Next Materials 9 (2025) 100947

Contents lists available at ScienceDirect

Next Materials

journal homepage: www.sciencedirect.com/journal/next-materials

Research article

A comparative study of machine learning techniques and data processing for predicting the compressive strength of pervious concrete with supplementary cementitious materials and chemical composition influence

Navaratnarajah Sathiparan [a,1,*], Pratheeba Jeyananthan [b], Daniel Niruban Subramaniam [c]

[a] Department of Civil Engineering, Faculty of Engineering, University of Jaffna, Sri Lanka
[b] Department of Computer Engineering, Faculty of Engineering, University of Jaffna, Sri Lanka
[c] Department of Civil Engineering, Faculty of Engineering, University of Jaffna, Sri Lanka

ARTICLE INFO

Keywords:
Machine Learning
Pervious Concrete
Compressive Strength
Data Processing
Supplementary Cementitious Materials
Extreme Gradient Boosting

ABSTRACT

This study introduces a novel approach that combines machine learning algorithms, such as Extreme Gradient Boosting (XGB) and Artificial Neural Network (ANN), with chemical composition analysis to predict the compressive strength of pervious concrete. By considering a wider range of supplementary cementitious materials (SCMs) and chemical oxides, such as calcium oxide (CaO) and silicon dioxide (SiO_2), this approach significantly improves prediction accuracy over traditional empirical models, providing a more robust solution for sustainable construction. A comprehensive dataset of 659 observations was compiled from various studies, emphasizing the significance of input variables such as calcium oxide (CaO), silicon dioxide (SiO_2), aluminium oxide (Al_2O_3), and curing period. Various data processing methods were employed to enhance model performance, including Max-Min normalization, Z-score normalization, robust scaling, log transformation, and sigmoid normalization. The study demonstrates that XGB outperformed other machine learning models, achieving a training R^2 of 0.99 and a testing R^2 of 0.92, with an RMSE of 2.85 MPa. This research highlights the significance of incorporating chemical composition analysis (CaO, SiO_2) into machine learning models to enhance the prediction accuracy of the compressive strength of pervious concrete. The novelty of the approach lies in combining advanced data processing techniques with a diverse dataset of SCMs, offering an innovative solution for optimizing concrete formulations in engineering. Sensitivity analysis highlighted the critical importance of CaO, SiO_2, and curing period in predicting compressive strength, while aggregate size had a minimal impact. This research contributes to international efforts in sustainable infrastructure development by integrating machine learning techniques with chemical composition analysis to predict the compressive strength of pervious concrete. This innovative approach offers global implications for optimizing concrete mix designs, reducing material waste, and enhancing the durability of urban infrastructure.

1. Introduction

Pervious concrete, characterized by its high porosity and permeability, is critical in modern construction, particularly sustainable urban development [1]. This innovative material enables efficient rainwater drainage, thereby reducing stormwater runoff and mitigating the risk of flooding in urban areas. By facilitating groundwater recharge and minimizing surface water accumulation, pervious concrete plays a vital role in enhancing infrastructure sustainability while improving environmental quality [2]. For several reasons, incorporating Supplementary Cementitious Materials (SCMs) as cement replacements is essential. First, SCMs can enhance the mechanical properties of concrete, including compressive strength, durability, and resistance to environmental degradation [3]. They also contribute to the sustainability of construction practices by reducing the overall cement content and lowering the carbon emissions associated with cement production [4]. Furthermore, SCMs can improve concrete workability and long-term performance, making them indispensable in developing high-quality pervious concrete [5].

Several factors significantly influence the compressive strength of

* Correspondence to: Department of Civil Engineering, Faculty of Engineering, University of Jaffna, Ariviyal Nager, Kilinochchi, Sri Lanka.
E-mail address: sakthi@eng.jfn.ac.lk (N. Sathiparan).
1 ORCID: 0000–0001-8570–0580

https://doi.org/10.1016/j.nxmate.2025.100947
Received 31 March 2025; Received in revised form 9 July 2025; Accepted 14 July 2025
Available online 22 July 2025

pervious concrete. These include the water-to-cement ratio, aggregate size and grading, the type and proportion of SCMs used, and the curing conditions. The balance between void space and solid material is crucial in pervious concrete, as excessive porosity may compromise structural integrity [6,7]. Additionally, the presence of oxides, such as calcium oxide (CaO), silicon dioxide (SiO_2), aluminium oxide (Al_2O_3), and iron oxide (Fe_2O_3), can further affect the hydration process and the overall mechanical performance of the concrete [8]. The interaction between these oxides and the other constituents of the mix can lead to variations in compressive strength, necessitating a detailed investigation. Predicting the compressive strength of pervious concrete is essential for various reasons. Accurate predictions enable engineers and architects to design structures that meet specific performance criteria, ensuring safety and longevity [9]. Moreover, understanding compressive strength allows for more effective material selection and mix design, ultimately leading to cost-effective and sustainable construction practices. As the demand for pervious concrete solutions grows, the ability to predict their performance becomes increasingly critical.

Machine learning has emerged as a powerful tool for predicting compressive strength due to its ability to handle large datasets and recognize complex non-linear relationships among variables [10]. Traditional empirical approaches often rely on simplified models that may not capture the intricate interactions between mixed constituents [11]. In contrast, machine learning algorithms can analyze vast amounts of data, identify patterns, and provide more accurate predictions, making them superior [12]. The integration of machine learning techniques enhances predictive accuracy and offers insights into the underlying mechanisms that affect compressive strength. Considering the role of oxides in the overall mix of cement and SCMs is crucial in predicting the compressive strength of pervious concrete. The interaction between these oxides and the hydration process can significantly impact the formation of microstructures within the concrete matrix [13]. By understanding how different oxides influence the mechanical properties, researchers can optimize mix designs that enhance the performance and longevity of pervious concrete. This comprehensive approach not only contributes to the theoretical understanding of concrete behavior but also supports the development of practical guidelines for sustainable construction practices, ultimately leading to more resilient infrastructure solutions in urban environments.

