

Article

TabPFN-Based Prediction of Concrete Compressive Strength

Zhihao Zhao ¹ , Jinjin Wang ^{2,*}, Guohui Ma ¹ and Mingjie Han ³

¹ School of Statistics and Data Science, Capital University of Economics and Business, Beijing 100070, China; zhzhao@cueb.edu.cn (Z.Z.); 22025123075@cueb.edu.cn (G.M.)

² Tongzhou Campus, Renmin University of China, Beijing 100872, China

³ Beijing Building Construction Research Institute Co., Ltd., Beijing 100039, China; hanmingjie-yjzx@jyy.bcegc.com

* Correspondence: wangjinjin02@ruc.edu.cn

Abstract

The use of supplementary cementitious materials such as fly ash can reduce environmental impacts and improve the sustainability of concrete construction. However, the nonlinear interactions among mixture design parameters make accurate prediction of concrete compressive strength challenging. In this study, TabPFN, a pre-trained foundation model for tabular data, was applied to predict the compressive strength of fly ash concrete and compared with tuned Random Forest, support vector regression, an artificial neural network, LightGBM, CatBoost, Ridge regression, and Abrams empirical regression. A dataset containing 1062 samples and eight mixture-level variables was used for model development and evaluation. Predictive performance was assessed using the coefficient of determination, mean absolute error, and root mean square error over 100 repeated random splits. The results showed that TabPFN achieved the best overall performance, with an average coefficient of determination of 0.9329, a mean absolute error of 3.2758 MPa, and a root mean square error of 4.6678 MPa. Compared with the strongest tuned gradient-boosting baseline, CatBoost, TabPFN reduced the mean absolute error and root mean square error by 0.8768 MPa and 0.8560 MPa, respectively. Furthermore, repeated-split conformal prediction demonstrated reliable uncertainty quantification, with an average prediction interval coverage probability of 0.9615 and a mean prediction interval width of 23.4554 MPa. SHAP analysis identified the water-to-cement ratio, mortar strength, and water-to-binder ratio as important variables, while additional multicollinearity and feature ablation analyses indicated that correlated ratio variables should be interpreted cautiously. The results indicate that TabPFN provides an accurate, robust, and uncertainty-aware framework for preliminary prediction of 28-day fly ash concrete compressive strength.

Keywords: fly ash concrete; compressive strength; TabPFN; machine learning; uncertainty quantification; interpretable artificial intelligence



Academic Editor: Hsuan-Teh Hu

Received: 3 June 2026

Revised: 27 June 2026

Accepted: 9 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Concrete is one of the most widely used construction materials in civil engineering, and its compressive strength is a critical parameter, governing structural safety, durability, and overall engineering performance [1]. However, the extensive use of cement in concrete production has raised significant environmental concerns because cement manufacturing contributes approximately 7–8% of global anthropogenic CO₂ emissions [2,3]. To promote sustainable construction and reduce the environmental burden associated with cement production, supplementary cementitious materials such as fly ash (FA) have been widely

incorporated into concrete mixtures [4,5]. The utilization of FA not only reduces cement consumption and greenhouse gas emissions but also improves resource utilization efficiency and supports the development of environmentally friendly construction materials.

Despite its environmental advantages, the incorporation of FA increases the complexity of concrete mixture design because compressive strength is influenced by highly nonlinear interactions among cementitious materials, aggregates, water content, and chemical admixtures [6,7]. Accurate prediction of concrete compressive strength is essential for ensuring structural reliability, optimizing material utilization, reducing construction costs, and improving engineering efficiency [8]. Consequently, the development of reliable prediction approaches for FA concrete compressive strength has become an important research topic in sustainable civil engineering.

Traditional prediction methods mainly rely on empirical equations and statistical regression models [9,10]. Although these approaches provide useful engineering guidance, they often exhibit limited capability in capturing the complex nonlinear relationships among mixture parameters, particularly in FA concrete systems [6,7]. With the rapid development of artificial intelligence (AI), machine learning (ML) techniques have increasingly been applied to concrete strength prediction. Numerous studies have demonstrated that ML models, including artificial neural networks (ANNs), Random Forest (RF), support vector regression (SVR), and gradient-boosting approaches, can achieve significantly improved prediction accuracy compared with conventional empirical models [11–13]. Recent studies have further confirmed the effectiveness of ML-based approaches for predicting the compressive strength of FA concrete under various material compositions and engineering conditions [14–17]. In particular, LightGBM and CatBoost have become strong baselines for structured tabular prediction because they combine nonlinear boosted decision trees with efficient optimization and regularization strategies [18,19].

Although existing ML models have achieved promising results, their predictive performance often depends on extensive hyperparameter tuning and model-specific optimization. Furthermore, the optimal model reported in the literature frequently varies across datasets, indicating that prediction performance is sensitive to data characteristics and experimental settings. These limitations are particularly relevant for engineering datasets, which are commonly characterized by limited sample sizes and complex feature interactions. Therefore, developing prediction frameworks that can achieve high accuracy and strong robustness with minimal tuning requirements remains an important challenge.

Recently, the Tabular Prior-Data Fitted Network (TabPFN) has emerged as a promising foundation model specifically designed for tabular data prediction. Unlike conventional ML algorithms that require dataset-specific training and extensive hyperparameter optimization, TabPFN performs prediction using a transformer-based framework pre-trained on a large number of synthetic tabular tasks. Through prior-data fitting and in-context learning, TabPFN can effectively leverage learned prior knowledge to perform predictions on previously unseen datasets with minimal model configuration. Previous studies have demonstrated that TabPFN achieves competitive performance on small- and medium-sized tabular datasets while substantially reducing the effort associated with model selection and hyperparameter tuning. These characteristics make TabPFN particularly attractive for engineering applications, where data availability is often limited. Nevertheless, its applicability to concrete compressive strength prediction, especially for FA concrete, has not yet been systematically investigated.

Motivated by the above considerations, this study investigates the potential of TabPFN for predicting the compressive strength of FA concrete. To provide a comprehensive evaluation, the predictive performance of TabPFN is systematically compared with tuned conventional ML models, gradient-boosting baselines, and simple empirical baselines.

In addition, conformal prediction is incorporated to quantify predictive uncertainty, while SHAP analysis is employed to improve model interpretability. The main contributions of this study are summarized as follows:

1. TabPFN is introduced and evaluated for FA concrete compressive strength prediction and its performance is systematically compared with tuned RF, SVR, ANN, LightGBM, CatBoost, Ridge, and Abrams baselines.
2. A conformal prediction framework is integrated to assess the reliability and uncertainty characteristics of model predictions.
3. A SHAP-based interpretability analysis is conducted to identify the most influential mixture parameters and reveal their effects on compressive strength prediction.
4. The predictive accuracy, robustness, uncertainty quantification capability, and interpretability of TabPFN are comprehensively assessed using repeated experiments and multiple statistical analyses.

The findings of this study are expected to provide new insights into the application of foundation models for civil engineering material prediction and contribute to the development of reliable and sustainable data-driven approaches for concrete performance assessment.

The remainder of this paper is organized as follows: Section 2 presents the dataset description and statistical characteristics. Section 3 introduces the comparative models used in this study, including the RF, SVR, ANN, LightGBM, CatBoost, Ridge regression, and Abrams regression empirical baselines. Section 4 describes the TabPFN methodology. Section 5 presents the experimental design, including data preprocessing, hyperparameter optimization, evaluation metrics, conformal prediction for uncertainty quantification, and SHAP-based model interpretation. Section 6 discusses the prediction results, statistical analysis, residual analysis, uncertainty quantification performance, and interpretability analysis. Finally, Section 7 summarizes the main conclusions of this study. Additional robustness and sensitivity analyses are provided in the Supplementary Materials.

2. Data Description

In this study, the FA concrete dataset reported by Wang et al. [20] is adopted to develop and evaluate the compressive strength prediction models. The dataset contains 1062 valid samples, each associated with an experimentally measured 28-day compressive strength value. The relatively large sample size provides sufficient data coverage for model development and performance evaluation, thereby enhancing the reliability of the statistical analyses and model comparisons.

For each sample, eight input variables describing the concrete mixture composition and material properties are collected: water content, cement content, admixture content, cement mortar strength (CMS), binder content, water-to-cement ratio (W/C), water-to-binder ratio (W/B), and aggregate-to-binder ratio (A/B). Binder content refers to the total cementitious materials in the mixture, consisting of cement and fly ash. The target variable is the 28-day compressive strength of the concrete. All variables are complete and consistently defined across the dataset, ensuring data integrity and minimizing uncertainties associated with missing or inconsistent measurements.

The variables used in this study and their corresponding units are summarized in Table 1, while the descriptive statistics of the input and output variables are presented in Table 2.

The compressive strength values in the dataset range from 14.6 MPa to 115.0 MPa, with most samples concentrated between approximately 30 MPa and 60 MPa. This indicates that the dataset covers both the normal-strength and the relatively high-strength concrete commonly encountered in engineering practice. In addition, the dataset contains a wide

range of mixture-proportion conditions, which is beneficial for evaluating model robustness and predictive performance across different mixture designs.

Table 1. Input and output variables used in this study.

Notation	Description	Unit
Water	Water content	kg/m ³
Cement	Cement content	kg/m ³
Admixture	Admixture content	kg/m ³
CMS	Cement mortar strength	MPa
Binder	Cement and fly ash content	kg/m ³
W/C	Water-to-cement ratio	–
W/B	Water-to-binder ratio	–
A/B	Aggregate-to-binder ratio	–
Compressive Strength	28-day compressive strength	MPa

Figure 1 illustrates the distribution of the 28-day compressive strength values. Most samples are concentrated in the medium-strength range, whereas relatively fewer samples are located in the high-strength region. The overall distribution exhibits moderate dispersion without obvious extreme clustering, which is favorable for machine learning model training and performance evaluation.

