Journal of Science and Mathematics Letters
ISSN 2462-2052 • eISSN 2600-8718

*Research Article*

# K-Nearest-Neighbours-Based Fuzzy Regression for Modelling Heterogeneous Strength Development in Cement-Based Materials

**Hazmira Yozza[1,2], Riswan Efendi[1,3]*, Nor Azah Samat[1], Izzati Rahmi[1,2], Ridho Saputra[2], Asko Mononen[4], Petrônio Cândido de Lima e Silva[5]**

1 Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
2 Department of Mathematics and Data Science, Faculty of Mathematics and Natural Sciences, Universitas Andalas, 25163, Padang, Indonesia
3 Department of Mathematics, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, 28293, Pekanbaru, Indonesia
4 Digital Living Lab, Laurea University of Applied Sciences, Espoo, 02650, Finland
5 Federal Institute of Northern Minas Gerais (IFNMG), Januária, Minas Gerais, Brazil
* Corresponding author: riswanefendi@fsmt.upsi.edu.my

**ARTICLE HISTORY**

**Received**
2 March 2026
**Revised**
20 June 2026
**Accepted**
5 July 2026
**Published**
25 July 2026

**KEYWORDS**

Compressive strength
Integrated KNN fuzzy regression
Possibilistic fuzzy regression
Predictive performance
Triangular fuzzy number

**ABSTRACT**

Compressive strength is a key indicator of cement performance. The 28-day compressive strength is a benchmark for compressive strength development. In most studies, compressive strength data are treated as precise values. However, compressive strength measurements inherently involve uncertainty due to experimental variability, material heterogeneity, and testing conditions. This uncertainty suggests that compressive strength may be appropriately represented as a fuzzy number and analysed using fuzzy regression. This study aims to investigate the use of integrated k-nearest-neighbour fuzzy regression (I-KNNFR) to predict the triangular fuzzy number of the 28-day compressive strength of cement mortar. For each prediction point, the method constructs a possibilistic fuzzy regression model using locally similar observations identified through the nearest-neighbour mechanism. The I-KNNFR was developed to address issues in possibilistic fuzzy regression related to global modelling, outliers, and over-constrained problems. Its predictive performance was evaluated using 10-fold range-preserved controlled cross-validation and compared with that of possibilistic fuzzy regression based on mean squared error (MSE) and mean relative distance (MRD). The best predictive performance was achieved when I-KNNFR used nine nearest neighbours and the Euclidean distance. Compared with possibilistic fuzzy regression, I-KNNFR reduced MSE and MRD by approximately 85% and 83%, respectively, indicating a substantial improvement in predictive performance. In conclusion, I-KNNFR more effectively captures local strength-development patterns through fuzzy modelling. It outperforms possibilistic fuzzy regression in predicting the triangular fuzzy number of the compressive strength. By using locally selected neighbours, it also represents the uncertainty associated with 28-day compressive-strength prediction.

https://doi.org/10.37134/jsml.vol14.3.9.2026

## 1. INTRODUCTION

Cement is an essential material for global infrastructure development. According to Wahyuni et al. (2025), the main purpose of cement is to bind aggregates into a solid and durable mass, which is necessary for the construction of buildings, roads, and other infrastructure projects. Compressive strength is an important parameter used in the cement industry to assess cement quality. It measures the maximum load per unit area that a hardened cement paste or concrete specimen can endure before failure under uniaxial compression (Siddique & Mehta, 2014). The compressive strength value indicates the qualities of concrete, encompassing quality and durability (Al Biajawi et al., 2022). This statistic indicates a concrete structure's capacity to endure the applied load. In a cement production facility, there exists a standard for this indicator that all cement factories are required to adhere to. In Indonesia, ordinary Portland cement (OPC) is governed by the standard SNI 2049:2020 (Maryoto et al., 2020).

Routine compressive-strength testing is a vital procedure before cement is sold to the market. This indicator is continuously monitored for every production batch to ensure it remains within an acceptable limits. Testing of compressive strength can be viewed as an estimation of the structure's reliability. This is conducted as part of the quality control procedure (Tur & Derechennik, 2018). Monitoring this indicator typically involves assessing the development of compressive strength at the $3^{rd}$, $7^{th}$, and $28^{th}$ days. The $28^{th}$ day serves as a benchmark in compressive strength evaluation because, during the first 28 days, strength develops significantly, as shown by Sajedi (2012) and Mohan et al. (2025). Onjure et al. (2023) also found that ordinary Portland cement concrete (OPC) reaches its maximum compressive strength after 28 days of hydration. Theoretically, cement's 28 day compressive strength is influenced by the complex interaction of its chemical and physical properties, including the proportions of major oxides (such as $SiO_2$, $Al_2O_3$, $Fe_2O_3$, CaO, MgO and $SO_3$), parameters related to burning efficiency (such as loss on Ignition/LoI) and fineness of cement particles (such as residue on 45-micron sieve). This interaction leads to unavoidable variability in compressive strength, even under controlled laboratory conditions. Therefore, modelling the 28-day compressive strength is important to understand these influences, predict performance consistently, and support quality control and process optimisation in cement manufacturing.

Some studies model the compressive strength of cement or concrete using multiple regression or other machine learning approaches, such as ANN, SVR, KNN-regressor, random forest, response surface, and hybrid artificial intelligence techniques (Safiuddin et al., 2016; Hameed & Al Omar, 2020; Shaaban et al., 2025; Zhang et al., 2025). Most of these studies assume that the relationship between compressive strength and its explanatory variables is deterministic and free of measurement uncertainty. However, in practice, the development of compressive strength may be heterogeneous. Cement specimens with similar chemical and physical characteristics may exhibit different compressive strength values, as observed in laboratory testing. This variation may arise from the inherent features of the hydration process, variations in raw materials, testing conditions, and other factors that may be difficult to observe. Therefore, there is uncertainty in the compressive strength. In this context, the compressive strength may not be sufficiently represented by a single crisp value but may be appropriately expressed as a range of plausible values. In addition, the relationship between cement characteristics and the compressive strength may be heterogeneous across local regions of the input space. For both reasons, a single deterministic regression cannot adequately capture these uncertainties and heterogeneous patterns arising from incomplete knowledge of the system.

In this context, fuzzy regression offers a conceptually and practically appropriate alternative because it explicitly incorporates uncertainty within the modelling framework. In the fuzzy regression framework, variables are represented as fuzzy numbers, one of which is in the triangular form. The fuzzy number is characterized by a membership function that represents the degree of possibility of a specific value. In fuzzy regression, model parameters are also represented as fuzzy numbers. As a result, fuzzy regression provides predictions as fuzzy intervals, indicating not only the most likely strength value but also its potential range, along with corresponding degrees of possibility. For industrial practitioners, including those in the cement industry, such information is more meaningful: for example, a predicted 28-day strength of 50 MPa with a fuzzy interval of [48, 52] MPa provides a clearer sense of tolerance and reliability than a single-point estimate.

