Machine learning based conformal predictors for uncertainty-aware compressive strength estimation of concrete

Construction and Building Materials 487 (2025) 141844

Contents lists available at ScienceDirect

Construction and Building Materials

journal homepage: www.elsevier.com/locate/conbuildmat

Machine learning based conformal predictors for uncertainty-aware compressive strength estimation of concrete

Pranjal Tamuly [a,*], Vincenzo Nava [b]

[a] Basque Center for Applied Mathematics, Alameda de Mazarredo, 14, Bilbao, 48009, Spain
[b] Politecnico di Torino, Corso Duca degli Abruzzi, 24, Turin, 19129, Italy

ARTICLE INFO

Keywords:
Conformal predictor
Machine learning
Compressive strength
Uncertainty quantification
Prediction interval

ABSTRACT

Estimating concrete compressive strength is crucial for accurately predicting its performance, optimising material usage, and ensuring the durability and safety of the structure. Traditional machine learning (ML) models have primarily focused on deterministic predictions of compressive strength, often overlooking the uncertainty associated with these estimates. However, concrete is a non-homogeneous material with complex and variable behaviour, making it inherently difficult to predict compressive strength with precision. Therefore, incorporating uncertainty into predictive modelling is essential for producing more reliable and practical results in real-world engineering applications. This study addresses this gap by proposing a comprehensive framework for uncertainty quantification in concrete strength estimation using conformal prediction methods. In this comprehensive study, eight distinct machine learning models are systematically integrated with six conformal prediction variants to construct statistically rigorous prediction intervals. To evaluate the performance of the models holistically in engineering contexts, a novel Efficiency Score (ES) is proposed, combining empirical coverage, mean interval width, and point prediction accuracy. The findings reveal notable trade-offs between predicted interval width and empirical coverage across the model spectrum. Among the tested combinations, LightGBM coupled with Jackknife+ emerges as the most effective configuration, demonstrating the highest efficiency score. Additionally, conformal predictors exhibit satisfactory adaptation to heteroscedasticity, which arises in the predictions of higher-grade concrete (> 40 MPa). Thus, the proposed framework empowers more informed decision-making in concrete design and quality control by providing robust uncertainty bounds advancing beyond traditional deterministic point predictions to support risk-aware infrastructure development.

1. Introduction

Compressive strength serves as the primary indicator of concrete quality, influencing the safety, durability, and performance of infrastructure projects. The complex behaviour of concrete depends upon numerous interrelated factors, including cement type, water-cement ratio, aggregate properties, admixtures, curing conditions, and age. Traditionally, the compressive strength of concrete is determined according to structural design codes, involving the casting and curing of specimens, followed by testing their compressive strength after a specified period. However, this process is highly inefficient as it is time-consuming, labour-intensive, and delays quality assessment until the curing period is complete. Predictive models overcome this limitation by estimating compressive strength based on mix composition and curing conditions, enabling engineers to optimise designs and adjust mixes proactively. From an economic and environmental perspective, prediction models minimise waste, optimise material usage, and prevent costly remediation. Thus, enhanced prediction models can significantly contribute to sustainability efforts by optimising concrete usage and reducing carbon emissions.

One of the pioneering studies on predictive models for concrete compressive strength was conducted by Abrams [1], who established the fundamental inverse exponential relationship between the water-cement ratio and strength. Subsequent researchers expanded this approach by using multiple linear regression models, incorporating additional factors such as cement content, aggregate properties, and age [2,3]. Another approach to predictive modelling utilised physics-based alternatives, linking physical processes to strength development [4,5]. However, these methods face significant limitations in capturing the complex, non-linear relationships inherent in concrete behaviour, especially when dealing with diverse mix compositions that include supplementary cementitious materials or chemical admixtures.

To address these challenges, machine learning algorithms were developed, demonstrating remarkable potential in accurately predicting

* Corresponding author.
E-mail address: ptamuly@bcamath.org (P. Tamuly).

https://doi.org/10.1016/j.conbuildmat.2025.141844
Received 15 April 2025; Received in revised form 8 May 2025; Accepted 15 May 2025

concrete properties and enhancing the precision of the mix design [6–8]. Unlike traditional approaches that rely on predefined mathematical relationships, machine learning algorithms learn directly from data patterns without requiring explicit knowledge of the physical mechanisms involved. This fundamental difference enables ML models to capture the complex, non-linear interactions among concrete constituents that influence compressive strength. Various machine learning techniques, including artificial neural networks (ANN) [9,10], gene expression programming (GEP) [11], support vector machines (SVM) [12], random forests (RF) [13], AdaBoost (AB) [14], multi-expression programming (MEP) [15], and gradient boosting machines (GBM) [16], were effectively utilised to assess the compressive strength of concrete. Similarly, machine learning models like linear regression (LR), nonlinear regression (NLR), multilinear regression (MLR), artificial neural networks, and M5P decision trees were successfully applied to predict the compressive strength of polymer-modified cement-grouted sands, considering factors such as sand size, water-to-cement ratio, polymer content, and curing time [17]. Further, ML models were also utilised to predict the compressive strength of lightweight structural concrete made with sustainable materials derived from oil palm by-products [18]. Apart from these models, various hybrid approaches have been proposed to enhance the prediction of concrete compressive strength. Al-Jamimi et al. [19] introduced a hybrid support vector machine-genetic algorithm (SVM-GA) model that effectively predicted concrete compressive strength with high precision, outperforming traditional regression models. Similarly, Wu and Zhou [20] developed a hybrid machine learning model that combines support vector regression with grid search optimisation (GS-SVR), demonstrating superior predictive capabilities for sustainable concrete. Ahmad et al. [21] utilised an adaptive neuro-fuzzy inference system (ANFIS) to predict the compressive strength of geopolymers synthesised from fly ash, achieving improved performance compared to the multivariate adaptive regression spline (MARS) model. Similarly, the neuro-fuzzy inference system combined with the particle swarm optimisation algorithm was reported to be effective in accurately predicting the axial stress–strain behaviour of FRP-confined concrete columns [22]. Furthermore, Das and Kashem [23] applied machine learning and hybrid methods to predict the compressive strength of ultra-high-performance concrete (UHPC), revealing that the hybrid XGBoost-LightGBM model significantly outperformed traditional models. Another investigation for forecasting the compressive strength of ground granulated blast furnace slag (GGBFS) concrete showed that the SVR-Particle Swarm Optimisation and SVR-Grey Wolf Optimisation (GWO) enhanced the predictions [24]. In a similar manner, an innovative study demonstrated that integrating the extreme learning machine (ELM) with the grey wolf optimiser significantly enhances the prediction accuracy of compressive strength in concrete with partial cement replacements [25].

Likewise, ensemble learning methods combine multiple models to improve prediction accuracy and robustness. Li and Song [26] investigated the compressive and tensile strength of high-performance concrete (HPC) incorporating fly ash and silica fume, and demonstrated that the Gradient Boosting Decision Tree (GBDT) based ensemble model outperformed other machine learning methods. The study on reactive powder concrete demonstrated that ensemble-learning techniques, particularly stacking, significantly outperform traditional machine learning methods in predicting compressive strength [27]. Another study developed a different stacking ensemble learning-based model to predict the compressive strength of rice husk ash (RHA) concrete, using RF and XGBoost as base learners and linear regression as the meta-learner [28]. The model outperformed mainstream machine learning methods, demonstrating superior predictive accuracy and effectively identifying cement and age as the most influential factors. Jia et al. [29] leveraged ensemble machine learning for concrete compressive strength prediction, identifying ensemble LightGBM as the best model with enhanced interpretability.

The aforementioned studies indicate that machine learning models and their variants exhibit superior performance in predicting concrete compressive strength. However, variability in prediction across different datasets remains a challenge. ML models are inherently sensitive to the distribution of the training data, which can result in biased or inaccurate predictions. Furthermore, models trained on limited datasets may exhibit reduced generalisability, particularly when applied to concrete mixtures with substantially different material compositions or curing conditions. Additionally, the effectiveness of these models often depends on careful selection and tuning of hyperparameters, which can significantly influence predictive performance.

Traditional machine learning models for concrete compressive strength have predominantly focused on deterministic point predictions, with model performance typically quantified by coefficient of determination (R^2) values and similar loss metrics. However, this approach faces two critical challenges in predicting concrete strength. First, the limited availability of high-quality testing data for the complete range of compressive strength. Additionally, curating these datasets is constrained by the resource-intensive and time-consuming nature of concrete testing. Second, the highly complex behaviour of concrete is influenced by numerous factors including the distribution of material properties, cement hydration kinetics, and curing conditions. These factors introduce significant variability even among specimens produced from nominally identical mixes. In this context, Mahmood et al. [30] showed that the compressive and flexural strength of polymer-modified cement grout varies significantly with sand grading, water-to-cement ratio, and standards used. Hence, it is necessary that prediction models should quantify uncertainty, acknowledging the inherent variability regardless of model sophistication. To address this gap, this study proposes a robust conformal prediction framework for uncertainty-aware compressive strength estimation. Conformal prediction offers a robust statistical framework that complements point predictions with valid prediction intervals, enabling informed decision-making that accounts for both the expected performance and the associated uncertainty inherent in it. While conformal prediction has gained significant popularity in fields such as medicine [31], finance [32] and biochemistry [33], its application to engineering problems remains extremely limited. A handful of studies have explored conformal prediction in engineering contexts such as structural health monitoring [34], railway engineering [35], and robotics [36]. However, within the specific domain of compressive strength estimation, the application of conformal prediction methods appears to be relatively unexplored. Thus, this study marks an important step towards practical advancements in concrete strength modelling, effectively bridging the gap between deterministic point predictions and the demands of real-world engineering applications. The specific noteworthy contributions of this study are:

- One of the first comprehensive implementations of conformal prediction methods for uncertainty-aware strength estimation of concrete, providing statistically valid prediction intervals beyond deterministic point prediction.
- Introduction of a novel efficiency score that holistically evaluates prediction models by integrating empirical coverage, interval width, and point prediction accuracy.
- Systematic comparison of six conformal prediction variants across eight machine learning models, revealing their relative strengths and limitations for concrete compressive strength estimation.
- Demonstration of heteroscedasticity in concrete strength predictions, particularly at higher compressive strength levels (>40 MPa).
- Developed a practical framework for uncertainty-aware concrete strength estimation that balances statistical validity with engineering utility.