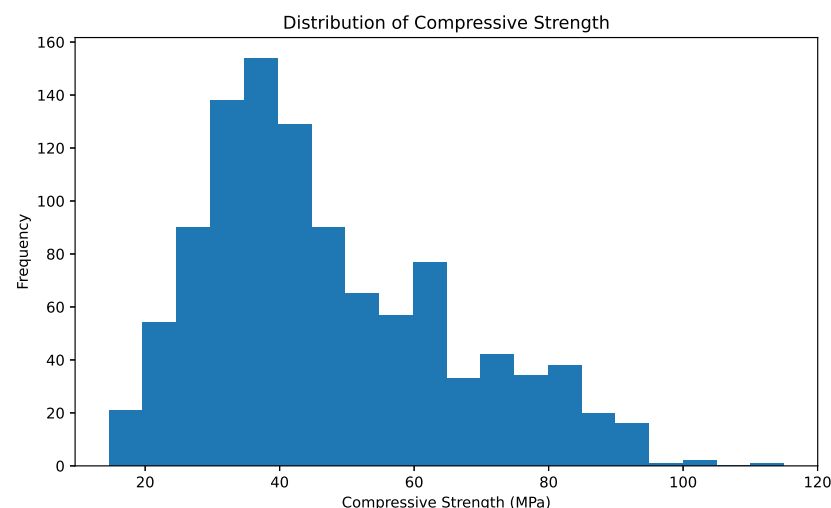


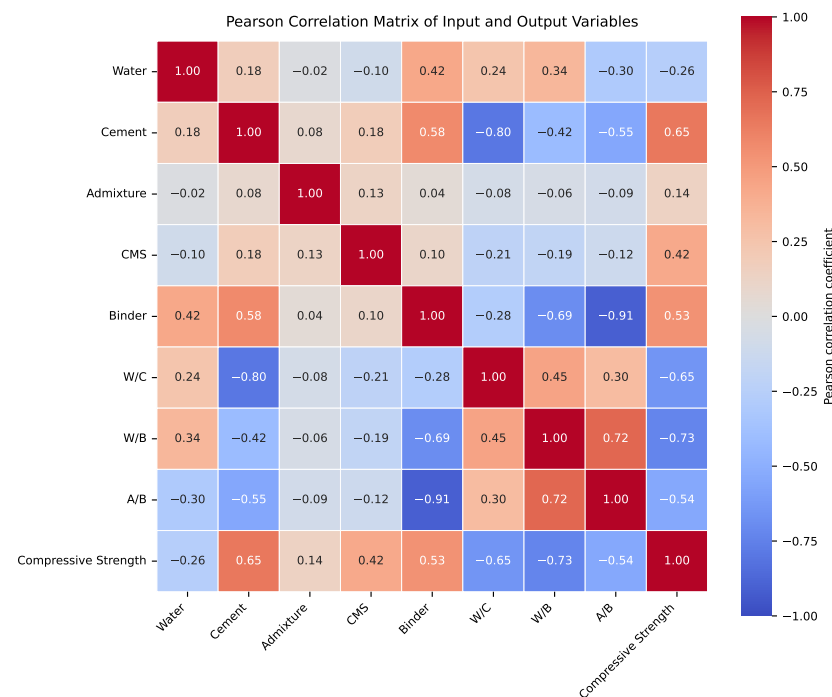
Figure 1. Distribution of 28-day compressive strength values.

To further characterize the statistical properties of the dataset, descriptive statistical analysis is performed for all input variables and the target variable. The corresponding descriptive statistics, including the minimum, maximum, mean, standard deviation, and median values, are summarized in Table 2. The results reveal substantial variability among the mixture parameters, indicating the presence of complex nonlinear relationships between mixture composition and compressive strength. Such characteristics provide a suitable basis for evaluating advanced machine learning models designed to capture nonlinear feature interactions.

To investigate the relationships among the variables, Pearson correlation analysis is conducted, and the corresponding heatmap is presented in Figure 2. The results show that ratio-based variables, particularly the W/C and W/B, exhibit relatively strong negative correlations with compressive strength, which is consistent with established concrete material theories. In contrast, cement mortar strength demonstrates a positive correlation with compressive strength, indicating its contribution to matrix strength development.

Table 2. Descriptive statistics of input and output variables.

Variable	Min	Max	Mean	Std	Median
Water (kg/m ³)	92.81	450.00	171.20	34.02	168.00
Cement (kg/m ³)	60.00	700.00	335.44	113.79	322.00
Admixture (kg/m ³)	0.00	106.00	4.61	9.39	2.08
CMS (MPa)	28.40	62.50	49.35	5.58	49.30
Binder (kg/m ³)	213.88	1000.00	457.57	109.55	444.00
W/C	0.20	2.25	0.58	0.24	0.53
W/B	0.20	0.63	0.39	0.09	0.38
A/B	0.93	8.02	4.11	1.14	3.99
Compressive Strength (MPa)	14.60	115.00	47.11	18.38	42.08

**Figure 2.** Pearson correlation heatmap of input variables and compressive strength.

The selection of input variables is guided by classical concrete strength theories, including Feret's formula [21] and the *Specification for Mix Proportion Design of Ordinary Concrete* (JGJ 55-2011) [22], as well as the underlying mechanisms governing concrete strength formation. From a physical perspective, compressive strength is strongly influenced by the compactness of the cementitious matrix and the load-bearing contribution of the aggregate skeleton. The *W/C* and *W/B* are key parameters controlling hydration degree, pore structure development, and matrix density, and they are generally negatively correlated with compressive strength.

Before model development, the numeric variables are checked for missing values and invalid records. Model-specific preprocessing is then performed within each training split to avoid information leakage. Tree-based models and TabPFN are fitted using the original numeric features, whereas SVR and the ANN use training set-only Z-score standardization; SVR further applies target scaling through a transformed-target regression pipeline.

3. Benchmark Models

To comprehensively evaluate the predictive performance of TabPFN for concrete compressive strength prediction, five machine learning models, namely, RF, SVR, an ANN,

LightGBM, and CatBoost, are selected as benchmark models for comparative analysis. RF, SVR, and ANNs are widely used conventional models in civil engineering material prediction, whereas LightGBM and CatBoost are strong gradient-boosting decision tree baselines for tabular regression. In addition, simple Ridge and Abrams regressions are included as empirical reference baselines in the experimental comparison.

3.1. RF

RF is an ensemble learning algorithm proposed by Breiman [23] which combines multiple decision trees to improve prediction accuracy and reduce overfitting. RF constructs numerous regression trees using bootstrap sampling and random feature selection, and the final prediction result is obtained by averaging the outputs of all decision trees.

For a given input sample, the prediction output of RF can be expressed as

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x) \quad (1)$$

where $\hat{y}$ denotes the predicted compressive strength, N represents the number of regression trees, and $T_i(x)$ denotes the prediction result generated by the i -th regression tree.

During the tree construction process, the optimal node splitting criterion is determined by minimizing the mean squared error (MSE), which can be expressed as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2)$$

where y_i and $\hat{y}_i$ represent the experimental and predicted values, respectively.

The ensemble averaging mechanism enables RF to effectively reduce prediction variance and improve model robustness. In addition, RF can capture complex nonlinear relationships and feature interactions among concrete mixture variables without requiring explicit assumptions regarding data distribution. Owing to these characteristics, RF has been widely used for compressive strength prediction in civil engineering materials. Previous studies have demonstrated that RF can effectively model the nonlinear relationships among mixture parameters and achieve reliable prediction accuracy for concrete compressive strength prediction [8,13,15].

3.2. SVR

SVR is a regression-based extension of a Support Vector Machine (SVM) developed by Vapnik [24]. SVR aims to establish an optimal regression hyperplane while minimizing structural risk and maintaining prediction errors within a predefined tolerance range.

For nonlinear regression problems, the original input variables are mapped into a higher-dimensional feature space through a kernel function. The regression function of SVR can be expressed as

$$f(x) = w^T \phi(x) + b \quad (3)$$

where $\phi(x)$ represents the nonlinear mapping function, w denotes the weight vector, and b is the bias term.

To improve nonlinear fitting capability, the radial basis function (RBF) kernel is adopted in this study. The RBF kernel function is expressed as

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (4)$$

where γ is the kernel parameter controlling the influence range of training samples.

The optimization objective of SVR can be formulated as

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (5)$$

where C denotes the penalty coefficient, and ξ_i and ξ_i^* are slack variables representing prediction deviations outside the tolerance interval.

In addition, the ε -insensitive loss function adopted in SVR can be expressed as

$$L_\varepsilon(y, f(x)) = \max(0, |y - f(x)| - \varepsilon) \quad (6)$$

where ε defines the insensitive error tolerance region.

SVR has demonstrated strong nonlinear regression capability and good generalization performance for engineering prediction problems, particularly under limited sample conditions. Numerous studies have successfully applied SVR to predict concrete compressive strength and other mechanical properties of cementitious materials [12,14,16]. These studies indicate that SVR can effectively capture nonlinear relationships among mixture variables and provide stable prediction performance for engineering datasets.

3.3. ANN

An ANN is a data-driven machine learning approach inspired by biological neural systems [25]. An ANN consists of interconnected neurons organized into multiple layers, including the input layer, hidden layers, and output layer.

In this study, a feedforward neural network architecture is adopted for compressive strength prediction. During the forward propagation process, the hidden neurons receive weighted input signals and subsequently perform nonlinear transformations using activation functions. The propagation process of neurons can be expressed as

$$a_j^{(l)} = f \left(\sum_i w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)} \right) \quad (7)$$

where $a_i^{(l-1)}$ and $a_j^{(l)}$ denote the neuron outputs of the previous and current layers, respectively, $w_{ij}^{(l)}$ represents the connection weight, $b_j^{(l)}$ is the bias term, and $f(\cdot)$ denotes the activation function.

During the training process, the network parameters are iteratively optimized using backpropagation and gradient-based optimization algorithms to minimize prediction errors. The loss function used in ANN training can be expressed as

$$L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

where y_i and $\hat{y}_i$ denote the experimentally measured and predicted compressive strength values, respectively.

Because of its multilayer nonlinear structure, an ANN can effectively learn complex nonlinear relationships and hidden feature interactions within concrete mixture datasets. Consequently, ANNs have been widely applied in compressive strength prediction and other civil engineering material prediction tasks. Previous studies have shown that ANNs can effectively model the complex interactions among concrete mixture parameters and provide accurate prediction results for concrete compressive strength [7,10,11]. In addition, ANNs have demonstrated strong adaptability and nonlinear approximation capability for complex engineering material datasets.

3.4. LightGBM

LightGBM is a gradient-boosting decision tree algorithm designed for efficient and accurate learning on structured tabular data [18]. Similar to other boosting methods, LightGBM constructs an additive ensemble of regression trees where each new tree is fitted to reduce the residual errors of the previous ensemble. The prediction function can be expressed as

$$\hat{y} = \sum_{m=1}^M f_m(x), \quad (9)$$

where $f_m(x)$ denotes the m -th regression tree and M is the total number of boosting iterations.