Possibilistic fuzzy regression is a widely used fuzzy regression method for capturing uncertainty in both the data and the relationships among variables. In the possibilistic framework, the regression parameters are estimated by solving a linear programming problem that minimizes the total spread of the model parameters. Several issues are encountered with this approach, including the widening the spread as more observations are included and the possibility of degeneracy and multiple optimal solutions. These issues motivate the exploration of a local fuzzy regression formulation based on the information from nearby observations. In this approach, the k-nearest-neighbour (KNN) algorithm serves as a neighbourhood-selection mechanism rather than directly averaging the neighbouring responses. For each prediction point, KNN identifies training observations with the most similar chemical and physical characteristics. These observations are then used to construct a local possibilistic fuzzy regression model. This procedure allows the model to adapt to local strength-development patterns, limits the influence of distant or atypical observations, and reduces the number of observations involved in each linear programming formulation. This approach is referred to as the integrated k-nearest-neighbour fuzzy regression (I-KNNFR).

This study aims to investigate the use of I-KNNFR to predict the triangular fuzzy number for the 28-day compressive strength of cement mortar. In addition, the predictive performance of I-KNNFR is compared with possibilistic fuzzy regression.

## 2. METHODOLOGY

### *2.1. Fuzzy Number*

Suppose *X* is a classical crisp set of objects whose generic element is denoted *x*. The membership of *x* in a subset *A* of *X* is a characteristic function from *X* to a valuation set {0, 1} such that

$$\mu_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{1}$$

(Dubois & Prade, 1980). If the valuation set is a real interval [0,1], *A* is called a *fuzzy set* (Zadeh, 1965).

**Definition 1** (Chaira, 2019)
If X is a collection of objects denoted by x, then a fuzzy set $\tilde{A}$ in X is a set of ordered pairs:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}|(x) x \epsilon X\} \tag{2}$$

$\mu_{\tilde{A}}(x)$ is called the membership function or degree of membership of x in $\tilde{A}$ that maps X to the membership space M. The degree of membership ranges from [0,1], where "0" denotes the absence of that element in $\tilde{A}$ and "1" denotes its complete existence. Intermediate values indicate partial presence of an element in a $\tilde{A}$ (Chaira, 2019).

**Definition 2** (Chaira, 2019)
A fuzzy number $\tilde{A}$ is a fuzzy set on real numbers $\mathbb{R}$, such that $\tilde{A}$ is piecewise continuous, convex, is normal, monotone ascending in the interval $(-\infty, m)$, and monotone descending in the interval $(m, \infty)$.

**Definition 3** (Bector & Chandra, 2005)
A fuzzy number $\tilde{A}$ is called a triangular fuzzy number (TFN) if its membership function $\mu_{\tilde{A}}(x)$ is given by:

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & , \ x < a_l \ or \ x > a_u \\ \dfrac{x - a_l}{a - a_l} & , \ a_l \le x \le a \\ \dfrac{a_u - x}{a_u - a} & , \ a < x \le a_u \end{cases} \tag{3}$$

The triplet $\tilde{A} = (a_l, a, a_u)$ is usually used to denote the TFN $\tilde{A}$. Let $c_L = (a - a_l)$ and $c_R = (a_u - a)$ are the left and right spreads of TFN $\tilde{A}$. If $c_L = c_R = c$, the TFN is said to be symmetrical, otherwise asymmetrical. A symmetrical TFN is commonly presented by $\tilde{A} = (a, c)$ with $c$ is the spread of symmetrical TFN.

### *2.2. Fuzzy Regression*

Ordinary least squares (OLS) Regression is a basic and powerful statistical method in data analytics. It continues to be widely applied in recent scientific studies, including earth geophysics, biological sciences, chemical sciences, environmental sciences, health sciences, engineering, marketing, social sciences, and education (Zakaria et al., 2023; Lam et al., 2023; Kalwa et al., 2023; Sunar et al., 2024; Gupta et al., 2024; Asanre et al., 2024; Rezaei, 2025; Jia, 2025; Nelson et al., 2026).

Using this analysis, data are treated as precise measurements. In many real-world problems, information is often fuzzy, imprecise, or even described linguistically. It is also often found that the relationships among variables are unclear. In this situation, single-value observations are inadequate, and the OLS regression becomes limited. Fuzzy regression was developed to model the relationship between a response variable and a set of explanatory variables in a fuzzy environment. The basic model of fuzzy regression is:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + \tilde{A}_2 X_2 + \cdots + \tilde{A}_q X_q \tag{4}$$

where $X_j$s are explanatory variables that may be crisp (nonfuzzy) or fuzzy. $\tilde{A}_j$s are fuzzy regression coefficients which are expressed as a fuzzy number. Accordingly, the model response, $\tilde{Y}$, is also a fuzzy number.

Two general approaches can be used to estimate the coefficients of a fuzzy regression model: the possibilistic and the least-squares approaches. Possibilistic regression was first proposed by Tanaka et al. (1982). This method integrates regression concepts with fuzzy set theory introduced by Zadeh (1965).

Using possibilistic regression, it is assumed that $\tilde{A}_j$ are symmetrical TFNs expressed as $\tilde{A}_j = (a_j, c_j)$. As a consequence, $\tilde{Y}_i = (m_i, w_i)$ $(i = 1, 2, \dots, n)$ with mode and spread are $m_i = \sum_{j=0}^{q} a_j\, x_{ij}$ and $w_i = c_0 + \sum_{j=1}^{q} c_j\, |x_{ij}|$, respectively. The fuzzy coefficients are estimated by solving a linear programming problem aimed at minimising the total spread of all coefficients, which is formulated as

$$Z = \min nc_o + \sum_{i=1}^{n} \sum\nolimits_{j=1}^{q} c_j |x_{ij}|$$

subject to

$$\begin{aligned}
&\left(a_0 + \sum\nolimits_{j=1}^{q} a_j x_{ij}\right) - (1-h)\left(c_0 + \sum\nolimits_{j=1}^{q} c_j |x_{ij}|\right) \le (y_i - (1-h)e_i \\
&\left(a_0 + \sum\nolimits_{j=1}^{q} a_j x_{ij}\right) + (1-h)\left(c_0 + \sum\nolimits_{j=1}^{q} c_j |x_{ij}|\right) \ge (y_i + (1-h)e_i \\
&c_j \ge 0 \\
&i = 1, 2, \dots, n; j = 0, 1, \dots, q
\end{aligned} \tag{5}$$

where $y_i$ and $e_i$ are mode and spread of the $i$th observed TFN, respectively; and $h$ is $h$-certain factor. This linear programming solution is determined using the simplex method.

