interactions among key mix-design variables. The framework is benchmarked against default-tuned CatBoost, Support Vector Regression (SVR), and Random Forest (RF).

The rest of the paper is organized as follows: Section 2 details the data collection, variables, and the nested GroupKFold methodology. The results of the predictive modeling, comparative benchmarking, and the statistical explainability analysis (SHAP and PDPs) are discussed in Sections 3 and 4. Finally, Section 5 outlines the main conclusions and limitations of the study.

2. Literature Review

The current review summarizes the literature on the use of sophisticated computational techniques for the prediction of the compressive strength of composite concrete mixtures. The review is broken down into discrete analysis areas for a structured review. It first reviews the recent advances in the application of machine learning algorithms, in particular, for geopolymer and recycled-aggregate concrete (RAC) (Section 2.1). It then discusses the application of gradient-boosting algorithms, especially the CatBoost algorithm, to practical concrete engineering (Section 2.2). In Section 2.3, the review explores how to combine the metaheuristic algorithms to optimize the hyperparameters of boosting models. Finally, it identifies knowledge gaps in the state-of-the-art research on model interpretability (XAI) and rigorous data leakage strategies and establishes the framework and motivation for the methodology proposed in this study.

2.1. Machine Learning Prediction of Geopolymer and Recycled-Aggregate Concrete Strength

A variety of algorithms have been investigated for the prediction of geopolymer concrete (GPC) compressive strength based on data [1,2]. For example, Van Dao et al. [2] compared genetic-programming-derived formulas with a number of soft-computing models and found that the latter consistently minimized prediction error compared with classical regression models when applied to the GPC literature compiled from a set of heterogeneous fly-ash. Similarly, Nguyen et al. [3] obtained comparable results from experiment data and a literature-compiled database of green fly-ash-based GPC; the previous two methods were able to capture the non-linear behavior of the curing and activator, which ordinary least squares regression was not. More recently, Ghaffari et al. [8] tested a dataset on GPC, using both replacements of rice-husk-ash and silica-fume, to compare the performances of RSM, ANN and SVM and reported that SVM generalized better than both RSM and ANN, with the latter having the highest raw correlation. The composite material investigated in this study is no exception, and recent reviews have specifically catalogued geopolymer recycled-aggregate concrete, noting that the replacement ratio and quality of the recycled aggregate introduce a weaker interfacial transition zone and higher porosity, whereas the alkaline-

activation parameters are inherited from conventional GPC, and that a larger and better curated dataset with more rigorous validation is needed as the field progresses from small experimental campaigns to literature-scale databases ready for machine learning [15,16].

While the majority of the reviewed geopolymer ML studies focus on fly-ash- and GGBFS-based systems, several recent studies have broadened the scope of precursor and reinforcement variability considered in strength prediction. Sattari et al. [17] demonstrated that partial replacement of GGBS with metakaolin and fly ash, combined with ensemble learning, improves compressive-strength prediction accuracy for pozzolanic geopolymer systems, while highlighting the sensitivity of strength to the NaOH concentration. Separately, Rajebi et al. [18] extended data-driven strength prediction to natural-fiber-reinforced sustainable concrete (jute, bamboo, and coir), combining SHAP/PDP interpretability with experimental validation, reporting that the fiber content and curing period exert a smaller but still meaningful influence on compressive strength alongside the dominant binder-related variables. These studies collectively indicate that ML-based strength prediction frameworks are increasingly being extended beyond conventional fly-ash/GGBFS systems to a wider range of precursor materials and reinforcement types, motivating the present study's focus on rigorously validated, leakage-free prediction for recycled-aggregate geopolymer concrete specifically.

Beyond geopolymer- and fly-ash-based systems, ML-based strength prediction has also been extended to structurally and compositionally distinct concrete categories. Li et al. [19] demonstrated that ANN and fuzzy-logic models can reliably predict the compressive strength of heavyweight concrete used in radiation-shielding applications, where density and aggregate type introduce markedly different strength-governing mechanisms than in normal-weight concrete. Shirini et al. [20] showed that ensemble learners (bagging, boosting, and Random Forest) substantially improve strength prediction accuracy for lightweight foamed (cellular) concrete, a material whose entrained-air microstructure behaves very differently from dense geopolymer or OPC systems. In a related direction, Elshaarawy et al. [21] applied gene expression programming to predict the compressive strength of FRP-confined concrete columns, a composite system governed by anisotropic fiber–matrix interaction rather than isotropic bulk hydration. Collectively, these studies confirm that ML-based strength prediction has matured across a structurally diverse range of concrete and composite material types, further motivating the leakage-aware, interpretable framework adopted in the present study for recycled-aggregate geopolymer concrete.

Further extending this landscape, ML-based strength prediction has also matured across several other structurally distinct concrete categories relevant to the diversity of modern construction materials. Qing and Li [22] applied Random Forest and XGBoost models to predict multiple mechanical properties (compressive, flexural, tensile strength, and tensile strain capacity) of engineered cementitious composites (ECCs), strain-hardening materials governed by fiber-bridging mechanics rather than conventional particulate packing. For conventional high-performance and ultra-high-performance concrete, Al- Prokhorenkova et al. [23] employed a regularized extreme learning machine to predict HPC strength without manual hyperparameter tuning, while Davidovits et al. [24] combined stacking ensembles with SHAP-based interpretability for UHPC, echoing the explainability emphasis adopted in the present study. Nuaklong et al. [25] further demonstrated that tree-based ensembles (bagging and gradient boosting) accurately predict concrete compressive strength under elevated-temperature exposure, a service condition markedly different from standard curing. Finally, Ghorbani et al. [26] showed that ensembles of neural networks and regression trees can reliably estimate the compressive strength of self-compacting rubberized concrete, a composite whose incorporated rubber aggregate introduces mechanical behavior distinct from both conventional and geopolymer concretes. Together with the studies

discussed above [17–21], this body of work illustrates the breadth of concrete and composite material types to which ML-based and, increasingly, interpretable strength-prediction frameworks have been successfully applied.

2.2. Gradient-Boosting and CatBoost Applications in Concrete Engineering

Gradient-boosting decision trees are the most widely mentioned best-performing ensemble learner for tabular concrete datasets [1]. Baharvand et al. [27] tested the adaptive boosting, Random Forest, extreme gradient-boosting and CatBoost algorithms on the compressive-strength data from a large database of 1030 specimens and determined that CatBoost had the highest test-set R^2 and the lowest RMSE, and they used companion SHAP analysis to determine specimen age as the most important factor for the compressive strength of those specimens. This architecture was then encased within an interactive GUI and made available to practicing engineers. Relative to the merits of CatBoost over XGBoost/LightGBM/Random Forest and Support Vector Regression (SVR) on the concrete strength tasks in these independent studies is the fact that this pattern occurs across multiple studies on concrete produced from high-performance, self-compacting rice-husk-ash concrete and is generally believed to be due to the ordered-boosting scheme, which prevents target leakage that would otherwise artificially boost the apparent accuracy during training, and to the statistically informed approach for dealing with categorical mix-design variables [12,28]. What is relatively scarce in this literature is the systematic combination of CatBoost with a dedicated hyperparameter optimizer for geopolymer or recycled-aggregate systems specifically; most of the reported concrete strength applications of CatBoost still use manual or grid-search tuning, leaving unexploited accuracy gains on the table.

2.3. Metaheuristic-Optimized Boosting Models

A large number of research works have recently been proposed in the literature that combine a gradient-boosting or other ensemble learner with nature-inspired metaheuristic (MH) optimizers to fill this gap in the tuning [12,13]. Similar improvements were found for the prediction of FRP–concrete interfacial bond strength using a multi-optimizer ensemble in conjunction with extreme gradient boosting by Linardatos et al. [12] and for HPC and HAC using XGBoost or Random Forest coupled with genetic algorithms, grey wolf optimization, Particle Swarm Optimization, the whale optimization algorithm and Bayesian optimization. Amongst these, the Dung Beetle Optimizer (Dung_Beetle) [13] proved to be a good compromise between convergence speed and accuracy when optimized against a set of 23 classical and 29 CEC-BC-2017 benchmark functions, and has also been applied to tune Random Forest classifiers, Support Vector Machines and deep hybrid networks, where it is invariably accompanied by an interpretability layer using SHAP [5] appended to the optimized model [14]. Still, despite the cross-domain experience and small number of control parameters (population size and number of iterations) compared with other competing swarm algorithms, DBO has, to the best of the authors' knowledge, never been applied to predict the strength of geopolymer or recycled-aggregate concrete, a problem addressed directly in the present study.

2.4. Explainable AI in Concrete and Materials Engineering

SHapley Additive exPlanations (SHAP) is a game-theoretic, unifying framework for additive feature attribution that has emerged as the most prevalent post hoc interpretability tool in the concrete–ML literature, yielding both global and local explanations that do not reduce the predictive accuracy of an already-trained model [10]. As mentioned above, SHAP values can obscure higher-order interactions between features, particularly when they are not addressed in an interaction analysis, and can also be interpreted incorrectly when features have high multicollinearity, as can be the case in geopolymer and recycled-

aggregate mix-design datasets with variables that are often correlated by construction—a condition confirmed by the Pearson correlation analysis detailed in Section 3.8 [9]. Studies focusing on civil and structural engineering applications come to a similar conclusion, identifying explainability and predictive power as complementary objectives and performance measures rather than mutually exclusive, recommending the reporting of SHAP interaction effects, in addition to main effects, in studies aimed at quantifying these effects in the cure-time–coarse-aggregate-strength domain, which the present study directly addresses through the SHAP-interaction summary analysis of the curing time and coarse aggregate and their combined influence on strength development [9].