LightGBM improves computational efficiency through histogram-based tree construction and leaf-wise tree growth with depth constraints. These characteristics allow it to capture nonlinear feature interactions while maintaining relatively low computational cost. Because concrete mixture datasets are typically structured, nonlinear, and moderate in size, LightGBM is a suitable strong baseline for evaluating whether the performance of TabPFN exceeds that of modern boosted-tree methods.

3.5. CatBoost

CatBoost is another gradient-boosting decision tree method that introduces ordered boosting and symmetric tree structures to reduce prediction shift and improve generalization [19]. Although CatBoost was originally developed with strong support for categorical features, it is also effective for numerical tabular regression tasks because of its robust regularization and boosting strategy.

In this study, CatBoost is used as an additional strong tabular baseline. Its objective is to minimize the regression loss over an ensemble of boosted trees:

$$\mathcal{L} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \Omega(F), \quad (10)$$

where $\Omega(F)$ denotes the regularization term associated with the ensemble. Including CatBoost and LightGBM helps ensure that the comparison does not rely only on conventional default baselines, but also considers competitive modern gradient-boosting models commonly used for tabular prediction.

3.6. Empirical Regression Baselines

To provide engineering reference baselines in addition to machine learning models, Ridge regression and an Abrams' law-inspired regression are also included in the comparison. Abrams' law is a classical empirical relationship in concrete technology which states that compressive strength is strongly governed by the water-to-cement ratio [9]. In this study, the Abrams' law-inspired baseline is fitted as a log-linear regression using W/C :

$$\log(f_c) = a + b \log(W/C), \quad (11)$$

where f_c denotes the 28-day compressive strength, and a and b are regression coefficients estimated from the training data. This baseline is not intended to compete with modern nonlinear models; instead, it provides a simple physically motivated reference for evaluating how much predictive improvement is obtained by using multivariable machine learning methods.

4. TabPFN Methodology

TabPFN is a transformer-based tabular foundation model specifically developed for small- and medium-sized tabular datasets [26]. Unlike conventional machine learning

models that require iterative optimization for each downstream dataset, TabPFN adopts an in-context learning paradigm in which the model is pre-trained on a large number of synthetic tabular datasets and subsequently performs prediction through a single forward inference process. The overall framework of TabPFN mainly consists of synthetic data generation, transformer-based pre-training, and downstream inference stages.

The core idea of TabPFN is to learn generalized prediction strategies from diverse synthetic tabular tasks. During the pre-training stage, millions of synthetic datasets containing different nonlinear relationships, feature distributions, noise conditions, and structural dependencies are generated using structural causal models (SCMs) [26]. These synthetic datasets are designed to mimic the heterogeneous characteristics commonly observed in real-world tabular data, enabling the model to learn generalized feature interaction patterns and prediction mechanisms.

Unlike traditional deep learning models trained on individual samples, TabPFN is trained across entire datasets. During inference, both training samples and testing samples are simultaneously provided to the pre-trained model, allowing prediction to be completed in a single forward pass through in-context learning. The prediction objective of TabPFN can be expressed as

$$p(\hat{y}_{test} \mid X_{train}, y_{train}, X_{test}) \quad (12)$$

where X_{train} and y_{train} denote the training features and labels, respectively, and X_{test} represents the testing features. The model directly estimates the posterior predictive distribution of the target variable conditioned on the observed training samples.

The neural architecture of TabPFN is based on a transformer encoder specifically adapted for tabular data. Different from standard transformers designed for sequential data, TabPFN treats each table cell as an individual representation and employs two-dimensional attention mechanisms to simultaneously model feature-wise and sample-wise dependencies. The inter-feature attention learns relationships among variables within the same sample, while the inter-sample attention captures dependencies among different samples. This mechanism enables TabPFN to effectively model the complex nonlinear feature interactions commonly observed in concrete mixture datasets.

The attention mechanism adopted in TabPFN can be generally expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

where Q , K , and V denote the query, key, and value matrices, respectively, and d_k represents the feature dimension.

Compared with conventional machine learning methods, TabPFN possesses several advantages for engineering tabular prediction tasks: First, the model can effectively capture highly nonlinear relationships and feature interactions without extensive feature engineering. Second, the pre-trained inference mechanism significantly reduces training time and hyperparameter tuning requirements. Third, TabPFN demonstrates strong predictive performance for small- and medium-sized datasets, which are common in civil engineering applications. According to the original study, TabPFN achieves state-of-the-art prediction performance on datasets with up to 10,000 samples and 500 features [26].

Figure 3 illustrates the overall workflow of the proposed TabPFN framework for concrete compressive strength prediction. The framework mainly consists of two stages, namely, the offline pre-training stage and the online inference stage.

During the offline pre-training stage, TabPFN is trained on a large number of synthetic tabular datasets containing diverse feature distributions, nonlinear relationships, missing values, and noise patterns. Through transformer-based attention mechanisms and in-context learning, the model learns generalized tabular prediction knowledge from large-scale synthetic tasks.

During the online inference stage, the concrete dataset is directly provided to the pre-trained TabPFN model without extensive hyperparameter optimization. The model jointly learns feature interactions and sample relationships from the contextual training samples and subsequently predicts the compressive strength of test samples.

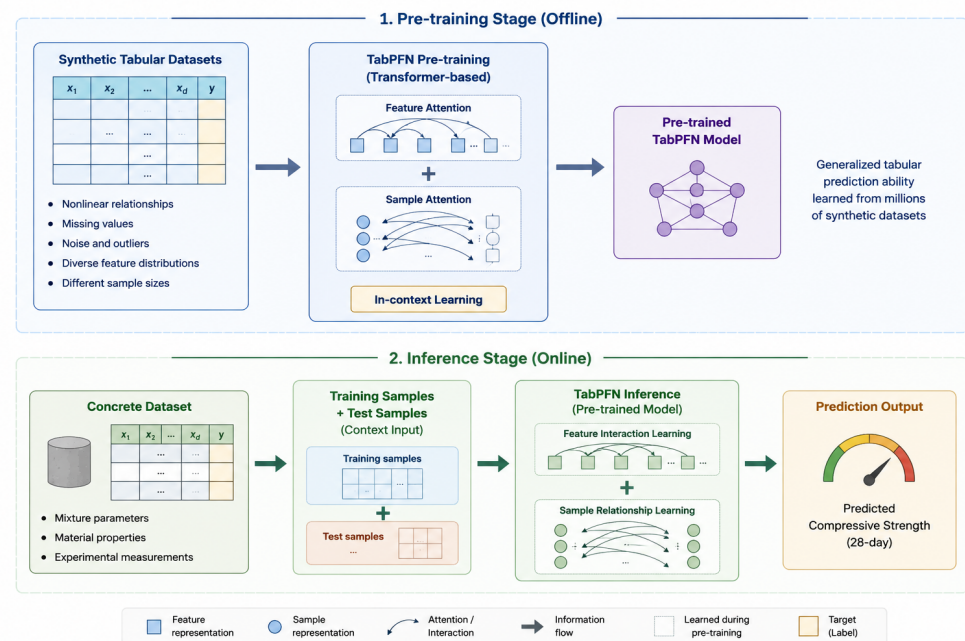


Figure 3. Overall workflow of the proposed TabPFN framework for concrete compressive strength prediction, including the offline pre-training stage and online inference stage.

In the present study, the concrete compressive strength dataset contains complex nonlinear relationships among mixture parameters, including the water-to-cement ratio, water-to-binder ratio, admixture dosage, and cementitious material content. Moreover, the dataset size falls within the typical application range where TabPFN has demonstrated strong predictive capability and computational efficiency. Therefore, TabPFN is considered suitable for predicting concrete compressive strength in the present work.

To evaluate the predictive capability of TabPFN, its prediction performance is compared with tuned machine learning benchmark models and empirical regression baselines using identical repeated data splits.

5. Experimental Design

5.1. Data Preprocessing and Experimental Setup

The original dataset is first preprocessed to improve data quality and numerical stability. Since the dataset does not contain missing values after removing non-numeric header rows, no missing-value imputation procedure is required. To avoid data leakage, all preprocessing parameters and hyperparameters are estimated only from the corresponding training subset in each repeated split. For SVR and the ANN, input variables are standardized using Z-score normalization fitted on the training data:

$$x' = \frac{x - \mu}{\sigma} \quad (14)$$

where x and x' denote the original and standardized feature values, respectively, and μ and σ represent the mean value and standard deviation of each feature.

For point prediction evaluation, the dataset is randomly divided into training and testing subsets using an 80%/20% split ratio. To reduce the influence of random data

partitioning and obtain reliable performance estimates, the random splitting procedure is repeated 100 times using random seeds 0–99. In each run, model fitting and hyperparameter optimization are performed on the training subset only, whereas model evaluation is conducted on the corresponding testing subset. The final prediction performance is reported in terms of the mean and standard deviation of the evaluation metrics across all repeated runs.

For uncertainty quantification using conformal prediction, the dataset is divided into a training subset (636 samples), a calibration subset (213 samples), and a testing subset (213 samples), corresponding to approximately 60%, 20%, and 20% of the entire dataset, respectively. The training subset is used to fit the prediction model, the calibration subset is employed to estimate residual quantiles for prediction interval construction, and the testing subset is used for final interval evaluation. This split conformal procedure is repeated over 100 random seeds using strictly non-overlapping training, calibration, and test subsets. A confidence level of 95% is adopted throughout the conformal prediction analysis.

The TabPFN model is implemented using the official ‘tabpfn’ Python package (version 8.0.1) and ‘TabPFNRegressor’. The model is run on a CPU using the official pre-trained TabPFN v3 regressor checkpoint with default inference settings, including $n_estimators = 8$, ‘model_path=auto’, ‘fit_mode=fit_preprocessors’, and ‘inference_precision=auto’. No built-in uncertainty output from TabPFN is used; all uncertainty intervals are produced externally through split conformal prediction. RF, SVR, the ANN, and Ridge and Abrams regressions are implemented using Scikit-learn, LightGBM using the official LightGBM package, and CatBoost using the official CatBoost package. To ensure fair comparison, tunable machine learning benchmark models are optimized by randomized hyperparameter search with 3-fold cross-validation inside the training subset, following common optimization practice in engineering ML studies [27], whereas Ridge and Abrams regressions are included as empirical reference baselines. The conformal dataset partition and hyperparameter search spaces adopted in this study are summarized in Tables 3 and 4, respectively.

All experiments are conducted in the Python environment using the TabPFN, Scikit-learn, LightGBM, CatBoost, SHAP, NumPy, Pandas, and Matplotlib libraries.