Several studies focus on the use of waste glass as a Supplementary material in concrete to enhance its sustainability and mechanical properties. Basil Ishaq and Salih Mohammed [14] investigate the effect of waste glass powder on compressive strength, emphasizing the role of chemical compositions such as silicon dioxide (SiO_2) and calcium oxide (CaO). It reveals that the optimal glass powder content and SiO_2/CaO ratio improve concrete performance, with curing time significantly influencing the final strength. Ishaq et al. [15] explore the chemical interactions between glass powder and cement, utilizing univariate, bivariate, and multivariate approaches to predict compressive strength and splitting tensile strength. It underscores the importance of chemical compositions, such as the silica modulus (SiO_2/Al_2O_3 ratio) and alumina modulus (Al_2O_3/Fe_2O_3 ratio), in influencing concrete properties, offering insights into the complex interactions that drive performance. Ishaq et al. [16] build on this by proposing models to predict compressive strength based on chemical analysis, including the alumina/iron oxide ratio in glass powder. It highlights the effectiveness of these chemical compositions in achieving both high strength and sustainable concrete. Together, these studies demonstrate the potential of waste glass as a sustainable alternative to cement in concrete, improving its environmental footprint while enhancing key mechanical properties.

Table 1 summarizes studies that predict the compressive strength of blended cement-based materials, particularly those that incorporate supplementary cementitious materials (SCMs). The studies encompass various types of concrete, including regular and pervious concrete, with diverse input variables influencing compressive strength, such as cement content, water-to-binder ratio, aggregate types, and specific chemical compositions, including calcium oxide and silicon dioxide. The research employs advanced machine learning models, including artificial neural networks (ANN), random forest regression (RF), and support vector regression (SVR), reflecting a growing trend in leveraging computational techniques for improved prediction accuracy. The volume of data used in these studies varies significantly, ranging from 236 to 659 data points, which can impact the reliability of the findings, as larger datasets typically enhance model training and validation. Notably, the emphasis on SCMs, such as fly ash, underscores their critical role in modern concrete formulations, highlighting the necessity for precise predictive models to evaluate material performance effectively.

Over the years, numerous studies have investigated the influence of various factors, including chemical composition, water-to-cement ratio, aggregate size, and curing conditions, on the compressive strength of pervious concrete. Recent advances in machine learning (ML) have shown promising results in predicting compressive strength by analyzing large datasets and capturing the non-linear relationships between different input variables. Studies utilizing machine learning techniques, such as Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGB), have made significant strides in improving prediction accuracy. However, many of these studies are limited by small datasets, a lack of comprehensive analyses involving diverse SCMs, and narrow material scopes, primarily focusing on specific types

Table 1
Relevant work on the prediction of compressive strength of SCM blended cement-based materials.

References	Year	Material used	SCMs used	Input variable[a]	Machine learning models used[b]	Data processing method	No of dataset
[17]	2023	Concrete	FA	C.C, FA/B, F.Agg, C.Agg, W, W/B, T, CaO, SiO_2, Al_2O_3, Fe_2O_3, Specific gravity of FA	LR, DT, RF, SVR, XGB, ANN, M5P, MLP, GP	Max-Min	455
[18]	2023	Concrete	FA	C.C, FA, FA/B, F.Agg, C.Agg, W/B, CaO, SiO_2, T	FQ, NLR, MLR, ANN	Max-Min	236
[19]	2023	Concrete	FA	C.C, FA, FA/B, F.Agg, C.Agg, W/B, SiO_2/CaO, T	LR, FQ, M5P	Max-Min	236
[20]	2024	Pervious concrete	FA	Agg/B, FA/B, W/B, T, CaO, SiO_2, Al_2O_3, Fe_2O_3	ANN, KNN, SVR, XGB	Raw data	315
[8]	2024	Pervious concrete	Several SCMs	C.Agg, W, Agg.Size, T, CaO, SiO_2, Al2O3, Fe_2O_3, Mgo, Eq.Na_2O,	LR, FQ, ANN	Raw data	659
[21]	2024	Pervious concrete	Several SCMs	Agg/B, SCM/B, W/B, CaO, SiO_2, Agg.Size, T	ANN, KNN, RF, BT, SVR, XGB	Raw data	576
[13]	2025	Cement mortar	Fly ash	Agg/B, FA/B, W/B, T, CaO, SiO_2, Al2O3, Fe_2O_3, Mgo, Eq.Na_2O,	ANN, KNN, SVR, XGB	Raw data, Max-Min, Z-Score	481

[a] C.C: cement content, FA: fly ash content, C.Agg: coarse aggregate content, F.Agg: fine aggregate content, Agg/B: aggregate to binder ratio, W: water content, W/B: water binder ratio, T: curing period, CaO: calcium oxide content, SiO_2: silicon dioxide content, Al_2O_3: aluminium oxide content, Fe_2O_3: ferric oxide content, Eq.Na_2O: equivalent alkalis, Agg.Size: coarse aggregate mean size.