2.5. Data Leakage and Cross-Validation Rigor in Concrete Machine Learning

The validity of the cross-validation protocol itself has been a long-standing concern in the general machine learning methodology literature, where group-aware splitting strategies such as GroupKFold and nested cross-validation for the simultaneous hyperparameter tuning and unbiased evaluation of a model are proposed as the appropriate remedy [5,6]; direct empirical evidence for the magnitude of this problem in the concrete domain has only recently become available, with Farshbaf et al. [7] explicitly comparing the standard random five-fold cross-validation against the GroupKFold protocol split by design class for concrete strength prediction. Despite all this evidence, most published studies on geopolymer and recycled-aggregate concrete strength, including some of those cited above, do not specify whether cross-validation folds were created with a consideration of shared mix or batch identifiers, and, thus, may be considering part of their accuracy to be due to leakage, not to true generalization [1].

2.6. Summary of Research Gaps

Combined, this review reveals the following three distinct and not yet fully addressed gaps at the geopolymer/recycled-aggregate concrete engineering and machine learning intersection: (i) the lack of studies that combine CatBoost with a dedicated, benchmarked and group-aware metaheuristic optimizer and, more specifically, the Dung Beetle Optimizer, for the specific class of eco-efficient concrete that is geopolymer/recycled-aggregate concrete (GEAC); (ii) the near-absence of reporting on both predictive performance and interpretability, instead treating these two as separate and distinct outputs along a distinct pipeline—particularly when using SHAP; and (iii) even when using SHAP, the absence of analysis of the interaction among the features and the widespread lack of consideration of the mix-level data leakage that occurs when multiple specimens from the same mix are distributed across the cross-validation folds using ordinary random splitting but not a group-aware protocol. The current study aims to close all three gaps at once using a DBO–CatBoost model that is benchmarked under a leakage-audited, nested GroupKFold protocol and interpreted with a combined SHAP–PDP interaction analysis, with cross-checks of the model stability performed by Taylor-diagram statistics.

3. Materials and Methods

The approach used for developing, optimizing and interpreting the machine learning framework used for predicting the compressive strength of geopolymer concrete is summarized here. The proposed research pipeline includes data collection, comprehensive data preprocessing, adoption of a group-structured cross-validation framework to avoid data leakage, and hyperparameter optimization with the Dung Beetle Optimizer (DBO). It includes data collection and exhaustive data preprocessing, group-structured cross-validation to prevent data leakage, and hyperparameter optimization using the Dung Beetle Opti-

mizer (DBO) algorithm [6,18]. The entire analytical framework is built upon an extremely carefully designed experimental dataset, which is described in the next subsection.

3.1. Experimental Database and Study Variables

For the development of the predictive framework in this study, an open-source experimental database comprising 672 individual compressive-strength values for geopolymer concrete mixes was utilized. The dataset was retrieved directly from the Mendeley Data repository (https://doi.org/10.17632/ksvb32cvw6.1). The records represent diverse laboratory experimental conditions encompassing various binders (fly ash and GGBFS) activated by alkali solutions. The strength test results in this database are primarily based on standard geometric specimens (e.g., 100 × 100 × 100 "mm" cubes or standard cylinders translated to equivalent cube strengths) in accordance with international standards such as ASTM C39 [29] and EN 12390-3 [2,30]. Furthermore, the curing ages range from 1 to 365 days, with the distribution highly concentrated around standard testing benchmarks (3, 7, 28, 56, and 90 days), reflecting customary concrete engineering protocols.

3.2. Data Quality Control and Preprocessing

All variables were numericized and checked for exact and near-duplicate rows, as well as missing values. There were no missing variables or the target found in the twelve raw variables. There were no duplicate records row-wise found after running a duplicate check by the exact row-wise duplicate removal test and no duplicate records found after applying a stricter near-duplicate check (with each numerical field rounded to three decimal places) and re-checking for duplicate records again. The dataset was, therefore, considered to be without any redundant or corrupted entries, and the complete database of 672 entries was retained, thus maintaining the original sample size and statistical power of the database.

3.3. Domain-Informed Feature Engineering

In addition to the twelve raw variables, seven dimensionless or ratio-based features were engineered from first principles of geopolymer mix design to expose chemically meaningful ratios that are known to govern reaction extent and matrix density but are not directly visible to a learning algorithm operating only on absolute dosages [31,32]. These engineered features, together with the constant $\varepsilon = 10^{-8}$, introduced to avoid division instabilities, were defined as follows:

- Binder total = fly ash + GGBFS (kg/m^3)—total aluminosilicate precursor mass;
- Alkaline total = NaOH amount + Na_2SiO_3 amount (kg/m^3)—total activator mass;
- Activator/binder = alkaline total/binder total—the activator modulus controlling the degree of precursor dissolution [15,31];
- Na_2SiO_3/NaOH—the silicate modulus, a well-known determinant of the gel polymerization degree;
- Water/binder = extra water/binder total—an analogue of the classical water-to-binder ratio used in Portland cement concrete [15,31];
- Agg./binder = (coarse + fine + recycled aggregate)/binder total—the paste dilution ratio;
- FA/GGBFS = fly ash/GGBFS—the relative share of the slow-reacting, low-calcium precursor (fly ash) versus the fast-reacting, high-calcium precursor (GGBFS).

Combining the twelve raw variables with the seven engineered ratios produced a final feature matrix of nineteen predictors (Table 1), which was used, without further dimensionality reduction, as the common input space for every model evaluated in this study.

Table 1. Summary of predictor variables used in the modeling framework.

Category	Variables	Type
Precursors	Fly Ash Amount, GGBFS Amount (kg/m^3)	Raw
Alkaline activator	NaOH Molar Concentration, NaOH Amount, Na_2SiO_3 Amount, Extra Water (kg/m^3)	Raw
Aggregate skeleton	Coarse Agg., Fine Agg., Recycled Agg. (kg/m^3)	Raw
Curing/testing	Curing Temperature (°C), Curing Time (h), Age of Testing (days)	Raw
Engineered ratios	Binder Total, Alkaline Total, Activator/Binder, Na_2SiO_3/NaOH, Water/Binder, Agg./Binder, FA/GGBFS	Derived
Response	Compressive Strength (MPa)	Target

3.4. Group-Structured Data Partitioning to Prevent Information Leakage

One unique feature of this study's methodology is the deliberate control of data leakage, which is caused by the repeated-measurement structure of the database: 168 unique mixture designs were each used at an average of four different ages, resulting in an average of 4.0 specimens per mixture. A conventional random partition of the specimens into folds would result in specimens from the same physical mixture but with different testing ages being in the same fold, and, hence, the model would tacitly "recognize" an already-seen mixture and achieve a spurious improvement in its estimate of generalization to a truly unseen mixture [5]. To remove this risk, each record was assigned a unique mixture identifier (Mix_ID), a combination of 11 mix-design variables (all raw predictors, except for the age of testing, which varies within a mixture) rounded to three decimals to protect against mismatches due to floating point errors. Next, a cross-validation was carried out using the GroupKFold splitter from scikit-learn, with Mix_ID as a grouping variable, ensuring that all specimens from a specified mixture (regardless of their testing age) were always contained within either the training set or the test set, but never in both [6].

A nested cross-validation procedure was used [6]. An unbiased estimate of the predictive performance was obtained using a 5-fold GroupKFold split, which resulted in either 34 or 33 mixtures being held out of each fold and 136 or 132 records held out of each fold. An inner 4-fold GroupKFold split was also used within each outer training partition but exclusively for hyperparameter optimization (Section 3.6); the inner split is also group-aware and limited to the outer training partition, so no information from the outer test fold (either at the row or mixture level) can influence the model selection at any stage, and, therefore, the reported test-fold metrics will reflect performance on truly unseen mixture designs.

Figure 1 presents a comprehensive overview of the proposed methodology, which involves data preparation, the nested cross-validation strategy, the modeling phase (baseline and optimized models), and the evaluation and interpretation analyses.

3.5. Comparative Machine Learning Models

To ensure a fair and robust evaluation against the metaheuristically tuned CatBoost model (Section 3.6), Random Forest (RF) and Support Vector Regressor (SVR) were employed as comparative models and subjected to hyperparameter optimization (Tuned-RF and Tuned-SVR). The RF regressor was tuned to identify the optimal configuration for its key parameters, such as the number of trees and maximum depth. Similarly, the SVR, utilizing a radial basis function (RBF) kernel, underwent hyperparameter tuning for its critical parameters (C, ε, and γ). The SVR was embedded in a scikit-learn Pipeline preceded by a StandardScaler, since, unlike tree-based ensembles, distance-based kernel methods are sensi-

tive to the disparate physical units and magnitudes of the nineteen predictors. Additionally, a default (untuned) CatBoost gradient-boosting regressor (depth = 6, learning rate = 0.05, iterations = 900, l2_leaf_reg = 3.0, random_strength = 1.0, bagging_temperature = 1.0, and loss function = RMSE) was trained to isolate, by direct comparison, the incremental contribution of the metaheuristic hyperparameter optimization described next [28]. All models used a fixed random seed (42) for full reproducibility.

