Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

Evaluation of Gaussian Process Regression, Support Vector Regression, and K-Nearest Neighbors for Predicting Concrete Compressive Strength at Various Curing Ages

Elmanani Simamora[1], Abil Mansyur[2], Muhammad Badzlan Darari[3], Suvriadi Panggabean[4], Rizki Habibi[5]
[1,2,3,4,5]Department of Mathematics, Universitas Negeri Medan, Medan, Indonesia
[1]elmanani_simamora@unimed.ac.id

ABSTRACT

Accurate prediction of concrete compressive strength is essential for quality control, structural evaluation, and concrete mix optimization. However, the relationships among mixture composition, curing age, and compressive strength are often nonlinear and heterogeneous. This study evaluates four prediction models—Native Gaussian Process Regression (NGPR), Gaussian Process Regression with Conformal Intervals (GPR-CI), Support Vector Regression (SVR), and K-Nearest Neighbors (KNN)—for predicting concrete compressive strength at various curing ages. The predictor set includes mixture-composition variables, curing age, and material-relevant derived features. Model performance is assessed using Leave-One-Age-Out validation and repeated stratified K-fold cross-validation by age, based on root mean square error, mean absolute error, coefficient of determination, empirical coverage, and average interval length. The results show that NGPR provides the most consistent overall performance across both evaluation schemes. Under the Leave-One-Age-Out scheme, NGPR achieves better point-prediction accuracy and higher empirical coverage than GPR-CI, whereas under repeated age-stratified cross-validation, GPR-CI achieves slightly higher coverage at the cost of wider prediction intervals. Overall, NGPR provides the most favorable balance among point-prediction accuracy, interval calibration, and interval efficiency for heterogeneous concrete compressive-strength data across curing ages.

Keywords: Concrete Compressive Strength, Gaussian Process Regression, Support Vector Regression, K-Nearest Neighbors, Prediction Interval, Curing Age

Corresponding Author:
Elmanani Simamora,
Department of Mathematics,
Faculty of Mathematics and Natural Sciences,
Universitas Negeri Medan
Jalan William Iskandar Ps. V, Medan, Indonesia, 20221
elmanani_simamora@unimed.ac.id

1. INTRODUCTION

One of the most important properties of concrete in structural design, quality control, and material performance evaluation is its compressive strength. This measure is used in the profession of civil engineering to assess the acceptability of the mixture, compliance with standards, and to make technical decisions in production and execution. However, the relationship between mixture composition and compressive strength is not simple. Cement, water, slag, fly ash, superplasticizer, aggregate, and curing age work together, so that the resulting response tends to be nonlinear and influenced by interactions between materials (Yeh, 1998; Yeh, 2006; Feng et al., 2020; Kaloop et al., 2020; Beskopylny et al., 2022). Classical and recent literature both show that curing age, water content, water-to-binder ratio, and additional cementitious materials remain the dominant variables in the formation of compressive strength (Yeh, 1998; Yeh, 2006; Beskopylny et al., 2022; Liu & Sun, 2023).

These developments have driven a shift from conventional empirical models to machine learning-based approaches. Various studies have used ensemble models, interpretable models, and kernel methods to predict concrete compressive strength (Feng et al., 2020; Kaloop et al., 2020; Beskopylny et al., 2022; Liu & Sun, 2023; Elhishi et al., 2023; Zhang et al., 2024; Angiulli et al., 2024; Miao et al., 2025). Boosting approaches can capture complex nonlinear relationships with high accuracy (Feng et al., 2020; Kaloop et al., 2020). Interpretable models such as the Explainable Boosting Machine and SHAP-based approaches have also shown that good predictions can still be achieved with explanations of variable contributions (Liu & Sun, 2023; Elhishi et al., 2023; Zhang et al., 2024). On the other hand, comparative studies using concrete data still rank Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) as important benchmarks because they represent two distinct strategies for interpreting nonlinear data structures (Beskopylny et al., 2022; Angiulli et al., 2024). Thus, the main question is no longer which model has the smallest error. The more appropriate question is which model remains stable when the data structure changes, and which model provides a reasonable balance between point accuracy and interval quality.

In recent years, attention has also shifted towards models that are not only accurate but also more informative in terms of prediction uncertainty. At this point, Gaussian Process Regression (GPR) becomes relevant. GPR is not only flexible in modeling nonlinear relationships but also naturally generates predictive distributions (Miao et al., 2025; Xie et al., 2025). This characteristic is important when the analysis objective is not limited to point predictions, but also includes prediction intervals and actual coverage. Recent literature on multi-binder concrete, non-destructive testing, and uncertainty-aware compressive strength estimation shows that uncertainty is increasingly difficult to ignore (Miao et al., 2025; Xie et al., 2025; Tamuly & Nava, 2025; Fu et al., 2025). Correspondingly, conformal prediction is increasingly used to generate more reliable intervals without relying heavily on distributional assumptions about the model (Tamuly & Nava, 2025; Fu et al., 2025).

However, one point is still often not discussed explicitly. Many studies report Root Mean Square Error **(**RMSE), Mean Absolute Error **(**MAE), Coefficient of Determination **(**R^2), and other metrics using a common validation scheme, but do not always treat the curing age as an explicit evaluation criterion. In fact, in concrete compressive strength data, age is often recorded at discrete points, and the number of observations is not evenly distributed across age groups. Recent literature on high-performance concrete, high-strength concrete, and multi-dataset analysis also shows that curing age remains a dominant variable, and its distribution can affect the model's behavior (Zhang et al., 2024; Shaaban et al., 2025; Mobasheri et al., 2025; Xiao et al., 2025; Olvera-Mayorga et al., 2025). If so, the usual random evaluation is likely too optimistic, especially when the dominant age group consistently appears in both the training and test data.

The need is particularly relevant to the data in this study. The exploratory results indicate that the compressive strength distribution is unimodal and right-skewed, while variation with curing age is not homogeneous. Median compressive strength tends to increase from early to intermediate ages, particularly up to approximately 28 days, but after that, the pattern of increase is no longer consistent across age groups. Furthermore, the ages in the data are recorded at discrete points rather than on a continuous scale, so the age–compressive strength relationship is better understood as gradual and clustered. The 28-day age group is particularly significant because it has the largest number of observations and is often regarded as the primary reference for assessing concrete compressive strength in civil engineering practice. But some of the earlier ages have significantly less data. The circumstances indicate that age in the study is not only a predictor column but also part of the evaluation structure that needs to be expressly addressed.

Based on this, we focus the comparison on Native GPR (NGPR), GPR with Conformal Intervals (GPR-CI), SVR, and KNN. These four models were chosen because they represent different approaches to nonlinear regression. The GPR family was placed more firmly from the outset because it is substantively closest to the needs of this study. This model is flexible to nonlinear relationships and supports the formation of prediction distributions. In this study, the GPR family is divided into two forms: NGPR and GPR-CI, because the mechanisms for forming prediction intervals differ. Meanwhile, SVR is a well-established kernel-based comparator in nonlinear regression, and KNN is a simple local proximity-based comparator that remains relevant to tabular data (Beskopylny et al., 2022; Miao et al., 2025; Xie et al., 2025; Tamuly & Nava, 2025; Fu et al., 2025).

The research problem is formulated in two directions. First, whether different prediction models will show the same stability when the data is evaluated by explicitly taking into account the curing age. Second, whether a model that is good in terms of point predictions also remains good in terms of prediction intervals. To solve these problems, we conduct assessments using two complementary frameworks: Leave-One-Age-Out (LOAO) and repeated stratified K-fold by Age. LOAO is used to examine the generalization capacity when one age group is not involved in the training at all. Conversely, repeated stratified K-fold by Age is used to obtain a more empirically stable picture of performance because all age groups are still represented in each fold. This choice is in line with this study, as the results show that model performance varies with the age structure encountered, and that the quality of the interval cannot be assessed solely from RMSE or R^2.