[b] ANN: artificial neural network, BT: boosted tree regression, DT: decision tree regression, FQ: full quadratic model, LR: linear regression, M5P: M5 model tree, NLR: non-linear regression, RF: random forest regression, SVR: support vector regression, XGB: eXtreme gradient boosting.

of SCMs such as fly ash. These limitations hinder the generalizability of the findings and prevent the development of robust predictive models that can be applied to a broader range of pervious concrete formulations. Additionally, while data processing techniques like normalization and feature extraction have been applied in previous studies, their effectiveness in enhancing the performance of machine learning models for compressive strength prediction has not been thoroughly evaluated.

Another significant limitation of past research, especially those focused on pervious concrete, is the narrow material scope; many studies have concentrated primarily on specific types of SCMs, such as fly ash, which may limit the generalizability of their findings to other SCM blends. This focus potentially overlooks the benefits of incorporating a broader range of materials that could enhance compressive strength. Data processing methods utilized in these studies present another area for improvement. Various preprocessing techniques, including normalization and feature extraction, were employed; however, the effectiveness of these methods should be evaluated in terms of the specific characteristics and variability of pervious concrete research. Several entries indicated the use of raw data for analysis. While raw data can be beneficial for discovering patterns, processed data often provides cleaner insights that can enhance model performance. The choice between using raw versus processed data should align with the predictive goals of the research. Moreover, the studies employed a range of machine learning models, including linear regression (LR), Artificial Neural Networks (ANN), and Support Vector Regression (SVR). Each model has inherent strengths and weaknesses in terms of bias, variance, and interpretability, which can impact its overall predictive accuracy and reliability.

1.1. Research gap

Recent advancements in the field have led to a deeper understanding of the influence of chemical compositions, particularly oxides such as calcium oxide (CaO), silicon dioxide (SiO_2), and aluminum oxide (Al_2O_3), on the hydration and mechanical performance of pervious concrete. However, predicting the compressive strength of pervious concrete remains a challenge due to the complex interactions between material compositions and mix design parameters. While empirical methods have been widely used, they often fail to account for the intricate non-linear relationships inherent in concrete strength prediction.

Machine learning (ML) techniques, particularly those capable of handling large datasets, have emerged as powerful tools for overcoming these limitations. These approaches allow for more accurate and nuanced predictions of compressive strength by analyzing complex data patterns that traditional models cannot capture. A growing body of research has explored various ML models such as Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGB) for this purpose. However, the majority of studies have been limited by small datasets and a narrow focus on specific types of SCMs, limiting their generalizability.

This study aims to bridge this gap by conducting a comparative analysis of several ML models and data processing methods for predicting the compressive strength of pervious concrete containing a diverse range of SCMs. Specifically, this research examines the impact of chemical compositions on compressive strength and assesses the efficacy of various preprocessing techniques, including normalization and transformation methods.

1.2. Significance of the study

This study contributes to the field of sustainable construction by optimizing the formulation of pervious concrete through advanced machine-learning techniques. By providing insights into the influence of various chemical components on compressive strength, the research supports the development of more resilient and environmentally friendly construction materials. Additionally, the findings highlight the importance of effective data processing methods, contributing to the knowledge on enhancing predictive accuracy in concrete performance assessments. Ultimately, this study aims to inform engineers and practitioners, facilitating better material selection and mix design, essential for improving infrastructure resilience and supporting sustainable urban development.

1.3. Objective of the study

Despite the growing body of literature on pervious concrete, a significant gap remains in understanding the complex interactions between chemical compositions, particularly Supplementary Cementitious Materials (SCMs), and compressive strength. Much of the existing research has focused on limited SCMs, such as fly ash, which restricts the generalizability of the findings. Moreover, traditional empirical methods often fail to account for the non-linear relationships between mix components. This study addresses these gaps by leveraging advanced machine learning techniques to predict compressive strength, using a comprehensive dataset that incorporates a wide range of SCMs. By evaluating the influence of various chemical compositions and employing effective data processing methods, this study provides a more robust framework for pervious concrete design, contributing to more sustainable and resilient infrastructure solutions. The primary objective of this study is to conduct a comparative analysis of various machine learning techniques and data processing methods for predicting the compressive strength of pervious concrete containing supplementary cementitious materials. Specifically, the study aims to:

- Evaluate the impact of different chemical compositions on the compressive strength of pervious concrete.
- Identify the most effective machine learning models for predicting compressive strength, with a focus on Extreme Gradient Boosting (XGB).
- Examine the impact of various data processing techniques on model performance, including normalization and transformation methods.
- Analyze the significance of input variables, particularly chemical compositions, in predicting compressive strength through sensitivity analysis.
- Provide practical recommendations for optimizing pervious concrete formulations to enhance sustainability and structural integrity in construction applications.

2. Methodology

Fig. 1 outlines a framework for analyzing the compressive strength of pervious concrete. The methodology includes data cleaning, various normalization techniques, and dataset preparation, split into 70 % training and 30 % testing for machine learning models. Hyperparameter tuning and K-fold validation optimize performance, while prediction errors are analyzed through Venn diagrams and regression error characteristic (REC) curves. The study concludes with insights on sensitivity analysis, limitations, and future research directions.