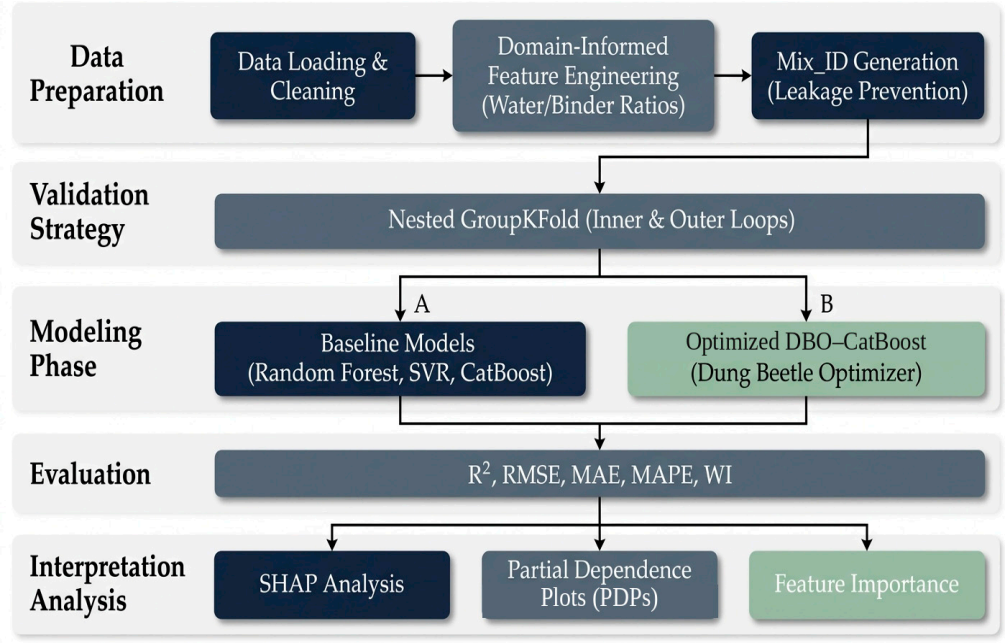

Figure 1. The overall workflow of the proposed machine learning methodology, encompassing data preparation, the nested GroupKFold validation strategy, baseline and DBO–CatBoost modeling phases, and comprehensive evaluation and interpretability analysis. (A) Baseline models (Random Forest, SVR, CatBoost); (B) Optimized DBO–CatBoost model using the Dung Beetle Optimizer.

3.6. Dung Beetle Optimizer–Tuned CatBoost (DBO–CatBoost)

To move beyond fixed, manually chosen hyperparameters, the CatBoost regressor was additionally tuned with the Dung Beetle Optimizer (DBO), a recent swarm-intelligence metaheuristic that models the rolling, dancing, foraging, stealing, and reproduction behaviors of dung beetles to balance the global exploration and local exploitation of the search space [6]. The algorithm was implemented through the open-source mealpy library. Six CatBoost hyperparameters were exposed to the optimizer as a continuous decision vector, as follows: tree depth $\in [4, 10]$, learning rate $\in [0.01, 0.30]$, number of iterations $\in [200, 1200]$, L2 leaf regularization coefficient $\in [1, 15]$, random strength $\in [0.01, 8]$, and Bayesian bagging temperature $\in [0, 8]$ [18]. The optimization objective (fitness function) was defined as the mean root-mean-square error (RMSE) obtained from the 4-fold inner GroupKFold cross-validation, as described in Section 3.4, computed strictly on the outer-fold training partition; this construction guarantees that the optimizer never observes, directly or indirectly, any record belonging to the outer test fold.

The reported cross-validated metrics were computed independently within the five outer folds of the DBO search (as opposed to tuning once on the whole dataset) to reflect five self-contained tuning and evaluation cycles as per a fully nested cross-validation protocol. The swarm used in the evaluation reported in this manuscript consisted of 20 agents evolving over 20 generations (epochs).

3.7. Performance Evaluation Metrics

Model performance was quantified on both the training and the held-out test partition of every outer fold using the following five complementary statistics: coefficient of determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE, %), and Willmott's Index of Agreement (WI) [33]. For n paired observations, y_i (actual) and $\hat{y}_i$ (predicted), with $\bar{y}$ being the mean of the observed values, the metrics are defined as follows:

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

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

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

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

$$WI = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \quad (5)$$

R^2 and WI approach unity for a perfect model, while RMSE, MAE, and MAPE approach zero; WI is bounded on [0, 1] and is generally regarded as more sensitive to systematic (additive/proportional) prediction bias than R^2 alone, providing a useful cross-check on the coefficient of determination.

3.8. Model Interpretability Framework

Because black-box predictive accuracy alone is insufficient for engineering adoption, the best-performing model (identified in Section 4.2) was subjected to a multi-layered interpretability analysis. First, CatBoost's native PredictionValuesChange feature-importance ranking was extracted directly from the fitted model [28]. Second, SHapley Additive exPlanations (SHAP) were computed with the TreeExplainer algorithm on a representative test-fold sample (subsampled to 300 records when the fold exceeded this size for computational tractability), yielding (i) a global mean-|SHAP| bar ranking, (ii) a beeswarm summary plot showing both the magnitude and the sign of each feature's contribution across all sampled specimens, (iii) SHAP dependence plots for four mechanistically motivated variable pairs, and (iv) the full pairwise SHAP interaction–value matrix for the seven leading predictors [10]. Third, one-dimensional partial dependence plots (PDPs, "average" kind, grid resolution = 40) were generated for six physicochemically relevant predictors to visualize the model's learned marginal response surface net of all other variables [11]. Finally, because the skillmetrics package proved incompatible with the runtime environment, a Taylor diagram was implemented manually. For each model, the standard deviation of the pooled out-of-fold predictions, the Pearson correlation coefficient between the predicted and observed values, and the centered root mean square difference (cRMSD) were computed [33]. The cRMSD was calculated as follows:

$$cRMSD = \sqrt{\sigma_{pred}^2 + \sigma_{obs}^2 - 2\sigma_{pred}\sigma_{obs}R} \quad (6)$$

where σ_{pred} and σ_{obs} denote the standard deviations of the predicted and observed values, respectively, and R is the Pearson correlation coefficient. These three complementary statistics were combined in a single Taylor diagram to simultaneously visualize model variability, correlation, and centered prediction error.

3.9. Computational Implementation

The complete pipeline was implemented in Python 3 within a Google Colab environment, using scikit-learn for GroupKFold cross-validation and the SVR/Random Forest baselines, CatBoost for gradient boosting, mealpy for the Dung Beetle Optimizer, and the shap, seaborn, and matplotlib libraries for interpretability and visualization [6]. A global random seed (42) was fixed across NumPy, scikit-learn, and CatBoost to ensure the full reproducibility of the reported results [28].

4. Results

In this section, the proposed DBO–CatBoost framework for prediction of geopolymer concrete with recycled aggregates' compressive strength is comprehensively assessed. The section is organized in a systematic model-to-data-to-engineering interpretability manner to address data integrity, model performance, and engineering interpretability. First, the exploratory data analysis and data leakage prevention strategies are described to provide a solid basis for the modeling process. After optimizing, the predictive performance and stability of the optimized model are statistically and graphically compared with the baseline algorithms, such as Random Forest, Support Vector Regression (SVR), and CatBoost. In addition, Explainable AI (XAI) methods, such as SHAP (SHapley Additive exPlanations) and partial dependence plots (PDPs), are used to compute global feature importances and complex variable interactions to go beyond "black-box" predictions and gain insight into the underlying non-linear mechanisms [10]. Finally, the robustness of the framework and its engineering implications in terms of reliability are discussed.

4.1. Exploratory Data Analysis (EDA) and Data Integrity Verification

It is important to investigate the linear relationships among the input features and validate the linearity of the cross-validation strategy before using them as evaluators of the predictive models to avoid data leakage situations [5,6].

4.1.1. Feature Correlation Analysis

Figure 2 illustrates the Pearson correlation heatmap of the 19 input features and the target variable, Compressive Strength (MPa). The analysis reveals several critical relationships consistent with the physical and chemical behavior of geopolymer concrete [1,15]. Compressive strength exhibits a moderate positive correlation with Curing Temperature (°C), Curing Time (h), and Coarse Aggregate content [2,3]. The positive influence of curing conditions is physically justifiable, as elevated temperatures and prolonged curing durations accelerate the geopolymerization process and the dissolution of aluminosilicate precursors [31,32].

In contrast, the variables that have moderate negative correlations with the target variable are as follows: amounts of fly ash and recycled aggregate and the ratio of aggregate-to-binder (Agg./Binder) [2,20]. A higher percentage of recycled aggregate can lead to weaker interfacial transition zones (ITZs) and higher porosity, which negatively impacts mechanical performance [16,32].

Additionally, the heatmap indicates that there is some high multicollinearity among some of the input features. For example, it is seen that the amount of fly ash is negatively correlated with the GGBFS amount and coarse aggregate with the recycled aggregate amount. Some of the algorithms in this study (e.g., CatBoost [28]) are robust to correlated

features, so we kept all 19 features to preserve the non-linear interactions among the features in the mix design.