Table 3. Dataset partitioning used for conformal prediction.

Subset	Samples	Proportion (%)
Training	636	59.9
Calibration	213	20.1
Testing	213	20.1
Total	1062	100.0

Table 4. Hyperparameter search spaces of the investigated models.

Model	Hyperparameter Settings or Search Space
TabPFN	Official pre-trained TabPFN v3 regressor checkpoint; CPU; $n_estimators = 8$; default inference configuration
CatBoost	iterations $\in \{200, 300, 500\}$; learning rate $\in \{0.03, 0.05, 0.1\}$; depth $\in \{4, 6, 8\}$; l2_leaf_reg $\in \{1, 3, 5, 10\}$; bagging_temperature $\in \{0, 0.5\}$; random_strength $\in \{0, 0.5\}$; $n_estimators \in \{100, 200, 400, 700\}$; learning rate $\in \{0.01, 0.03, 0.05, 0.1\}$;
LightGBM	num_leaves $\in \{15, 31, 63\}$; max_depth $\in \{-1, 3, 5, 8\}$; min_child_samples $\in \{5, 10, 20, 40\}$
RF	$n_estimators \in \{200, 400, 600\}$; max_depth $\in \{\text{None}, 8, 12, 20\}$; min_samples_split $\in \{2, 5, 10\}$; min_samples_leaf $\in \{1, 2, 4\}$; max_features $\in \{0.5, 0.75, 1.0, \sqrt{p}\}$
ANN	hidden layers $\in \{(64), (128), (64, 32), (100, 50), (128, 64)\}$; $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$; learning rate $\in \{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$; batch size $\in \{32, 64, 128, \text{auto}\}$; early stopping enabled

Table 4. Cont.

Model	Hyperparameter Settings or Search Space
SVR	RBF kernel; $C \in \{0.3, 1, 3, 10, 30, 100, 300\}$; $\epsilon \in \{0.01, 0.03, 0.05, 0.1, 0.2, 0.5\}$; $\gamma \in \{\text{scale}, 0.01, 0.03, 0.1, 0.3, 1.0\}$
Ridge	RidgeCV with $\alpha \in \{10^{-4}, 10^{-3}, \dots, 10^3, 10^4\}$ selected by cross-validation within the training subset
Abrams	Log-linear empirical regression using W/C ; no tunable hyperparameters

5.2. Evaluation Metrics

To comprehensively evaluate prediction performance, three widely used statistical indicators, namely, the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE), are adopted in this study.

The coefficient of determination (R^2) measures the goodness of fit between predicted and experimental values and can be expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (15)$$

where y_i , $\hat{y}_i$, and $\bar{y}$ denote the experimental value, predicted value, and mean experimental value, respectively.

The MAE evaluates the average absolute prediction error and can be expressed as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (16)$$

where a smaller MAE value indicates higher prediction accuracy.

The RMSE measures the square root of the average squared prediction error and is more sensitive to large deviations. The RMSE can be expressed as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (17)$$

In general, larger R^2 values and smaller MAE and RMSE values indicate better predictive performance and stronger generalization capability of the developed models.

5.3. Conformal Prediction for Uncertainty Quantification

To further evaluate prediction reliability and uncertainty characteristics, conformal prediction (CP) is incorporated into the present study. Unlike conventional point prediction methods that only provide deterministic outputs, CP generates statistically valid prediction intervals without requiring strong assumptions regarding the underlying probability distribution of the data.

In this study, prediction intervals are constructed based on calibration residuals obtained from the calibration dataset. For a given confidence level $(1 - \alpha)$, the prediction interval can be expressed as

$$[\hat{y} - q_\alpha, \hat{y} + q_\alpha] \quad (18)$$

where $\hat{y}$ denotes the predicted compressive strength and q_α represents the empirical quantile of calibration residuals corresponding to the selected confidence level.

To quantitatively evaluate interval reliability and interval compactness, the prediction interval coverage probability (PICP) and mean prediction interval width (MPIW) are adopted in this study.

PICP measures the proportion of experimental values that fall within the generated prediction intervals and can be expressed as

$$\text{PICP} = \frac{1}{N} \sum_{i=1}^N c_i \quad (19)$$

where N denotes the total number of testing samples, and c_i is a binary indicator defined as

$$c_i = \begin{cases} 1, & y_i \in [L_i, U_i] \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

where y_i denotes the experimentally measured compressive strength value, and L_i and U_i represent the lower and upper bounds of the prediction interval, respectively.

MPIW evaluates the average width of prediction intervals and can be expressed as

$$\text{MPIW} = \frac{1}{N} \sum_{i=1}^N (U_i - L_i) \quad (21)$$

A reliable uncertainty quantification framework should achieve PICP values close to the nominal confidence level while maintaining relatively narrow MPIW values. PICP values below the nominal level indicate undercoverage, whereas values far above the nominal level may indicate overly conservative intervals. Smaller MPIW values imply more informative and compact prediction intervals when coverage is close to the target level. Therefore, the combination of PICP and MPIW enables comprehensive evaluation of both the statistical validity and practical usefulness of the generated uncertainty intervals.

5.4. SHAP-Based Model Interpretation

To investigate the influence of mixture variables on compressive strength prediction and improve model interpretability, SHAP (SHapley Additive exPlanations) analysis is conducted in this study.

SHAP is a game theory-based interpretation method that quantifies the contribution of each input variable to the prediction result. The prediction output can be expressed as

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (22)$$

where ϕ_0 denotes the base prediction value and ϕ_i represents the contribution of the i -th feature.

In this study, SHAP summary plots and feature importance rankings are used to analyze the influence of concrete mixture parameters on compressive strength prediction. Through SHAP analysis, the contributions of key variables such as the water-to-cement ratio, water-to-binder ratio, and cementitious material content can be quantitatively interpreted, thereby improving the transparency and engineering interpretability of the developed prediction framework.

6. Results and Discussion

6.1. Point Prediction Performance Analysis

The point prediction performance of TabPFN, the machine learning benchmark models, and the empirical baselines was comprehensively evaluated based on 100 repeated random splitting experiments. The average values and standard deviations of the R^2 , MAE, and RMSE metrics are summarized in Table 5.

Table 5. Comparison of point prediction performance among different models. Values are reported as mean (standard deviation). Bold values indicate the best predictive performance.

Model	R^2	MAE	RMSE
TabPFN	0.9329 (0.0116)	3.2758 (0.2097)	4.6678 (0.3667)
CatBoost	0.9061 (0.0142)	4.1526 (0.2919)	5.5238 (0.4003)
LightGBM	0.9011 (0.0153)	4.2173 (0.3167)	5.6704 (0.4205)
RF	0.8818 (0.0179)	4.6997 (0.3000)	6.1984 (0.4283)
ANN	0.8509 (0.0263)	5.3065 (0.4404)	6.9537 (0.5629)
SVR	0.8334 (0.0252)	5.4890 (0.4143)	7.3607 (0.5346)
Ridge	0.7635 (0.0212)	7.0004 (0.2956)	8.7855 (0.3531)
Abrams	0.4866 (0.0306)	10.1365 (0.5425)	12.9654 (0.6758)

As shown in Table 5, TabPFN achieved the best prediction performance among all investigated models in terms of R^2 , MAE, and RMSE. The mean R^2 value of TabPFN reached 0.9329, which was higher than those of the tuned CatBoost (0.9061), LightGBM (0.9011), RF (0.8818), ANN (0.8509), SVR (0.8334), Ridge regression (0.7635), and Abrams regression (0.4866). Meanwhile, the MAE and RMSE values of TabPFN were 3.2758 MPa and 4.6678 MPa, respectively, both of which were lower than those of the machine learning benchmark models and empirical baselines.

Among the non-TabPFN baselines, CatBoost and LightGBM achieved the strongest performance, confirming that gradient-boosting decision tree models are highly competitive for the present tabular dataset. However, TabPFN still improved R^2 by 0.0267 relative to CatBoost and reduced MAE and RMSE by 0.8768 MPa and 0.8560 MPa, respectively. Compared with the simple Ridge and Abrams regressions, all nonlinear machine learning models performed substantially better, indicating that nonlinear interactions among mixture variables are important for accurate strength prediction.

From the perspective of prediction stability, TabPFN also demonstrated small performance fluctuations during repeated random experiments. The standard deviation of R^2 for TabPFN was 0.0116, which was comparable to or lower than those of the tuned machine learning benchmark models. These results suggest that TabPFN retains strong prediction accuracy and stability even when compared with optimized modern tabular baselines and simple empirical baselines instead of only default conventional models.

Figure 4 presents the boxplots of prediction performance for different models. The R^2 distribution of TabPFN is concentrated within a higher range, whereas its MAE and RMSE distributions are concentrated within lower ranges. In addition, the relatively compact box and whisker distributions indicate that TabPFN maintained stable prediction performance across repeated experiments.

Figure 5 further illustrates the average prediction performance and corresponding uncertainty ranges of different models using error bar plots. TabPFN achieved the highest mean R^2 and the lowest MAE and RMSE values, as shown by relatively short error bars, indicating lower uncertainty and stronger robustness.

Compared with the strongest non-TabPFN baseline, CatBoost, the MAE and RMSE values of TabPFN were reduced by approximately 0.8768 MPa and 0.8560 MPa, respectively. Since RMSE is more sensitive to large prediction deviations, the lower RMSE value of TabPFN indicates its stronger capability in controlling large prediction errors. The performance gap is smaller than that observed against untuned conventional baselines, but the repeated-split results still support the practical advantage of TabPFN for this dataset.

The superior performance of TabPFN can be attributed to its pre-trained tabular learning mechanism. Through large-scale synthetic dataset pre-training, TabPFN learns generalized inductive biases for tabular prediction tasks and can effectively capture nonlin-

ear feature interactions under limited sample conditions. Concrete compressive strength is jointly influenced by multiple coupled variables, including the water-to-cement ratio, water-to-binder ratio, cementitious material system, admixture dosage, and matrix strength. Under limited dataset conditions, conventional machine learning models are more sensitive to data partitioning and hyperparameter selection, whereas the pre-trained prior knowledge of TabPFN helps reduce modeling variance and improve generalization capability. Therefore, the point prediction results demonstrate that TabPFN possesses superior prediction accuracy and stability for concrete compressive strength prediction.