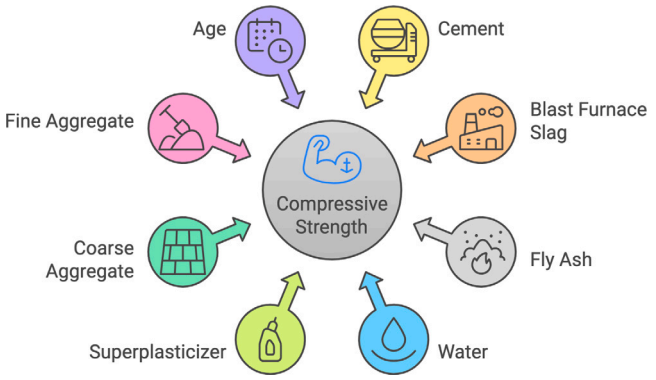

Fig. 1. Feature space for concrete compressive strength prediction.

The structure of the remainder of the paper is as follows: Section 2 presents the dataset, detailing its source, features, preprocessing steps, and outlier analysis. Section 3 discusses the research methodology, including hyper-parameter optimisation of machine learning models and the implementation of conformal prediction variants for uncertainty quantification. Section 4 provides comprehensive results and discussion, analysing the performances and trade-offs. Finally, Section 5 summarises the key conclusions, with limitations and suggestions for future research.

2. Dataset

The dataset is sourced from the work of Yeh [37] and consists of laboratory-measured HPC compressive strength from various design mixes. It is publicly accessible through the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu). It contains 1030 instances, each characterised by eight features: cement content, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age, along with their corresponding compressive strength values. This dataset has been widely utilised in machine learning-based prediction studies, as summarised in Table 1. These studies primarily employed ML-based regression models to predict compressive strength using the given features, as illustrated in Fig. 1.

Data preprocessing is a critical step in concrete strength prediction, as it ensures data consistency, and improves the accuracy of predictive models. Table 2 presents the minimum, maximum, and median values of the features, illustrating the range covered for each concrete constituent and the compressive strength, while the median provides insight into the central tendency of the dataset. This dataset represents parameter ranges typical of conventional to moderately high-performance concretes, with compressive strengths reaching up to approximately 83 MPa. It includes specimen ages ranging from 1 to 365 days, allowing for analysis of early-age, standard, and long-term strength development. However, the dataset does not fully encompass the broader range of parameters, such as higher dosages of superplasticizers and supplementary cementitious materials like blast furnace slag and fly ash. This limitation is common in experimental datasets due to the significant time and cost associated with testing and data acquisition. Nevertheless, this dataset remains a widely recognised benchmark in the literature (refer to Table 1) and continues to serve as a foundational resource for data-driven modelling and analysis in concrete research. To better understand the relationships between input features and concrete compressive strength, a scatter plot of the feature space is presented in Fig. 2. The figure suggests that the data points are distributed in an approximately random manner with no recognisable pattern, thereby supporting the development of ML-based models. To assess multicollinearity in the dataset, pairwise correlations between variables are analysed using Pearson correlation coefficients (r) as shown in Fig. 3. The correlation matrix presents linear associations between predictor variables, with correlations falling below the $|r| = 0.80$ threshold recommended to maintain variable independence in regression analyses [47].

In addition, outlier detection is performed on the concrete compressive strength dataset using the Interquartile Range (IQR) method to enhance model reliability [48]. Unlike z-score approaches, which assume normal distribution and are sensitive to extreme values themselves, the IQR method makes no distributional assumptions. The outliers are imputed with the median value of their respective variables rather than removed [49]. This approach maintained the original sample size of 1030 instances, ensuring more samples for subsequent machine learning.

Table 1
Past studies on Yeh [37] dataset.

Model	References
Linear Regression	Yeh [38] Chou et al. [39] Mandal [40]
Decision Tree	Chou et al. [39] Farooq et al. [41] Mandal [40]
Support Vector Machine	Chou et al. [39] Zhao et al. [42] Mandal [40]
Artificial Neural Network	Yeh [38] Erdal et al. [43] Asteris et al. [44] Zhao et al. [42] Mandal [40]
Random Forest	Farooq et al. [41] Zhao et al. [42] Mandal [40]
Gradient Boosting Regression	Varma et al. [45] Li et al. [46] Mandal [40]
XGBoost	Farooq et al. [41] Zhao et al. [42] Mandal [40]

Table 2
Feature ranges in the dataset.

Feature	Minimum	Median	Maximum
Cement (kg/m^3)	102.000	272.900	540.000
Slag (kg/m^3)	0.000	22.000	359.400
Fly ash (kg/m^3)	0.000	0.000	200.100
Water (kg/m^3)	121.750	185.000	247.000
Superplasticizer (kg/m^3)	0.000	6.350	32.200
Coarse aggregate (kg/m^3)	801.000	968.000	1145.000
Fine aggregate (kg/m^3)	594.000	779.510	992.600
Age (days)	1.000	28.000	365.000
Compressive strength (MPa)	2.332	34.443	82.599

3. Research methodology

A systematic approach is utilised to enable uncertainty-aware estimation of concrete compressive strength by selecting eight machine learning models commonly used in concrete strength prediction studies. Each model undergoes rigorous hyper-parameter optimisation to enhance robustness and minimise over-fitting. Following model optimisation, six conformal prediction variants are implemented to ensure valid uncertainty quantification. Performance evaluation is conducted through a comprehensive framework that assesses both point prediction accuracy and uncertainty quantification using traditional metrics. Additionally, a novel efficiency score (ES) is introduced as a domain-specific metric to reward prediction accuracy while penalising insufficient empirical coverage. This balanced approach is crucial for the

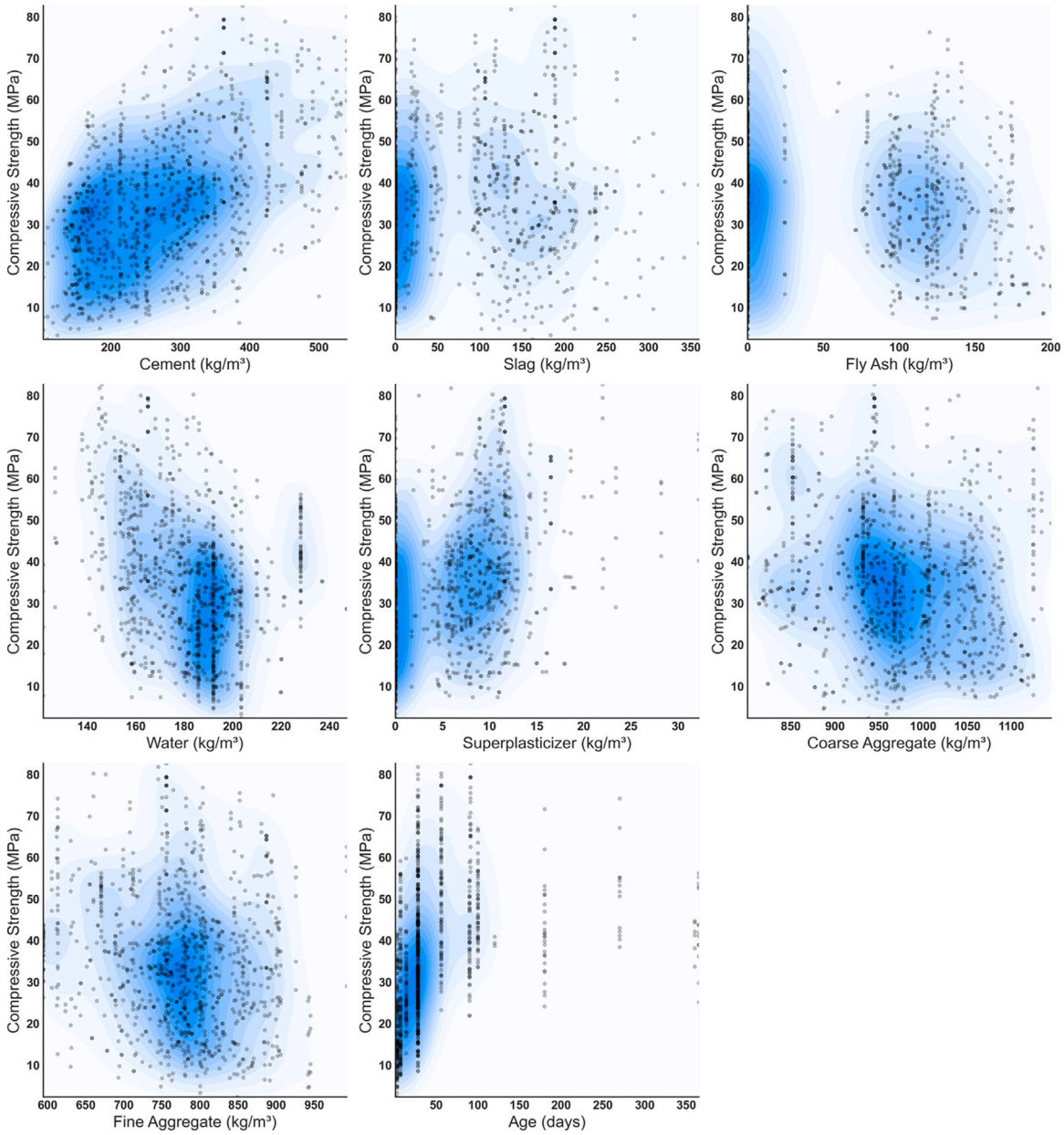

Fig. 2. Scatter plots showing the relationships between concrete compressive strength (MPa) and eight features: cement, slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age. The blue coloured gradients highlight data distribution patterns.

safety-critical nature of concrete strength estimation, where reliable uncertainty bounds are essential for engineering applications. The overall research methodology section is organised into three sequential stages: first, the description of machine learning models; second, the introduction of conformal prediction followed by its variants; and third, the description of the comprehensive performance matrices both traditional and the proposed efficiency score.