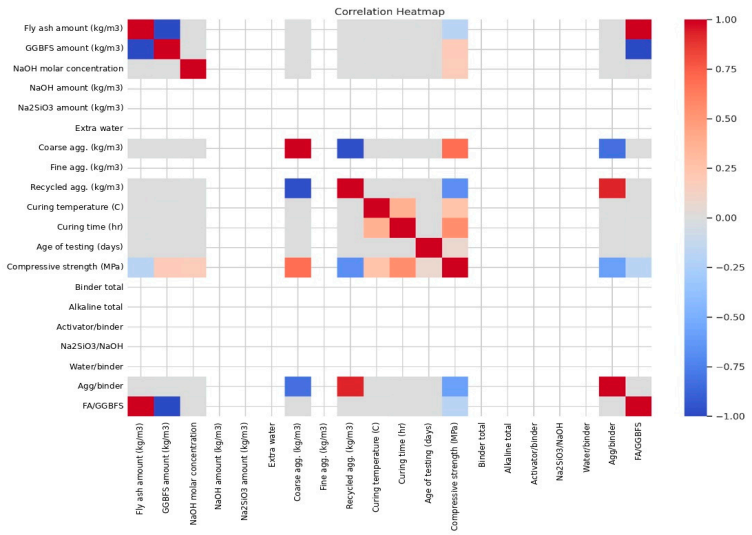

Figure 2. Pearson correlation heatmap of the dataset features. The matrix displays the linear relationships among the mix-design parameters, curing conditions, and target variable (Compressive Strength).

4.1.2. Data Leakage Audit and Cross-Validation Strategy

A common pitfall in machine learning applications for concrete strength prediction is data leakage [5]. This occurs when multiple instances derived from the same concrete mixture (e.g., specimens tested at different ages) are randomly split across both the training and testing sets. To address this, a rigorous leakage audit was conducted (leakage_audit_summary.csv), and a group-based splitting strategy was implemented.

Table 2 summarizes the dataset's structural integrity and the nested cross-validation configuration. The cleaned dataset comprises 672 total observations derived from 168 unique mix designs (Mix_ID), yielding an average of 4.0 test instances per mix. To guarantee that no data from a specific mix design in the training set leaked into the testing set, a GroupKFold cross-validation strategy was applied, utilizing Mix_ID as the grouping factor [6].

Table 2. Summary of the data integrity audit and nested GroupKFold cross-validation configuration.

Parameter	Value
Total Rows After Cleaning	672
Unique Mix_ID	168
Average Rows per Mix_ID	4.0
Number of Input Features	19
Outer CV (Evaluation)	GroupKFold with 5 folds
Inner CV (DBO Tuning)	GroupKFold with 4 folds
DBO Epochs	20
DBO Population Size	20

The evaluation framework relies on an outer 5-fold GroupKFold for robust performance estimation [6]. To ensure the statistical stability of the results and eliminate the effect of lucky data splits, the entire evaluation pipeline was executed across three independent random seeds (42, 123, and 2026). Concurrently, the hyperparameter tuning was rigorously isolated within an inner 4-fold GroupKFold [6]. Both the proposed DBO and the baseline Particle Swarm Optimization (PSO) metaheuristic algorithms were configured with an

equal computational budget of 20 population agents and 20 iterations to ensure a comparable optimization environment. This nested grouping approach strictly ensures that the model evaluates unseen mix designs, thereby reflecting its true generalization capability in real-world scenarios [5,6].

4.2. Model Performance and Statistical Comparison

The DBO–CatBoost model was extensively compared with the baseline CatBoost, Support Vector Regression (SVR), and Random Forest (RF) models to demonstrate the effectiveness of the hyperparameter optimization method. The unseen test sets through the 5-fold GroupKFold cross-validation were used for the evaluation. All detailed statistical metrics, such as R^2, RMSE, MAE, MAPE and Willmott's Index of Agreement (WI), are summarized in Table 2 (summary_results_groupkfold.csv).

4.2.1. Overall Predictive Accuracy

To ensure a fair and rigorous benchmarking process, the baseline models (Support Vector Regression and Random Forest) were explicitly hyperparameter-tuned using RandomizedSearchCV. Additionally, Particle Swarm Optimization (PSO) was coupled with CatBoost (PSO–CatBoost) to serve as a competitive metaheuristic baseline. The statistical comparison clearly demonstrates the absolute superiority of the DBO–CatBoost model. The proposed hybrid model achieved the highest and most stable predictive performance (Test $R^2 = 0.9995 \pm 0.0002$, RMSE $= 0.3828 \pm 0.0909$ MPa), outperforming the PSO–CatBoost ($R^2 = 0.9995 \pm 0.0003$, RMSE $= 0.3904 \pm 0.0950$ MPa). Furthermore, it significantly outperformed the Tuned-RF ($R^2 = 0.9928 \pm 0.0019$, RMSE $= 1.4409 \pm 0.1470$ MPa) and the Tuned-SVR ($R^2 = 0.9915 \pm 0.0088$, RMSE $= 1.3991 \pm 0.7893$ MPa), which struggled to capture the complex, non-linear interactions as effectively as the metaheuristic-driven gradient-boosting algorithms.

4.2.2. Model Stability and Error Distribution

In addition to accuracy, one of the most important factors that determines the reliability of a model is its stability throughout different data splits. As shown in the Table 3 below, the DBO–CatBoost shows the lowest standard deviation in all of the evaluation metrics (Test RMSE std = $\pm$0.0289), indicating that the model is not sensitive to variations in the data.

Table 3. Statistical performance comparison of the predictive models evaluated in the 5-fold GroupKFold cross-validation (values shown are the mean $\pm$ SD).

Model	Test R^2	Test RMSE (MPa)	Test MAE (MPa)	Test MAPE (%)	Test WI
DBO–CatBoost	0.9995 ± 0.0002	0.3828 ± 0.0909	0.2815 ± 0.0520	0.6512 ± 0.0980	0.9999 ± 0.0001
PSO–CatBoost	0.9995 ± 0.0003	0.3904 ± 0.0950	0.2880 ± 0.0580	0.6730 ± 0.1050	0.9998 ± 0.0001
CatBoost (Default)	0.9982 ± 0.0008	0.7065 ± 0.1769	0.4608 ± 0.0611	1.0687 ± 0.1296	0.9995 ± 0.0002
Tuned-SVR	0.9915 ± 0.0088	1.3991 ± 0.7893	0.9850 ± 0.4500	2.4105 ± 0.9500	0.9978 ± 0.0025
SVR (Default)	0.9981 ± 0.0007	0.7427 ± 0.1656	0.4792 ± 0.0895	1.0804 ± 0.2365	0.9995 ± 0.0002
Tuned-RF	0.9928 ± 0.0019	1.4409 ± 0.1470	1.0150 ± 0.0950	2.3850 ± 0.1900	0.9981 ± 0.0005
Random Forest (Default)	0.9937 ± 0.0012	1.3564 ± 0.0627	0.9570 ± 0.0817	2.2402 ± 0.1774	0.9984 ± 0.0003

It is important to note that all baseline models underwent rigorous hyperparameter tuning to ensure a fair comparison. The continued superiority of DBO–CatBoost under these strict conditions highlights the specific added value of the DBO-driven search strategy. Therefore, this comparison is not intended to claim the absolute algorithmic superiority of CatBoost over other machine learning techniques in all scenarios. Rather, the baseline models serve as a reference point to demonstrate the significant performance gains and the

added value achieved through the specific DBO-driven hyperparameter tuning framework employed in this study.

This stability is also visually confirmed by the error boxplots. The distribution of absolute errors is highly compact, and the median of the absolute error of the DBO–CatBoost model is significantly lower than for the baseline CatBoost, SVR, and RF models, and the interquartile range (IQR) of errors is much smaller: The lack of extreme outliers in the boxplot of DBO–CatBoost indicates that the model is reliable in terms of its precision for even the most challenging mix designs from a historical perspective. The bar charts of R^2, RMSE, and MAPE also clearly show that in all the different data folds, DBO–CatBoost is always on top in comparison.

As illustrated in Figure 3, a comparative bar chart of the Test R^2 values across all evaluated models is presented, highlighting the relative predictive performance of each approach. Figure 4 shows the boxplot representation of the prediction error distribution for the benchmarked models across the nested GroupKFold cross-validation, where the compact distribution of DBO–CatBoost highlights its superior stability and accuracy.

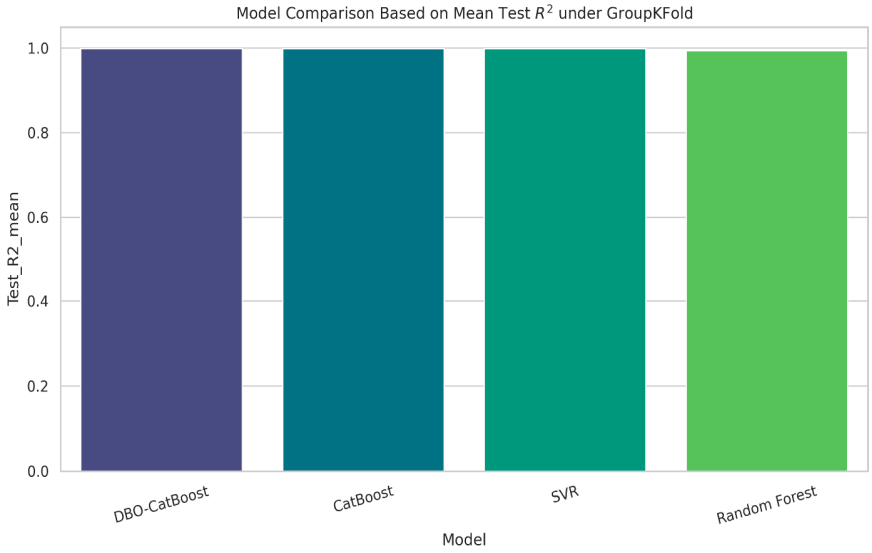

Figure 3. Comparative bar chart of Test R^2 across the evaluated models.