### *2.3. k-Nearest Neighbour Regressor*

The *k*-nearest neighbours (KNN) regression is a supervised machine learning technique typically used to predict the response for a new observation. KNN regression infers a function (learner) from a collection of training samples called a training dataset that consists of pairs of an input vector and a desired response value. This approach works by exploring the entire training dataset to identify the *k* most similar instances, which serve as the basis for predicting the output of a new observation (Song et al., 2017; Kohli et al., 2021).

Let $X = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$ be a training dataset that is formed by *N* samples. Each sample $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{iq})$ is a vector of explanatory variables that contains *q* features, and $y$ denotes the response variable. Given a new query sample denoted as ***z***. The predicted response value for ***z*** is determined by identifying the *k* most similar observations to ***z***. These observations define a set of ***z***'s nearest neighbours, which are used in the subsequent step to predict *z*'s response. The predicted response value for ***z*** is computed as a weighted-average of the responses of its selected nearest neighbours. In general, the weight assigned to the *i*th observation is inversely proportional to the distance between z and the *i*th observation (Kumbure et al., 2020; Fathabadi et al., 2022).

The distance metric is crucial to the KNN algorithm. The Minkowski distance is a widely used metric for measuring the closeness of two observations. Suppose the *i*th and *j*th observations are denoted as $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, respectively, where $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \dots, x_{iq}) \in \mathbb{R}^q$ and $\boldsymbol{X}_j = (x_{j1}, x_{j2}, \dots, x_{jq}) \in \mathbb{R}^q$. The Minkowski distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is defined as:

$$d_{MD}(X_i, X_j) = \left(\sum_{l=1}^{q} \left|x_{il} - x_{jl}\right|^p\right)^{\frac{1}{p}} \qquad p \geq 1 \tag{6}$$

Different distance measures are obtained by setting different *p*. For instance, the Euclidean distance can be considered as the Minkowski distance with *p* = 2.

### *2.4. Integrated KNN Fuzzy Regression*

Using possibilistic fuzzy regression, regression coefficients are estimated by solving the linear programming problem formulated in Eq. (5). Increasing the number of data points increases the number of constraints, potentially leading to an over-constrained problem. To address this issue, the number of observations involved in the linear programming formulation must be reduced. A strategy is to remove inconsistent data using rough set theory, as suggested by Rahmi et al. (2025). For the same purpose, this study introduced the integrated *k*-nearest neighbour fuzzy regression (I-KNNFR). This approach integrates Tanaka's possibilistic fuzzy regression with the k-nearest neighbours algorithm.

Suppose a training dataset consists of $n$ tuples of data, $(\boldsymbol{X_i}, \tilde{Y}_i);\ i = 1, 2, \dots, n$ where $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \dots, X_{iq})$ is a vector of crisp explanatory variables and $\tilde{Y}_i$ is a fuzzy response that is presented as a symmetrical TFN, denoted as $\tilde{Y}_i = (y_l, y, y_u) = (y - e, y, y + e)$. Suppose there is a new query with an unknown fuzzy response, denoted by ***z***. The proposed algorithm for predicting TFN for z, $(\hat{y}_l, \hat{y}, \hat{y}_u) = (\hat{y} - \hat{e}, \hat{y}, \hat{y} + \hat{e})$ is as follows.

a) Calculate the appropriate distance between ***z*** and all observations in the training data.
b) Set the number of nearest neighbours (*k*) and identify the *k* nearest neighbours, which are the observations in the training dataset that are closest in distance to ***z***.
c) Based on its *k* nearest neighbours, estimate the fuzzy linear regression model between the fuzzy/crisp dependent variable and the crisp independent variables, using possibilistic regression explained in Section 2.2.
d) Use the resulting model to predict the TFN for the new observation. The predicted TFN of the dependent variable for the new observation *z* is
e)

$$\hat{\tilde{Y}}_z = \hat{\tilde{A}}_{0,z} + \hat{\tilde{A}}_{1,z} X_1 + \hat{\tilde{A}}_{2,z} X_2 + \dots + \hat{\tilde{A}}_{q,z} X_q \tag{7}$$

It is important to note that Tanaka's fuzzy regression model provides only a fuzzy prediction for a new observation with predictor values within the range of the dataset used to generate the fuzzy model. Extrapolation beyond the X range in the training dataset carries the risk of producing an invalid TFN (Škrabánek & Martinková, 2021).

### *2.5. Performance Measure*

The accuracy of the predicted fuzzy outcome is calculated based on the distance between membership functions of two outcome TFNs, namely the observed outcome's TFN, denoted by $\tilde{Y}$ = ($y_l$, $y$, $y_u$), and the predicted outcome's TFN, denoted by $\hat{\tilde{Y}} = \hat{y}_l, \hat{y}, \hat{y}_u)$. Suppose $\mu_{\tilde{y}_i}(x)$ and $\mu_{\hat{\tilde{y}}_i}(x)$ are membership functions of $\tilde{Y}$ and $\hat{\tilde{Y}}$. The general formula for the relative distance between those fuzzy membership functions is

$$D_r = \frac{\int_{S_{\tilde{Y}} \cup S_{\hat{\tilde{Y}}}} \left|\mu_{\tilde{y}_i}(x) - \mu_{\hat{\tilde{y}}_i}(x)\right| dx}{\int_{S_{\tilde{Y}}} \mu_{\tilde{y}_i}(x) dx} \tag{8}$$

where $S_{\tilde{Y}}$ and $S_{\hat{\tilde{Y}}}$ are supports of $\tilde{Y}$ and $\hat{\tilde{Y}}$ (Kim & Bishu, 1998; Škrabánek & Martinková, 2021).