In the context of prediction intervals, this study also starts from the view that the best model cannot be judged solely by the narrowest interval or the highest coverage. Intervals that are too narrow may appear sharp but fail to adequately cover the actual values. Conversely, intervals that are too wide are indeed safe, but less efficient. The literature on probabilistic prediction emphasizes that interval quality needs to be assessed from two perspectives simultaneously: calibration and sharpness (Gneiting et al., 2007; Tamuly & Nava, 2025; Dong & Zhang, 2025; Almutairi, 2025). The principle is very relevant in this study, because the results show a clear trade-off between interval length and actual coverage, both in the LOAO scheme and in the repeated stratified K-fold by Age.

Based on these factors, the purpose of this work is to evaluate the performance of NGPR, GPR-CI, SVR, and KNN in predicting concrete compressive strength at various curing ages, with an emphasis on two aspects: point prediction accuracy and prediction interval quality. The proposed working hypothesis is that the GPR family will display more stability than SVR and KNN when evaluated on a scheme sensitive to age-group shifts. At the same time, the difference between NGPR and GPR-CI will predominantly show as in trade-off between interval efficiency and coverage conservatism. The expected result is a stronger empirical basis for selecting concrete compressive strength prediction models that are not only good in aggregate but also remain reliable as the curing age structure changes and when prediction uncertainty needs to be read explicitly.

The novelty of this study lies not in the four algorithms themselves, which are well established, but in the integrated age-structured framework used to evaluate both predictive accuracy and uncertainty. The framework combines Leave-One-Age-Out validation to assess generalization to curing-age groups excluded from model training with repeated age-stratified cross-validation to assess performance across resampled partitions that preserve the curing-age distribution as far as the group sizes permit. This complementary design enables the four models to be compared under two distinct age-based validation settings in terms of point-prediction accuracy, interval calibration as measured by empirical coverage, and interval sharpness.

2. RESEARCH METHOD

A. Data and predictor formation

We use concrete compressive strength data that have also been used in several regression analyses, both in multiple linear regression and in robust regression and bootstrap studies (Rana et al., 2012; Simamora et al., 2025), with one response variable, namely concrete compressive strength, and several predictor variables representing the mixture composition and curing age. In general, the data is expressed as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where y_i represents the concrete compressive strength at the i-th observation, while $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ represents the predictor vector. The relationship between the response and predictors is written in equation (2.1) as

$$y_i = f(x_i) + \varepsilon_i, \tag{2.1}$$

with $f(\cdot)$ the unknown regression function and ε_i representing random errors. Curing age is retained as the main predictor. In addition to the basic variables, materially relevant derived features are also formed, namely total binder, water to binder ratio, indicators of the presence of slag, fly ash, and superplasticizer, and log(age) transformation. The formation of these features is used to enrich the representation of the nonlinear relationship between mixture composition, curing age, and concrete compressive strength.

Before modeling, data exploration was conducted using compressive strength histograms, boxplots by curing age, and scatterplots between curing age and compressive strength. This stage was used to interpret the response distribution, the imbalance in the number of observations between ages, and the initial pattern of the relationship between age and compressive strength. This arrangement was maintained because it was consistent with the research results, which interpreted the data based on the global distribution, variation between ages, and the pattern of the relationship between age and compressive strength.

B. Preprocessing and validation design

All predictors are standardized using formula (2.2) so that models are compared on the same input scale. Standardization is performed only based on the training data statistics for each partition, so that information from the test data does not enter the training stage. If X_{tr} and X_{te} represent the training and test predictor matrices, respectively, then the transformation is performed with

$$Z_{\text{tr}} = \frac{X_{\text{tr}} - \bar{X}_{\text{tr}}}{s_{\text{tr}}},\ Z_{\text{te}} = \frac{X_{\text{te}} - \bar{X}_{\text{tr}}}{s_{\text{tr}}}, \tag{2.2}$$

where $\bar{X}_{\text{tr}}$ and s_{tr} denote the mean vector and standard deviation vector computed from the training predictors. The same training statistics are then applied to the test predictors to prevent information leakage from the test data. The evaluation is conducted within a multilevel validation framework. Two schemes are used in the outer layer. The first scheme is LOAO. In this scheme, all observations from a single curing age are removed from the training data and used as test data. If the set of unique ages is written as $\mathcal{A} = \{a_1, \dots, a_G\}$, then at the g-th iteration,

$$\mathcal{D}_{\text{test}}^{(g)} = \{(x_i, y_i) : t_i = a_g\},\ \mathcal{D}_{\text{train}}^{(g)} = \mathcal{D} \setminus \mathcal{D}_{\text{test}}^{(g)}, \tag{2.3}$$

where t_i denotes the curing age of the i-th observation. The scheme in (2.3) was chosen because it provides a rigorous generalization test when the model has to predict age groups that do not appear at all in the training.

The second scheme is a repeated stratified K-fold by Age. In this approach, the data is divided into five folds, stratified by curing age, and then the division is repeated five times with different random partitions. Within each outer partition, five-fold cross-validation is performed on the training data to select the best hyperparameters. This design is appropriate for the data structure with an unbalanced age distribution. It is also in line with research findings, as stratified evaluation by Age yields a more stable performance picture than LOAO.

C. Prediction models and prediction intervals

The study compares four models: NGPR, GPR-CI, SVR, and KNN. Conceptually, these four models represent three families of approaches. The Gaussian process family is divided into two because the mechanisms for forming prediction intervals differ. In NGPR, point predictions and prediction intervals are both derived from the model's predictive distribution. In contrast, in GPR-CI, fixed-point predictions are obtained from a Gaussian process model, whereas prediction intervals are formed via a conformal approach. Meanwhile, SVR and KNN are used as comparison models to assess the stability of the kernel- and local-proximity-based approaches on concrete compressive strength data.

In the GPR model, the latent function $f_{GP}(x)$ in equation (2.4) is assumed to follow a Gaussian process,

$$f_{GP}(x) \sim GP(\mu_{GP}(x), k(x, x')), \tag{2.4}$$

where $\mu_{GP}(x)$ represents the mean function and $k(x, x')$ represents the covariance or kernel function. The mean function is assumed to be constant. The observation model is then expressed in equation (2.5) as

$$y_i = f_{GP}(x_i) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_n^2),\ i = 1,2, \dots, n, \tag{2.5}$$

where y_i denotes the concrete compressive strength at the i-th observation, $f_{GP}(x_i)$ denotes the latent function at point x_i and σ_n^2 denotes the variance of the observation noise. For a new prediction point x_*, the posterior mean of the latent function is given by

$$\hat{\mu}_{f|\mathcal{D}}(x_*) = \mu_{GP}(x_*) + k_*^{\top}(K + \sigma_n^2 I)^{-1}(y - \mu), \tag{2.6}$$

and the posterior variance of the latent function is given by

$$\hat{\sigma}_{f|\mathcal{D}}^2(x_*) = k(x_*, x_*) - k_*^{\top}(K + \sigma_n^2 I)^{-1} k_*. \tag{2.7}$$

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

Because this study evaluates the prediction interval against the actual observed value, the predictive variance for the new observation y_* is written as

$$\hat{\sigma}^2_{y|\mathcal{D}}(x_*) = \hat{\sigma}^2_{f|\mathcal{D}}(x_*) + \sigma_n^2. \tag{2.8}$$

In Equations (2.6)–(2.8), $K = [k(x_i, x_j)]_{i,j=1}^{n}$ is the $n \times n$ covariance matrix containing the kernel values for all pairs of training points, while $k_* = (k(x_1, x_*), k(x_2, x_*), \dots, k(x_n, x_*))^\top$ denotes the covariance vector between the new point x_* and all training points. Furthermore, $y = (y_1, y_2, \dots, y_n)^\top$ denotes the observed response vector and $\mu = (\mu_{GP}(x_1), \mu_{GP}(x_2), \dots, \mu_{GP}(x_n))^\top$ denotes the mean prior vector over all training points. Thus, $\hat{\mu}_{f|\mathcal{D}}(x_*)$ in Equation (2.6) represents the point prediction, while $\hat{\sigma}^2_{f|\mathcal{D}}(x_*)$ and $\hat{\sigma}^2_{y|\mathcal{D}}(x_*)$ in Equations (2.7) and (2.8) represent the prediction uncertainty for the latent function and the new observation.