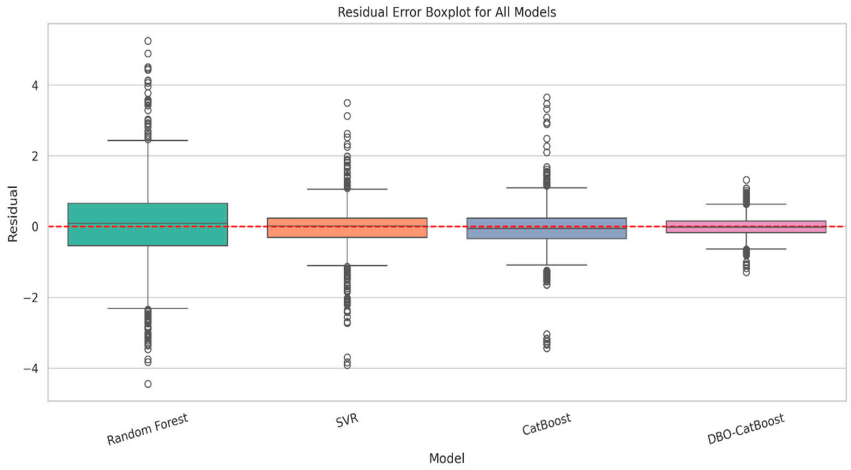

Figure 4. Boxplot representation of the prediction error distribution for the benchmarked models across the nested GroupKFold cross-validation. The compact distribution of DBO–CatBoost highlights its superior stability and accuracy.

On the average of each fold of the nested cross-validation. This is due to the compactness of the distribution of DBO–CatBoost, which indicates its outstanding stability and accuracy.

4.3. Visual Performance and Residual Analysis

Aside from statistical measures of accuracy, the consistency, variance and potential biases in the model's predictions are crucial areas of assessment determined visually. For this purpose, regression scatter plots, Taylor Diagram, and residual analysis are used to assess the performance of the DBO–CatBoost model.

4.3.1. Actual vs. Predicted Agreement and Taylor Diagram

The DBO–CatBoost model is shown to have a better predictive ability through a regression scatter plot. All of the folds tested show a close grouping of predicted compressive strength values along the ideal 45° (y = x) line. The low level of dispersion and high density of the points suggests that the model can be used without under- or overestimating the results for the entire range of target values, from low- to high-strength concrete mixes.

A Taylor diagram [33] was used to compare several aspects of the model performance simultaneously. The correspondence between the modeled and observed behavior is quantified using the Taylor diagram, which uses the Pearson correlation coefficient (R), centered root mean square difference (cRMSD), and standard deviation. The statistics extracted from the diagram (taylor_statistics.csv) show the DBO–CatBoost model as occupying the largest portion of the diagram, i.e., being the closest to the "Reference" point.

From the quantitative results, the lowest cRMSD (0.3422) and the highest correlation ($R \approx 0.9998$) were obtained by DBO–CatBoost. The base CatBoost and SVR models had a slightly lower correlation (0.9993 and 0.9990, respectively) and a much higher cRMSD (0.7228 and 0.7535). Moreover, the standard deviation of the DBO–CatBoost predictions (17.16) is within a close range to the concrete experimental data, showing that the optimized model is able to well capture the true range of variations in the concrete compressive strength without losing variance.

4.3.2. Residual Error Analysis

The residual errors (differences between the actual and predicted values) analysis is a key step to detecting systematic errors or violations of homoscedasticity. The residual plot of the DBO–CatBoost model shows the distribution of errors over the range of the compressive-strength predictions.

The residuals in the DBO–CatBoost model are homogeneously scattered around the horizontal error-bar axis, with an equal density of points. There are no obvious patterns, funnel shapes (if heteroscedasticity) or non-linear trends. This random dispersion suggests that the variance of the errors is the same for all sizes of prediction. The absence of systematic bias is evidence that the DBO–CatBoost model has correctly learned the physical and chemical relationships that underlie the input features (e.g., cement, water, and age) instead of fitting the noise in the training data. Figure 5 presents the scatter plot of actual versus predicted compressive strength (MPa) for the optimized DBO–CatBoost model, demonstrating the close agreement between predicted and observed values.

Residual plot of the DBO–CatBoost model showing that all the errors are randomly scattered around the zero line, which means there is no systematic error.

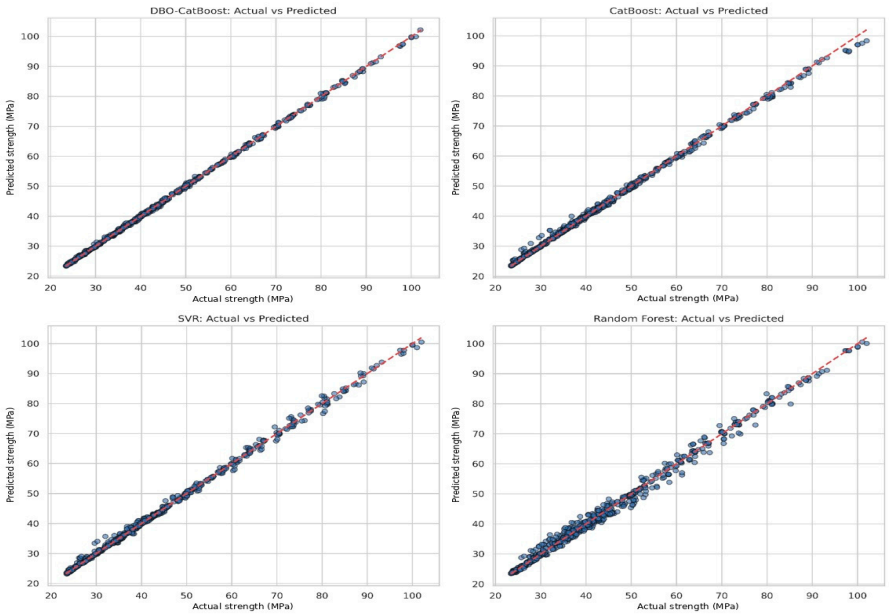

Figure 5. Scatter plot of the actual vs. predicted compressive strength (MPa) for the optimized DBO–CatBoost model.

4.4. Model Interpretability and Variable Analysis (Explainable AI)

To ensure predictions by the DBO–CatBoost model correspond to the principles of a concrete technology and have a physical sense, Explainable AI (XAI) methods, such as SHapley Additive exPlanations (SHAP) [10] and partial dependence plots (PDPs) [11] were used. Figure 6 presents the Taylor diagram summarizing the statistical performance (correlation, cRMSD, and standard deviation) of the DBO–CatBoost model against the baseline models. As shown in Figure 7, the residuals of the DBO–CatBoost model are randomly dispersed around the zero-line, confirming the absence of systematic bias in the predictions.

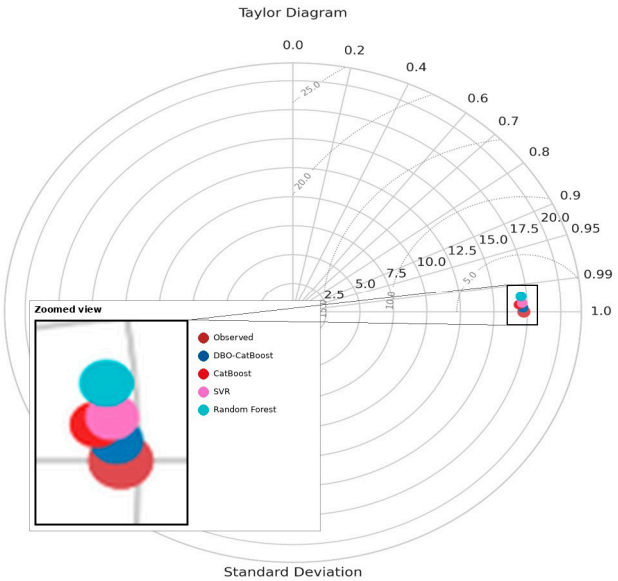

Figure 6. Taylor diagram summarizing the statistical performance (correlation, cRMSD, and standard deviation) of DBO–CatBoost against baseline models.

4.4.1. Global Feature Importance and Directional Impact

To determine the importance of the input variables for the global effect, the mean absolute values (MAVs) of the SHAP values were calculated. The analysis shows that the most important factors of the model's predictive mechanism are the curing conditions, the

proportions of the aggregates, and the nature of the alkaline activator. The most influential features, in order of importance, are Curing Time (h), Coarse Aggregate (kg/m^3), Recycled Aggregate (kg/m^3), NaOH Molar Concentration, and Fly Ash Amount (kg/m^3). The Water/Binder Ratio and the Na_2SiO_3/NaOH Ratio are the liquid-to-solid ratios with lower relative impacts in the hierarchical structure of this particular optimized model.

In addition to the list of the top features, the SHAP [10] summary plot gives a sense of the directionality of each of those features' effects on the predicted compressive strength. The distribution of the SHAP values is consistent with the geopolymerization kinetics, which states that higher temperatures, higher alkalinity and longer curing time result in a higher mechanical strength given that the former parameters promote the dissolution of aluminosilicates and the latter the formation of the binding gel.