2.1. Data collection

In the present study, the dataset was derived from the authors' previous research [8], which comprehensively examined the characteristics of pervious concrete incorporating various supplementary cementitious materials (SCMs). The dataset was compiled through a systematic literature review, following the PRISMA methodology, which focused on studies that reported the chemical composition of SCMs and their corresponding mix designs. The initial literature screening collected 659 observations from 20 published articles [22–41], which met the criteria of providing detailed information on the admixtures and

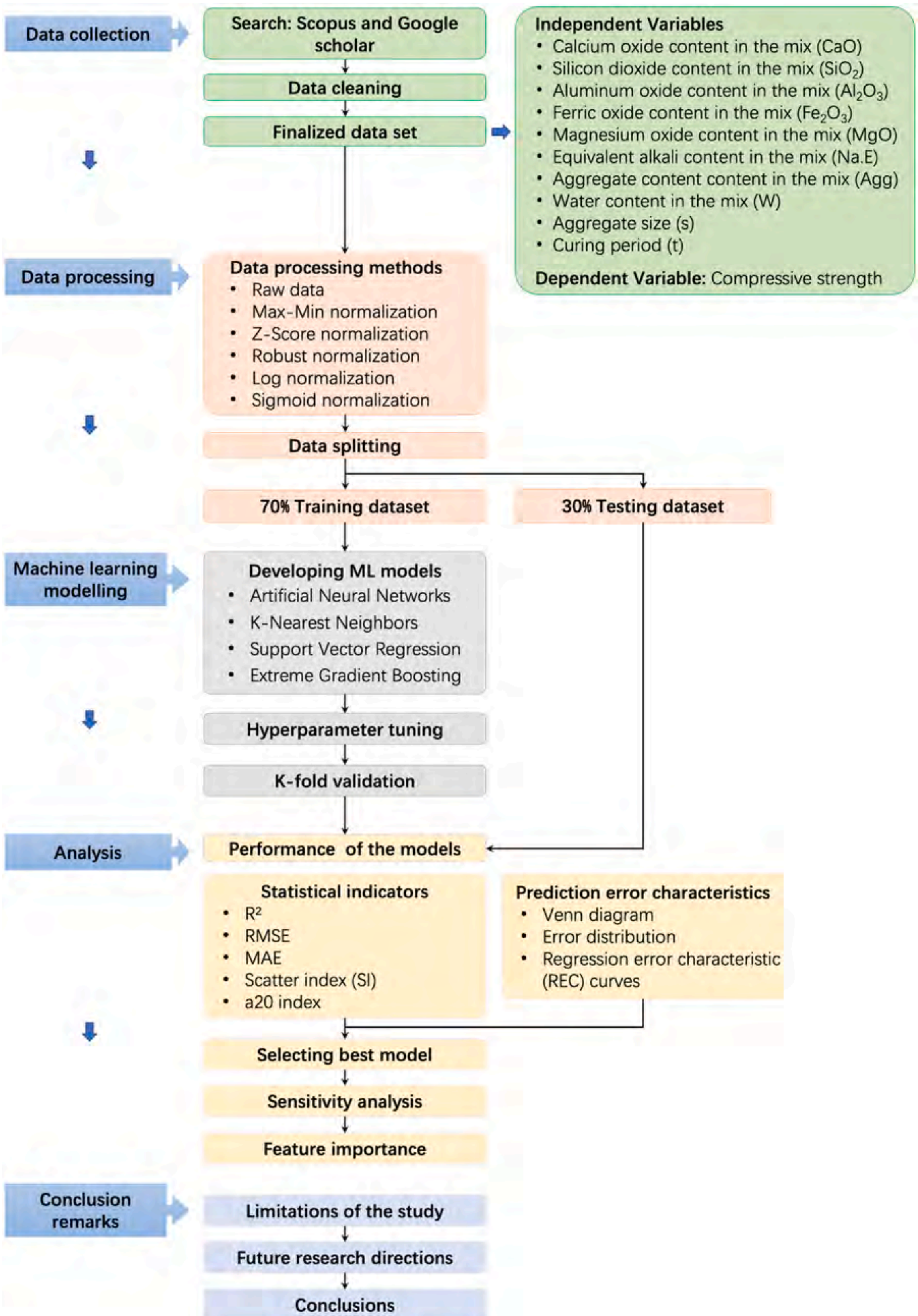

Fig. 1. Methodology flowchart used in the study.

mix designs used. Key parameters extracted included the chemical compositions of calcium oxide (CaO), silicon dioxide (SiO_2), ferric oxide (Fe_2O_3), aluminium oxide (Al_2O_3), magnesium oxide (MgO), equivalent alkali content, aggregate content, water content, aggregate size, and curing period. The studies included in this dataset were selected based on their relevance to the hydration processes and microstructural formation of pervious concrete, which are critical for understanding the compressive strength outcomes. Exclusions were made for studies lacking detailed descriptions of SCMs and those that did not provide sufficient information on the effects of aggregate size on strength characteristics. The final dataset is a robust foundation for developing predictive models to estimate the compressive strength of pervious concrete mixtures. For the present study, the following input variables are used:

- Calcium oxide content in the mix (CaO) in kg/m^3
- Silicon dioxide content in the mix (SiO_2) in kg/m^3
- Aluminium oxide content in the mix (Al_2O_3) in kg/m^3
- Ferric oxide content in the mix (Fe_2O_3) in kg/m^3
- Magnesium oxide content in the mix (MgO) in kg/m^3
- Equivalent alkali content in the mix (Na.E) in kg/m^3
- Aggregate content in the mix (Agg) in kg/m^3
- Water content in the mix (w) in kg/m^3
- Aggregate size (s) in mm
- Curing period (t) in days

Eq. (1) [8] calculates calcium oxide content in the mix.

$$CaO\left(in\frac{kg}{m^3}\right) = (C_{cement} \times CaO_{cement}/100) + \sum_{i=1}^{n}(C_{SCM_i} \times CaO_{SCM_i}/100) \tag{1}$$

Where, C_{cement}: cement content in the mix (kg/m^3), C_{SCM}: one particular SCM content in the mix (kg/m^3), i: ith number of SCM, n: number of SCMs in the mix, CaO_{cement}: CaO content in cement in percentage, and CaO_{SCM}: CaO content in SCM in percentage.