3.1. Machine learning models

The selection of machine learning models for predicting concrete compressive strength is based on their widespread use in the literature and their distinct advantages and disadvantages. Decision Trees and Random Forests are preferred for their interpretability and ability to handle non-linear relationships. Gradient Boosting and its variants, XGBoost and LightGBM, are chosen for their strong predictive performance and efficiency in handling large datasets. Support Vector Regression (SVR) is included for its effectiveness in high-dimensional spaces and its robustness against over-fitting. Multi-Layer Perceptrons (MLP) are selected for their flexibility in capturing complex patterns through deep learning architectures. Additionally, Linear Regression is included as a baseline model to provide a simple and interpretable benchmark for comparison.

While these models offer advantages such as improved accuracy and the ability to model complex relationships, they need extensive hyper-parameter tuning. The hyper-parameter optimisation process uses both grid search and randomised search techniques, depending on the complexity of the model. For more complex models such as Random Forest, Gradient Boosting, XGBoost, and LightGBM, randomised search efficiently explores a broader hyper-parameter space. The hyper-parameters of each model are fine-tuned using cross-validation to ensure robust performance across different data splits. The optimal hyperparameters for each model, as shown in Table 3, are determined based on their ability to minimise the root mean squared error (RMSE) and maximise the coefficient of determination (R^2) on the

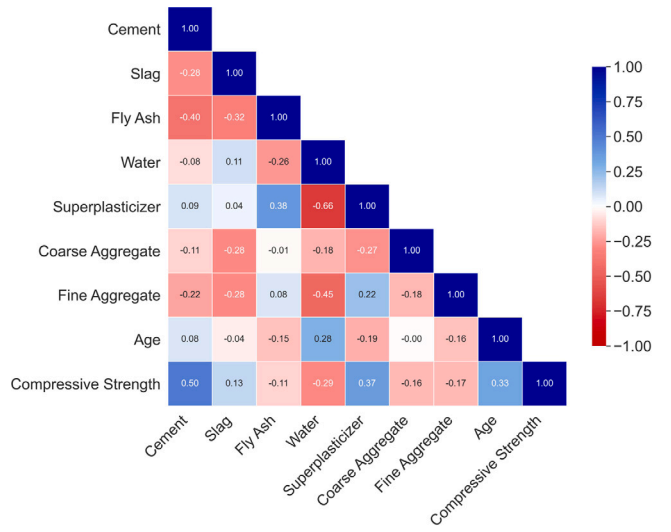

Fig. 3. Pearson correlation matrix showing the relationships between features and compressive strength.

Table 3
Optimal hyperparameters for machine learning models.

Model	Optimal hyperparameters
Decision Tree	max_depth = 15, max_features = auto, min_samples_leaf = 2, min_samples_split = 2
Random Forest	n_estimators = 200, max_depth = 20, min_samples_split = 2, min_samples_leaf = 1, max_features = auto
Gradient Boosting	n_estimators = 100, learning_rate = 0.05, max_depth = 5, min_samples_split = 2, min_samples_leaf = 4, subsample = 0.8
XGBoost	n_estimators = 200, learning_rate = 0.05, max_depth = 7, min_child_weight = 5, gamma = 0, subsample = 0.8, colsample_bytree = 0.9
LightGBM	n_estimators = 200, learning_rate = 0.05, num_leaves = 70, max_depth = 9, min_child_samples = 20, subsample = 1.0, colsample_bytree = 0.8
SVR	kernel = rbf, C = 100, gamma = 0.1, epsilon = 0.5
MLP	hidden_layer_sizes = (200, 100, 50), activation = relu, solver = adam, alpha = 0.01, learning_rate = adaptive

test dataset. This comprehensive approach ensures that the models are well-calibrated to capture the complex relationships in the concrete dataset.

3.2. Conformal prediction

Conformal Prediction (CP) is an innovative approach for constructing prediction intervals for independent and identically distributed (i.i.d.) or exchangeable data, which was introduced by Vovk et al. [50]. The core idea involves assessing how well a new data point, x_{p+1} (with corresponding outcome y_{p+1}), "conforms" to the observed sample $\{(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)\}$ using a predefined nonconformity measure. Detailed tutorials on the principles and implementation of conformal prediction can be found in [51,52]. The basic concept of conformal prediction is shown in Fig. 4.

To construct prediction intervals using CP, the available dataset is first divided into two parts: a proper training set, which is used to fit the predictive model $\hat{f}$, and a calibration set, denoted as $\{(x_i, y_i)\}_{i=1}^{m}$. The conformal algorithm begins by computing nonconformity scores as absolute residuals for each point in the calibration set:

$$s_i = \left| y_i - \hat{f}(x_i) \right| \quad \text{for } i = 1, \ldots, m. \tag{1}$$

These scores (s_i) capture the difference between the predictions and the actual observed values in the calibration data. For a new input x_{n} from the test set $\{(x_i, y_i)\}_{i=1}^{n}$, a prediction interval is then constructed around the prediction $\hat{f}(x_{\mathrm{n}})$. Specifically, the interval is given by

$$C_{1-\alpha}(x_{\mathrm{n}}) = \hat{f}(x_{\mathrm{n}}) \pm q_{1-\alpha}\{s_1, \ldots, s_m\}, \tag{2}$$

The value $q_{1-\alpha}\{s_1, \ldots, s_m\}$ corresponds to the empirical $(1-\alpha)$-quantile. The parameter $\alpha \in (0, 1)$ specifies the desired error level for the prediction interval. For example, setting $\alpha = 0.1$ corresponds to constructing intervals that cover the true response value with at least 90% probability. Based on this quantile, the method constructs a prediction interval $C_{1-\alpha}(x_{\mathrm{n}})$ that satisfies the following marginal coverage guarantee:

$$\mathbb{P}\left(y_{\mathrm{n}} \in C_{1-\alpha}(x_{\mathrm{n}})\right) \geq 1 - \alpha, \tag{3}$$

A key strength of conformal prediction lies in its distribution-free guarantee regardless of the complexity of the underlying predictive model $\hat{f}$, the resulting prediction intervals are guaranteed to contain the true target value with probability at least $(1-\alpha)$, assuming exchangeability of the data. This robust property makes conformal prediction highly attractive for uncertainty quantification in real-world applications. However, despite its strong theoretical foundations and conceptual simplicity, conformal prediction can become computationally intensive, particularly when applied to large datasets or complex models. The computational cost becomes especially pronounced with full conformal approaches, which may require repeated model evaluations across candidate outputs or data subsets. As the size of the calibration set increases, computing the empirical quantile of nonconformity scores becomes a potential bottleneck. These scalability concerns motivate the adoption of more efficient and scalable conformal variants.

The study focuses on three main classes of conformal predictors, chosen for their complementary strengths: **Naive** conformal, **Cross-Validation** (CV)-based conformal, and **Jackknife** conformal. Naive conformal offers a computationally lightweight baseline that is well-suited for large datasets. CV leverages K-fold cross-validation to make better use of limited data, while Jackknife works on the leave-one-out principle providing strong finite-sample validity and adapting effectively to heteroscedasticity. By selecting these methods, it is aimed to explore the interplay between computational cost and prediction reliability for compressive strength estimation tasks. A brief overview of these three methods is presented in the following subsection.

3.2.1. Naive conformal method

Also known as Split Conformal Prediction and is among the most widely used approaches in conformal prediction due to its simplicity and computational efficiency [53]. The Naive method divides the data into a proper training and calibration set. A model is trained exclusively on the training set, and the absolute residuals from the calibration set are used to form prediction intervals for new data. While computationally efficient, requiring only one model to be trained, its performance depends heavily on the representativeness of the calibration data and may not adapt well to heteroscedastic data. Nonetheless, it serves as a benchmark for computational efficiency, given its minimal resource requirements and provides a baseline for evaluating the improvements offered by more advanced conformal variants.

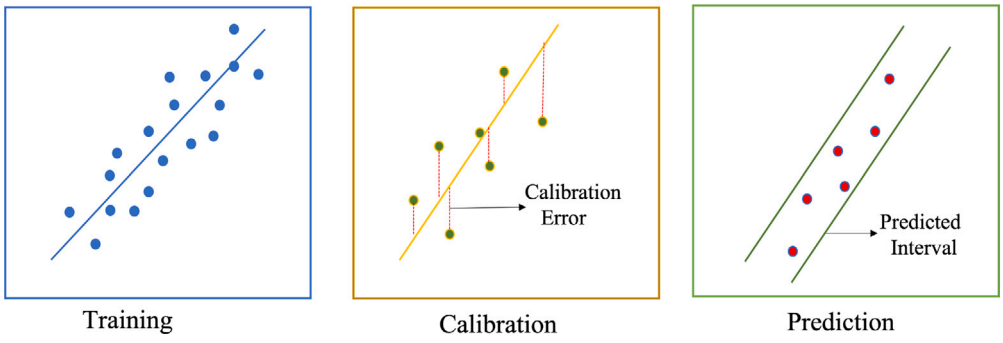

Fig. 4. Basic concept of conformal prediction.

3.2.2. Cross-validation-based conformal methods

Cross-validation conformal methods enhance traditional conformal prediction by making more efficient use of the available data through repeated training and calibration [54]. In the basic CV method, the dataset is partitioned into K folds; K models are trained, each using $(K-1)$ folds for training and the remaining fold for calibration. This yields K sets of residuals and prediction intervals, which are averaged to produce final intervals. The CV+ method improves upon basic CV by aggregating all residuals from the data folds to construct a single pooled conformity score distribution [55]. This pooling enhances the statistical efficiency and stability of the prediction intervals, particularly under data scarcity. CV-minmax builds on CV+ by selecting the maximum width across all K prediction intervals. While this results in wider intervals, it offers stronger coverage guarantees. In summary, while CV offers balanced efficiency, CV+ improves statistical robustness, and CV-minmax prioritises coverage reliability at the cost of increased conservativeness.