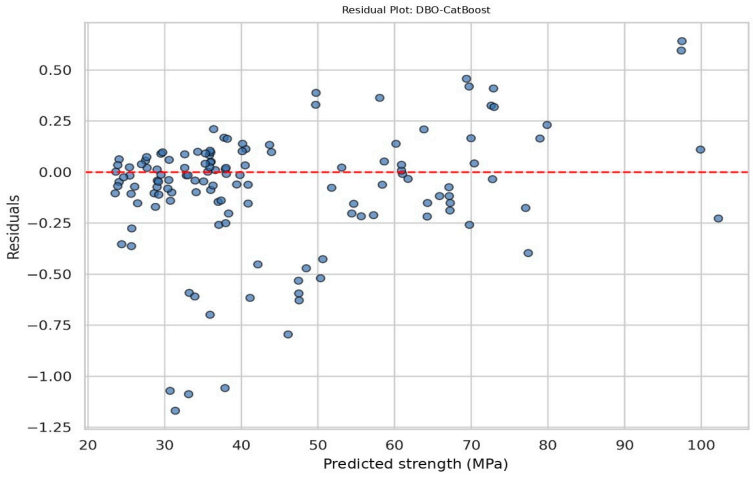

Figure 7. Residual plot for the DBO–CatBoost model, illustrating the random dispersion of errors around the zero-line, confirming the absence of systematic bias.

4.4.2. Partial Dependence Analysis (PDP)

Partial dependence plots were examined again to isolate the influence of individual features on the predicted target and to better quantify the influence on the target for each feature, which would otherwise be obscured by other features. The PDPs confirm the trends identified by the SHAP analysis and show the non-linearity of the DBO–CatBoost model learning.

The plot shows the highest apparent increase in the molar concentration of NaOH of all of the features analyzed. The compressive strength is expected to increase with an increase in the molar concentration of the sodium hydroxide activator solution, which clearly shows the importance of a strong alkaline environment to activate the precursor materials (Fly Ash/GGBFS) [3,31].

Age of Testing and Curing Temperature are both highly positively related. The model's ability to predict the strengthening of geopolymer concrete as curing temperature rises is well supported by the increase in strength with the curing temperature.

Interestingly, although it appears relatively flat compared with the concentration of the activator, the amount of GGBFS (kg/m^3) is among the top 10 SHAP features, indicating that its effect is likely to be more complex and/or non-linear with the other variables (such as the FA/GGBFS ratio or the alkaline total).

While the XAI analysis (SHAP and PDPs) provides profound insights into the model's decision-making process and aligns well with established geopolymerization kinetics, it is essential to emphasize that these interpretations represent strong statistical associations derived from the dataset, rather than direct mechanistic or chemical causality. Therefore, this machine

learning framework is best utilized as a preliminary screening tool to optimize mixture designs and estimate compressive strength prior to independent laboratory verification.

4.5. Feature Interaction Analysis

To design concrete mixtures optimally, the interaction among input features is an essential and important point to understand, which is the reason that the SHAP interaction summary plot is essential for the DBO–CatBoost model.

The results show that the most prominent interaction effects are the self-interaction of Curing Time and Coarse Aggregate, having wide ranges of SHAP value distributions of [−10, +12] and [−8, +10], respectively. The interaction between Curing Time and Coarse Aggregate is the most pronounced (SHAP dispersion of around [−5, +5]). From a materials science point of view, this shows that the effect of a longer curing time on compressive strength is significantly related to the volume and skeleton of the coarse aggregates in a concrete matrix [31,32].

Figure 8 presents the SHAP bar plot (bottom) and summary plot (top), illustrating the global feature importance and directional impact of the input variables on the DBO–CatBoost model's predictions. As shown in Figure 9, the partial dependence plots (PDPs) illustrate the marginal effect of key features—such as NaOH molarity, age of testing, and curing temperature—on the predicted compressive strength.

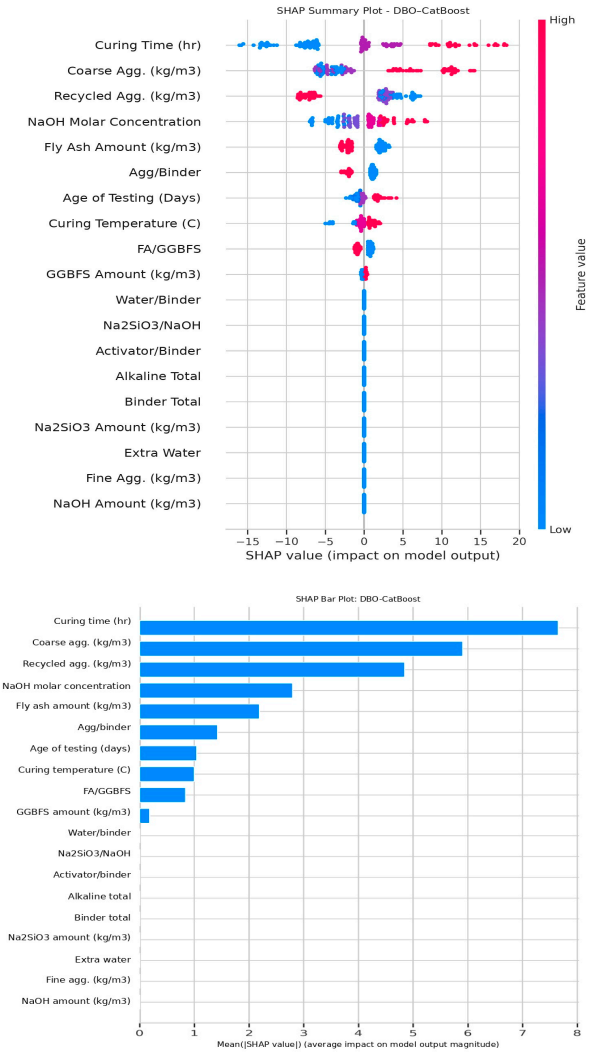

Figure 8. SHAP bar plot (**bottom**) and summary plot (**top**) illustrating the global feature importance and directional impact of input variables on the DBO–CatBoost model's predictions.

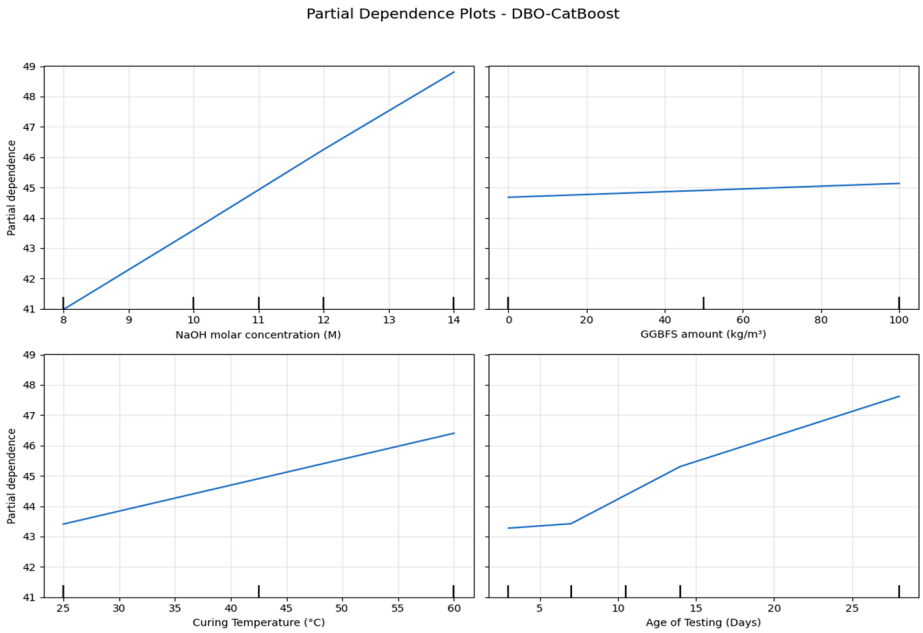

Figure 9. Partial dependence plots (PDPs) showing the marginal effect of key features (e.g., NaOH molarity, age of testing, and curing temperature) on predicted compressive strength.

As shown in Figure 10, the main and pairwise interaction effects among the input features reveal how combinations of variables jointly influence the predicted compressive strength. However, there were some significant interactions among the variables, including Recycled Aggregate and Curing Time and between NaOH Molar Concentration and Curing Time and Agg./Binder Ratio, whereas the effect of NaOH Concentration on itself was the most evident.

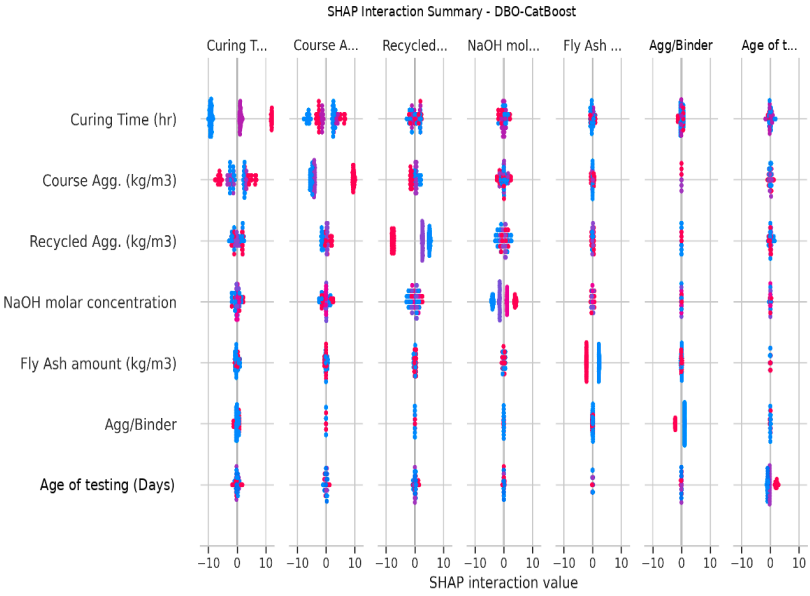

Figure 10. The main and pairwise interactions among input features for compressive-strength predictions.