During model tuning, several standard kernels were considered, namely the Squared Exponential (SE), Matérn 3/2, Matérn 5/2, and Rational Quadratic (RQ) kernels.For two predictor vectors x and x', and $r = \|x - x'\|$, the four kernels are defined as

$$\begin{aligned} k_{\mathrm{SE}}(x, x') &= \sigma_k^2 \exp\left(-\frac{r^2}{2\ell^2}\right), \\ k_{\mathrm{Mat\,3/2}}(x, x') &= \sigma_k^2 \left(1 + \frac{\sqrt{3}\,r}{\ell}\right) \exp\left(-\frac{\sqrt{3}\,r}{\ell}\right), \\ k_{\mathrm{Mat\,5/2}}(x, x') &= \sigma_k^2 \left(1 + \frac{\sqrt{5}\,r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}\,r}{\ell}\right), \text{ and} \\ k_{\mathrm{RQ}}(x, x') &= \sigma_k^2 \left(1 + \frac{r^2}{2\gamma\ell^2}\right)^{-\gamma}. \end{aligned} \tag{2.9}$$

In Equation (2.9), σ_k^2 represents the kernel amplitude parameter that governs the magnitude of the latent function variation, ℓ is the scale length parameter, and γ is the shape parameter in the rational quadratic kernel. The squared exponential kernel produces a very smooth function. The Matérn 3/2 and Matérn 5/2 kernels provide flexibility with a more moderate level of smoothness, while the rational quadratic kernel can accommodate variations in scale length in the data. The best kernel is then selected based on the smallest average RMSE value in the inner cross-validation.

In NGPR, point predictions and prediction intervals are both derived from the model's predictive distribution. For a nominal confidence level of $100(1-\alpha)\%$, the prediction interval for a new observation y_* is written as

$$\hat{\mu}_{f|\mathcal{D}}(x_*) \pm z_{1-\alpha/2}\hat{\sigma}_{y|\mathcal{D}}(x_*), \text{ where } \hat{\sigma}_{y|\mathcal{D}}(x_*) = \sqrt{\hat{\sigma}^2_{y|\mathcal{D}}(x_*)}, \tag{2.10}$$

with α representing the miscoverage level of the prediction interval.

In contrast, in GPR-CI, point predictions are derived from a Gaussian process, whereas prediction intervals are formed via split conformal prediction. Suppose the external training data is divided into the main training data and the calibration data. If the absolute residuals on the calibration data are expressed as

$$s_j = |y_j^{\mathrm{cal}} - \hat{y}_j^{\mathrm{cal}}|, \; j = 1,2, \dots, m, \tag{2.11}$$

then the conformal quantile is determined as

$$q_{1-\alpha} = s_{(\lceil (m+1)(1-\alpha) \rceil)}, \tag{2.12}$$

with m representing the number of observations in the calibration data, α denoting the miscoverage level and $s_{(1)} \le s_{(2)} \le \dots \le s_{(m)}$ are the ordered calibration residuals. The conformal prediction interval for the new point x_* is then written as

$$\hat{y}(x_*) \pm q_{1-\alpha} \tag{2.13}$$

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

Thus, the difference between NGPR and GPR-CI lies in the source of the prediction intervals. In the first model, the intervals are derived from the Gaussian process predictive distribution as in equation (2.10). In the second model, the intervals are derived from residual calibration through conformal prediction as in equation (2.13).

SVR is used as a kernel-based comparator. In general, SVR forms a prediction function

$$\hat{f}_{SVR}(x_*) = w^\top \phi(x) + b, \tag{2.14}$$

where $\phi(x)$ represents the feature transformation, w the coefficient vector, and b a constant. The penalty parameters, kernel scale, and ε_{SVR}, which controls the width of the epsilon-insensitive tube in SVR, are selected through inner cross-validation. Since the SVR in equation (2.14) does not produce a native predictive distribution, its prediction interval is formed using the same conformal procedure as in Equations (2.11)–(2.13).

KNN is used as a local proximity-based comparator. If weighting is used, the point prediction for a new point x_* is written as

$$\hat{f}_{KNN}(x_*) = \sum_{j=1}^{k_{KNN}} w_j \, y_{(j)}, \tag{2.15}$$

with $y_{(j)}$ denoting the response of the j-th neighbor and the weight

$$w_j = \frac{1/d_j}{\sum_{h=1}^{k} 1/d_h}. \tag{2.16}$$

In equation (2.16), the calculation of d_j is the distance between x and the j-th neighbor. Meanwhile, the number of neighbors k_{KNN}, the distance metric, and the use of weights are selected via inner cross-validation. Just like SVR, the KNN prediction interval in equation (2.15) is formed using the conformal procedure in equations (2.11)–(2.13).

D. Performance measures

Model performance is assessed from two perspectives: point prediction and prediction interval. For point prediction, RMSE, MAE, and R^2 are used:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{2.17}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{2.18}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{2.19}$$

For the prediction interval, coverage and Average Interval Length (AIL) are used. If the prediction interval of the i-th observation is $[L_i, U_i]$, then

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(L_i \le y_i \le U_i), \tag{2.20}$$

$$\text{AIL} = \frac{1}{n} \sum_{i=1}^{n} (U_i - L_i). \tag{2.21}$$

where $\mathbb{I}$ is the indicator function, L and U are the lower and upper prediction limits, respectively. RMSE, MAE, and R^2 were selected because they provide complementary information about point-prediction performance for a continuous engineering response measured in MPa. RMSE assigns a larger penalty to large prediction errors because the residuals are squared before averaging. This property is relevant to concrete compressive-strength prediction because substantial overestimation or underestimation may have more serious implications for quality-control and engineering decisions. MAE measures the typical absolute prediction error in the original MPa scale and is less strongly influenced by a small number of large residuals. Therefore, reporting both RMSE and MAE helps distinguish models that perform consistently from models whose average performance is affected by occasional large errors. The coefficient of determination, R^2, measures the proportion of the observed

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

variation in compressive strength explained by the model relative to a mean-based prediction. However, R^2 is interpreted jointly with RMSE and MAE rather than as a stand-alone measure, particularly for curing-age groups with small sample sizes, where R^2 may be unstable or negative. These three metrics are widely used in machine-learning studies of concrete compressive-strength prediction because they jointly describe error magnitude and explanatory performance (Beskopylny et al., 2022; Zhang et al., 2024).

Because point-prediction accuracy alone does not quantify predictive uncertainty, coverage and Average Interval Length (AIL) were also evaluated. Coverage measures the empirical proportion of observed compressive-strength values contained within the prediction intervals and therefore indicates interval calibration relative to the nominal coverage level. AIL measures interval sharpness or efficiency, with shorter intervals providing more precise predictions, provided that adequate coverage is maintained. Coverage and AIL must consequently be interpreted together: a narrow interval is not useful when it systematically undercovers the observations, whereas a very wide interval may achieve high coverage but provide limited practical precision. This joint assessment follows the calibration–sharpness principle for probabilistic prediction, according to which prediction intervals should be as narrow as possible subject to satisfactory calibration (Gneiting et al., 2007). The combined use of RMSE, MAE, R^2, coverage, and AIL is therefore appropriate for this study because it evaluates both the accuracy of predicted concrete compressive strength and the reliability and efficiency of the associated prediction intervals.