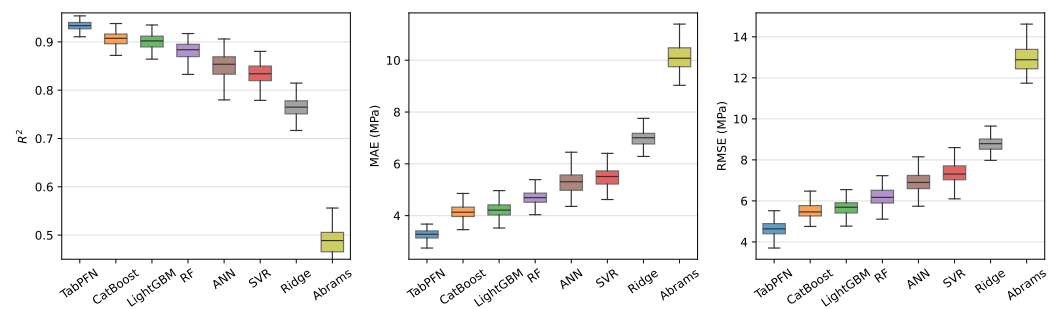


Figure 4. Boxplots of prediction performance for different models.

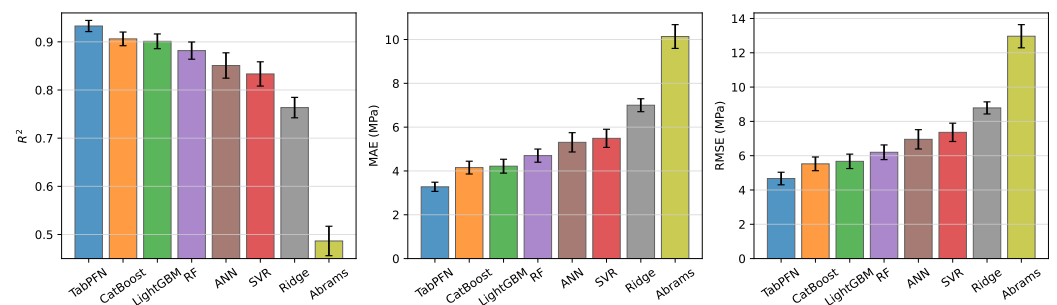


Figure 5. Error bar plots of prediction performance for different models.

6.2. Statistical Significance Analysis

To further verify the significance of prediction performance differences among models, Wilcoxon signed-rank tests were conducted between TabPFN and each comparison model, including both machine learning benchmarks and empirical baselines, based on the paired repeated-split results. The Wilcoxon test was adopted as the primary statistical test because it does not require a normality assumption for paired performance differences. The statistical significance analysis results are summarized in Table 6.

As shown in Table 6, all Wilcoxon p -values and Holm-adjusted p -values were substantially smaller than 0.001 for all evaluation metrics and comparison models. These results indicate that the prediction performance improvements achieved by TabPFN are statistically significant rather than caused by random fluctuations or occasional data partitions.

The significance analysis further strengthens the reliability of the comparative results. Although average evaluation metrics can reflect overall model performance, they may still be influenced by individual data splits or outlier samples. In contrast, paired nonparametric testing based on repeated experiments can determine whether performance differences consistently exist from a probabilistic perspective. In the present study, TabPFN achieved statistically significant improvements over conventional ML models, strong gradient-boosting baselines, Ridge regression, and the Abrams empirical baseline for all evaluation metrics, providing statistical support for the effectiveness of the proposed prediction framework.

Table 6. Statistical significance analysis of TabPFN and comparison models. Difference represents the performance difference calculated as TabPFN minus the corresponding comparison model. Holm-adjusted p -values were calculated across all 21 paired comparisons.

Metric	Compared Model	Mean Difference	Wilcoxon p	Holm-Adjusted p
R^2	CatBoost	+0.0267	4.67×10^{-18}	8.18×10^{-17}
	LightGBM	+0.0318	4.02×10^{-18}	8.18×10^{-17}
	RF	+0.0511	3.90×10^{-18}	8.18×10^{-17}
	ANN	+0.0820	3.90×10^{-18}	8.18×10^{-17}
	SVR	+0.0995	3.90×10^{-18}	8.18×10^{-17}
	Ridge	+0.1694	3.90×10^{-18}	8.18×10^{-17}
	Abrams	+0.4463	3.90×10^{-18}	8.18×10^{-17}
MAE	CatBoost	−0.8768	3.90×10^{-18}	8.18×10^{-17}
	LightGBM	−0.9415	3.90×10^{-18}	8.18×10^{-17}
	RF	−1.4239	3.90×10^{-18}	8.18×10^{-17}
	ANN	−2.0306	3.90×10^{-18}	8.18×10^{-17}
	SVR	−2.2132	3.90×10^{-18}	8.18×10^{-17}
	Ridge	−3.7246	3.90×10^{-18}	8.18×10^{-17}
	Abrams	−6.8607	3.90×10^{-18}	8.18×10^{-17}
RMSE	CatBoost	−0.8560	4.67×10^{-18}	8.18×10^{-17}
	LightGBM	−1.0026	4.02×10^{-18}	8.18×10^{-17}
	RF	−1.5306	3.90×10^{-18}	8.18×10^{-17}
	ANN	−2.2859	3.90×10^{-18}	8.18×10^{-17}
	SVR	−2.6929	3.90×10^{-18}	8.18×10^{-17}
	Ridge	−4.1177	3.90×10^{-18}	8.18×10^{-17}
	Abrams	−8.2976	3.90×10^{-18}	8.18×10^{-17}

6.3. Scatter Plot and Residual Analysis

To further evaluate the prediction consistency and error distribution characteristics of different models, scatter plots of experimentally measured and predicted compressive strength values are presented in Figure 6, while the corresponding residual distributions are shown in Figure 7.

As shown in Figure 6, the prediction results of TabPFN are distributed more closely around the reference line ($y = x$) than those of the comparison models, indicating superior prediction consistency. CatBoost and LightGBM also have relatively compact scatter patterns and thus were clearly stronger than the conventional RF, ANN, and SVR base-lines, which is consistent with the repeated-split results in Table 5. Ridge and Abrams regressions display larger deviations because of their limited ability to represent nonlinear mixture interactions.

The scatter plots also indicate that prediction errors become more visible in the high-strength range where fewer samples are available. Therefore, an additional supplementary analysis was conducted for selected high-strength samples across their test set appearances in the 100 repeated splits. The corresponding Supplementary Figure S1 and Supplementary Table S1 help verify that the representative split shown in Figure 6 is consistent with the repeated-split behavior.

As shown in Figure 7, the residuals of TabPFN are primarily concentrated around zero without obvious systematic bias, indicating strong generalization capability and stable prediction behavior across different compressive strength ranges. CatBoost and LightGBM also have relatively compact residual distributions, whereas RF, ANN, SVR, Ridge regression, and Abrams regression show larger residual dispersion.

The residual plots further show that the largest deviations tend to occur for high-strength samples. This trend was expected because high-strength mixtures were less frequent in the dataset and often involve more complex mixture-proportion effects.

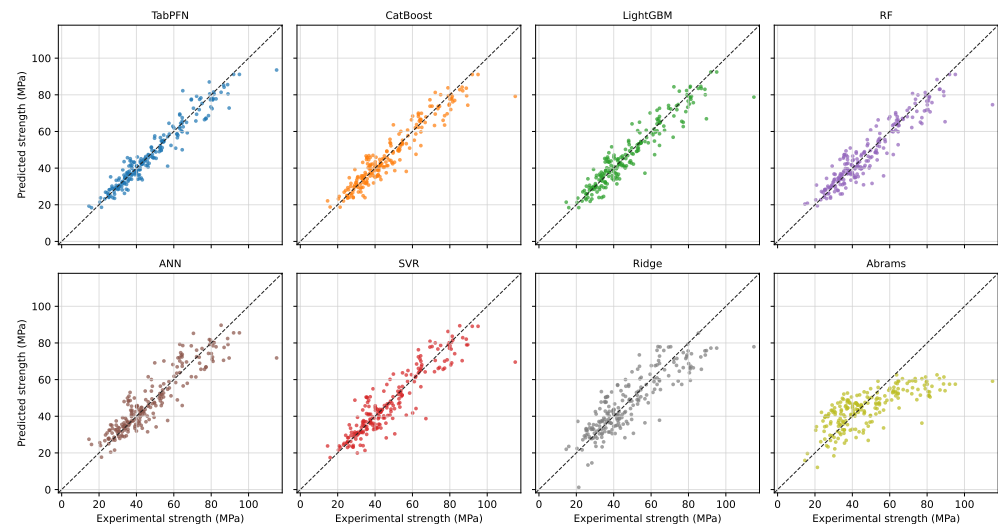


Figure 6. Scatter plots of experimental and predicted compressive strength values for different models in one representative repeated split.

To further investigate prediction errors, residual distributions of different models are illustrated in Figure 7.

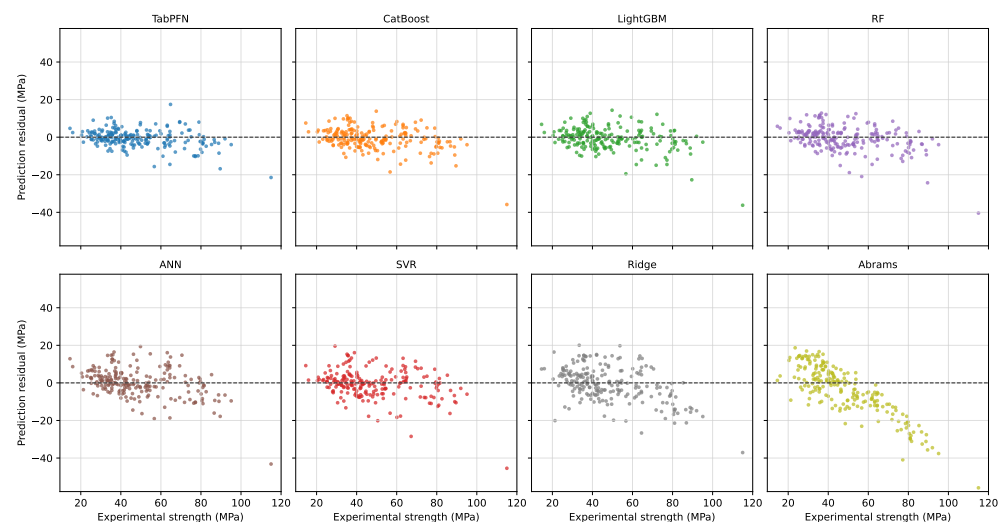


Figure 7. Residual distributions of different prediction models.