Analogously, the proportions of other chemical constituents within the mixture were determined using Eq. (1), substituting CaO with the respective chemical component. The equivalent alkali content was then calculated by summing the sodium oxide content and the potassium oxide content, converted using a factor of 0.658 ($Na_2O + 0.658\ K_2O$).

The distribution of the dataset utilized in the present study is analyzed through Fig. 2. The CaO ranges from a minimum of 123.62 kg/m^3 to a maximum of 384.55 kg/m^3 , with an average of 223.68 kg/m^3 . The skewness of 0.52 indicates a slight positive skew, while the kurtosis of −0.70 suggests a flatter distribution. The SiO_2 varies from 47.00 kg/m^3 to 178.66 kg/m^3 , averaging 105.86 kg/m^3 , with a skewness of 0.15 indicating near symmetry and a kurtosis of −0.62 reflecting a slightly flatter distribution. The Al_2O_3 exhibits a range from 8.54 kg/m^3 to 99.23 kg/m^3 , with an average of 25.05 kg/m^3 . The skewness of 1.93 indicates a significant positive skew, suggesting a concentration of lower values, while the high kurtosis of 6.06 indicates a sharp peak in the distribution. The Fe_2O_3 ranges from 1.95 kg/m^3 to 65.97 kg/m^3 , with an average of 15.25 kg/m^3 , a skewness of 1.83, and a kurtosis of 6.62, both indicating a strong positive skew and a peaked distribution, respectively. The MgO ranges from 2.51 kg/m^3 to 21.21 kg/m^3 , with an average of 9.54 kg/m^3 , displaying a skewness of 0.72 and a kurtosis of −0.40, suggesting a slight positive skew and a somewhat flatter distribution. Alkali content ranges from 0.00 kg/m^3 to 12.91 kg/m^3 , with an average of 3.43 kg/m^3 , a skewness of 1.03, and a kurtosis of 1.29, indicating a positive skew and a relatively normal distribution.

Coarse aggregate content values range from 689.0 kg/m^3 to 1833.6 kg/m^3 , averaging 1455.0 kg/m^3 , with a skewness of −0.83, reflecting a negative skew, and a kurtosis of 4.75, indicating a sharp peak. The water content ranges from 75.6 kg/m^3 to 265.0 kg/m^3 , with an average of 137.0 kg/m^3 , exhibiting a skewness of 1.44 and a kurtosis of 1.60, both of which indicate a positive skew and a moderate peak. The aggregate size ranges from 5 mm to 20 mm, with an average of 10.1 mm, a skewness of 0.82, and a kurtosis of −0.68, indicating a slight positive skew and a flatter distribution. The curing period ranges from one day to 120 days, with an average of 31.9 days, a skewness of 1.39, indicating a positive skew, and a kurtosis of 0.87, suggesting a distribution that is close to normal. Lastly, compressive strength values range from 4.01 MPa to 64.57 MPa, with an average of 16.09 MPa, exhibiting a skewness of 1.78 and a kurtosis of 4.43, which reflects a strong positive skew and a peaked distribution. This distribution analysis reveals significant variability in the dataset, with several parameters exhibiting skewed distributions. Such variability is crucial for understanding the behavior of pervious concrete mixtures and their compressive strength.

The distribution characteristics of these variables have significant implications for the predictive modeling process:

- Impact on Data Processing: The skewness and kurtosis of certain variables, particularly CaO and Fe_2O_3, suggest that normalization techniques such as Max-Min or Z-score normalization may be crucial in reducing the influence of extreme values or outliers. For instance,

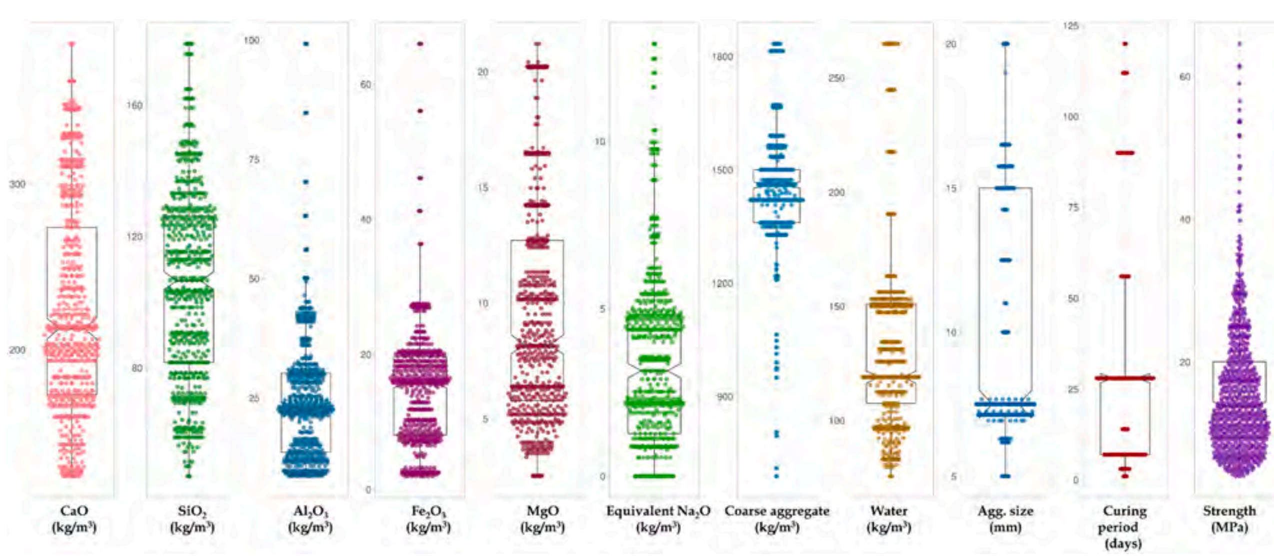

Fig. 2. Data distribution for the raw dataset.

features like CaO, which exhibit positive skewness, could potentially dominate the prediction model unless correctly scaled.