Another metric is mean squared error (MSE), formulated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_{li} - \hat{y}_{li})^2 + (y_i - \hat{y}_i)^2 + (y_{ui} - \hat{y}_{ui})^2 \tag{9}$$

## 3. RESULTS AND DISCUSSION

The data used are secondary, sourced from the laboratory's daily records at a cement plant in Indonesia. The response variable is the 28-day compressive strength of cement, defined as the maximum load per unit area applied to cement mortar or concrete until the material fails 28 days after hydration. It is calculated by applying a load to a test specimen, typically in the form of mortar or concrete moulded into a cylinder or cube. The load is increased continuously until the test specimen fails, and the load at failure determines the compressive strength. The dataset comprises records from 108 days, with 2 records per day. The explanatory variables are:

a) $X_1$= Sulfur trioxide ($SO_3$) content, which represents the amount of $SO_3$ contained in cement. It is generally governed by the addition of gypsum (Tsamatsoulis et al., 2023).
b) $X_2$ = Loss on Ignition (LoI), which represents the percentage mass lost when cement is heated to approximately 900°C–1000°C (Bernal et al., 2017; Lee et al., 2025).
c) $X_3$ = residue on 45-micron sieve, which represents the fineness of cement particles. It measures the percentage of cement particles retained on a 45-micron sieve (Ferraris & Garboczi, 2013).
d) $X_4$ = 7-day compressive strength, which represents cement compressive strength at day 7.

These variables are selected using OLS regression from a set of variables directly linked to clinker quality, sulfate balance, fineness, and hydration kinetics. Data were analysed using possibilistic fuzzy regression and I-KNNFR. Since the data is originally recorded as a single point (crisp), it needs to be fuzzified into a symmetrical TFN. This study adopts the fuzzification strategy employed by Türksen (2021) and Urbański (2023). Using this strategy, the mode ($m$) and spread ($c$) of TFN are calculated as the arithmetic mean and half-range of two data points, respectively. These statistics result in a symmetric TFN, $\tilde{Y} = (m - c, m, m + c)$.

### *3.1. Possibilistic Fuzzy Regression*

The possibilistic fuzzy regression model for cement 28-day compressive strength is:

$$\hat{\tilde{Y}} = (121.418,0) + (49.405, 13.483)X_1 - (31.326, 0)X_2 - (3.296, 0)X_3 + (0.786, 0)X_4 \qquad (10)$$

This model shows that $SO_3$ content and 7-day compressive strength positively influence 28-day compressive strength. The increases in $SO_3$ content and 7-day compressive strength are associated with an increase in the cement mortar compressive strength. On the contrary, loss on ignition and residue on the 45-micron sieve have a negative impact on 28-day compressive strength, indicating that increasing those variables decreases 28-day compressive strength. This model also exhibits that $SO_3$ content is the only explanatory variable with a non-zero spread. This indicates that the influence of $SO_3$ on the 28-day compressive strength is not fully deterministic but subject to uncertainty arising from complex chemical interactions. On the contrary, the coefficients for loss on ignition, residue on a 45-micron sieve, and 7-day compressive strength show crisp values, indicating more stable, monotonic relationships with 28-day compressive strength. Suppose there are five new cement mortar specimens with values of $X_1 - X_4$, as shown in columns 2-5 of Table 1. Using the estimated model presented in Eq. (10), the predicted 28-day compressive strength of these five specimens is shown in the last column of Table 1. Note that the predicted compressive strength for the fifth specimen is not available because the $X_1$ value for this specimen ($X_1 = 2.12$) is outside the range of $X_1$ in the training dataset.

**Table 1.** Predicted TFN of 28-day Compressive Strength for New Cement Mortar.

| No | $SO_3$ content | Loss on Ignition | Residue on 45-micron sieve | 7-day compressive strength | Predicted TFN of 28-day compressive strength $(\boldsymbol{m}, \boldsymbol{c}) = (\boldsymbol{m} - \boldsymbol{c}, \boldsymbol{m}, \boldsymbol{m} + \boldsymbol{c})$ |
|---|---|---|---|---|---|
| 1 | 1.69 | 4.42 | 5.67 | 74.3 | (106.16, 22.79) = (83.38,106.16, 128.95) |
| 2 | 1.68 | 4.39 | 4.61 | 77.5 | (112.62, 22.65) = (89.97, 112.62, 135.27) |
| 3 | 2.00 | 4.73 | 9.45 | 78.8 | (102.85, 26.97) = (76.88, 102.85, 129.81) |
| 4 | 1.67 | 4.81 | 6.74 | 76.0 | ( 90.77, 22.52) = (68.25, 90.77, 113.28) |
| 5 | 2.12 | 4.93 | 6.12 | 77.1 | Not available |

Note that the estimated compressive strength value of the fifth specimen cannot be obtained because the $X_1$ value of this test specimen ($X_1$ = 2.12) is outside the range of $X_1$ ($SO_3$ content) in the training data used to form the model, which is [1.47, 2.02]. Meanwhile, Tanaka's fuzzy regression does not allow for extrapolation because it may produce an invalid TFN. The TFN prediction provides more information than a single-point prediction. It represents the most plausible compressive strength and also the range of plausible values associated with material heterogeneity, experimental variability, and incomplete knowledge of the strength-development process. Therefore, the TFN output provides information related to expected strength and the uncertainty surrounding it.

The model's predictive performance is evaluated using 10-fold range-preserved controlled cross-validation. Predictive performance is measured by mean squared error (MSE) and mean relative distance (MRD). Predictive models and their corresponding MSE and MRD, obtained from the ten training-testing dataset pairs, are presented in Table 2. The overall MSE and MRD are computed by averaging the MSE and MRD across all pairs of training and testing datasets. It is found that the overall MSE is 965.95 and the overall MRD is 34.37.

**Table 2.** Models and Predictive Performance Measures (Compressive Strength Dataset).

| Pair | Fuzzy Model | MSE | MRD |
|---|---|---|---|
| 1 | $\hat{Y}$ = (130.56, 0) + (48.030, 11.645) $X_1$ – (32.128, 0) $X_2$ - (3.56, 0) $X_3$ + (0.759, 0) $X_4$ | 872.03 | 62.28 |
| 2 | $\hat{Y}$ = (130.56, 0) + (48.033, 11.645) $X_1$ – (32.128, 0) $X_2$ - (3.557, 0) $X_3$ + (0.759, 0) $X_4$ | 1032.28 | 26.77 |
| 3 | $\hat{Y}$ = (76.331, 0) + (9.137, 0) $X_1$ – (0.775, 1.479) $X_2$ + (0.238, 0) $X_3$ + (0.539, 0) $X_4$ | 1028.78 | 8.08 |
| 4 | $\hat{Y}$ = (19.604, 0) + (75.732, 11.472) $X_1$ – (24.121, 0) $X_2$ - (1.604, 0) $X_3$ + (1.001, 0) $X_4$ | 1122.97 | 23.20 |
| 5 | $\hat{Y}$ = (130.56, 0) + (48.033, 11.645) $X_1$ – (32.128, 0) $X_2$ - (3.557, 0) $X_3$ + (0.759, 0) $X_4$ | 862.09 | 23.59 |
| 6 | $\hat{Y}$ = (130.56, 0) + (48.033, 11.645) $X_1$ – (32.128, 0) $X_2$ - (3.557, 0) $X_3$ + (0.759, 0) $X_4$ | 888.04 | 28.99 |
| 7 | $\hat{Y}$ = (59.539, 0) + (63.1, 11.388) $X_1$ – (22.14, 0) $X_2$ - (4.084, 0) $X_3$ + (0.785, 0) $X_4$ | 981.37 | 27.99 |
| 8 | $\hat{Y}$ = (130.56, 0) + (48.033, 11.645) $X_1$ – (32.128, 0) $X_2$ - (3.557, 0) $X_3$ + (0.759, 0) $X_4$ | 965.14 | 51.97 |
| 9 | $\hat{Y}$ = (69.401, 0) + (66.445, 11.509) $X_1$ – (28.09, 0) $X_2$ - (3.604, 0) $X_3$ + (0.889, 0) $X_4$ | 948.26 | 43.61 |
| 10 | $\hat{Y}$ = (130.56, 0) + (48.033, 11.645) $X_1$ – (32.128, 0) $X_2$ - (3.557, 0) $X_3$ + (0.759, 0) $X_4$ | 958.54 | 47.20 |