The relatively concentrated residual distribution of TabPFN indicates that the model can effectively reduce both underestimation and overestimation errors. Furthermore, no obvious heteroscedasticity trend was observed for TabPFN, suggesting that its prediction uncertainty remained relatively stable across different strength levels. These results further confirm the superior prediction robustness and generalization capability of TabPFN for concrete compressive strength prediction.

6.4. Uncertainty Quantification Using Conformal Prediction

To further evaluate prediction reliability and uncertainty characteristics, conformal prediction (CP) was incorporated into the present study to generate statistically valid prediction intervals for concrete compressive strength prediction. Figure 8 compares the repeated-split conformal prediction performance of different models in terms of PICP and MPIW, while the corresponding numerical uncertainty evaluation metrics are summarized in Table 7.

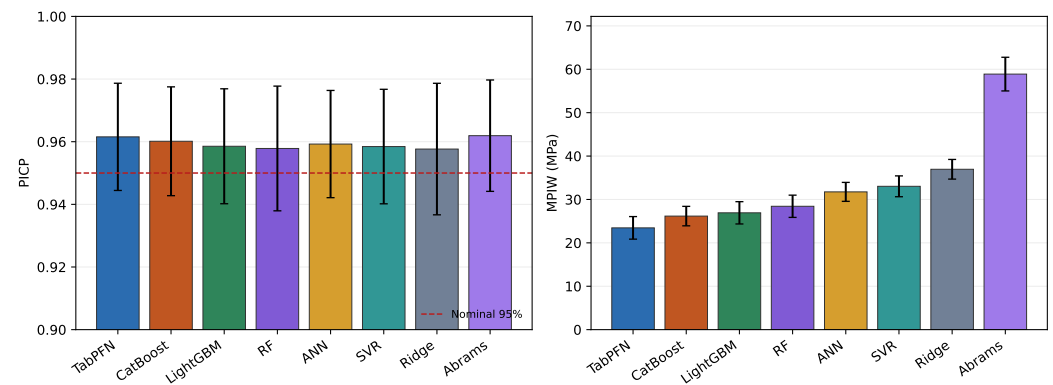


Figure 8. Comparison of repeated-split conformal prediction performance among different models.

Table 7. Comparison of conformal prediction performance among different models over 100 repeated 60/20/20 train/calibration/test splits. Values are reported as mean (standard deviation). PICP values closer to the nominal confidence level (95%) and lower MPIW values indicate better uncertainty quantification performance; bold indicates the narrowest MPIW among models with coverage close to the nominal level.

Model	PICP	MPIW
TabPFN	0.9615 (0.0171)	23.4554 (2.5996)
CatBoost	0.9601 (0.0174)	26.1682 (2.2428)
LightGBM	0.9585 (0.0183)	26.9239 (2.5686)
RF	0.9578 (0.0199)	28.4303 (2.5627)
ANN	0.9592 (0.0171)	31.7489 (2.1837)
SVR	0.9585 (0.0183)	33.0322 (2.4015)
Ridge	0.9577 (0.0210)	36.9571 (2.2665)
Abrams	0.9619 (0.0178)	58.8899 (3.8793)

As shown in Figure 8, most experimental compressive strength values were successfully covered by the generated prediction intervals for all investigated models, indicating satisfactory interval reliability. Nevertheless, noticeable differences were observed in interval width and uncertainty quantification efficiency among the models.

According to Table 7, all models achieved average PICP values close to the nominal confidence level of 95%, demonstrating that the generated prediction intervals provided reliable marginal coverage of the experimental observations. Among them, TabPFN attained an average PICP value of 0.9615, indicating that the prediction intervals successfully covered approximately 96.15% of the testing samples on average and provided slightly conservative but reliable uncertainty estimates.

Although the comparison models also achieved PICP values close to the nominal level, their MPIW values were larger than that of TabPFN. CatBoost, the strongest non-TabPFN conformal baseline in this analysis, obtained an average MPIW of 26.1682 MPa, whereas TabPFN achieved a narrower MPIW of 23.4554 MPa. The intervals generated by SVR, the ANN, Ridge regression, and Abrams regression were substantially wider, indicating that they maintained coverage reliability primarily by producing less informative intervals.

Overall, TabPFN achieved the narrowest prediction intervals while maintaining a PICP value close to the nominal confidence level. These results indicate that TabPFN provided the most efficient uncertainty quantification among the investigated models by achieving a favorable balance between interval reliability and interval compactness. It should be noted that split conformal prediction provides marginal coverage under exchangeability; individual random splits and specific strength ranges may still show coverage fluctuations due to a finite calibration set size. Therefore, the repeated-split results

are reported to characterize empirical stability rather than to imply identical coverage for every single split.

6.5. SHAP-Based Interpretability Analysis

To further investigate the influence of mixture parameters on concrete compressive strength prediction and enhance model interpretability, SHAP analysis was conducted for the developed TabPFN framework. Since TabPFN is a transformer-based foundation model rather than a tree ensemble, TreeSHAP is not applicable. Therefore, a model-agnostic KernelSHAP-style explainer was used with a fixed background subset sampled from the training data to control computational cost. Figure 9 presents the global feature importance ranking based on the mean absolute SHAP values.

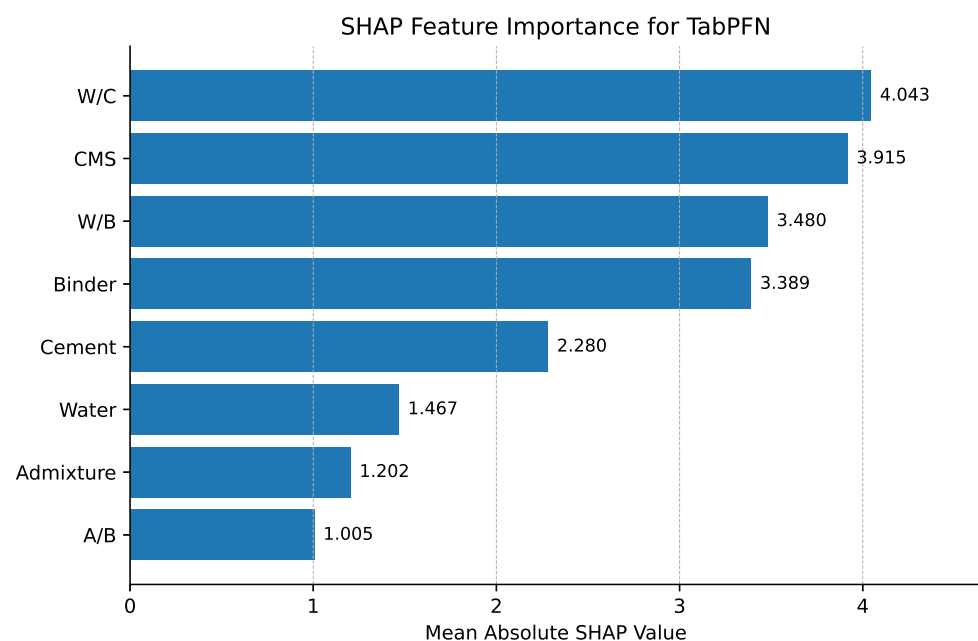


Figure 9. SHAP feature importance ranking.

As shown in Figure 9, W/C was identified as the most influential variable affecting concrete compressive strength prediction, followed by cement mortar strength and W/B. Binder content also exhibited a relatively high contribution, whereas water content, admixture content, and A/B had comparatively lower importance. These findings are consistent with established concrete material theories, in which water-related parameters strongly influence hydration reactions, pore structure development, and, ultimately, compressive strength.

Figure 10 illustrates the SHAP summary plot, which simultaneously reflects the importance, direction, and distribution of feature contributions.

The SHAP summary plot reveals that W/C exhibited the widest SHAP value distribution, indicating the strongest influence on model predictions. Lower W/C values generally contributed positively to compressive strength prediction, whereas higher W/C values tended to produce negative contributions. Similar trends could be observed for W/B. In contrast, higher mortar strength values were predominantly associated with positive SHAP values, suggesting a beneficial effect on compressive strength development. The relatively narrow SHAP distributions observed for admixture content and A/B indicate that their influence on model predictions was comparatively limited.

To further investigate the effects of the most important variables, SHAP dependence plots were generated for W/B, W/C, and mortar strength, as shown in Figures 11–13.

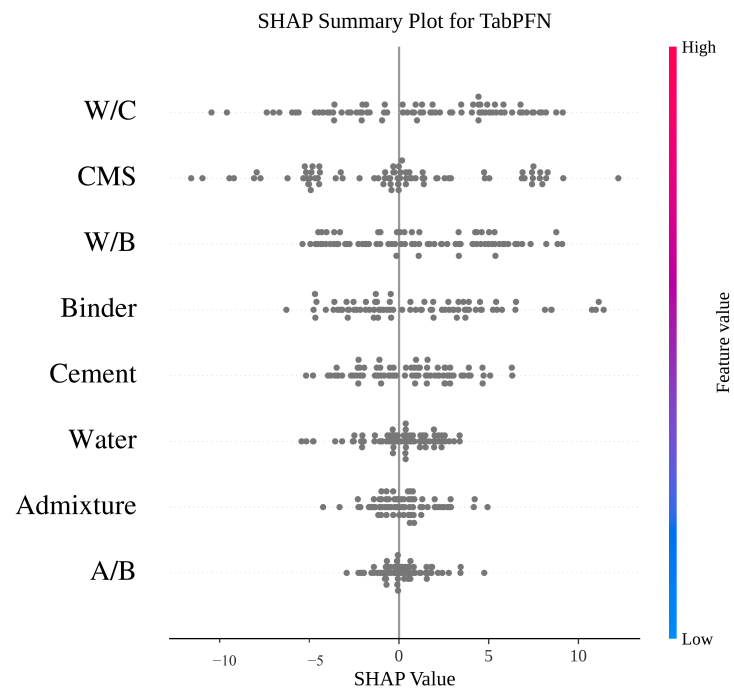


Figure 10. SHAP summary plot.