3.2.3. Jackknife-based conformal methods

Jackknife conformal methods are based on leave-one-out cross-validation, where multiple models are trained, each omitting one data point [56]. The basic Jackknife method utilises these models to estimate prediction intervals but does not guarantee finite-sample coverage. Jackknife+ addresses this limitation by adjusting prediction intervals using quantiles of leave-one-out residuals, offering rigorous finite-sample validity. It achieves a favourable trade-off between reliability and interval tightness, adapting well to the data distribution. Similar to CV-minmax, Jackknife-minmax adopts the same minmax principle, which selects the widest interval. Compared to CV methods, Jackknife-based methods tend to be more computationally intensive due to the large number of models trained but offer finer adaptation to uncertainty in the data.

The dataset is partitioned using a structured strategy to ensure rigorous and fair evaluation of conformal prediction methods, following established practices in the literature [57]. A random split is initially applied to divide the dataset into training (80%) and test (20%) sets, with these partitions held fixed throughout all experiments. Six distinct conformal prediction variants are evaluated: Naive, Jackknife+, Jackknife-minmax, CV, CV+, and CV-minmax, all calibrated to maintain a 90% confidence level ($\alpha = 0.1$). The choice of $\alpha = 0.1$ (90% confidence level) is made with careful consideration of both statistical robustness and practical engineering relevance. In concrete compressive strength prediction, a 90% confidence level represents a pragmatic balance between safety and practicality. Selecting a higher confidence level (e.g., 95% with $\alpha = 0.05$) would result in wider prediction intervals, which may limit practical applicability. Conversely, lower confidence (e.g., 80% with $\alpha = 0.2$) would yield narrower intervals but potentially underestimate uncertainty in safety-critical applications. For the Naive method, a 70/30 training-calibration split is implemented on the initial training set, with nonconformity scores computed on the calibration set. This ratio achieves a balance for the dataset used in this study by allocating sufficient data for effective model training while reserving a calibration set large enough to ensure stable quantile estimation for uncertainty quantification. This partition is in accordance with standard practices commonly adopted in the literature [57]. The cross-validation methods (CV, CV+, and CV-minmax) are implemented with 10-fold cross-validation, each utilising the training data from the initial split. The Jackknife variants (Jackknife+ and Jackknife-minmax) utilise the leave-one-out approach with the same training set from the initial split. In this framework, for each model, the calibration set consists of a single held-out observation, while the model is trained on the remaining points. In summary, this comprehensive implementation systematically evaluates the trade-off between computational requirements and the quality of prediction intervals across the spectrum of commonly used conformal methods.

3.3. Performance matrices

The performance of conformal prediction methods is evaluated using a set of complementary metrics, each capturing a distinct aspect of performance. These metrics assess point prediction accuracy, coverage reliability, predicted interval width, and a domain-specific efficiency score tailored for concrete compressive strength estimation.

3.3.1. Point prediction matrices

Standard regression metrics are used to evaluate the accuracy of point predictions, providing insight into the performance of the underlying predictive models. The coefficient of determination (R^2) measures the proportion of variance in the dependent variable that is explained by the independent variables in the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where y_i denotes the true value, $\hat{y}_i$ is the corresponding predicted value, $\bar{y}$ is the mean of the observed values, and n is the total number of samples in test set.

In addition to R^2, the root mean squared error (RMSE) is used to evaluate the average magnitude of prediction errors as follows

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

The mean absolute error (MAE) is also computed to provide a linear measure of error magnitude as given below

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

The A20 metric evaluates the proportion of predicted values that fall within ±20% of the corresponding true values ($\hat{y}$). Mathematically, it is defined as:

$$\mathrm{A20} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{F}\left(0.8 \cdot y_i \le \hat{y}_i \le 1.2 \cdot y_i\right) \tag{7}$$

where $\mathcal{F}$ is the indicator function that equals 1 if the condition holds and 0 otherwise.

3.3.2. Empirical coverage

Empirical coverage (EC) quantifies the proportion of predictions that lie within the corresponding predicted intervals. For a conformal prediction method calibrated to a confidence level of $1-\alpha$, the empirical coverage should ideally align with this target. It is computed as:

$$\mathrm{EC} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}(y_i \in [L_i, U_i]) \tag{8}$$

where U_i and L_i represent the upper and lower bounds of the prediction interval for the ith instance, respectively. $\mathcal{F}(y_i \in [L_i, U_i])$ is an indicator function that returns 1 if the true value y_i falls within the prediction interval $[L_i, U_i]$, and 0 otherwise. Ensuring that empirical coverage closely matches the nominal confidence level is critical for the reliability of predictive intervals, as it reflects the ability of the models to appropriately characterise uncertainty in its estimations.

3.3.3. Mean interval width

The mean interval width (MIW) of the prediction intervals serves as a measure of the precision of uncertainty quantification. It is defined as the average distance between the upper and lower bounds of the intervals across all test samples:

$$\mathrm{MIW} = \frac{1}{n}\sum_{i=1}^{n}(U_i - L_i) \tag{9}$$

A smaller average width reflects a more precise estimation, assuming that the empirical coverage remains aligned with the nominal confidence level. Striking a balance between compact interval width and reliable coverage is essential for generating informative and trustworthy uncertainty estimates.

3.3.4. Efficiency score

Traditional evaluation matrices mentioned above consider empirical coverage (EC), interval width (IW) and point prediction matrices as separate entities while the interplay between prediction accuracy and interval efficiency in engineering contexts is often overlooked. In this study, a novel efficiency score (ES) is introduced to evaluate prediction intervals specifically for concrete strength estimation problems covering all aspects of the model's performance. The proposed efficiency score is defined as

$$\mathrm{ES} = (\mathrm{C_{score}} \times R^2 \times \mathrm{IW_{score}}) - \mathrm{C_{penalty}} \tag{10}$$

where $\mathrm{C_{score}}$ represents the coverage score, rewarding models that maintain high coverage, $\mathrm{IW_{score}}$ denotes the interval width score, rewarding narrower, more informative intervals, and R^2 represents the coefficient of determination, rewarding higher predictive accuracy. $\mathrm{C_{penalty}}$ introduces a penalty for under-coverage, ensuring that models failing to adhere to the coverage guarantee are penalised accordingly. The multiplicative form in ES ensures that low performance in any single component (coverage, accuracy, or interval width) significantly impacts the overall score, reflecting the practical utility that requires adequate performance across all dimensions simultaneously. Here, $\mathrm{C_{score}}$ is defined as follows

$$\mathrm{C_{score}} = \min\left(\frac{\mathrm{EC}}{\mathrm{EC_{target}}}, 1.0\right) \tag{11}$$

with EC representing the empirical coverage achieved, and $\mathrm{EC_{target}}$ denoting the intended level of predefined confidence. The interval width score $\mathrm{IW_{score}}$ evaluates the efficiency of prediction intervals using a domain-specific approach appropriate for concrete strength estimation. First, each estimated interval width is divided by its corresponding model's point predictions to obtain the normalised interval width (NIW):

$$\mathrm{NIW}_i = \frac{\mathrm{U}_i - \mathrm{L}_i}{\hat{y}_i} \tag{12}$$

where $\hat{y}_i$ is the point prediction from the model. This normalisation acknowledges that higher-strength concretes naturally exhibit greater absolute variability, making the relative width more meaningful than the absolute width. The mean of normalised interval width (NIW_{mean}) is then calculated:

$$\mathrm{NIW}_{mean} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{NIW}_i \tag{13}$$

Finally, the interval width score is calculated using a hyperbolic transformation that converts an unbounded positive value into a bounded score between 0 and 1:

$$\mathrm{IW_{score}} = \frac{1}{1+\mathrm{NIW}_{mean}} \tag{14}$$

This simple transformation ensures that narrower intervals receive higher scores while wider intervals receive lower scores. Importantly, this normalisation approach naturally adapts to the heteroscedasticity of concrete strength prediction, where uncertainty increases with strength magnitude. As estimates of concrete strength rise from normal ranges to high-strength, prediction intervals widen. This may occur due to increased sensitivity to variations in material composition, water-cement ratio, and curing conditions. The scoring mechanism accommodates this behaviour, ensuring models are not penalised disproportionately at higher strengths.

Another key feature of the proposed metric can be seen in the incorporation of a coverage penalty that imposes a cost for under-coverage. This is formalised as

$$\mathrm{C_{penalty}} = \max\left(0, \frac{\mathrm{EC_{target}} - \mathrm{EC}}{\mathrm{EC_{target}}}\right) \tag{15}$$

In civil constructions, underestimating uncertainty can violate structural integrity, leading to potentially catastrophic outcomes. Therefore, models are appropriately penalised if they fail to meet the empirical coverage requirement. Thus, the design of the efficiency score is driven by the practical needs of civil engineering, where the coverage penalty addresses the risk of underestimating uncertainty. This is especially important in concrete strength prediction, as insufficient confidence in predictions could lead to unsafe construction.

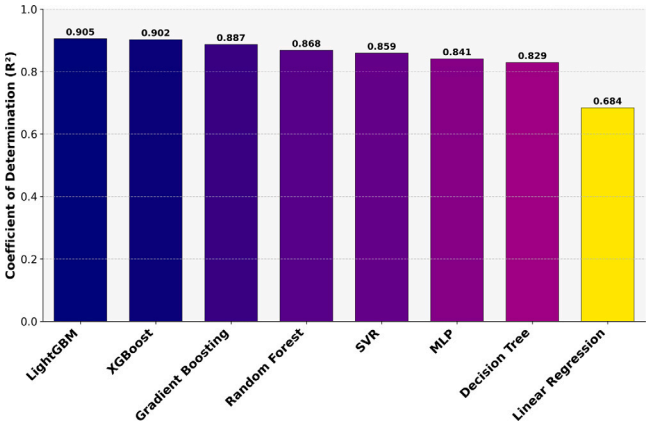

Fig. 5. Comparison of coefficient of determination (R^2) across different machine learning models.