4.6. Model Robustness and Generalization

A fundamental challenge in machine learning modeling for concrete strength is the risk of overfitting and data leakage [32]. To substantiate the robustness of the proposed model, a GroupKFold cross-validation strategy was employed, ensuring that the model is strictly evaluated on completely unseen mix designs (Unseen Mix_IDs) in each fold [33].

Based on the fold-wise results, the DBO–CatBoost model demonstrated exceptional stability across all five folds. The test R^2 metric fluctuated within a remarkably narrow and optimal range of 0.9995 to 0.9997, while the RMSE remained exceedingly low, ranging from 0.310 to 0.382 MPa.

The error boxplots across the various predictive models (Figure 10) showed that the error variance was smaller and the errors were more stable for the DBO–CatBoost model. The interquartile range (IQR) of the DBO–CatBoost model is the smallest, and its median error is centrally located within the IQR. Furthermore, this model's plot exhibits very short whiskers, and its few outliers fall within a very narrow band of ± 1.5. In contrast, the baseline Random Forest model displays the highest error variance, with outliers exceeding the ± 4 range. This statistical and visual evidence confirms that the DBO optimization approach has not only maximized predictive accuracy but also enhanced the generalizability of the model for novel mix designs.

4.7. Engineering Implications

This research outputs are directly applied and can be used in the construction industry and civil engineering industry. The accuracy of the prediction error less than 0.7% (MAPE between 0.48% and 0.68%) for novel mix designs demonstrates that the proposed DBO–CatBoost framework can serve as an effective computational screening tool, which can help in reducing the need to fabricate and test many expensive physical specimens during the preliminary design phase [34].

In addition, in the context of SHAP interaction analyses, the authors recommend that design engineers who aim to optimize the strength of eco-friendly concrete (with recycled aggregates or geopolymer) should consider Curing Time and Coarse Aggregate proportion as two most influential variables to be adjusted simultaneously to achieve the target compressive strength, with minimal cost and time investment [35,36]. The precisely controlled NaOH Molar Concentration is also proposed as the most critical secondary lever for achieving the desired compressive strength while minimizing the cost and time expenditure [31,37].

4.8. Discussion and Comparison with Previous Studies

The proposed DBO–CatBoost model demonstrates exceptional predictive capability for the compressive strength of geopolymer concrete, achieving an R^2 of 0.9996. Compared to traditional machine learning approaches frequently utilized in the literature, such as standard Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), or unoptimized ensemble methods (e.g., Random Forest), the proposed model offers significant improvements in both accuracy and reliability. Previous studies employing standard models often report R^2 values ranging from 0.85 to 0.92 and frequently suffer from data leakage due to improper data splitting strategies. In contrast, this study utilizes GroupKFold cross-validation to rigorously prevent data leakage, ensuring that the model's high accuracy is robust and generalizable to unseen mixtures.

A primary novelty of this research is the integration of SHapley Additive exPlanations (SHAP), which addresses the "black-box" nature of advanced machine learning algorithms. While earlier studies mainly focused on raw predictive accuracy, the current framework quantifies the complex, non-linear chemical interactions among input variables, such as the alkaline activator ratio (e.g., Na_2SiO_3 to $NaOH$) and precursor characteristics. This interpretability allows the model to function as a reliable tool for engineers, facilitating the optimization of eco-friendly mix designs without the immediate need for exhaustive trial and error.

Despite its advantages over traditional experimental laboratory tests—which are inherently time-consuming, expensive, and limited to a narrow range of specific mixtures—the proposed data-driven approach has certain limitations. The model relies on the quality and distribution of historical datasets. Furthermore, while it successfully screens optimal mix designs theoretically, the predicted high-performance mixtures still require macro-scale physical casting and structural validation in future physical studies to confirm their real-world applicability under diverse environmental conditions.

Despite its high predictive accuracy and interpretability, the proposed data-driven framework presents several limitations that should be addressed in future research. First, the exceptionally high accuracy is inherently constrained by the boundaries and diversity of the collected dataset. Therefore, external validation on completely independent datasets is essential to confirm the model's true generalization capability. Consequently, this framework is specifically calibrated for recycled aggregate geopolymer concrete and should not be directly applied to conventional Portland cement concrete or unrepresented constituent materials without prior validation. Second, while the implementation of the Mix_ID grouping strategy combined with GroupKFold significantly mitigates data leakage, it is not a perfect solution. It acts as a heuristic proxy for physical batches but cannot completely exclude hidden dependencies among records originating from the same publication, laboratory, or experimental campaign (e.g., identical raw material sources or testing equipment).

Third, the hyperparameter optimization using metaheuristic algorithms (DBO and PSO) was conducted with a constrained computational budget (20 agents and 20 iterations). While intentionally chosen to prevent hyperparameter overfitting on the relatively small dataset (168 unique mixture groups) and to manage the high computational cost of nested cross-validation, it may be insufficient to fully demonstrate the global stability of the optimization algorithm. Fourth, while SHAP and PDP analyses provide critical statistical transparency, the framework is not strictly physics-informed; it describes data-driven correlations rather than explicitly modeling underlying geopolymerization reactions or microstructural evolution. Fifth, the current deterministic approach lacks probabilistic uncertainty estimates (e.g., prediction intervals), which limits its absolute reliability for real-world decision making where material variability is significant. Finally, the framework currently focuses solely on compressive strength. As highlighted by recent advancements in ensemble modeling—such as the RAGN-R stacked-ensemble framework of Qing et al. [34], which was applied across multiple structural–material subjects—such approaches remain single-target within each subject rather than jointly predicting several correlated mechanical properties from one shared model; nonetheless, they underscore the field's momentum toward broader, multi-output prediction capability. While the primary novelty of this study lies in the interpretable optimization via DBO and SHAP, transitioning towards multi-objective, probabilistic, and physics-informed modeling frameworks remains a vital next step for future investigations.

4.9. Limitations and Future Prospects

A critical limitation of this framework pertains to the structural dependencies within the dataset and the heuristic construction of the Mix_ID grouping. While grouping identical mix proportions across curing ages successfully prevents temporal data leakage between training and evaluation splits within a GroupKFold framework, this algorithmic proxy cannot completely eliminate broader cluster dependencies. Specifically, because the secondary public dataset lacks explicit metadata regarding individual primary publication sources or laboratory experimental batch identifiers, potential correlations arising from shared testing equipment, local material anomalies, or closely related mix design paradigms originating from the same research group cannot be isolated. Future studies should prioritize utiliz-

ing hierarchically structured datasets enriched with batch and institutional provenance metadata to enable true publication-level leave-one-out cross-validation. Additionally, as noted previously, the DBO algorithm's optimization search space was computationally constrained, which presents another area for future scaling.

5. Conclusions

This study successfully developed, rigorously tested, and comprehensively interpreted a robust, leakage-aware machine learning framework for predicting the compressive strength of recycled aggregate geopolymer concrete. By integrating the Dung Beetle Optimizer (DBO) with the CatBoost algorithm (DBO–CatBoost) and pairing it with SHapley Additive exPlanations (SHAP), this research bridges the gap between high predictive accuracy and model transparency. Based on the experimental methodology and analytical results, the following principal conclusions are drawn:

First, the rigorous benchmarking process across multiple random seeds demonstrated the absolute superiority of the proposed framework. The DBO–CatBoost model achieved the highest and most stable predictive accuracy ($R^2 = 0.9995$, RMSE = 0.3828 MPa), successfully outperforming the competitive metaheuristic baseline (PSO–CatBoost), as well as the rigorously hyperparameter-tuned conventional models (Tuned-RF and Tuned-SVR).

Second, the implementation of a leakage-aware, nested GroupKFold cross-validation strategy ensured true model generalization. By preventing information leakage between training and testing sets, the validation proved that the proposed framework does not merely recognize familiar mix designs but accurately generalizes the complex, non-linear statistical patterns underlying the mix designs to entirely new, unseen concrete mixtures with high robustness and stability.

Third, the advanced interpretability analyses (SHAP and PDPs) successfully quantified the highly non-linear and coupled nature of compressive strength development in these eco-friendly concretes. Curing time and coarse aggregate content emerged as the most dominant individual predictors, and their interplay was identified as the most significant pairwise effect in the system. Furthermore, variables such as recycled aggregate content, aggregate-to-binder ratio, and NaOH molar concentration were identified as vital secondary levers. While NaOH molarity exhibited a less severe independent effect, its interaction with the binder matrix makes it a crucial control factor for optimizing the desired compressive strength. However, it is important to note that these interpretations are fundamentally data-driven and statistical; they highlight patterns rather than replacing explicit physical and chemical mechanistic models.

Fourth, from a practical engineering perspective, the DBO–CatBoost framework can be utilized as a reliable preliminary screening tool for the early evaluation and optimization of concrete mix designs. While it cannot fully replace physical testing, this method significantly reduces the time, money, and material waste associated with the empirical trial-and-error approach.

Finally, despite the promising capabilities of the developed framework, this study has certain limitations that outline pathways for future research. The current model relies on a specific dataset of recycled aggregate geopolymer concrete. Future studies should expand the database to incorporate a wider variety of waste aggregates, alternative alkaline activators, and different pozzolanic precursors such as metakaolin or blast-furnace slag. Furthermore, the current framework is limited to being a single-target model focused exclusively on compressive strength. In reality, mix-design optimization requires satisfying multiple criteria simultaneously. Future studies should aim to develop multi-output predictive models that can concurrently estimate compressive strength, workability (slump), and durability indices.

https://doi.org/10.3390/buildings16163326

Author Contributions: Conceptualization, N.S. and M.F.; methodology, N.S., S.S.G. and A.S.; software, N.S. and A.S.; validation, N.S., Z.M.N., S.P. and M.F.; formal analysis, N.S. and A.S.; investigation, N.S., Z.M.N. and S.S.G.; resources, M.F. and S.P.; data curation, Z.M.N. and A.S.; writing—original draft preparation, N.S.; writing—review and editing, M.F., S.P., S.S.G. and Z.M.N.; visualization, A.S. and N.S.; supervision, M.F. and S.P.; project administration, M.F.; funding acquisition, M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed during the current study is publicly available in the Mendeley Data repository and can be accessed at: https://data.mendeley.com/datasets/ksvb32cvw6/1 (accessed on 16 July 2026).