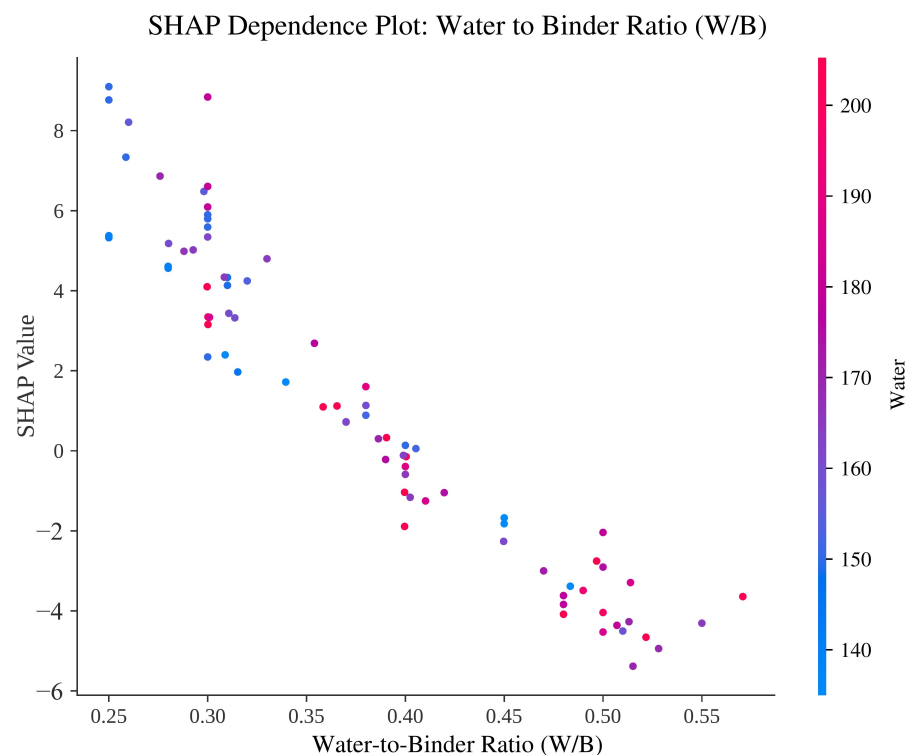


Figure 11. SHAP dependence plot of W/B .

Figure 11 demonstrates a clear negative relationship between W/B and SHAP values. As W/B increased, the SHAP contribution gradually decreased from positive to negative values, indicating that larger water-to-binder ratios adversely affect compressive strength prediction. This observation is consistent with the well-established understanding that excessive water content increases porosity and reduces matrix densification. The color distribution in the dependence plot is used to visualize the selected interacting variable and should be interpreted as model-attributed association rather than direct causal interaction.

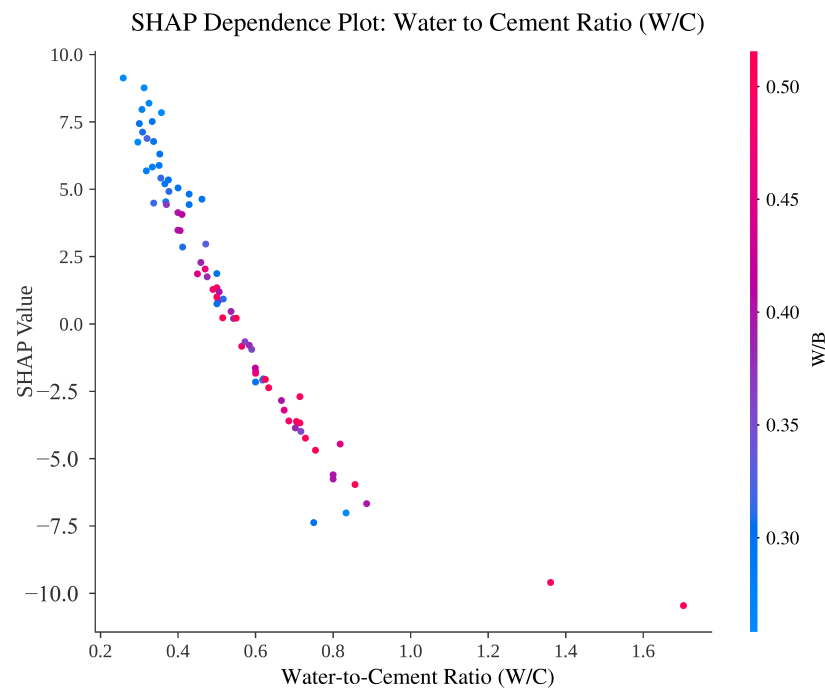


Figure 12. SHAP dependence plot of W/C .

A similar trend can be observed in Figure 12. The SHAP value decreased almost monotonically with increasing W/C , indicating that W/C is a dominant governing factor in compressive strength prediction. Lower W/C values produced positive contributions to the predicted strength, whereas higher W/C values generated increasingly negative contributions. This trend closely agrees with Abrams' law and confirms that the TabPFN model successfully captured physically meaningful relationships within the dataset.

Figure 13 shows a strong positive correlation between mortar strength and SHAP values. As mortar strength increased, the corresponding SHAP contribution rose substantially, indicating that higher mortar strength enhances the predicted compressive strength of concrete. This result highlights the important role of matrix mechanical properties in determining overall concrete performance.

Finally, a representative SHAP waterfall plot is presented in Figure 14 to visualize the contribution of individual variables to a specific prediction sample.

The waterfall plot provides a detailed explanation of how individual features influenced the final prediction. In the waterfall plot, $E[f(X)]$ represents the baseline prediction, corresponding to the average model output across all samples. The final prediction was obtained by adding the SHAP contributions of individual features to this baseline value.

For the representative sample shown in Figure 14, the baseline prediction was 45.43 MPa. The feature contributions collectively reduced the prediction to 37.21 MPa, which is very close to the experimentally measured value of 37.34 MPa. This result indicates that the developed TabPFN model can accurately capture the combined effects of multiple mixture variables for individual concrete samples.

Among all variables, CMS exhibited the largest negative contribution (−4.97 MPa), followed by Binder (−2.50 MPa) and W/B (−1.16 MPa). In contrast, W/C (+0.84 MPa) and Admixture (+0.76 MPa) contributed positively to the prediction and partially offset the negative effects of other variables. The remaining variables, including A/B , Cement, and Water, made relatively smaller contributions. These results demonstrate that the prediction was determined by the combined interactions among multiple mixture parameters rather than by a single dominant factor.

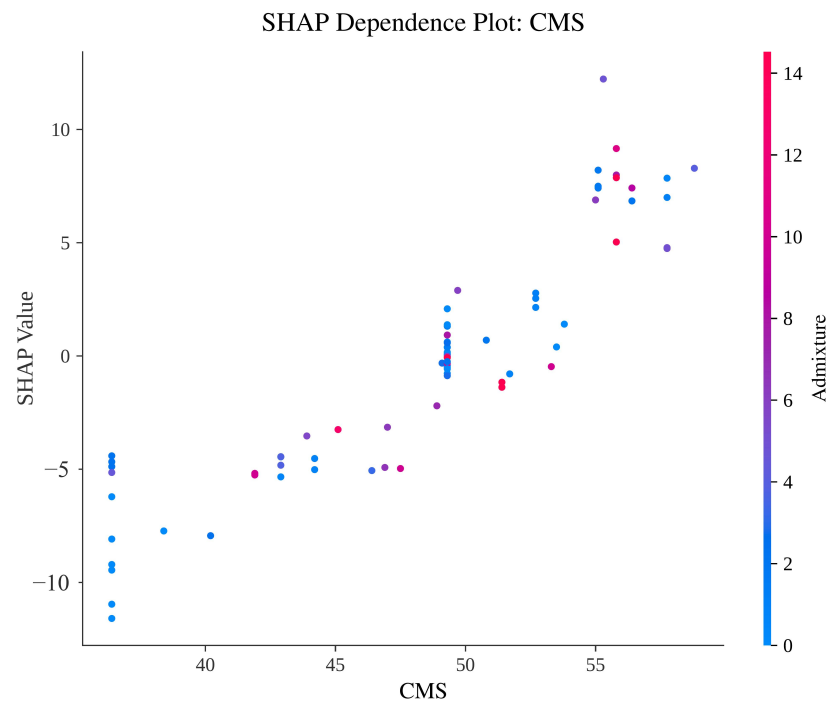


Figure 13. SHAP dependence plot of cement mortar strength.

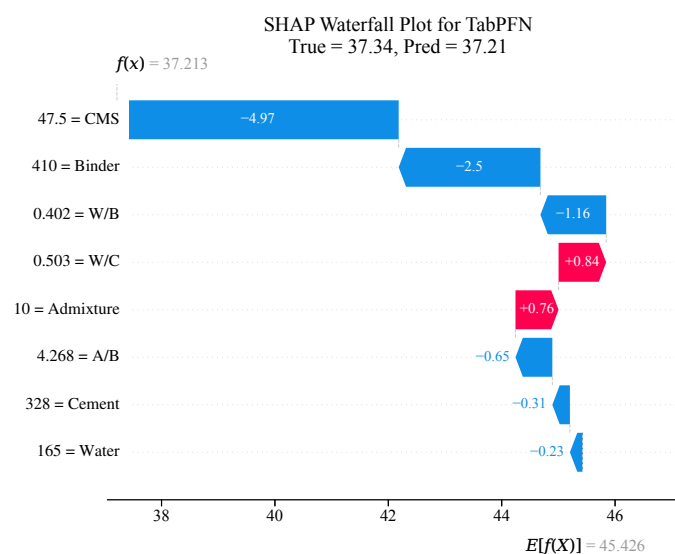


Figure 14. Representative SHAP waterfall plot.

The contribution patterns shown in the waterfall plot are consistent with the global SHAP analysis results presented in Figures 9 and 10. Variables identified as highly influential in the global analysis, particularly W/C, CMS, W/B, and Binder, also exhibited substantial contributions at the individual-sample level. This consistency further confirms the reliability of the SHAP interpretation framework.

Overall, the SHAP analysis demonstrates that the developed TabPFN framework not only achieves high prediction accuracy but also provides physically meaningful insights into the influence mechanisms of mixture variables. The global feature importance results identified W/C, CMS, W/B, and Binder as the most influential factors governing compressive strength prediction. However, W/C and W/B are physically related ratio variables. Their Pearson correlation in the present dataset was moderate ($r = 0.452$), and the variance inflation analysis indicates that W/B and binder content had high redundancy

with other mixture variables. Therefore, the SHAP values of correlated ratio variables should be interpreted as model-attributed contributions rather than independent causal effects. Additional feature ablation experiments showed that removing W/C or W/B did not materially reduce the performance of the strongest tree-based baselines; for example, LightGBM achieved $R^2 = 0.9047$ without W/C and $R^2 = 0.9023$ without W/B , compared with $R^2 = 0.9017$ using the full feature set. These results suggest that W/C and W/B provide partly redundant but physically interpretable information.