4. Results and discussions

A comprehensive comparative analysis is conducted to evaluate the performance of different machine learning models integrated with various conformal prediction methods for concrete compressive strength estimation. The evaluation framework encompasses eight distinct machine learning models, with each model being evaluated against six conformal prediction variants. Multiple performance metrics mentioned in the previous section are systematically assessed, which includes predictive accuracy measures (R^2, RMSE, MAE and A20), uncertainty

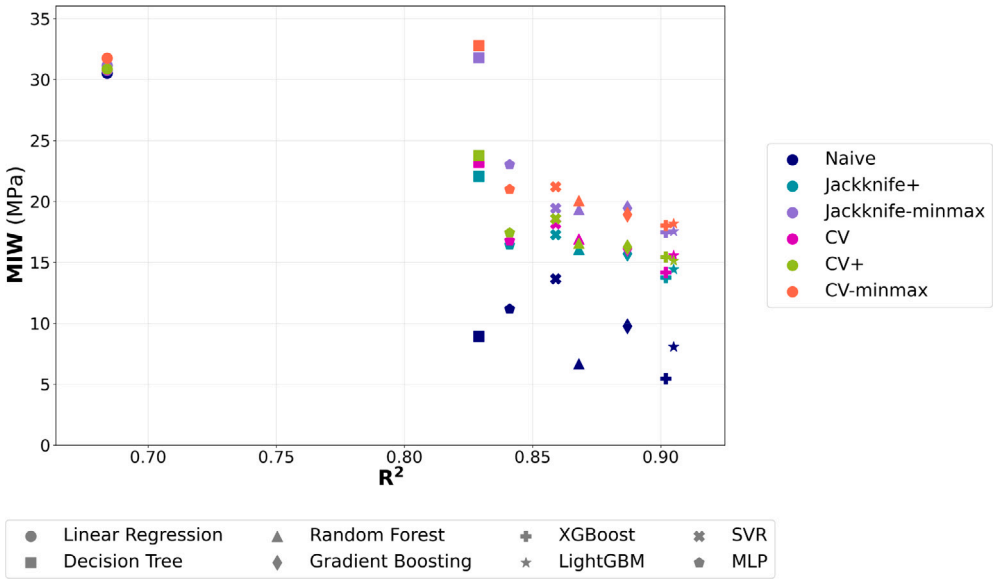

Fig. 6. Relationship between model accuracy (R^2) and mean interval width across (MIW) different conformal prediction methods and machine learning models.

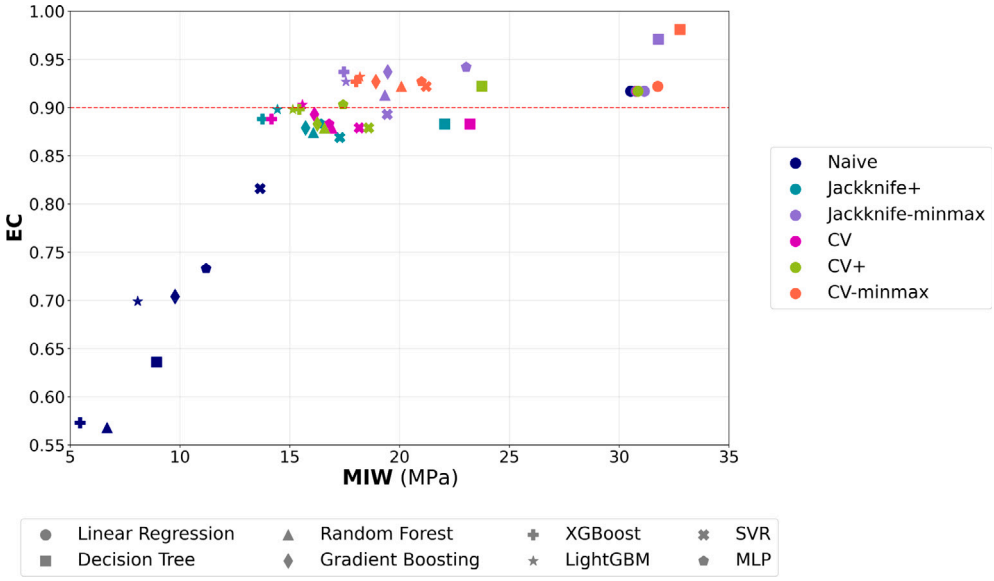

Fig. 7. Trade-off between empirical coverage (EC) and mean interval width (MIW) across different conformal prediction variants and machine learning models with the target empirical coverage of 90% indicated by a red dashed line.

quantification (EC and MIW), and proposed efficiency score (ES). This facilitates a deep understanding of the strengths and limitations inherent to each combination in the context of concrete strength estimation. Detailed results for each combination of ML models with conformal prediction variants are provided in Table 4 of the Appendix A.

Fig. 5 shows the coefficient of determination (R^2) for various machine learning models reflecting their deterministic point prediction accuracy. LightGBM and XGBoost achieve the highest R^2 values of 0.905 and 0.902, respectively, followed by Gradient Boosting and Random Forest. SVR and MLP demonstrate moderate accuracy, while Decision Tree and Linear Regression show the lowest R^2 values. The evaluation of RMSE, MAE and A20 further supports these observed trends across these models, as presented in Table 4. These results highlight the superior performance of tree-based models (LightGBM and XGBoost) in deterministic point predictions by capturing complex data patterns.

The coefficient of determination provides a stand-alone deterministic performance overview of these machine learning models. However, for a more comprehensive understanding of uncertainty-aware estimates, it should be evaluated alongside the mean interval width (MIW), as illustrated in Fig. 6. It shows a distinct pattern where models with higher R^2 values generally produce narrower prediction intervals. Advanced tree-based methods (XGBoost, LightGBM) achieve superior accuracy ($R^2 > 0.90$) while maintaining comparatively narrow intervals, whereas Linear Regression exhibits consistently wider intervals (>30 MPa) across all conformal variants. In general, Jackknife+, CV, and CV+ generate similarly wide intervals, whereas Jackknife-minmax and CV-minmax tend to produce wider ones. Across all ML models, Naive method generates significantly narrower intervals, with this gap widening as model accuracy increases. Although this suggests that the Naive method provides the best performance, however, another crucial factor to consider is empirical coverage. If this is not satisfied, the resulting interval becomes unreliable, as the probability of capturing the actual value within this interval will be very low.

Fig. 7 is plotted to examine the relationship between empirical coverage (EC) and mean interval width (MIW). The figure illustrates

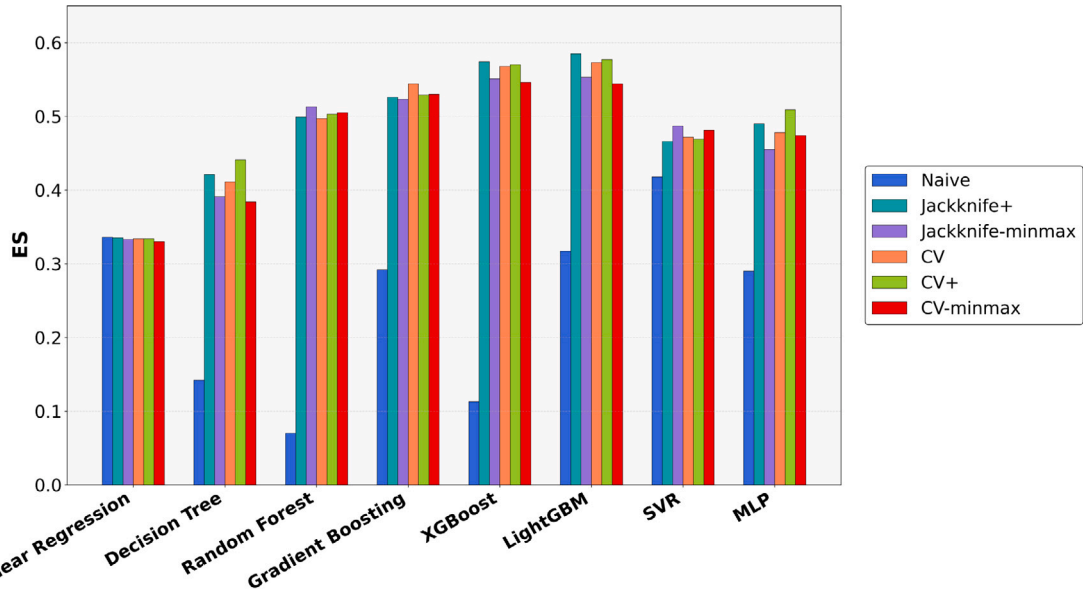

Fig. 8. Comparative analysis of efficiency scores (ES) across different machine learning models and conformal prediction variants.

this relationship across various conformal prediction methods, with the target empirical coverage of 90% marked by a red dashed line. It effectively captures the fundamental trade-off between coverage reliability and interval precision. Despite producing compact intervals, the Naive method systematically underperforms in coverage guarantee in tree based models. In contrast, minmax variants (Jackknife-minmax and CV-minmax) achieve superior empirical coverage (>90%), positioning themselves consistently above the target threshold. Further, they also produce more conservative estimates of prediction intervals compared to the plus (+) variants. This is due to the minmax approach selecting the maximum width of interval across all K-fold predictions for each test point, which typically leads to conservative estimates. Hence, Jackknife and CV minmax demonstrate exceptional coverage for tree-based models but at the cost of substantial interval width. Jackknife+, CV and CV+ aim to strike a balance between mean interval width and empirical coverage, positioning themselves closest to the target line. Further, model complexity also significantly influences this trade-off where advanced ensemble methods achieve more favourable balances, with LightGBM and XGBoost implementations achieving excellent empirical coverage with relatively reasonable mean interval widths. Finally, the plot demonstrates that high predictive accuracy from the underlying ML model does not automatically translate to coverage guarantee; even top-performing models like LightGBM and XGBoost exhibit poor empirical coverage when paired with the Naive method.

While Figs. 6 and 7 provide valuable insights into specific performance dimensions, they do not offer a holistic assessment in practical engineering applications. The proposed efficiency score addresses this limitation by integrating empirical coverage, mean interval width, and R^2 into a unified metric, as shown in Fig. 8. It highlights notable performance differences between the conformal variants across various model architectures. Notably, the Naive method demonstrates consistent underperformance across the majority of models, with the exception of linear regression, largely due to its insufficient coverage. The proposed efficiency score effectively incorporates this by assigning a high penalty to the overall efficiency score. Similarly, the below-average efficiency score becomes particularly evident in tree-based architectures, where, despite generating narrower intervals, the efficiency score decreases even further with the Naive method. On the other hand, the same tree-based models (Gradient Boosting, XGBoost, LightGBM) demonstrated the highest efficiency scores when combined with other conformal methods. Particularly, the LightGBM model with Jackknife+ demonstrates the highest overall efficiency (ES = 0.585), closely followed by its CV+ implementation (ES = 0.577). The optimal balance between empirical coverage and mean interval width is also reflected in Fig. 7, where these two align at the leftmost point on the target line. Thus, the proposed efficiency score effectively rewards models that strike an excellent balance, while penalising models proportionately that fail to meet the coverage guarantee.