E. Analysis procedures

The analysis was conducted in six stages. First, the data was cleaned, and a predictor set was formed. Second, data exploration was conducted to examine the distribution of compressive strength, variation with curing age, and the initial pattern of the relationship between curing age and compressive strength. Third, for each outer partition, the data were split into training and testing sets, and the predictors were standardized using the training data, as in Equation (2.2). Fourth, cross-validation was performed on the inner layer to select the best hyperparameters from NGPR, GPR-CI, SVR, and KNN. Fifth, the best model was retrained on the full outer training data and used to form point predictions and prediction intervals on the test data. Sixth, the prediction results were summarized into RMSE, MAE, R^2, coverage, and AIL as defined in equations (2.17)–(2.21), both overall and by curing age group.

3. RESULTS AND DISCUSSION

A. Results

Figure 1 presents the distribution of concrete compressive strength, its distribution by curing age, and the relationship between curing age and concrete compressive strength. This visualization provides an initial overview of the data characteristics, variations between age groups, and the trend of the relationship patterns that underlie further analysis. Figure 1(a) shows a unimodal distribution of compressive strength with a rightward skew. The majority of observations fall within the 20–50 MPa range, with the highest concentration around 30–40 MPa. Very low values are present but not dominant, while high values above 60 MPa are relatively few and form a long right tail up to around 80 MPa. This means that the data does not appear perfectly symmetrical and does not appear roughly normal. There is considerable variation, but the main mass remains in the center. Thus, descriptively, this dataset is dominated by medium-strength concrete, with a small proportion of high-strength mixtures.

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

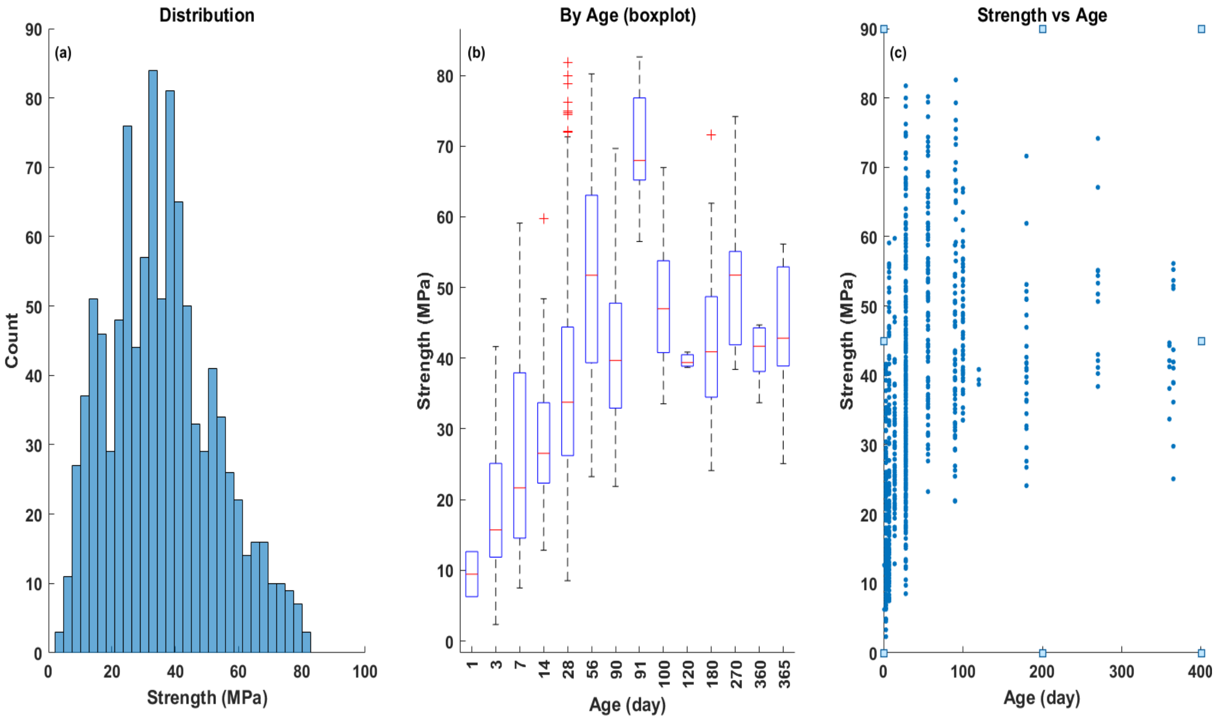

Figure 1: Characteristics of concrete compressive strength distribution and its relationship to curing age.

Figure 1(b) presents the distribution of concrete compressive strength across curing-age groups. The median strength generally increases from early to intermediate curing ages, particularly at 1, 3, 7, 14, and 28 days. However, the distribution within each age group is characterized not only by its median but also by its dispersion, range, and sample size. The comparatively wider interquartile ranges and longer whiskers observed at 28, 56, and 91 days indicate greater within-age variability, whereas the narrow distribution at 120 days should be interpreted cautiously because this age group contains only three observations. Points located beyond the boxplot whiskers represent unusually high or low compressive-strength observations and should not automatically be interpreted as statistically significant differences. Moreover, the observations at different curing ages do not represent repeated measurements of a single concrete mixture. Consequently, the differences among age groups reflect the combined effects of curing age, mixture composition, and unequal group sizes rather than the isolated effect of curing age alone.

Figure 1(c) shows that the relationship between age and compressive strength is positive but not linear. At low ages, particularly up to about 28 days, increasing age corresponds to an increase in the center of the compressive strength distribution. After that, the points do not form a neat ascending line, but rather a vertical distribution at discrete ages. This shows that the ages in this dataset were measured on certain test days rather than across a period. The link between age and compressive strength is more realistically described as a clustered, progressive relationship rather than a single smooth curve. Also, the point density is much higher for the younger ages, while the older ages (180, 270, 360, and 365 days) are much less frequent. Hence, the statistical information across large age groups is quite weak.

Beyond the median patterns shown in Figure 1, Table 1 summarizes the location, dispersion, and range of all variables in the dataset. The substantial differences in scale and variability across variables support the use of fold-specific predictor standardization and confirm that the data cover a broad range of mixture compositions and curing ages.

Table 1. Descriptive statistics of the concrete compressive strength dataset

Variable	Unit	Mean	SD	Minimum	Maximum
Cement	kg/m³	281.17	104.51	102.00	540.00
Blast-furnace slag	kg/m³	73.90	86.28	0.00	359.40
Fly ash	kg/m³	54.19	64.00	0.00	200.10
Water	kg/m³	181.57	21.36	121.75	247.00
Superplasticizer	kg/m³	6.20	5.97	0.00	32.20
Coarse aggregate	kg/m³	972.92	77.75	801.00	1,145.00
Fine aggregate	kg/m³	773.58	80.18	594.00	992.60
Curing age	day	45.66	63.17	1.00	365.00
Compressive strength	MPa	35.82	16.71	2.33	82.60

The concrete compressive strength has a mean of 35.82 MPa and a standard deviation of 16.71 MPa, with observed values ranging from 2.33 to 82.60 MPa. Curing age has a mean of 45.66 days and a standard deviation of 63.17 days, with values ranging from 1 to 365 days. The broad ranges and relatively large variations observed for several variables indicate substantial heterogeneity in mixture composition and curing-age structure. Pearson correlation analysis shows that compressive strength is positively associated with cement content ($\mathrm{r} = 0.498$), superplasticizer ($\mathrm{r} = 0.366$), curing age ($\mathrm{r} = 0.329$), and blast-furnace slag ($\mathrm{r} = 0.135$). Negative correlations are observed for water ($\mathrm{r} = -0.290$), fine aggregate ($\mathrm{r} = -0.167$), coarse aggregate ($\mathrm{r} = -0.165$), and fly ash ($\mathrm{r} = -0.106$). All correlations between the predictors and compressive strength were statistically significant ($p < 0.001$).