6.6. Sensitivity Analyses and Limitations

Several additional analyses were conducted to assess the robustness and limitations of the proposed framework. First, the influence of the extreme W/C value was examined using an IQR-based outlier sensitivity test. Removing 42 samples with extreme W/C values reduced the maximum W/C from 2.25 to 1.0563, and LightGBM remained the strongest evaluated baseline in this supplementary outlier check. This indicates that the reported model comparison was not driven solely by a small number of extreme-ratio samples.

Second, the split conformal prediction procedure was evaluated under alternative train/calibration/test ratios. For the 50/25/25 split, LightGBM achieved an average PICP of 0.9620 and an MPIW of 27.5470 MPa; for the 70/15/15 split, it achieved an average PICP of 0.9644 and an MPIW of 27.1289 MPa. These results, together with the 60/20/20 repeated CP results in Table 7, indicate that the average coverage remained close to the nominal 95% level across different split ratios. The 60/20/20 ratio was retained because it provided a balanced allocation between model fitting, calibration quantile estimation, and final testing.

Third, the present study focused on 28-day compressive strength prediction using eight mixture-level variables. Important factors such as fly ash chemical composition, curing regime, aggregate mineralogy and gradation, specimen geometry, long-term age effects, permeability, shrinkage, carbonation, chloride resistance, and service life indicators were not included in the dataset. Therefore, the proposed model should be regarded as a data-driven screening and decision support tool for preliminary mixture-property prediction rather than a substitute for laboratory validation or code-compliant structural assessment. Recent experimental-numerical studies on durability-related components, such as composite EPDM gaskets for underground structures [28], illustrated that multi-physics testing and numerical modeling remain essential for service life evaluation. Such approaches are complementary to the present data-driven strength prediction framework and should be integrated in future work when extending the method toward durability-oriented and life cycle-informed concrete design.

7. Conclusions

In this study, TabPFN was applied to predict the 28-day compressive strength of fly ash concrete, and its performance was systematically evaluated against tuned conventional machine learning models, gradient-boosting baselines, and simple empirical baselines. Based on the experimental results, the following conclusions can be drawn:

1. TabPFN achieved the best overall predictive performance among all investigated models. Across 100 repeated 80/20 train/test splits, it obtained an average R^2 of 0.9329, an MAE of 3.2758 MPa, and an RMSE of 4.6678 MPa, demonstrating strong accuracy for fly ash concrete compressive strength prediction.
2. Compared with tuned CatBoost and LightGBM, TabPFN still achieved lower prediction errors. Relative to CatBoost, TabPFN improved R^2 by 0.0267 and reduced MAE and RMSE by 0.8768 MPa and 0.8560 MPa, respectively. These results indicate that TabPFN remained competitive even against strong modern tabular baselines.

3. Statistical significance analysis based on Wilcoxon signed-rank tests confirmed that the performance improvements achieved by TabPFN were statistically significant. Scatter plots and residual analyses further demonstrated that TabPFN produced concentrated predictions and limited systematic prediction bias.
4. The integration of repeated-split conformal prediction enabled reliable uncertainty quantification. TabPFN achieved an average PICP of 0.9615 and an MPIW of 23.4554 MPa, indicating a favorable balance between interval reliability and interval compactness.
5. SHAP analysis revealed that W/C , CMS, W/B , and Binder were important factors governing compressive strength prediction. Because W/C and W/B are correlated ratio variables, their SHAP values should be interpreted as model-attributed contributions rather than independent causal effects.

Overall, the results demonstrate that TabPFN provides an accurate, robust, interpretable, and uncertainty-aware framework for preliminary prediction of fly ash concrete compressive strength. The proposed approach shows potential for data-driven material design and performance screening in sustainable civil engineering applications. Future work should validate TabPFN using larger and more diverse concrete databases, incorporate fly ash chemistry, curing conditions, aggregate characteristics, and age effects, and extend the framework to durability-related and service life-oriented properties.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/buildings16142781/s1>, Figure S1: Prediction error distributions for selected high-strength test samples; Tables S1–S6: Supplementary summary statistics, conditional coverage, split-ratio sensitivity, outlier sensitivity, and runtime details.

Author Contributions: Z.Z. contributed to the study conceptualization, methodology development, and data analysis. J.W. provided conceptual supervision, writing support, critical review, and manuscript revision. G.M. participated in data analysis and result interpretation. M.H. was responsible for the literature review, data curation, and manuscript finalization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Youth Academic Innovation Team Construction project of the Capital University of Economics and Business, grant number QNTD202303.

Data Availability Statement: The data presented in this study are openly available in GitHub at <https://github.com/zhzhao07/TabPFNforCCS> (accessed on 8 July 2026).

Conflicts of Interest: Author Mingjie Han was employed by Beijing Building Construction Research Institute Co., Ltd. The authors declare that this commercial affiliation did not constitute a conflict of interest and did not influence the design of the study; the collection, analyses, or interpretation of data; the writing of the manuscript; or the decision to publish the results.

References

1. Neville, A.M. *Properties of Concrete*; Pearson: Harlow, UK, 2011.
2. Andrew, R.M. Global CO₂ emissions from cement production, 1928–2018. *Earth Syst. Sci. Data* **2019**, *11*, 1675–1710. [CrossRef]
3. International Energy Agency. *Cement*; Technical Report; IEA: Paris, France, 2025. Available online: <https://www.iea.org/reports/cement> (accessed on 8 July 2026).
4. Mehta, P.K. Reducing the environmental impact of concrete. *ACI Concr. Int.* **2001**, *23*, 61–66. Available online: <https://ecosmartconcrete.com/docs/trmehta01.pdf> (accessed on 8 July 2026).
5. Siddique, R. Utilization of industrial by-products in concrete. *Procedia Eng.* **2014**, *95*, 335–347. [CrossRef]
6. Nandurkar, B.P.; Raut, J.M.; Hinge, P.K.; Bahoria, B.V.; Patil, T.R.; Upadhye, S.; Vairagade, V.S.; Shelare, S.D. Optimization and predictive performance of fly ash-based sustainable concrete using integrated multitask deep learning framework with interpretable machine learning techniques. *Sci. Rep.* **2025**, *15*, 30820. [CrossRef] [PubMed]
7. Sichani, M.N.; Dehkordi, O.M.; Khorshidi, M.; Teimortashlu, A.; Nejati-pour, P. Machine learning based optimization of fly ash content for improving geopolymers concrete compressive strength. *Sci. Rep.* **2025**, *15*, 44785. [CrossRef] [PubMed]

8. Asteris, P.; Kolovos, K.; Douvika, M.; Plevris, V. Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cem. Concr. Res.* **2021**, *145*, 106449. [\[CrossRef\]](#)
9. Abrams, D.A. *Design of Concrete Mixtures*; Structural Materials Research Laboratory: Chicago, IL, USA; Lewis Institute: Chicago, IL, USA, 1919.
10. Yeh, I.C. Modeling of strength of high-performance concrete using artificial neural networks. *Cem. Concr. Res.* **1998**, *28*, 1797–1808. [\[CrossRef\]](#)
11. Muliauwana, H.N.; Prayogo, D.; Gaby, G.; Harsono, K. Prediction of concrete compressive strength using artificial intelligence methods. *J. Phys. Conf. Ser.* **2020**, *1625*, 012018. [\[CrossRef\]](#)
12. Paudel, S.; Pudasaini, A.; Shrestha, R.K.; Kharel, E. Compressive strength of concrete material using machine learning techniques. *Clean. Eng. Technol.* **2023**, *15*, 100661. [\[CrossRef\]](#)
13. Liu, J.; Guan, D.; Liu, X. Comparative performance analysis of machine learning models for compressive strength prediction in concrete mix design. *Math. Comput. Appl.* **2025**, *30*, 128. [\[CrossRef\]](#)
14. Jiang, Y.; Li, H.; Zhou, Y. Compressive strength prediction of fly ash concrete using machine learning techniques. *Buildings* **2022**, *12*, 690. [\[CrossRef\]](#)
15. Khan, A.Q.; Naveed, M.H.; Rasheed, M.D.; Miao, P. Prediction of compressive strength of fly ash-based geopolymer concrete using supervised machine learning methods. *Arab. J. Sci. Eng.* **2024**, *49*, 4889–4904. [\[CrossRef\]](#)
16. Padavala, S.S.A.B.; Avudaiappan, S.; Prathipati, S.R.R.T.; Paluri, Y.; Sainath, V.; Kumara, A.M. Integrated experimental, machine learning, and life-cycle assessment of fly ash–silica fume based self-compacting geopolymer concrete. *Sci. Rep.* **2026**, *16*, 12845. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Kumar, A.; Sen, S.; Sinha, S. Support vector machine-based prediction model for the compressive strength for concrete reinforced with waste plastic and fly ash. *Asian J. Civ. Eng.* **2025**, *26*, 1429–1447. [\[CrossRef\]](#)
18. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 3146–3154.
19. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31, pp. 6638–6648.
20. Wang, J.; Wang, Y.; Shi, C.; Liu, Y. Strength-based life cycle CO₂ of fly ash concrete: Variation with common mix parameters. *Constr. Build. Mater.* **2024**, *411*, 134365. [\[CrossRef\]](#)
21. Feret, R. *Sur la Compacité des Mortiers Hydrauliques*; C. Dunod: Paris, France, 1892. Available online: <https://books.google.ca/books?id=gPmcPQAACAAJ> (accessed on 8 July 2026).
22. MOHURD. *Specification for Mix Proportion Design of Ordinary Concrete (JGJ 55-2011)*; China Architecture & Building Press: Beijing, China, 2011. Available online: <https://max.book118.com/html/2019/0103/8033030043001143.shtml> (accessed on 8 July 2026).
23. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
24. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: Berlin/Heidelberg, Germany, 1995. [\[CrossRef\]](#)
25. Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 1999.
26. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. *Nature* **2025**, *637*, 319–326. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Asgarkhani, N.; Kazemi, F.; Jankowski, R. Machine-learning based tool for seismic response assessment of steel structures including masonry infill walls and soil-foundation-structure interaction. *Comput. Struct.* **2025**, *317*, 107918. [\[CrossRef\]](#)
28. Wang, Z.; Shen, S.L.; Zhou, A. Performance of Composite EPDM Gaskets for Underground Structures: Experimental and Numerical Investigation. *Constr. Build. Mater.* **2025**, *500*, 143951. [\[CrossRef\]](#)

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