Note: $X_1$ = $SO_3$ content; $X_2$ = loss on ignition; $X_3$ = residue on 45 micron sieve; $X_4$ = 7-day compressive strength

### *3.2. Integrated KNN Fuzzy Regression*

After presenting the results of Tanaka's possibilistic fuzzy regression, this study proceeds with the analysis of the I-KNNFR. This approach is intended to predict the TFN of the cement mortar's 28-day compressive strength. Unlike possibilistic fuzzy regression, I-KNNFR does not produce a single mathematical equation applicable to all prediction points. Instead, this method yields a set of rules to predict the TFN of 28-day compressive strength for a newly produced cement mortar, given known explanatory variables. These rules include the number of nearest neighbours ($k$) and the Minkowski distance parameter ($p$) that are selected from $k \in \{9, 10, \dots 70\}$ and $p \in \{1.6, 1.7, \dots, 2.0\}$. I-KNNFR with the best $k$ and $p$ is employed to construct a local fuzzy regression model for each query observation. The resulting model is then used to predict the TFN of compressive strength for that observation.

To ensure methodological consistency, the predictive performance of the proposed I-KNNFR is evaluated using the same 10-fold range-preserved controlled cross-validation procedure used in the previous section. For each fold, MSE and MRD are computed. The reported values are overall predictive performance measures, computed as the average across the 10 cross-validation iterations. The optimal values of $k$ and $p$ are tuned sequentially rather than through a joint search over all possible combinations of $k$ and $p$. This sequential tuning strategy reduces the risk of excessive parameter search. It also preserves the interpretability of the tuning process by allowing each parameter to be assessed independently. The effect of $k$ is first evaluated, and then the effect of $p$ is evaluated.

Figure 1 displays the relationship between MSE and the number of nearest neighbours for $k$ = 9, 10, …, 70, presented in a line chart. The figure is displayed in five panels, each corresponding to a specific value of the Minkowski distance parameter, $p$. Within each panel, MSE values are obtained by averaging results from the ten cross-validation iterations. Across all considered values of $p$, the results show the same pattern in MSE as the number of nearest neighbours increases. In general, the MSE increases in line with the increase in the number of nearest neighbours used to model the 28-day compressive strength. The results also indicate that the MSE attains its minimum at $k$ = 9. This consistent behaviour suggests that $k$ = 9 yields the most favourable performance for the I-KNNFR model under the given experimental setting.

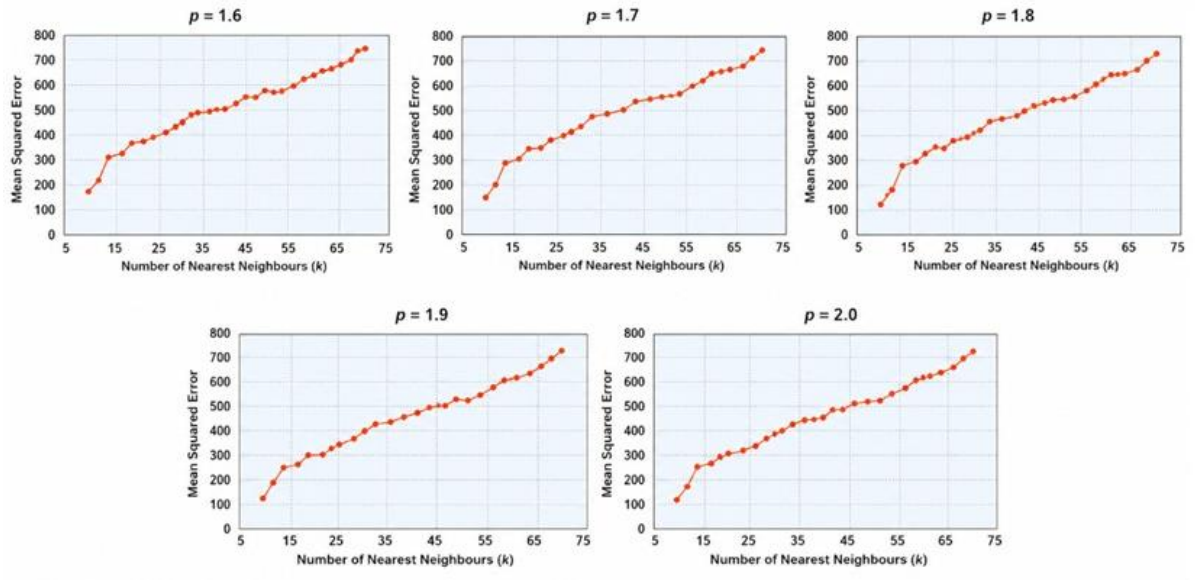


**Figure 1.** Relationship between MSE and k for Various p

In addition, Figure 2 presents line charts of the relationship between the MRD and the number of nearest neighbours. As with MSE, these relationships are presented in five panels, one for each value of the Minkowski distance parameter. The MRD values are also computed by averaging results from the ten cross-validation iterations. This figure shows a pattern similar to that in Fig. 1, where the MRD increases as the number of nearest neighbours increases. It is also found that the MRD attains its minimum at $k = 9$ for all values of $p$. This behaviour is consistent with the relationship between MSE and k. As in the previous conclusion, $k = 9$ provides the most favourable performance for the I-KNNFR model. Consequently, this value of $k$ is selected for the subsequent analysis.