The predictions of LightGBM models paired with Jackknife+, Jackknife-minmax, CV+, and CV-minmax are illustrated in Fig. 9. To support reproducibility, the results of the LightGBM model using various conformal methods are provided as supplementary files. Fig. 9 presents the comparison of the actual compressive strength to the estimates on the test data by the four best scoring combinations. These plots demonstrate that nearly all four implementations of the LightGBM models produce similar results with minor differences. In all of these cases, the point estimates closely track the actual values reflecting the high efficiency of the underlying LightGBM models. Also, the prediction interval successfully encompasses the majority of the actual data points, confirming the high empirical coverage achieved by these models. In general, the predicted interval generally remains wide enough to capture the true values with some occasional fluctuations.

One key finding from Fig. 9 is that pronounced heteroscedasticity exists in predictions noticeably depending on the strength level. This heteroscedasticity is most evident in the higher strength region (>40 MPa), where predictions vary considerably compared to the lower range. The heteroscedasticity observed in the predictions is further investigated using a formal statistical test. The Breusch–Pagan [58] test is conducted on the residuals from the LightGBM model, dividing the data into low-strength (≤40 MPa) and high-strength (>40 MPa) groups. The results exposed significant heteroscedasticity in the high-strength concrete set (p-value = 0.001740) while showing no significant heteroscedasticity in the low-strength samples (p-value = 0.367584). The observed heteroscedasticity for high-strength concretes likely arises from more complex interactions among mix constituents and a heightened sensitivity to curing conditions. The LightGBM with these selected conformal variants demonstrates satisfactory adaptation to this heteroscedasticity for most of the estimates, providing appropriately wider intervals to accommodate uncertainty in the predictions.

In addition to numerical efficiency, the computational cost is also crucial, as the demands can grow considerably with the complexity of the underlying machine learning model and the chosen conformal variant. Fig. 10 provides a visual representation of the computational requirements (using a MacBook Pro with an Intel Core i5 processor)

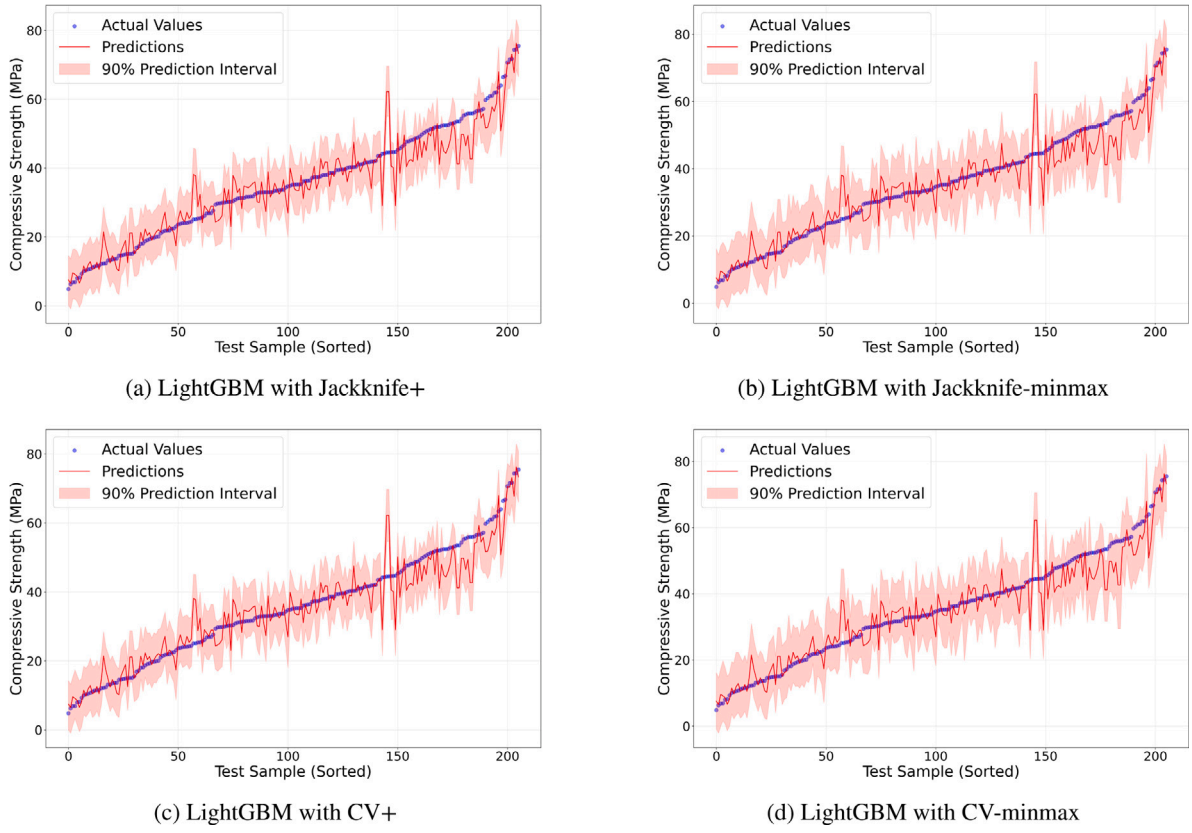

Fig. 9. Demonstration of conformal predictions for LightGBM models with different variants of conformal methods.

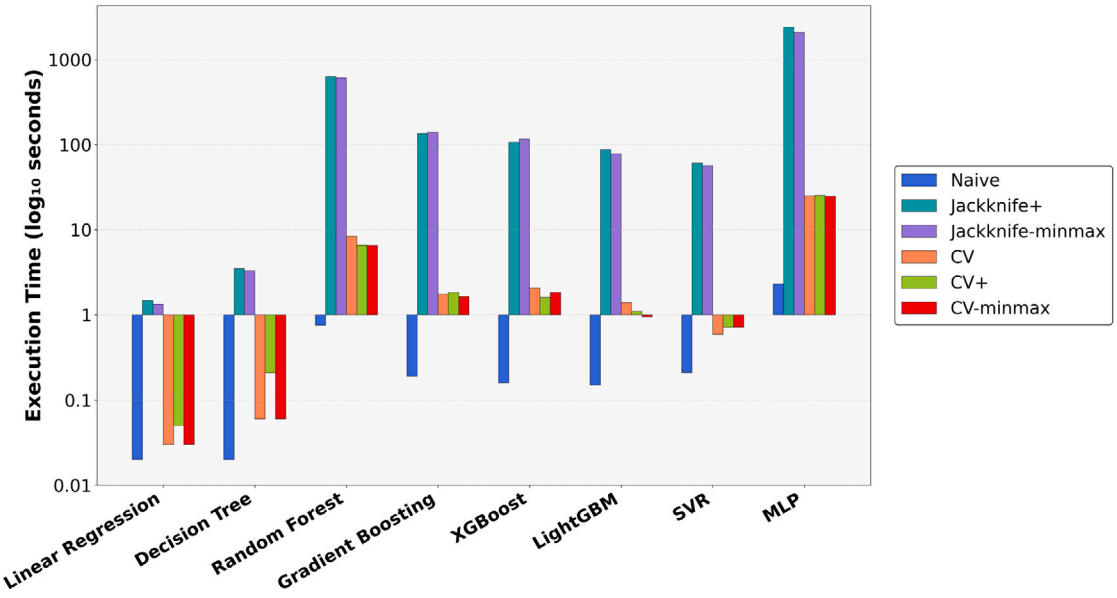

Fig. 10. Computational cost comparison across different combinations of machine learning models and conformal methods (logarithmic scale).

across various conformal prediction methods for 824 training instances and 206 testing instances. The computational cost, quantified by the time taken for each model–method combination, is presented in Table 4 (Appendix A). Here, Jackknife-based approaches demand significantly higher resources compared to CV-based variants. The elevated cost of Jackknife methods originates from their leave-one-out approach requiring a large number of model refits. This becomes particularly evident for computationally expensive models like MLP and Random Forest. Advanced CV-based methods achieve slightly inferior performance with reduced computational overhead by utilising K-fold cross-validation. Particularly, the CV-minmax represents a practical alternative that preserves strong coverage guarantees while significantly reducing computational time. Although it yields slightly wider prediction intervals,

its consistent ability to meet coverage requirements makes it particularly suitable for safety-critical applications. These characteristics make it a suitable substitute for Jackknife+ in scenarios involving limited computational resources, large-scale datasets, or situations requiring rapid deployment, such as on-site quality control.

5. Conclusions

This study presents a comprehensive assessment of uncertainty quantification methods for the compressive strength estimation of concrete through the novel application of conformal prediction techniques. The research addresses a critical gap in construction materials modelling by moving beyond deterministic predictions to provide statistically valid uncertainty estimates that account for the inherent variability in concrete properties. The novel efficiency score introduced in this study offers a practical metric for ML models in concrete strength estimation, balancing statistical coverage with interval precision. This metric acknowledges the safety-critical nature of concrete applications, where both coverage guarantee and accurate uncertainty bounds are essential. The pointwise key conclusions for this study are presented below:

- **Advanced tree-based models** (XGBoost, LightGBM, Gradient Boosting) demonstrate superior performance in both deterministic prediction and uncertainty quantification, which stems from their capacity to model complex non-linear interactions between concrete constituents and strength development.
- The proposed **efficiency score** has shown its excellent ability to assess the performance of ML models by rewarding the balanced trade-off between empirical coverage and mean interval width while applying proportional penalties to models that do not satisfy the coverage guarantee.
- **LightGBM with Jackknife+** represents the optimal model-method combination, achieving the highest efficiency score striking a perfect balance between empirical coverage and mean interval width.
- Significant **heteroscedasticity** is observed in the predictions for high-grade concrete (>40 MPa). Due to the leave-one-out strategy, Jackknife+ is especially effective for handling such heteroscedastic data.
- **The Naive method**, despite its computational efficiency, exhibits systematic under-coverage across most models. This under-coverage is model-dependent, with simpler models like Linear Regression achieving excellent coverage, while complex tree-based models such as XGBoost and LightGBM show substantial under-coverage.
- **Computational efficiency** analysis reveals that Jackknife-based methods require substantially higher computational resources due to their leave-one-out approach. CV-minmax method can be adopted as an alternative, offering strong coverage guarantees and substantial savings in computational cost, while yielding marginally lower predictive performance.