- Model Sensitivity: The variability in the distribution of input variables, especially SiO_2 and Al_2O_3, impacts how machine learning models, such as ANN and SVR, process these features. For example, models that rely heavily on distance-based metrics, like KNN, could benefit from normalization methods that ensure each feature contributes equally to the model. The data distribution also implies that the machine learning models used in this study may have varying degrees of difficulty in learning from this data, depending on the preprocessing method employed.
- Model Performance: Understanding the distribution of the data provides insight into model performance, as models trained on datasets with high skewness may face challenges in generalizing to unseen data. The tendency of the compressive strength data to cluster at the lower range could also result in models that perform well on predicting lower-strength values but struggle with high-strength predictions. This is something we address through various data processing methods to mitigate these biases.

2.2. Data processing methods

Normalization is a crucial preprocessing step, particularly for machine learning models that rely on distance metrics or assume normally distributed data. For instance, models like K-Nearest Neighbors (KNN) and Support Vector Regression (SVR) are susceptible to the scale of input features. If one feature has a much larger range than others (e.g., CaO content ranging from 0 to 1000 kg/m^3 versus aggregate size ranging from 5 to 20 mm), it can dominate the model's behavior, leading to biased predictions. For these models, normalization techniques like Max-Min normalization, Z-score normalization, or Robust normalization ensure that each feature contributes equally to the model's performance. This prevents certain features from unduly influencing the model, allowing algorithms to learn from the relationships among all input variables.

2.2.1. Max-min normalization

Max-Min normalization is a technique used to rescale data to a specific range, usually [0, 1]. This is particularly useful when preserving the relationships among data points while ensuring that all features have the same scale. The mathematical formula is given by Eq. (2) [42].

$$X\prime = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

Where:

X′ is the normalized value.
X is the original value.
Xmin is the minimum value in the dataset.
Xmax is the maximum value in the dataset.

2.2.2. Z-score normalization

Z-score normalization (standardization) transforms the data into a distribution with a mean of 0 and a standard deviation of 1. This method is useful for algorithms that assume a normally distributed dataset. The mathematical formula is given by Eq. (3) [42].

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

where

Z is the z-score normalized value.
X is the original value.
μ is the mean of the dataset.
σ is the standard deviation of the dataset.

2.2.3. Robust normalization

Robust normalization uses robust statistics (median and interquartile range) to mitigate the influence of outliers. This makes it a preferred choice in datasets where outliers are present. The mathematical formula is given by Eq. (4) [43].

$$X\prime = \frac{X - median(X)}{IQR} \tag{4}$$

Where:

X′ is the normalized value.
median(X) is the median of the dataset.
IQR=Q3 −Q1 is the interquartile range, with Q1 being the first quartile and Q3 being the third quartile.

2.2.4. Log normalization

Log normalization involves applying a logarithmic transformation to the data, which helps manage skewed distributions by compressing the scale of large values. The mathematical formula is given by Eq. (5) [44].

$$X' = \log(X + c) \tag{5}$$

where:

X′ is the log-normalized value.
c is a constant to avoid taking the log of zero.

2.2.5. Sigmoid normalization

Sigmoid normalization transforms the original values into a range between 0 and 1 using the sigmoid function, which has a characteristic S-shaped curve. The mathematical formula is given by Eq. (6) [45].

$$X\prime = \frac{1}{1 + e^{-X}} \tag{6}$$

Where:

X′ is the normalized value.
e is the base of the natural logarithm (approximately 2.71828).

Table 2 outlines the significance of data normalization in machine learning. Normalization is a crucial preprocessing step that adjusts the range and distribution of data features, ensuring that they contribute equally to the model's performance. The choice of normalization method can significantly impact the efficiency and accuracy of algorithms, tailoring the data to meet the specific requirements of various machine learning techniques [46]. By understanding the advantages and disadvantages of different normalization methods, practitioners can optimize model performance, enhance interpretability, and mitigate issues related to scale and outliers in data.

2.2.6. Distribution of data after normalization

The dataset analyzed in the present study employs various data processing techniques, as illustrated in Fig. 3 and the accompanying Table 3. The statistical parameters presented reflect the distribution of key components that affect the compressive strength of pervious concrete, processed through various data analysis techniques. In the Max-Min scaling technique, values for all parameters range from 0.00 to 1.00, with averages showing variability across components. For example, the average CaO content is 0.38, while the average C-Agg is significantly higher at 0.67. The skewness values indicate slight positive skewness for most parameters, particularly for Al_2O_3 (1.93) and Fe_2O_3 (1.83), suggesting that lower values are more prevalent. The kurtosis values for these components are notably high, indicating sharp peaks in their distributions.