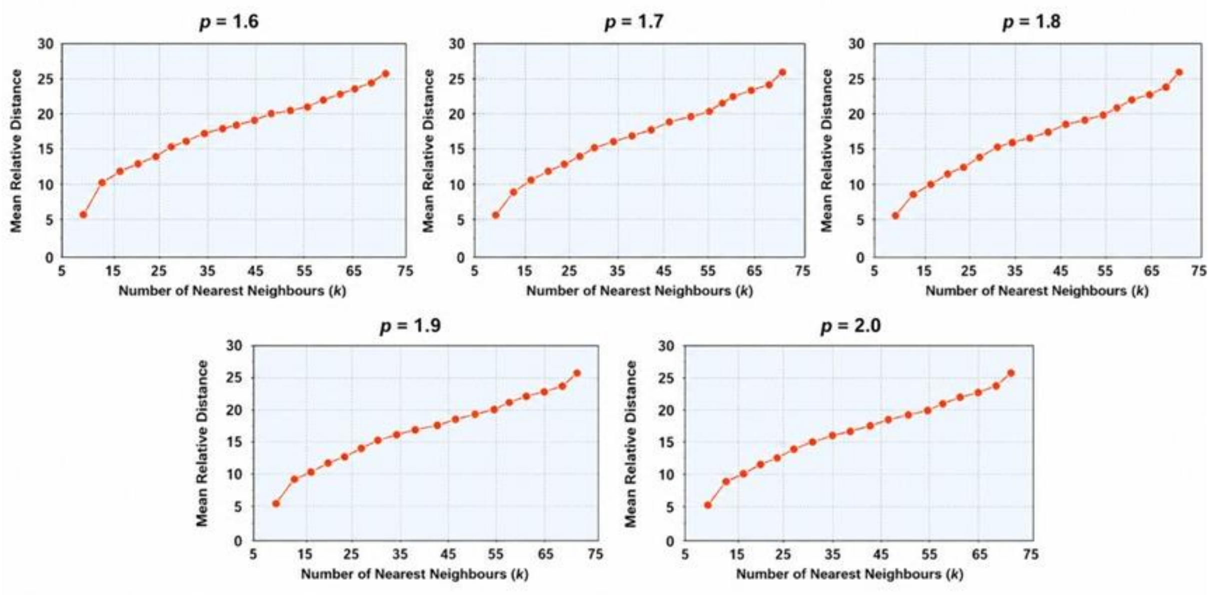


**Figure 2.** Relationship between MRD and k for Various p

After selecting $k = 9$ as the optimal number of nearest neighbours, the influence of the Minkowski distance parameter is examined. The average MSE and MRD for fixed $k = 9$, corresponding to $p = 1.6$, 1.7, 1.8, 1.9, and 2.0, are summarized in Table 3. The value of $p$ producing the lowest MSE is then selected as the optimal distance parameter for the I-KNNFR. Overall, the differences in MSE and MRD across all considered values of $p$ are marginal. This indicates that, in this case, the I-KNNFR is relatively insensitive to the choice of Minkowski distance parameter within the examined range. The minimum MSE is attained at $p = 2.0$, whereas the lowest MRD is observed for $p = 1.8$ and $p = 1.9$. However, given that the observed differences are minor and not practically significant, the Euclidean distance (Minkowski distance with $p = 2.0$) is selected for subsequent analysis. This choice is motivated by its slightly superior MSE and its stable behaviour across both evaluation metrics. Furthermore, this distance measure is widely used in *k*-nearest-neighbour-based methods.

**Table 3.** Average MSE and MRD for Different Minkowski Distance Parameters with k = 9.

| No | *p* | MSE | MRD |
|---|---|---|---|
| 1 | 1.6 | 141.7039 | 5.9748 |
| 2 | 1.7 | 141.7101 | 5.9729 |
| 3 | 1.8 | 141.8415 | 5.9623 |
| 4 | 1.9 | 141.8415 | 5.9623 |
| 5 | 2.0 | 141.7007 | 5.9690 |

Finally, the TFN for the 28-day compressive strength of a new cement mortar can be predicted using the I-KNNFR by following the rules; (a) Use the Euclidean distance to measure the closeness of a given observation to all observations in the training dataset; and (b) Use $k$ = 9 neighbours as a basis to construct the model that will be used in the subsequent step to predict the TFN of the 28-day compressive strength for a new cement mortar.

### *3.3. Comparative Analysis of the I-KNNFR and Tanaka's Fuzzy Regression*

The previous section illustrated the use of possibilistic fuzzy regression and I-KNNFR to model the 28-day compressive strength of cement mortar for ordinary Portland-type cement. The predictive performance of the two methods was evaluated using the same 10-fold range-preserved controlled cross-validation. The reported predictive performance measures (MSE and MRD) are averaged over the 10 folds. This ensures that the predictive performance comparison is conducted under an identical evaluation procedure. Under this procedure, possibilistic fuzzy regression yielded an MSE of 965.95 and an MRD of 34.37. Refer back to Table 3. This table shows that all variants of I-KNNFR yield lower MSE and MRD than those provided by possibilistic fuzzy regression. This indicates that, in general, possibilistic fuzzy regression yields poorer predictive performance than I-KNNFR. Specifically, the I-KNNFR with $k$ = 9 and $p$ = 2 achieved an average MSE of 141.70 and an MRD of 5.97. Thus, relative to possibilistic fuzzy regression, the I-KNNFR reduces the MSE by approximately 85%. This method also reduces the MRD by approximately 83%.

## 4. CONCLUSION

In this study, two fuzzy regression approaches were implemented for predicting the TFN of the 28-day compressive strength of cement mortar based on variables assumed to influence it. The two approaches were possibilistic fuzzy regression and I-KNNFR. The predictive performance of both methods was compared using MSE and MRD obtained through a 10-fold range-preserved controlled cross-validation. The possibilistic fuzzy regression yielded a model with an MSE of 965.95 and an MRD of 34.37. This model revealed that the $SO_3$ content is the only variable that contributes to uncertainty in 28-day cement compressive strength. On the other hand, the best I-KNNFR performance was achieved when it was performed using $k$ = 9 nearest neighbours and Euclidean distance. In this setting, I-KNNFR yielded lower performance measures: 141.70 for MSE and 5.97 for MRD. Thus, compared to possibilistic fuzzy regression, the I-KNNFR reduced the MSE and MRD by approximately 85% and 83%, respectively. The lower MSE and MRD indicate improved predictive performance.