Among the predictors, the largest absolute pairwise correlations occur between water and superplasticizer ($\mathrm{r} = -0.657$), water and fine aggregate ($\mathrm{r} = -0.451$), cement and fly ash ($\mathrm{r} = -0.397$), and fly ash and superplasticizer ($\mathrm{r} = 0.377$). Nevertheless, no predictor pair has an absolute correlation coefficient exceeding 0.80, indicating the absence of extremely high pairwise linear correlation. These findings suggest that some mixture components exhibit moderate interrelationships and that their associations with compressive strength should be evaluated jointly within a multivariable nonlinear modeling framework. Pearson correlations represent marginal linear associations only and should not be interpreted as isolated causal effects or as complete descriptions of potentially nonlinear and interactive relationships.

Table 2 shows the overall evaluation results of the LOAO scheme. In this scheme, NGPR achieves the best performance among the four models compared, with $\mathrm{RMSE} = 5.6259$, $\mathrm{MAE} = 3.9271$, and $\mathrm{R}^2 = 0.8865$. These values are slightly better than GPR-CI, which produces $\mathrm{RMSE} = 5.7086$, $\mathrm{MAE} = 4.0063$, and $\mathrm{R}^2 = 0.8831$. Meanwhile, SVR and KNN show a more obvious performance decline, with RMSE of 8.1760 and 9.7415, respectively. In terms of prediction intervals, NGPR achieved coverage of 0.9350 with an average interval length of 21.2339, while GPR-CI produced a shorter interval, namely 16.8252, but coverage dropped to 0.8379. The results in Table 2 show that in this most stringent evaluation, NGPR provides a better balance between point prediction accuracy and interval coverage.

Table 2. Overall performance of the concrete compressive strength prediction model in the LOAO scheme.

Model	RMSE	MAE	R^2	Coverage	AIL
NGPR	5.625882979	3.927106634	0.886479325	0.934951456	21.23388036
GPR-CI	5.708635369	4.006284048	0.883115161	0.837864078	16.82524068
SVR	8.176031413	5.774550776	0.760238924	0.831067961	21.10809974
KNN	9.741517809	7.646083052	0.65963348	0.842718447	27.80885379

Table 3 shows that the LOAO results are not completely uniform across age groups. At 3 days, GPR-CI slightly outperforms NGPR, with an RMSE of 3.9543 versus 4.1322 and an R^2 of 0.8380 versus 0.8231. At 7 days, the two GPR models remain relatively close, while SVR and KNN are weaker. At 14 days, NGPR again outperforms SVR and KNN, although the difference with respect to GPR-CI

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

is not significant. The 28-day age group, containing 425 observations, shows that NGPR remains superior with an RMSE of 7.3793 and an R^2 of 0.7478, while the other models produce larger errors. At 56 days, GPR-CI slightly outperforms NGPR. At 90 days, GPR-CI also produces a smaller RMSE. Interestingly, at 180 days, KNN actually produces the highest R^2, at 0.8271. The pattern in Table 3 shows that in the LOAO scheme, model performance is highly influenced by the age excluded from the training data.

Table 3. Performance of concrete compressive strength prediction model according to curing age in the LOAO scheme.

Age	Model	n	RMSE	MAE	R^2	Coverage	AIL
1	NGPR	2	6.400516256	6.343538972	-3.03745553	1	22.50918255
1	GPR-CI	2	7.519051348	7.499915119	-4.57190679	1	17.70379403
1	SVR	2	8.859054561	8.826766236	-6.73486066	1	20.11373996
1	KNN	2	18.91734521	17.59533536	-34.2694039	0.5	21.32541504
3	NGPR	134	4.132208113	3.051034952	0.823140308	1	24.99697648
3	GPR-CI	134	3.954289854	2.992640319	0.83804234	0.992537313	21.5595886
3	SVR	134	11.82054693	10.74994536	-0.44723692	0.686567164	25.37258378
3	KNN	134	12.3938025	10.97089942	-0.59101265	0.686567164	27.88975092
7	NGPR	126	4.694078886	3.004009566	0.895562246	0.952380952	17.77026331
7	GPR-CI	126	4.719718925	3.092082261	0.894418208	0.888888889	14.84231081
7	SVR	126	5.194697675	3.682498143	0.872098007	0.952380952	21.64695142
7	KNN	126	8.066518471	6.329659075	0.691589595	0.928571429	29.82957525
14	NGPR	62	3.601492561	2.338695673	0.823324116	0.935483871	15.21554608
14	GPR-CI	62	3.714520134	2.467831452	0.812060676	0.983870968	24.09361069
14	SVR	62	5.596681991	4.089793951	0.573348291	0.935483871	26.50410707
14	KNN	62	7.005962587	5.720680464	0.331428506	0.951612903	27.06714355
28	NGPR	425	7.379342969	5.541208422	0.747790392	0.875294118	22.55796799
28	GPR-CI	425	7.462315962	5.608726656	0.742086839	0.670588235	13.97951047
28	SVR	425	9.470698955	6.51654082	0.584577227	0.741176471	17.40244768
28	KNN	425	11.12693054	8.783810373	0.426574515	0.795294118	28.37514101
56	NGPR	91	3.221612006	2.499508566	0.948742494	0.978021978	15.85505454
56	GPR-CI	91	3.180838439	2.539028575	0.95003174	1	18.17260401
56	SVR	91	3.823191862	2.847826368	0.927812303	0.989010989	25.53891775
56	KNN	91	6.748646595	5.197067696	0.775071559	0.978021978	31.02860142
90	NGPR	54	3.153289236	2.474682148	0.894911879	1	21.59861262
90	GPR-CI	54	2.969166158	2.24123507	0.906825941	1	19.32221929
90	SVR	54	4.459627234	3.13236889	0.789804639	0.962962963	21.1863094
90	KNN	54	8.617634442	6.921105357	0.21512202	0.759259259	24.51574625
91	NGPR	22	2.737659695	2.180084417	0.867486536	1	19.66742811
91	GPR-CI	22	4.758917957	3.082166707	0.599578379	0.954545455	19.45172484
91	SVR	22	4.240409697	3.669158899	0.682080826	1	21.49752152
91	KNN	22	4.83279205	4.248715643	0.587050162	1	22.81423444
100	NGPR	52	3.611419687	2.859559809	0.811615935	1	20.46160645
100	GPR-CI	52	4.088141321	3.241494292	0.758598457	0.942307692	15.15699509
100	SVR	52	4.215264474	3.531477459	0.743351992	1	25.76873389
100	KNN	52	6.810538797	5.823809469	0.330036436	0.942307692	24.88612772
120	NGPR	3	2.205798157	1.615609408	-4.99847246	1	14.37497339
120	GPR-CI	3	2.163190842	1.568814044	-4.76897692	1	17.98186435
120	SVR	3	1.991846265	1.656231849	-3.89126014	1	19.03680728
120	KNN	3	1.766010558	1.627551376	-2.84499449	1	28.79251776
180	NGPR	26	4.975081908	3.349038748	0.784516049	0.923076923	23.50176905
180	GPR-CI	26	4.809464801	3.177578452	0.798623884	0.884615385	20.41129718
180	SVR	26	6.683170419	5.056489368	0.611152303	0.884615385	19.54922251