This study is conducted on a single dataset with 1030 samples, potentially limiting the generalisability of the findings to different concrete types or mix designs and the results or claims presented in this study are specific to this dataset's characteristics. This limitation is particularly important for conformal prediction methods, as their statistical guarantees apply only within the distribution represented by the training data. Extrapolation to concrete mixtures with constituent proportions or properties outside these ranges may result in prediction intervals with compromised coverage properties. Future research should explore the applicability of these methods across diverse concrete datasets, including those incorporating advanced admixtures, alternative cementitious materials, and sustainable options such as geopolymers, to validate the broader utility of the approach. Additionally, conducting feature space analysis in relation to machine learning models in future research may help to identify key input variables for compressive strength estimation, thereby enhancing the interpretability and practical relevance of the proposed framework. The proposed efficiency score has shown strong potential in assessing the performance of ML models; however, its robustness needs to be established through validation across multiple datasets. Furthermore, a systematic investigation of alternative formulations of efficiency metrics for uncertainty-aware concrete strength prediction would further contribute to advancing the field, particularly when guided by practical considerations and decision-making logic specific to the behaviour of concrete.

Overall, this research establishes a robust framework for uncertainty-aware concrete strength estimation that balances statistical validity with practical utility. The proposed approach enables engineers and practitioners to make more informed decisions about concrete mix design, quality control, and structural safety by quantifying uncertainty in a reliable and interpretable manner. This study improves the reliability of decision-making in concrete design and application by transitioning from deterministic point predictions to statistically sound prediction intervals.

CRediT authorship contribution statement

Pranjal Tamuly: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vincenzo Nava:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors also would like to acknowledge the "BCAM Severo Ochoa" accreditation of excellence CEX2021-001142-S/MICIN/AEI/10.13039/501100011033; and the Basque Government through the BERC 2022–2025 program. The authors also acknowledge funding from the ELKARTEK project RUL-ET by the Basque Government (KK-2024/00086). Vincenzo Nava's work is funded within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan – NRRP, Mission 4, Component 2, Investment 1.3 – D.D. 1243 2/8/2022, PE0000005)– SPOKE TS 2. The authors would like to thank UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/) and Professor I-Cheng Yeh for sharing the experimental data set.

Appendix A

See Table 4.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.conbuildmat.2025.141844.

Data availability

Data will be made available on request.

Table 4
Detailed comparison of conformal prediction methods across ML models.

Model	Method	R^2	RMSE (MPa)	MAE (MPa)	A20	EC	MIW (MPa)	ES	Time (s)
Linear Regression	Naive	0.684	9.017	7.051	0.558	0.917	30.527	0.336	0.01
	Jackknife+	0.684	9.017	7.051	0.558	0.917	30.762	0.335	1.46
	Jackknife-minmax	0.684	9.017	7.051	0.558	0.917	31.148	0.333	1.32
	CV	0.684	9.017	7.051	0.558	0.917	30.795	0.334	0.02
	CV+	0.684	9.017	7.051	0.558	0.917	30.858	0.334	0.04
	CV-minmax	0.684	9.017	7.051	0.558	0.922	31.754	0.330	0.02
Decision Tree	Naive	0.829	6.642	4.682	0.743	0.636	8.936	0.142	0.01
	Jackknife+	0.829	6.642	4.682	0.743	0.883	22.056	0.421	3.50
	Jackknife-minmax	0.829	6.642	4.682	0.743	0.971	31.793	0.391	3.29
	CV	0.829	6.642	4.682	0.743	0.883	23.208	0.411	0.05
	CV+	0.829	6.642	4.682	0.743	0.922	23.750	0.441	0.20
	CV-minmax	0.829	6.642	4.682	0.743	0.981	32.771	0.384	0.05
Random Forest	Naive	0.868	5.839	4.081	0.782	0.568	6.675	0.070	0.75
	Jackknife+	0.868	5.839	4.081	0.782	0.874	16.075	0.499	633.32
	Jackknife-minmax	0.868	5.839	4.081	0.782	0.913	19.328	0.513	610.39
	CV	0.868	5.839	4.081	0.782	0.879	16.913	0.497	8.38
	CV+	0.868	5.839	4.081	0.782	0.879	16.577	0.503	6.57
	CV-minmax	0.868	5.839	4.081	0.782	0.922	20.077	0.505	6.54
Gradient Boosting	Naive	0.887	5.402	3.894	0.806	0.704	9.775	0.292	0.18
	Jackknife+	0.887	5.402	3.894	0.806	0.879	15.714	0.526	135.23
	Jackknife-minmax	0.887	5.402	3.894	0.806	0.937	19.459	0.523	139.63
	CV	0.887	5.402	3.894	0.806	0.893	16.120	0.544	1.74
	CV+	0.887	5.402	3.894	0.806	0.883	16.250	0.529	1.81
	CV-minmax	0.887	5.402	3.894	0.806	0.927	18.918	0.530	1.63
XGBoost	Naive	0.902	5.031	3.387	0.835	0.573	5.454	0.113	0.15
	Jackknife+	0.902	5.031	3.387	0.835	0.888	13.754	0.574	107.37
	Jackknife-minmax	0.902	5.031	3.387	0.835	0.937	17.464	0.551	116.21
	CV	0.902	5.031	3.387	0.835	0.888	14.169	0.568	2.06
	CV+	0.902	5.031	3.387	0.835	0.898	15.435	0.570	1.59
	CV-minmax	0.902	5.031	3.387	0.835	0.927	18.022	0.546	1.82
LightGBM	Naive	0.905	4.960	3.563	0.845	0.699	8.074	0.317	0.14
	Jackknife+	0.905	4.960	3.563	0.845	0.898	14.434	0.585	87.72
	Jackknife-minmax	0.905	4.960	3.563	0.845	0.927	17.555	0.553	78.19
	CV	0.905	4.960	3.563	0.845	0.903	15.568	0.573	1.38
	CV+	0.905	4.960	3.563	0.845	0.898	15.139	0.577	1.09
	CV-minmax	0.905	4.960	3.563	0.845	0.932	18.191	0.544	0.94
SVR	Naive	0.859	6.038	4.088	0.820	0.816	13.646	0.418	0.20
	Jackknife+	0.859	6.038	4.088	0.820	0.869	17.271	0.466	61.11
	Jackknife-minmax	0.859	6.038	4.088	0.820	0.893	19.434	0.487	57.17
	CV	0.859	6.038	4.088	0.820	0.879	18.164	0.472	0.58
	CV+	0.859	6.038	4.088	0.820	0.879	18.587	0.469	0.71
	CV-minmax	0.859	6.038	4.088	0.820	0.922	21.212	0.481	0.71
MLP	Naive	0.841	6.402	4.465	0.767	0.733	11.182	0.290	2.29
	Jackknife+	0.841	6.402	4.465	0.767	0.883	16.429	0.490	2394.37
	Jackknife-minmax	0.841	6.402	4.465	0.767	0.942	23.028	0.455	2089.84
	CV	0.841	6.402	4.465	0.767	0.883	16.792	0.478	24.88
	CV+	0.841	6.402	4.465	0.767	0.903	17.430	0.509	25.30
	CV-minmax	0.841	6.402	4.465	0.767	0.927	20.998	0.474	24.78

References

[1] D.A. Abrams, Design of Concrete Mixtures, vol. 1, Structural Materials Research Laboratory, Lewis Institute, 1919.

[2] S. Popovics, Analysis of concrete strength versus water-cement ratio relationship, Mater. J. 87 (5) (1990) 517–529.

[3] F. De Larrard, Concrete Mixture Proportioning: a Scientific Approach, CRC Press, 1999.

[4] R. Feret, Sur la compacité des mortiers hydrauliques, Ann. Pntas Chaussees, Mem Doc 4 (1892) 5–164.

[5] T.C. Powers, Structure and physical properties of hardened portland cement paste, J. Am. Ceram. Soc. 41 (1) (1958) 1–6.

[6] R. Kumar, E. Althaqafi, S.G.K. Patro, V. Simic, A. Babbar, D. Pamucar, S.K. Singh, A. Verma, Machine and deep learning methods for concrete strength prediction: A bibliometric and content analysis review of research trends and future directions, Appl. Soft Comput. (2024) 111956.

[7] J. Sobhani, M. Najimi, A.R. Pourkhorshidi, T. Parhizkar, Prediction of the compressive strength of no-slump concrete: A comparative study of regression, neural network and ANFIS models, Constr. Build. Mater. 24 (5) (2010) 709–718.

[8] Z. Shen, A.F. Deifalla, P. Kamiński, A. Dyczko, Compressive strength evaluation of ultra-high-strength concrete by machine learning, Materials 15 (10) (2022) 3523.

[9] H. Eskandari-Naddaf, R. Kazemi, ANN prediction of cement mortar compressive strength, influence of cement strength class, Constr. Build. Mater. 138 (2017) 1–11.

[10] A. Kheyroddin, H. Naderpour, M. Ahmadi, Compressive strength of confined concrete in CCFST columns, J. Rehabil. Civ. Eng. (2014).

[11] H. Alabduljabbar, M. Khan, H.H. Awan, S.M. Eldin, R. Alyousef, A.M. Mohamed, Predicting ultra-high-performance concrete compressive strength using gene expression programming method, Case Stud. Constr. Mater. 18 (2023) e02074.

[12] M. Azimi-Pour, H. Eskandari-Naddaf, A. Pakzad, Linear and non-linear SVM prediction for fresh properties and compressive strength of high volume fly ash self-compacting concrete, Constr. Build. Mater. 230 (2020) 117021.