The findings demonstrate the importance of incorporating local neighbourhood information when modelling heterogeneous strength development under uncertainty associated with material heterogeneity, experimental variability, and incomplete knowledge of the strength-development process. The TFN prediction is more informative than a single-point prediction because it provides both the most plausible compressive strength value and a range of plausible values reflecting this uncertainty. A global fuzzy regression model produced by the possibilistic fuzzy regression applies a common relationship to all observations, although cement specimens may exhibit different local strength-development patterns. By selecting observations that are most similar to each prediction point, I-KNNFR constructs a local fuzzy regression model that is more responsive to these heterogeneous patterns and less influenced by distant observations. Incorporating the KNN mechanism substantially improves the predictive accuracy of possibilistic fuzzy regression. Consequently, the I-KNNFR can be considered a superior alternative to possibilistic fuzzy regression for modelling the TFN of cement's 28 day compressive strength.

**ACKNOWLEDGEMENT**
The authors are grateful to Universitas Andalas, Padang, Indonesia, for providing financial support throughout this study.

**CONFLICT OF INTEREST**
The authors declare no conflicts of interest.

**AUTHOR CONTRIBUTION**
Hazmira Yozza: Conceptualization, Methodology & Writing-original draft; Riswan Efendi: Writing-Review and Editing; Nor Azah Samat: Writing, Review & Editing; Izzati Rahmi: Writing, Review & Editing; Ridho Saputra: Software & investigation.

**DATA AVAILABILITY**
The data that support the findings of this study are available from a cement manufacturing company in Indonesia, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of that cement company.

**DECLARATION OF GENERATIVE AI**
During the preparation of this work, the authors used QuillBot, Grammarly, and ChatGPT for translating, paraphrasing, grammar checking, and enhancing the readability of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

**ETHICS**
Not applicable.

## REFERENCES

Al Biajawi MI, Embong R, Muthusamy K, Ismail N, Obianyo I. (2022). Recycled coal bottom ash as sustainable materials for cement replacement in cementitious composites: A review. *Construction and Building Materials*, 338, 127624. doi:10.1016/j.conbuildmat.2022.127624

Asanre AA, Ifamuyiwa AS, Abiodun TO. (2024). Dimensions of students' academic engagement as predictors of senior secondary school mathematics achievement. *Journal of Science and Mathematics Letters*, 12(1), 27-33. doi:10.37134/jsml.vol12.1.4.2024

Bector CR, Chandra S.(2005). Fuzzy Mathematical Programming and Fuzzy Matrix Games. Springer-Verlag, Germany. doi:10.1007/3-540-32371-6

Bernal SA, Juenger MCG, Ke X, Matthes W, Lothenbach B, De Belie N, Provis JL. (2017). Characterization of supplementary cementitious materials by thermal analysis. *Materials and Structures*, 50, article 26. doi:10.1617/s11527-016-0909-2

Chaira T. (2019). *Fuzzy Set and Its Extension: The Intuitionistic Fuzzy Set*. Wiley. USA. p. 1-42. doi:10.1002/9781119544203

Dubois D, Prade H. (1980). *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, USA. p. 9-11.

Fathabadi A, Seyedian SM, Malekian A. (2022). Comparison of Bayesian, k-nearest neighbor and Gaussian process regression methods for quantifying uncertainty of suspended sediment concentration prediction. *Science of the Total Environment*, 818, article 151760. doi:10.1016/j.scitotenv.2021.151760.

Ferraris CF, Garboczi EJ. (2013). Identifying improved standardized tests for measuring cement particle size and surface area. *Transportation Research Record: Journal of the Transportation Research Board*, 2342(1), 10-16. doi:10.3141/2342-02

Gupta M, Nirola M, Sharma A, Dhungel P, Singh H, Gupta A. (2024). Geospatial analysis of contagious infection growth and cross-boundary transmission in non-vaccinated districts of North-East Indian states during the COVID-19 pandemic. *The Lancet Regional Health - Southeast Asia*, 28, 100451. doi:10.1016/j.lansea.2024.100451

Hameed MM, AlOmar MK. (2020). Prediction of compressive strength of high-performance concrete: hybrid artificial intelligence technique. In Khalaf M, Al-Jumeily D, Lisitsa A (eds.), *Applied Computing to Support Industry: Innovation and Technology*, p. 323-335. Sringer, Switzerland. doi:10.1007/978-3-030-38752-5_26

Jia X. (2025). Digital marketing and brand competitiveness: The mediating role of marketing capability and operational capability. *Journal of Retailing and Consumer Services*, 87, 104429. doi:10.1016/j.jretconser.2025.104429

Kalwa V, Ullegaddi MM, Kittali P, Thottempudi B. (2023). Ordinary least square regression technique for predicting wear rate of EN8 carbon steel. *Materials Today: Proceedings*, advance online publication. doi:10.1016/j.matpr.2023.03.388

Kim B, Bishu RR. (1998). Evaluation of fuzzy linear regression models by comparing membership functions. *Fuzzy Sets and Systems*, 100(1-3), 343-352. doi:10.1016/S0165-0114(97)00100-0

Kohli S, Godwin GT, Urolagin S. (2021). Sales Prediction Using Linear and KNN Regression. In Patnaik S, Yang X, Sethi IK. Advances in Machine Learning and Computational Intelligence, p. 321-329. Springer, Singapore. doi:10.1007/978-981-15-5243-4_29

Kumbure MM, Luukka P, Collan M. (2020). A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern Recognition Letters*, 140, 172-178. doi:10.1016/j.patrec.2020.10.005.

Kumbure MM, Luukka P. (2022). A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. Granular Computing, 7, 657-671. doi:10.1007/s41066-021-00288-w

Lam J, Broccatelli C, Baxter J. (2023). Diversity of strong and weak ties and loneliness in older adults. *Journal of Aging Studies,* 64, 101097, doi:10.1016/j.jaging.2022.101097.

Lee GS, Erkinov FB, Aripova MK. (2025). Additional loss on ignition of cement raw mixes at high temperatures. *Chemistry and Chemical Engineering*, *2025* (3), 15-20. doi:10.70189/1992-9498.1670

Maryoto A, Setijadi R, Widyaningrum A, Waluyo S. (2020). Drying shrinkage of concrete containing calcium stearate, $Ca(C_{18}H_{35}O_2)_2$, with ordinary Portland cement (OPC) as a binder: Experimental and modelling studies. *Molecules*, 25(21), Article 4880. doi:10.3390/molecules25214880.