The Z-score normalization technique standardizes the parameters, resulting in a mean of 0 and a standard deviation of 1. The skewness values remain consistent with the Max-Min scaling, particularly for Al_2O_3 and Fe_2O_3, which continue to show strong positive skewness. The kurtosis values also reflect similar characteristics, indicating peaked distributions. Robust scaling, which is less sensitive to outliers, exhibits

Table 2
Comparison between data processing methods for machine learning prediction.

Data Processing Method	Advantages	Disadvantages	Use Cases	Suitable ML Algorithms
Raw Data	Retains original values and relationships. Simple to implement.	Different scales can lead to biased results. Sensitive to outliers.	When data is already on the same scale. Exploratory data analysis.	Any algorithm that does not require scaling.
Max-Min Normalization	Scales data to a fixed range for straightforward interpretation. Preserves relationships.	Sensitive to outliers; can distort results. Requires known min/max values.	Image processing. Neural networks.	K-Means, KNN, Neural Networks.
Z-Score Normalization	Standardizes data with a mean of 0 and a standard deviation of 1. Handles Gaussian distributions well.	Not robust to outliers; they can skew mean/std.	When the data follows a normal distribution.	SVR, Linear Regression, Logistic Regression.
Robust Normalization	Reduces the effect of outliers. Uses median and IQR for better robustness.	May not center the data around 0. More complex to implement.	Datasets with significant outliers.	Tree-based models, Robust regression.
Log Normalization	Handles skewed distributions effectively. Compresses large ranges.	Cannot handle zero or negative values without adjustments. Interpretation can be less intuitive.	Financial data, where values can span orders of magnitude.	Regression models, Neural Networks.
Sigmoid Normalization	Maps values between 0 and 1, suitable for probabilities. Smoothly compresses extreme values.	Sensitive to outliers; can squash data into a small range. Non-linear transformation may distort relationships.	Probabilistic models, binary classification.	Neural Networks, Logistic Regression.

a maximum range of values similar to that of Z-score normalization. The averages across components vary slightly, with an average of 0.11 for CaO and 0.41 for Water content. The skewness and kurtosis values remain consistent with previous methods, particularly for Al_2O_3 and Fe_2O_3, highlighting their distribution characteristics. Log transformation adjusts the distribution of the data, particularly for parameters that might exhibit exponential characteristics. The average values for components such as CaO (2.33) and SiO_2 (2.01) indicate a shift towards higher values. Skewness values are lower for most components, with some becoming negative, indicating a more symmetric distribution compared to previous techniques. The kurtosis values reflect a flatter distribution for most parameters, except for Alkali content, which shows a high kurtosis value of 5.41, suggesting a peaked distribution.

The Sigmoid scaling technique results in values that range between 0.50 and 1.00, indicating a compression of the data towards the middle of the scale. The average values hover around 0.75 for most components, with the aggregate content demonstrating a higher average of 0.98. The skewness values are mostly negative, particularly for Al_2O_3 (-0.89) and Fe_2O_3 (-0.47), suggesting a concentration of higher values. The kurtosis values indicate a high peak for aggregate content (62.77), highlighting the influence of this parameter. Overall, the analysis of distributions across different data processing techniques reveals significant variability and shifts in the dataset's characteristics. Each technique impacts the skewness and kurtosis of the distributions, influencing how the parameters relate to the predicted compressive strength of pervious concrete. These insights are crucial for selecting the appropriate data processing method for modeling and predicting outcomes in the study.

2.3. Machine learning algorithms

The effectiveness of machine learning models depends not just on the model itself but also on the quality of the data and the preprocessing techniques applied. In this study, the selection of XGB, ANN, KNN, and SVR was based on their theoretical ability to model non-linear relationships and handle complex interactions between multiple input variables (e.g., chemical composition, aggregate size, curing period). Each of these models has strengths that make them suitable for predicting compressive strength:

- ANNs are selected for their ability to capture complex, non-linear relationships between input features and output predictions. The flexibility of neural networks allows them to learn intricate patterns in the data that simpler models may miss.
- KNN is chosen for its simplicity and ability to predict based on local patterns. However, it can suffer from the curse of dimensionality when data includes too many features, making normalization essential to ensure meaningful distances between points.
- SVR offers robustness against overfitting and is well-suited for high-dimensional data. It is beneficial when the relationship between the input variables and the target (compressive strength) is complex and non-linear.
- XGB is chosen due to its ability to handle large datasets, capture non-linear interactions, and its inherent robustness against overfitting. The gradient boosting technique ensures that each tree corrects the errors of the previous one, leading to high accuracy in predictions.

2.3.1. Artificial neural network (ANN)

The ANNs are computational models inspired by the neural networks in the human brain, consisting of interconnected nodes (neurons) organized in layers [47]. One of the primary advantages of using ANNs to predict compressive strength is their ability to model complex and non-linear relationships between input features and output targets, which is often essential in such predictions. Additionally, ANNs are adaptable, improving their predictions as more data becomes available, and they can automatically learn relevant features from raw data, reducing the need for manual feature extraction [48,49]. Their robustness allows them to handle noise and missing values effectively, and once trained, they can generalize well to unseen data. However, ANNs typically require large amounts of data for training to achieve good performance, and they are prone to overfitting the training data if not properly regularized [50]. Furthermore, ANNs often operate as "black boxes," making it challenging to interpret their decision-making processes. The computational intensity of training these networks can also be a drawback, requiring significant processing power and time. The ANN architecture used for predicting compressive strength is shown in Fig. 4.