[13] M. Khodaparasti, A. Alijamaat, M. Pouraminian, Prediction of the concrete compressive strength using improved random forest algorithm, J. Build. Pathol. Rehabil. 8 (2) (2023) 92.

[14] D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei, Z.-M. Jiang, Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach, Constr. Build. Mater. 230 (2020) 117000.

[15] M. Khan, M. Ali, T. Najeh, Y. Gamil, Computational prediction of workability and mechanical properties of bentonite plastic concrete using multi-expression programming, Sci. Rep. 14 (1) (2024) 6105.

[16] T. Nguyen-Sy, J. Wakim, Q.-D. To, M.-N. Vu, T.-D. Nguyen, T.-T. Nguyen, Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method, Constr. Build. Mater. 260 (2020) 119757.

[17] W. Mahmood, A.S. Mohammed, P. Sihag, P.G. Asteris, H. Ahmed, Interpreting the experimental results of compressive strength of hand-mixed cement-grouted sands using various mathematical approaches, Arch. Civ. Mech. Eng. 22 (1) (2021) 19.
[18] A. Ashrafian, E. Panahi, S. Salehi, M. Karoglou, P.G. Asteris, Mapping the strength of agro-ecological lightweight concrete containing oil palm by-product using artificial intelligence techniques, Structures 48 (2023) 1209–1229, http://dx.doi.org/10.1016/j.istruc.2022.12.108, URL: https://www.sciencedirect.com/science/article/pii/S2352012422012966.
[19] H.A. Al-Jamimi, W.A. Al-Kutti, S. Alwahaishi, K.S. Alotaibi, Prediction of compressive strength in plain and blended cement concretes using a hybrid artificial intelligence model, Case Stud. Constr. Mater. 17 (2022) e01238.
[20] Y. Wu, Y. Zhou, Hybrid machine learning model and Shapley additive explanations for compressive strength of sustainable concrete, Constr. Build. Mater. 330 (2022) 127298.
[21] M. Ahmad, K. Rashid, Z. Tariq, M. Ju, Utilization of a novel artificial intelligence technique (ANFIS) to predict the compressive strength of fly ash-based geopolymer, Constr. Build. Mater. 301 (2021) 124251.
[22] D.-P.N. Kontoni, M. Ahmadi, 8 - practical prediction of ultimate axial strain and peak axial stress of FRP-confined concrete using hybrid ANFIS-PSO models, in: Artificial Intelligence Applications for Sustainable Construction, in: Woodhead Publishing Series in Civil and Structural Engineering, Woodhead Publishing, 2024, pp. 225–255, http://dx.doi.org/10.1016/B978-0-443-13191-2.00015-8, URL: https://www.sciencedirect.com/science/article/pii/B9780443131912000158.
[23] P. Das, A. Kashem, Hybrid machine learning approach to prediction of the compressive and flexural strengths of UHPC and parametric analysis with shapley additive explanations, Case Stud. Constr. Mater. 20 (2024) e02723.
[24] H.U. Ahmed, R.R. Mostafa, A. Mohammed, P. Sihag, A. Qadir, Support vector regression (SVR) and grey wolf optimization (GWO) to predict the compressive strength of GGBFS-based geopolymer concrete, Neural Comput. Appl. 35 (3) (2023) 2909–2926.
[25] M. Shariati, M.S. Mafipour, B. Ghahremani, F. Azarhomayun, M. Ahmadi, N.T. Trung, A. Shariati, A novel hybrid extreme learning machine–grey wolf optimizer (ELM-gwo) model to predict compressive strength of concrete with partial replacements for cement, Eng. Comput. (2022) 1–23.
[26] Q.-F. Li, Z.-M. Song, High-performance concrete strength prediction based on ensemble learning, Constr. Build. Mater. 324 (2022) 126694, http://dx.doi.org/10.1016/j.conbuildmat.2022.126694.
[27] M.I. Khan, Y.M. Abbas, Intelligent data-driven compressive strength prediction and optimization of reactive powder concrete using multiple ensemble-based machine learning approach, Constr. Build. Mater. 404 (2023) 133148, http://dx.doi.org/10.1016/j.conbuildmat.2023.133148.
[28] Q. Li, Z. Song, Prediction of compressive strength of rice husk ash concrete based on stacking ensemble learning model, J. Clean. Prod. 382 (2023) 135279, http://dx.doi.org/10.1016/j.jclepro.2022.135279.
[29] J.-F. Jia, et al., An interpretable ensemble learning method to predict the compressive strength of concrete, Structures 46 (2022) http://dx.doi.org/10.1016/j.istruc.2022.11.090.
[30] W. Mahmood, A.S. Mohammed, P.G. Asteris, R. Kurda, D.J. Armaghani, Modeling flexural and compressive strengths behaviour of cement-grouted sands modified with water reducer polymer, Appl. Sci. 12 (3) (2022) 1016.
[31] J. Vazquez, J.C. Facelli, Conformal prediction in clinical medical sciences, J. Heal. Inform. Res. 6 (3) (2022) 241–252.
[32] J.A. Bastos, Conformal prediction of option prices, Expert Syst. Appl. 245 (2024) 123087.
[33] M.G. de Lomana, et al., ChemBioSim: enhancing conformal prediction of in vivo toxicity by use of predicted bioactivities, J. Chem. Inf. Model. 61 (7) (2021) 3255–3272.
[34] A. Plesner, A. Engsig-Karup, H. True, Detecting railway track irregularities using conformal prediction, in: International Conference on Artificial Neural Networks, Springer Nature Switzerland, Cham, 2024, pp. 295–309.
[35] L. Andéol, et al., Confident object detection via conformal prediction and conformal risk control: an application to railway signaling, in: Conformal and Probabilistic Prediction with Applications, PMLR, 2023.
[36] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, M. Pavone, Sample-efficient safety assurances using conformal prediction, Int. J. Robot. Res. 43 (9) (2024) 1409–1424.
[37] I.-C. Yeh, Concrete compressive strength, 1998, http://dx.doi.org/10.24432/C5PK67, UCI Machine Learning Repository.
[38] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cem. Concr. Res. 28 (12) (1998) 1797–1808, http://dx.doi.org/10.1016/S0008-8846(98)00165-3.
[39] J.-S. Chou, C.-K. Chiu, M. Farfoura, I. Al-Taharwa, Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques, J. Comput. Civ. Eng. 25 (3) (2011) 242–253, http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000088.
[40] A. Mandal, Predicting compressive strength of concrete using advanced machine learning techniques: a combined dataset approach, Asian J. Civ. Eng. 26 (2025) 1225–1241, http://dx.doi.org/10.1007/s42107-024-01247-x.
[41] F. Farooq, W. Ahmed, A. Akbar, F. Aslam, R. Alyousef, Predictive modeling for sustainable high-performance concrete from industrial wastes: A comparison and optimization of models using ensemble learners, J. Clean. Prod. 292 (2021) 126032, http://dx.doi.org/10.1016/j.jclepro.2021.126032.
[42] Y. Zhao, D. Goulias, S. Saremi, Enhancing prediction accuracy of concrete compressive strength using stacking ensemble machine learning, Comput. Concr. 32 (3) (2023) 233.
[43] H.I. Erdal, O. Karakurt, E. Namli, High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform, Eng. Appl. Artif. Intell. 26 (4) (2013) 1246–1254, http://dx.doi.org/10.1016/j.engappai.2012.10.014.
[44] P.G. Asteris, A.D. Skentou, A. Bardhan, P. Samui, K. Pilakoutas, Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models, Cem. Concr. Res. 145 (2021) 106449.
[45] B.V. Varma, E.V. Prasad, S. Singha, Study on predicting compressive strength of concrete using supervised machine learning techniques, Asian J. Civ. Eng. 24 (7) (2023) 2549–2560, http://dx.doi.org/10.1007/s42107-023-00662-w.
[46] D. Li, Z. Tang, Q. Kang, X. Zhang, Y. Li, Machine learning-based method for predicting compressive strength of concrete, Processes 11 (2) (2023) 390, http://dx.doi.org/10.3390/pr11020390.
[47] A.H. Gandomi, A. Faramarzifar, P. Ghanad Rezaee, A. Asghari, S. Talatahari, New design equations for elastic modulus of concrete using multi expression programming, J. Civ. Eng. Manag. 21 (6) (2015) 761–774, http://dx.doi.org/10.3846/13923730.2014.893910.
[48] N. Cangussu, P. Milheiro-Oliveira, A.M. Matos, F. Aslani, L. Maia, Comparison of outlier detection approaches for compressive strength of cement-based mortars, J. Build. Eng. 95 (2024) 110276.
[49] U. Asif, S.A. Memon, Interpretable predictive modeling, sustainability assessment, and cost analysis of cement-based composite containing secondary raw materials, Constr. Build. Mater. 473 (2025) 140924.
[50] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, vol. 29, Springer, 2005.
[51] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (3) (2008).
[52] T. Cordier, V. Blot, L. Lacombe, T. Morzadec, A. Capitaine, N. Brunel, Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library, in: Conformal and Probabilistic Prediction with Applications, 2023.
[53] H. Papadopoulos, K. Proedrou, V. Vovk, A. Gammerman, Inductive confidence machines for regression, in: Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13, Springer, 2002, pp. 345–356.
[54] S. Khaki, D. Nettleton, Conformal prediction intervals for neural networks using cross validation, 2020, arXiv preprint arXiv:2006.16941.
[55] Y. Romano, M. Sesia, E. Candes, Classification with valid and adaptive coverage, Adv. Neural Inf. Process. Syst. 33 (2020) 3581–3591.
[56] R.F. Barber, E.J. Candes, A. Ramdas, R.J. Tibshirani, Predictive inference with the jackknife+, Ann. Statist. 49 (1) (2021) 486–507.
[57] N. Gauraha, L. Carlsson, O. Spjuth, Conformal prediction in learning under privileged information paradigm with applications in drug discovery, in: Conformal and Probabilistic Prediction and Applications, PMLR, 2018, pp. 147–156.
[58] T.S. Breusch, A.R. Pagan, A simple test for heteroscedasticity and random coefficient variation, Econ.: J. Econ. Soc. (1979) 1287–1294.