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

Age	Model	n	RMSE	MAE	R^2	Coverage	AIL
180	KNN	26	4.456103032	3.400025481	0.8271279	1	21.99888091
270	NGPR	13	4.757151418	3.001782079	0.783632525	1	25.25367921
270	GPR-CI	13	4.955724042	3.442736112	0.765192342	0.846153846	20.74053654
270	SVR	13	7.717243649	4.91143938	0.43059346	0.846153846	21.8554607
270	KNN	13	5.082671417	3.787306578	0.75300845	0.923076923	20.32563229
360	NGPR	6	2.521886956	2.06945403	0.560944722	1	27.81737742
360	GPR-CI	6	2.6864393	2.372149648	0.501778992	1	22.24125934
360	SVR	6	4.471023338	4.351320496	-0.38000876	1	23.13387127
360	KNN	6	4.283804117	4.070400486	-0.26685578	1	24.62128823
365	NGPR	14	3.696772201	3.12623899	0.840973889	1	32.25250892
365	GPR-CI	14	4.526943186	3.52761279	0.761530332	0.857142857	14.24686546
365	SVR	14	7.169589991	5.964053642	0.401847807	0.714285714	18.93156876
365	KNN	14	6.167385349	5.350879898	0.557385917	0.928571429	25.24024364

Overall, LOAO provides a fairly rigorous evaluation because the model was tested on age groups not used in training. Substantively, these results indicate that GPR-based models are more adaptable to shifts in the curing age structure, while SVR and especially KNN are more prone to accuracy losses when certain age groups are completely removed from the training process. Thus, the results in Tables 2 and 3 indicate that NGPR provides the most consistent overall performance under the stringent LOAO generalization scheme.

The next evaluation was conducted using repeated stratified K-fold by Age. Unlike LOAO, this scheme provides a more empirically stable performance because all age groups are still represented in the fold distribution. Table 4 shows that in the overall evaluation, NGPR is again the best model with $\text{RMSE} = 4.2931$, $\text{MAE} = 2.8890$, and $R^2 = 0.9339$. This result is better than GPR-CI, which has $\text{RMSE} = 4.5314$ and $R^2 = 0.9264$, and were considerably better than the results obtained by SVR and KNN. In terms of prediction interval, NGPR also produces the shortest interval, namely 16.2586, with $\text{coverage} = 0.9437$. In comparison, GPR-CI has the widest coverage, 0.9631, but it comes with a wider interval, i.e., 23.3587. SVR and KNN still have longer gaps with more prediction errors. The results showed that overall, NGPR was the most efficient, but GPR-CI was more conservative in maintaining interval coverage.

Table 4. Overall performance of the concrete compressive strength prediction model in the repeated stratified K-fold by Age scheme.

Model	RMSE	MAE	R^2	Coverage	AIL
NGPR	4.293053597	2.889005007	0.933896235	0.94368932	16.25859249
GPR-CI	4.531445577	3.066529298	0.926350956	0.963106796	23.35873683
SVR	5.272841968	3.283625019	0.900279812	0.957281553	26.58946668
KNN	6.084619556	4.164952494	0.867211509	0.949514563	27.38362591

The age-specific results in Table 5 are largely consistent with the overall performance pattern reported in Table 4. NGPR resulted in an RMSE of 3.4744 and an R^2 of 0.8750 at 3 days, which was somewhat better than GPR-CI. NGPR fared better again at 7 days, with an RMSE of 4.6671 and an R^2 of 0.8968. At 14 days, GPR-CI produced a lower RMSE of 3.3742 than NGPR, which produced an RMSE of 3.5778. The most representative findings were observed in the 28-day age group (425 observations): NGPR achieved an RMSE of 5.0611, an MAE of 3.3804, and an R^2 of 0.8814, the best among all models. But the coverage of this model was 0.9129, slightly below the theoretical 95% objective, with an interval length of 16.1021. SVR obtains $\text{coverage} = 0.9624$ and an interval length of 29.3704, whereas GPR-CI is average with $\text{coverage} = 0.9412$ and an interval length of 24.5457. Hence, there is a trade-off between the sharpness and the coverage of the interval at 28 days.

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 3, September 2026, pp. 215~230
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i3.31258

Table 5. Performance of concrete compressive strength prediction model according to curing age in the repeated stratified K-fold by Age scheme.

Age	Model	n	RMSE	MAE	R^2	Coverage	AIL
1	NGPR	2	5.836028421	5.790401551	-2.35670007	1	22.37556045
1	GPR-CI	2	6.334354751	6.213020728	-2.95441738	1	22.45900519
1	SVR	2	11.46222335	11.46039498	-11.9483811	0.5	21.42375693
1	KNN	2	17.83164518	16.69141132	-30.3372274	0.5	25.14554067
3	NGPR	134	3.474448142	2.546921972	0.874963687	0.97761194	17.00544122
3	GPR-CI	134	3.579946096	2.578511439	0.867255215	0.985074627	19.44695182
3	SVR	134	4.733539701	3.245754621	0.767920628	0.910447761	19.50417411
3	KNN	134	6.608379495	4.854100708	0.54767099	0.962686567	31.80203254
7	NGPR	126	4.667082205	2.904653906	0.896760081	0.936507937	15.6272964
7	GPR-CI	126	4.913944815	3.088551887	0.885549602	0.976190476	32.79586784
7	SVR	126	5.254090889	3.236714138	0.869156571	0.952380952	32.7412578
7	KNN	126	6.022365915	3.933360646	0.828094	0.952380952	26.85072975
14	NGPR	62	3.577792482	2.641966387	0.825641742	0.935483871	14.66353836
14	GPR-CI	62	3.374205788	2.517562768	0.844920154	0.983870968	20.73363325
14	SVR	62	3.589837346	2.477996558	0.824465789	0.983870968	23.76332447
14	KNN	62	5.515843742	4.249832857	0.58558436	0.983870968	25.43800489
28	NGPR	425	5.061110032	3.380401153	0.881363645	0.912941176	16.10205715
28	GPR-CI	425	5.311994609	3.581071408	0.869310262	0.941176471	24.54569296
28	SVR	425	6.082973928	3.617408947	0.828620786	0.962352941	29.37040572
28	KNN	425	6.395414353	4.044290243	0.810563522	0.929411765	28.11547887
56	NGPR	91	3.439117467	2.653613515	0.941587604	0.967032967	15.2008838
56	GPR-CI	91	3.898021172	2.950704614	0.924958866	0.956043956	17.79923511
56	SVR	91	4.330422081	3.271502315	0.907387105	0.945054945	23.76336901
56	KNN	91	5.460526549	4.357173724	0.85274161	0.956043956	23.70487497
90	NGPR	54	2.847551926	1.866293	0.914302271	1	17.15167246
90	GPR-CI	54	3.297501139	2.131285725	0.88507992	0.962962963	20.84762437
90	SVR	54	2.744053578	2.005837192	0.920418672	1	23.70412175
90	KNN	54	5.103265915	3.842006267	0.724753119	0.981481481	25.35984981
91	NGPR	22	3.509094668	2.376101779	0.782283546	0.954545455	18.28407647
91	GPR-CI	22	3.681874406	2.634174001	0.760316011	1	20.65284691
91	SVR	22	5.39793418	3.841180073	0.484823253	0.954545455	23.2646945
91	KNN	22	4.682392867	2.995859796	0.612352683	1	25.26992619
100	NGPR	52	3.924832743	3.040242893	0.777499726	0.961538462	15.7749814
100	GPR-CI	52	3.994584019	3.248920883	0.769521	1	20.82891111
100	SVR	52	4.179880465	3.256219067	0.747642646	0.980769231	23.55110862
100	KNN	52	6.167200713	4.912548973	0.450630674	0.942307692	25.40804999
120	NGPR	3	1.98611982	1.492976586	-3.86317638	1	13.9404428
120	GPR-CI	3	1.958097511	1.555201445	-3.72691465	1	19.63350878
120	SVR	3	1.767124435	1.289921819	-2.84984633	1	25.11116653
120	KNN	3	1.848550484	1.826097033	-3.21280892	1	25.60055682
180	NGPR	26	2.021643391	1.55180396	0.964418557	1	17.72870964
180	GPR-CI	26	3.423724816	2.142046056	0.897950199	0.961538462	20.58941814
180	SVR	26	2.89015598	2.064527797	0.927279472	1	23.49514001
180	KNN	26	4.432033845	3.342881841	0.828990359	1	25.54046688
270	NGPR	13	2.170557896	1.775340683	0.954955603	1	18.95099663
270	GPR-CI	13	2.217977519	1.804021656	0.952965958	1	20.50289229
270	SVR	13	9.840905892	4.517163429	0.074091258	0.923076923	23.97657896
270	KNN	13	5.233875221	3.450071467	0.738094417	0.923076923	25.46055185
360	NGPR	6	1.061546727	0.899888944	0.922205982	1	21.67988276
360	GPR-CI	6	1.230239436	0.949446992	0.895516609	1	20.00845415