Mohan S, Ajmal M, Drewniok MP. (2025). Prediction of 28-day compressive strength of high-slag concrete by establishing accelerated oven curing regimes for rapid quality control. *Journal of Materials Science: Materials in Engineering*, 20, 111. doi:10.1186/s40712-025-00333-5

Mouawad F, Homsi F, Geara F, Mina R. (2025). Predicting compressive strength of sustainable concrete using machine learning and artificial neural networks. *Construction Materials*, 5(3), article 56. doi:10.3390/constrmater5030056

Nelson BR, Jamaludin AA, Shahimi S, Lian CJ, Meilana L, Kumaran JV, Raj Kumar HC, Moqbel RGA, Sharon K, Chowdhury AJK, Sheikh HI. (2026). Outlining probable fundamental doctrine and relations to disasters in the continental crust. *Journal of Science and Mathematics Letters*, 14(1), 15-28. doi:10.37134/jsml.vol14.1.2.2026

Onjure EO, Kabubo CK, Okumu V. (2023). Use of Plant-Based Accelerator to Enhance Rate of Gain of Strength of Kenyan Blended Cement. *Open Journal of Civil Engineering*, 13, 399-410. doi:10.4236/ojce.2023.133030

Rahmi I, Efendi R, Yanuar F, Yozza H, Wahyudi M, Burney SMA, Eğrioğlu E. (2025). Essentials of preprocessing data in improving logistic regression performance based on rough sets theory: A case study of stunting in West Sumatra, Indonesia. *Journal of Science and Mathematics Letters*, 13(1), 124-139. doi:10.37134/jsml.vol13.1.12.2025

Rezaei A, Areshenkoff CN, Gale DJ, De Brouwer AJ, Nashed JY, Flanagan JR, Gallivan JP. (2025). Transfer of motor learning is associated with patterns of activity in the default mode network. *Plos Biology*, 23(8), e3003268. doi:10.1371/journal.pbio.3003268

Safiuddin M, Raman SN, Salam MA, Jumaat MZ. (2016). Modeling of compressive strength for self-consolidating high-strength concrete incorporating palm oil fuel ash. *Materials,* 9(5), Article 396. doi:10.3390/ma9050396

Saha D, Paul B, Sarkar B. (2025). Revolutionizing sustainable construction through predictive modeling of green concrete. *Developments in the Built Environment*, 23, 100740. doi:10.1016/j.dibe.2025.100740

Sajedi F. (2012). Strength Development of Cement-Slag Mortars. *Journal of Civil Engineering and Architecture*, 6(9), 1169-1183. doi:10.17265/1934-7359/2012.09.009

Schack T, Strybny B, Haist M. (2024). Improving the early age strength of eco-efficient mortar with low clinker content considering binder granulometry and chemical additives. *Materials*, 17(18), 4509. doi:10.3390/ma17184509

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

Siddique R, Mehta A. (2014). Effect of carbon nanotubes on properties of cement mortars. *Construction and Building Materials,* 50, 116-129. doi:10.1016/j.conbuildmat.2013.09.019

Škrabánek P, Marek J, Pozdílková A. (2021). Boscovich Fuzzy Regression Line. *Mathematics*, 9, 685. doi:10.3390/math9060685

Škrabánek P, Martinková N. (2021). Algorithm 1017: fuzzyreg: An R package for fitting fuzzy regression models. *ACM Transactions on Mathematical Software*, 47(3), Article 29, 1-8. doi:10.1145/3451389

Solano-Vinueza G, Albuja-Sánchez J. (2025). Development and field based validation of a predictive model for concrete compressive strength using fresh state properties in a large scale construction project. *Case Studies in Construction Materials*, 24, e05705. doi:10.1016/j.cscm.2025.e05705

Song Y, Liang J, Lu J, Zhao X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, 251, 26-34. doi:10.1016/j.neucom.2017.04.018

Sunar F, Dervisoglu A, Yagmur N, Atabay H, Donertas A. (2024). Comparison of conventional and machine learning regression models for accurate prediction of selected optical active components - A case study: The Gulf of Izmit. *Marine Pollution Bulletin*, 208, Article 116942. doi:10.1016/j.marpolbul.2024.116942.

Tanaka H, Uejima S, Asai, K. (1982) Linear regression analysis with fuzzy model. *IEEE Transaction on System, Man, and Cybernetics*, 12(6), 903-907. doi:10.1109/TSMC.1982.4308925

Tsamatsoulis DC, Korologos CA, Tsiftsoglou DV. (2023). Optimizing the sulfate content of cement using neural networks and uncertainty analysis. *ChemEngineering*, 7, 58. doi:10.3390/chemengineering7040058

Tur VV, Derechennik SS. (2018). Assessment of the Concrete compressive strength in existing structures based on core test results. *Solid State Phenomena*, 272, 238-243. doi:10.4028/www.scientific.net/ssp.272.238

Türkşen Ö. (2021). A comparison analysis of fuzzy and Bayesian linear model parameter estimates for replicated response measures. *Gazi University Journal of Science*, 34(1), 251-270. doi:10.35378/gujs.660845

Urbański MK, Wójcicka K, Wójcicki P. (2023). Fuzzy approach with t-norm arithmetic to analyze measurement uncertainty-concept and algorithm. *Przegląd Elektrotechniczny*, 99(11), 216-220. doi:10.15199/48.2023.11.39

Wahyuni, Abdullah, Mulyati S, Hayati Y. (2025). Performance analysis of ordinary Portland cement and Portland pozzolan cement in improving sea sand and sea water concrete with zeolite powder. *Journal of Ecological Engineering*, 26(4), 264-274. doi:10.12911/22998993/200278

Yankelevsky DZ. (2024). The uniaxial compressive strength of concrete: revisited. Materials and Structures, 57, Article 144. doi:10.1617/s11527-024-02422-x

Zadeh LA. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353. doi:10.1016/S0019-9958(65)90241-X

Zakaria AS, Shafi MA, Zim MAM, Razali SNAM. (2023). The use of fuzzy linear regression modeling to predict high-risk symptoms of lung cancer in Malaysia. *International Journal of Advanced Computer Science and Applications*, 14(5), 586-593 doi:10.14569/ijacsa.2023.0140562

Zhang W, Guo J, Ning C, Cheng R, Liu Z. (2024). Prediction of concrete compressive strength using a Deepforest-based model. Scientific Reports, 14, 18918. doi:10.1038/s41598-024-69616-9

Zhang Y, Li J, Li X. (2025). Research on compressive strength prediction and formula optimization based on random forest and response surface method. SSRN preprint, p. 17. doi:10.2139/ssrn.5262134

Zhao N, Zhang H, Xie P, Chen X, Wang X. (2025). Prediction of compressive strength of multiple types of fiber-reinforced concrete based on optimized machine learning models. *Engineering Applications of Artificial Intelligence*, 152, 110714. doi:10.1016/j.engappai.2025.110714