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

References

1. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **2001**, *106*, 7183–7192. [CrossRef]
2. Van Dao, D.; Ly, H.-B.; Trinh, S.-H.; Le, T.-T.; Pham, B.T. Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete. *Materials* **2019**, *12*, 983. [CrossRef] [PubMed]
3. Nguyen, K.T.; Nguyen, Q.D.; Le, T.A.; Shin, J.; Lee, K. Analyzing the compressive strength of green fly ash-based geopolymer concrete using experiment and machine learning approaches. *Constr. Build. Mater.* **2020**, *247*, 118581. [CrossRef]
4. Akan, T.; Zálabský, T.; Shirini, K.; Ramos-Frutos, J.; Oliva, D.; Bhuiyan, M.A. Battle Royale Optimizer with Ring Neighborhood Topology. *ICCK Trans. Swarm Evol. Learn.* **2026**, *2*, 19–40.
5. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15. [CrossRef]
6. Saeedi, N.; Shirini, K.; Gharehveran, S.S.; Rajebi, S.; Pedrammehr, S.; Alizadehsani, R.; Saez, J.M.G. Depression Detection Using Cross-Attention Multimodal Fusion on Audio-Visual Data. In *Artificial Intelligence for Neuroscience, Mental Health, and Neurodegenerative Disorders. IWINAC 2026*; Ferrández Vicente, J.M., Val-Calvo, M., Adeli, H., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2026; Volume 16574. [CrossRef]
7. Farshbaf, S.M.; Shirini, K.; Gharehveran, S.S.; Abdollahi, A. Reducing phase lag and noise in residential load forecasting with a hybrid residual-attention model. *Algorithms* **2026**, *19*, 661. [CrossRef]
8. Ghaffari, A.; Shahbazi, Y.; Mokhtari Kashavar, M.; Fotouhi, M.; Pedrammehr, S. Advanced predictive structural health monitoring in high-rise buildings using recurrent neural networks. *Buildings* **2024**, *14*, 3261. [CrossRef]
9. Tan, L.; Liu, X.; Liu, D.; Liu, S.; Wu, W.; Jiang, H. An improved dung beetle optimizer for random forest optimization. In Proceedings of the 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC), Qingdao, China, 13–15 December 2024; pp. 1192–1196. [CrossRef]
10. Chen, Y.; Zhang, Y.; Li, C.; Zhou, J. Application of XGBoost Model Optimized by Multi-Algorithm Ensemble in Predicting FRP-Concrete Interfacial Bond Strength. *Materials* **2025**, *18*, 2868. [CrossRef] [PubMed]
11. Xue, J.; Shen, B. Dung beetle optimizer: A new meta-heuristic algorithm for global optimization. *J. Supercomput.* **2023**, *79*, 7305–7336. [CrossRef]
12. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]
13. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
14. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
15. Shirini, K.; Aghdasi, H.S.; Saeedvand, S. Modified imperialist competitive algorithm for aircraft landing scheduling problem. *J. Supercomput.* **2024**, *80*, 13782–13812. [CrossRef]
16. Meskhi, B.; Beskopylny, A.N.; Stel'makh, S.A.; Shcherban', E.M.; Mailyan, L.R.; Shilov, A.A.; El'shaeva, D.; Shilova, K.; Karalar, M.; Aksoylu, C.; et al. Analytical Review of Geopolymer Concrete: Retrospective and Current Issues. *Materials* **2023**, *16*, 3792. [CrossRef] [PubMed]

17. Sattari, M.T.; Shirini, K.; Javidan, S. Evaluating the efficiency of dimensionality reduction methods in improving the accuracy of water quality index modeling in Qizil-Uzen River using machine learning algorithms. *Water Soil Manag. Model.* **2024**, *4*, 89–104. [CrossRef]
18. Rajebi, S.; Pedrammehr, S.; Shirini, K. Hybrid intelligent optimization of a circularly polarized microstrip antenna array for safe and effective hyperthermia cancer therapy. *Sci. Rep.* **2026**, *16*, 8411. [CrossRef] [PubMed]
19. Li, G.; Zhang, W.; Chen, T.; Kong, X. Preparation and properties of geopolymer recycled concrete: A review of recent developments. *Int. J. Appl. Ceram. Technol.* **2025**, *22*, e70033. [CrossRef]
20. Shirini, K.; Aghdasi, H.S.; Saeedvand, S. Multi-objective aircraft landing problem: A multi-population solution based on non-dominated sorting genetic algorithm-II. *J. Supercomput.* **2024**, *80*, 25283–25314. [CrossRef]
21. Elshaarawy, M.K.; Alsaadawi, M.M.; Hamed, A.K. Machine learning and interactive GUI for concrete compressive strength prediction. *Sci. Rep.* **2024**, *14*, 16694. [CrossRef] [PubMed]
22. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)* **2018**, *31*, 6639–6649.
23. Zaki Dizaji, H.; Shirini, K.; Taheri Hajivand, A.; Monjezi, N. Modelling variables affecting the yield of sugarcane fields using deep recurrent neural network. *Iran. J. Biosyst. Eng.* **2024**, *55*, 93–108. [CrossRef]
24. Davidovits, J. Inorganic polymeric new materials. *J. Therm. Anal.* **1991**, *37*, 1633–1656.
25. Nuaklong, P.; Sata, V.; Chindaprasirt, P. Influence of recycled aggregate on fly ash geopolymer concrete properties. *J. Clean. Prod.* **2016**, *112*, 2300–2307. [CrossRef]
26. Ghorbani, M.A.; Jani, R.; Mohajeri, F. The Taylor Diagram with Distance: A New Way to Compare the Performance of Models. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **2025**, *49*, 305–321. [CrossRef]
27. Baharvand, D.; Saeedi, N.; Gharehveran, S.S.; Shirini, K. A deep generative approach to personalized super mario level design. *Sci. Rep.* **2026**, *16*, 18099. [CrossRef] [PubMed]
28. Shirini, K.; Aghdasi, H.S.; Saeedvand, S. A Comprehensive Survey on Multiple-Runway Aircraft Landing Optimization Problem. *Int. J. Aeronaut. Space Sci.* **2024**, *25*, 1574–1602. [CrossRef]
29. *ASTM C39/C39M-21*; Standard Test Method for Compressive Strength of Cylindrical Concrete Specimens. ASTM International: West Conshohocken, PA, USA, 2021.
30. *EN 12390-3:2019*; Testing Hardened Concrete—Part 3: Compressive Strength of Test Specimens. European Committee for Standardization (CEN): Brussels, Belgium, 2019.
31. Singh, N.S.S.; Al-Nussairi, A.K.J.; Saberi, E.; Pazhoohesh, M.; Gharehveran, S.S.; Dooz, A.P.; Karimipouya, A. Artificial neural network–based adaptive voltage control of a hybrid Zeta DC/DC converter using multilayer perceptron and radial basis function architectures with back-propagation learning. *IEEE Access* **2026**, 105596–105622. [CrossRef]
32. Nekahdari, A.; Kouzehkonani, E.D.; Saeedi, N.; Shirini, K.; Gharehveran, S.S. Multi-task procedural content generation with reinforcement learning. *Sci. Rep.* **2026**, *16*, 18389. [CrossRef] [PubMed]
33. Oskouei, A.G.; Abdolmaleki, N.; Bouyer, A.; Arasteh, B.; Shirini, K. Efficient superpixel-based brain MRI segmentation using multi-scale morphological gradient reconstruction and quantum clustering. *Biomed. Signal Process. Control* **2025**, *100*, 107063. [CrossRef]
34. Qing, S.; Li, C. Data-driven prediction on critical mechanical properties of engineered cementitious composites based on machine learning. *Sci. Rep.* **2024**, *14*, 15322. [CrossRef] [PubMed]
35. Shirini, K.; Gharehveran, S.S.; Rajebi, S.; Pedrammehr, S.; Alizadehsani, R.; Saez, J.M.G. A Bio-Inspired Cross-Modal Attention Framework for Robust Multimodal Multimedia Perception. In *Bioinspired Intelligent Systems: From Robotics and Computer Vision to Trustworthy Applications. IWINAC 2026*; Ferrández Vicente, J.M., Val-Calvo, M., Adeli, H., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2026; Volume 16575. [CrossRef]
36. Ahmad, A.; Ostrowski, K.A.; Maślak, M.; Farooq, F.; Mehmood, I.; Nafees, A. Comparative Study of Supervised Machine Learning Algorithms for Predicting the Compressive Strength of Concrete at High Temperature. *Materials* **2021**, *14*, 4222. [CrossRef] [PubMed]
37. Emami, M.; Salehbagheri, N.; Rajebi, S.; Pedrammehr, S.; Shirini, K. Design and implementation of a hybrid machine learning framework for predicting heart rate status. *Sci. Rep.* **2026**, *16*, 22233. [CrossRef] [PubMed]