Age	Model	n	RMSE	MAE	R^2	Coverage	AIL
360	SVR	6	1.725376888	1.336416609	0.794488597	1	23.02493109
360	KNN	6	3.347220707	2.946722351	0.226542115	1	25.94685528
365	NGPR	14	2.6839525	2.172706781	0.916175165	1	20.72053718
365	GPR-CI	14	2.982361465	2.383556917	0.896499221	1	21.08738729
365	SVR	14	3.801746221	2.948636112	0.831814208	1	23.89817623
365	KNN	14	4.48844503	3.42108774	0.765569086	1	25.19213396

At intermediate ages, the dominance of NGPR remains quite clear. At 56 days, the model yields an RMSE of 3.4391, an R^2 of 0.9416, and a mean interval of 15.2009. At 90 days, there is an interesting exception: the SVR yields the best results for point predictions in the previous analysis, but overall, the GPR family remains more stable. At 91 and 100 days, NGPR again yields the best or most efficient results. Meanwhile, at later ages, such as 180, 270, 360, and 365 days, NGPR also remains strong, although some small-sample groups require caution in interpretation. The 1-day, 120-day, and 360-day age groups, for example, have very limited observations, making the R^2 values less stable. Therefore, the main conclusions are more appropriately emphasized on age groups with sufficient data, especially 3, 7, 14, 28, 56, 90, 100, and 180 days.

B. DISCUSSION

The results obtained in the previous section reveal that predicting concrete compressive strength from this data cannot be described by a simple linear relationship. The curing age is certainly important, but it is not independent of other factors. The pattern of age-compressive strength is more likely to be a gradual, nonlinear relationship rather than a clean, rising line from start to finish, depending on the mix composition, water-cement ratio, slag content, fly ash, and interactions among materials. This has long been noted in the data-driven concrete literature, and more recent studies have shown the same thing: age, water/cement ratio, slag, and water continue to emerge as the most influential variables when more flexible and interpretable models are used.

In line with the character of the relationship, the Matérn 3/2 kernel is the best GPR kernel in this study because it is most often selected in the inner cross-validation, namely 92.857% in the LOAO scheme and 100% in the repeated stratified K-fold by Age scheme. This finding indicates that the relationship among mix composition, curing age, and concrete compressive strength is better modeled with a flexible kernel and exhibits a moderate level of smoothness.

In that context, NGPR's comparatively consistent performance across the two evaluation schemes is consistent with both its probabilistic formulation and the empirical results. As a Gaussian process model, NGPR is flexible in representing nonlinear relationships and naturally provides a predictive distribution, allowing point-prediction accuracy and prediction uncertainty to be evaluated jointly (Rasmussen & Williams, 2006; Miao et al., 2025; Xie et al., 2025). This advantage is particularly relevant for concrete compressive-strength data, in which the relationships among mixture composition, curing age, and strength are complex and nonlinear. The LOAO scheme provides a stringent generalization assessment because the model must predict an entire curing-age group that is not represented in the training data. Under this setting, NGPR achieved empirical coverage of 0.9350 with an average interval length of 21.2339, whereas GPR-CI produced a shorter average interval length of 16.8252 but substantially lower coverage of 0.8379. These results indicate that NGPR provided a better empirical balance between point-prediction performance, interval calibration, and interval efficiency in the LOAO analysis. Thus, a narrower prediction interval should not be considered superior when it is accompanied by substantial undercoverage.

The same pattern is again observed in the repeated stratified K-fold analysis by Age. However, the results appear more stable because each fold still contains representatives from all age groups. In this evaluation, NGPR again demonstrated the most efficient performance with an average interval length of 16.2586. On the other hand, GPR-CI provided higher coverage (0.9631) but wider intervals (23.3587). These results demonstrate a common trade-off in evaluating prediction intervals. High coverage is usually associated with longer intervals, while too narrow intervals tend to reduce coverage. Therefore, the best model should not be judged solely by the shortest interval or the highest coverage, but by its ability to balance calibration and interval sharpness. In these results, this balance is more often demonstrated by NGPR.

The 28-day age group is of exceptional interest because it has the largest number of observations. Also, in civil engineering practice, the 28-day age is often selected as the main reference for determining concrete compressive strength. In this group, NGPR has the shortest interval (16.1021), but its coverage is 0.9129, just below the nominal 95% limit. In contrast, SVR produces a much wider interval, at 29.3704, while GPR-CI falls in the middle with an interval length of 24.5457. These results indicate that at the 28-day age, the issue is not simply choosing the most accurate model, but rather choosing one that strikes a reasonable balance between predictive efficiency and inferential caution. From an application perspective, these findings provide a fairly clear direction. If the primary goal is to obtain efficient yet informative predictions, then NGPR is a better priority. However, if more emphasis is placed on more conservative intervals, GPR-CI remains a reasonable choice.

From an industrial perspective, the proposed prediction framework can support preliminary concrete mix screening, quality-control planning, and the prioritization of specimens requiring confirmatory laboratory testing. By providing both predicted compressive strength and the associated prediction interval, the framework enables engineers to distinguish between mixtures that are likely to satisfy the required strength, mixtures that are unlikely to meet the requirement, and mixtures whose performance remains uncertain. Laboratory resources can therefore be directed more efficiently toward mixtures with high predictive uncertainty or predicted strength close to the specified threshold. In this context, the model may reduce unnecessary preliminary and repetitive laboratory testing, shorten the initial mix-evaluation process, and support more efficient use of materials, testing equipment, and technical personnel. Nevertheless, the model is intended as a decision-support tool rather than a replacement for mandatory standardized compressive-strength testing, particularly for final mix approval, regulatory compliance, and safety-critical structural decisions. Moreover, because the present study did not include a direct cost-effectiveness analysis or prospective industrial implementation, actual reductions in laboratory testing costs have not yet been empirically demonstrated.

Another important consideration is how to interpret the results for age groups with small data sets. At 1 day, 120 days, and 360 days, for example, the limited sample size makes performance measures more susceptible to change. Under such conditions, a model that appears superior in one age group may not necessarily be superior in general. Therefore, the main conclusions are more appropriate for age groups supported by more adequate data sets, specifically 3, 7, 14, 28, 56, 90, 100, and 180 days. Based on this, the results of this study generally indicate that the GPR family, particularly NGPR, is better able to adapt to the heterogeneity of concrete compressive strength data across various curing ages than SVR and KNN. This finding aligns with recent literature emphasizing that concrete compressive strength modeling should combine nonlinear flexibility, rigorous evaluation, and well-explained interpretation.

4. CONCLUSION

The analysis results show that the compressive strength of concrete at various curing ages exhibits a non-linear, inhomogeneous pattern. The response distribution tends to be right-skewed; variations across age groups differ, and the relationship between curing age and compressive strength is better understood as a gradual pattern rather than a single linear relationship. The condition emphasizes that predicting concrete compressive strength requires a flexible model and an evaluation design that explicitly considers the curing age structure.

The GPR family showed more consistent overall performance than SVR and KNN across both evaluation schemes. Under LOAO, NGPR achieved the lowest RMSE (5.6259) and MAE (3.9271), the highest R^2 (0.8865), and empirical coverage of 0.9350 with an AIL of 21.2339. In comparison, GPR-CI produced a slightly higher RMSE (5.7086) and a shorter AIL (16.8252), but its coverage decreased to 0.8379, while SVR and KNN yielded substantially higher RMSE values of 8.1760 and 9.7415, respectively. Under repeated age-stratified K-fold cross-validation, NGPR again achieved the lowest RMSE (4.2931) and MAE (2.8890), the highest R^2 (0.9339), and the shortest AIL (16.2586), while maintaining coverage of 0.9437. GPR-CI achieved slightly higher coverage (0.9631), but with a wider AIL (23.3587). These findings indicate that NGPR provides the most favorable overall balance among point-prediction accuracy, interval calibration, and interval sharpness, whereas GPR-CI provides more conservative intervals under repeated age-stratified validation.

Further work could explore the use of external data, more balanced distributions of curing age, and comparisons with other uncertainty techniques, such as more adaptive conformal models, bootstrapping, or richer Bayesian models. Extension to more particular concrete kinds or to shortfall risk analysis is also crucial to provide broader relevance of the results to engineering contexts.

ACKNOWLEDGEMENTS

This research was funded by the PNBP Fund of Universitas Negeri Medan through the 2023 Applied Research Scheme, Contract Number: 0038/UN33.8/PPKM/PPT/2023.

REFERENCES

Almutairi, A. (2025). Explainable machine learning-based prediction of compressive strength in sustainable recycled aggregate self-compacting concrete using SHAP analysis. *Sustainability, 17*(24), 11334. https://doi.org/10.3390/su172411334

Angiulli, G., Calcagno, S., La Foresta, F., & Versaci, M. (2024). Concrete compressive strength prediction using combined non-destructive methods: A calibration procedure using preexisting conversion models based on Gaussian process regression. *Journal of Composites Science, 8*(8), 300. https://doi.org/10.3390/jcs8080300

Beskopylny, A. N., Stel'makh, S. A., Shcherban', E. M., Mailyan, L. R., Meskhi, B., Razveeva, I., Chernil'nik, A., & Beskopylny, N. (2022). Concrete strength prediction using machine learning methods CatBoost, k-nearest neighbors, support vector regression. *Applied Sciences, 12*(21), Article 10864. https://doi.org/10.3390/app122110864

Dong, S., & Zhang, Z. (2025). Hybrid deep learning with conformal prediction for recycled aggregate self-compacting concrete strength prediction. *Buildings, 15*(24), 4419. https://doi.org/10.3390/buildings15244419

Elhishi, S., Elashry, A. M., & El-Metwally, S. (2023). Unboxing machine learning models for concrete strength prediction using XAI. *Scientific Reports, 13*, 19892. https://doi.org/10.1038/s41598-023-47169-7

Feng, D.-C., Liu, Z.-T., Wang, X.-D., Chen, Y., Chang, J.-Q., Wei, D.-F., & Jiang, Z.-M. (2020). Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. *Construction and Building Materials, 230*, 117000. https://doi.org/10.1016/j.conbuildmat.2019.117000

Fu, H., Zhou, X., Xu, P., & Sun, D. (2025). Prediction of compressive strength of concrete using explainable machine learning models. *Materials, 18*(21), 5009. https://doi.org/10.3390/ma18215009

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Kaloop, M. R., Kumar, D., Samui, P., Hu, J. W., & Kim, D. (2020). Compressive strength prediction of high-performance concrete using gradient tree boosting machine. *Construction and Building Materials, 264*, 120198. https://doi.org/10.1016/j.conbuildmat.2020.120198

Liu, G., & Sun, B. (2023). Concrete compressive strength prediction using an explainable boosting machine model. *Case Studies in Construction Materials, 18*, e01845. https://doi.org/10.1016/j.cscm.2023.e01845

Miao, Q., Gao, Z., Zhu, K., Guo, Z., Sun, Q., & Zhou, L. (2025). Interpretable machine learning model for compressive strength prediction of self-compacting concrete with recycled concrete aggregates and SCMs. *Journal of Building Engineering, 108*, 112965. https://doi.org/10.1016/j.jobe.2025.112965

Mobasheri, F., Hosseinpoor, M., Yahia, A., & Pourkamali-Anaraki, F. (2025). Predicting the compressive strength of self-consolidating concrete using machine learning and conformal inference. *Computer Modeling in Engineering & Sciences, 145*(3), 3309–3347. https://doi.org/10.32604/cmes.2025.072271

Olvera-Mayorga, C. E., López-Martínez, M. d. J., Rodríguez-Rodríguez, J. A., Vázquez-Reyes, S., Solís-Sánchez, L. O., de la Rosa-Vargas, J. I., Duarte-Correa, D., González-Aviña, J. V., & Olvera-Olvera, C. A. (2025). AI-based inference system for concrete compressive strength: Multi-dataset analysis of optimized machine learning algorithms. *Applied Sciences, 15*(23), Article 12383. https://doi.org/10.3390/app152312383

Rana, S., Midi, H., & Imon, A. H. M. R. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering, 2012*, Article 730328. https://doi.org/10.1155/2012/730328

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press. https://doi.org/10.7551/mitpress/3206.001.0001

Shaaban, M., Amin, M., Selim, S., & Riad, I. M. (2025). Machine learning approaches for forecasting compressive strength of high-strength concrete. *Scientific Reports, 15*, 25567. https://doi.org/10.1038/s41598-025-10342-1

Simamora, E., Mansyur, A., Firdaus, M., & Habibi, R. (2025). Residual bootstrap analysis on multiple linear regression: Case study of concrete compressive strength. *AIP Conference Proceedings, 3316*(1), 040003. https://doi.org/10.1063/5.0290027

Tamuly, P., & Nava, V. (2025). Machine learning based conformal predictors for uncertainty-aware compressive strength estimation of concrete. *Construction and Building Materials, 487*, 141844. https://doi.org/10.1016/j.conbuildmat.2025.141844

Xiao, K., Zhang, H., Wei, S., Zhu, C., He, J., Zhu, S., & Yang, X. (2025). Prediction of compressive strength of high-performance concrete based on multiple machine learning models. *Frontiers in Materials, 12*, 1698248. https://doi.org/10.3389/fmats.2025.1698248

Xie, L., He, J., Wang, C., Zhu, S., Zhang, H., & Yang, X. (2025). Enhanced Bayesian Gaussian process regression for compressive strength prediction of multi-binder concrete. *Journal of Building Engineering, 111*, Article 113308. https://doi.org/10.1016/j.jobe.2025.113308

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research, 28*(12), 1797–1808. https://doi.org/10.1016/S0008-8846(98)00165-3

Yeh, I.-C. (2006). Analysis of strength of concrete using design of experiments and neural networks. *Journal of Materials in Civil Engineering, 18*(4), 597–604. https://doi.org/10.1061/(ASCE)0899-1561(2006)18:4(597)

Zhang, Y., Ren, W., Chen, Y., Mi, Y., Lei, J., & Sun, L. (2024). Predicting the compressive strength of high-performance concrete using an interpretable machine learning model. *Scientific Reports, 14*, 28346. https://doi.org/10.1038/s41598-024-79502-z