Mathematically, an ANN can be represented as starting with an input vector, as shown in Eq. (7).

$$X = [x_1, x_2, \ldots, x_n] \tag{7}$$

where n is the number of input features, each connection between neurons has associated weights *(wij) and biases (bj)*. The output of a hidden neuron h_j is given by Eq. (8) [51].

$$h_j = f\left(\sum_i^n w_{ij}x_i + b_j\right) \tag{8}$$

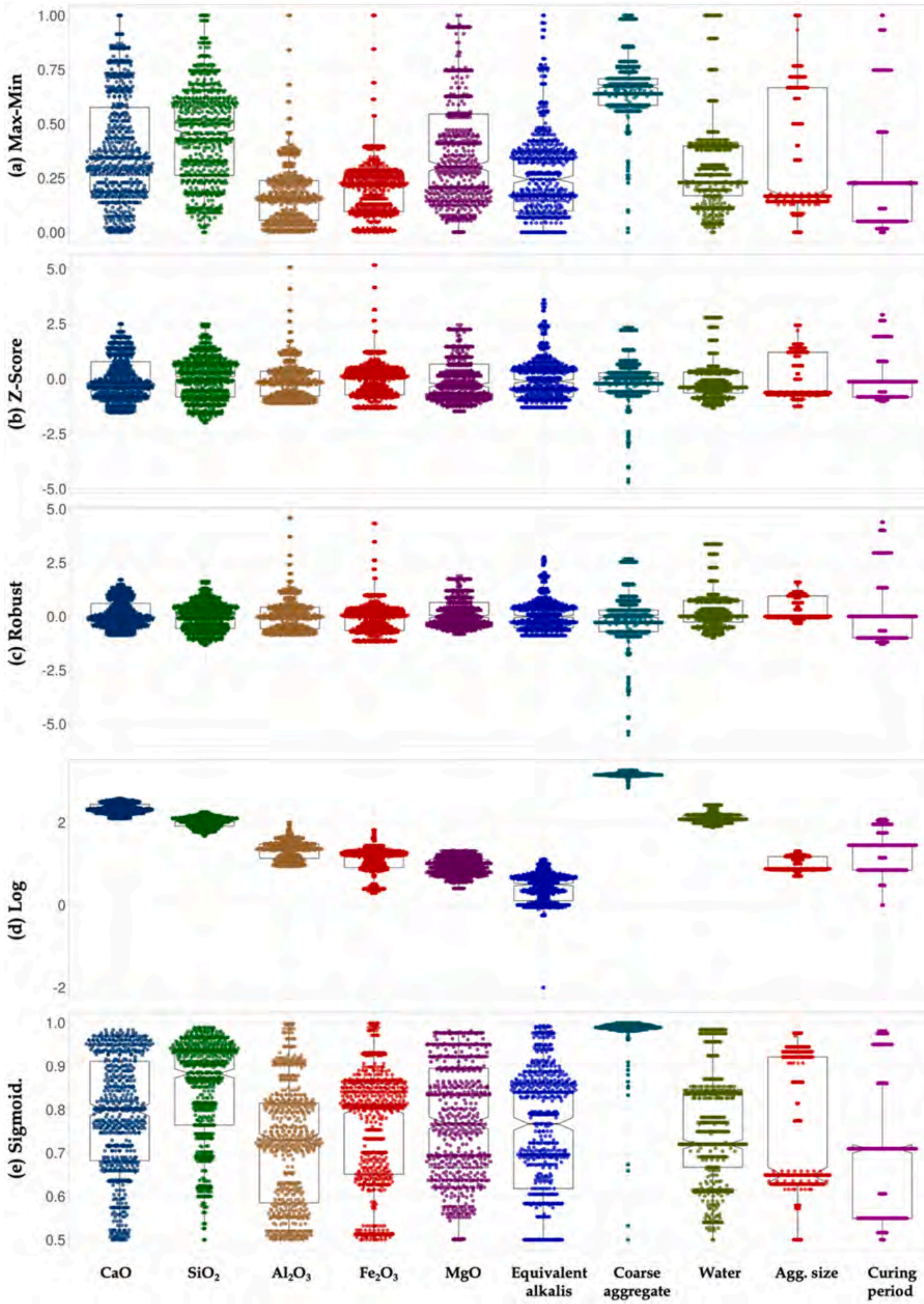

Fig. 3. Data distribution of the dataset after various data processing methods.

where f is an activation function (such as sigmoid or ReLU), the final output, y (the compressive strength prediction), is calculated using Eq. (9).

$$y = \sum_{i}^{m} w_j h_j + b \tag{9}$$

where m is the number of hidden neurons, and w_j and b are weights and

Table 3
Statistical indicators for different machine learning methods and data processing methods.

Data processing	Statistical values	CaO	SiO_2	Al_2O_3	Fe_2O_3	MgO	Alkali Eq.	C-Agg	Water	Agg.Size	Curing days	Strength
Raw	Minimum	123.62	47.00	8.54	1.95	2.51	0.00	689.00	75.57	5.00	1.00	4.01
	Maximum	384.55	178.66	99.23	65.97	21.21	12.91	1833.60	265.00	20.00	120.00	64.57
	Average	223.68	105.86	25.05	15.25	9.54	3.43	1455.05	137.01	10.08	31.93	16.09
	Standard Dev.	64.52	29.44	14.60	9.81	4.75	2.64	162.55	46.00	4.07	30.24	9.70
	Skewness	0.52	0.15	1.93	1.83	0.72	1.03	−0.83	1.44	0.82	1.39	1.78
	kurtosis	−0.70	−0.62	6.06	6.62	−0.40	1.29	4.75	1.60	−0.68	0.87	4.43
Max-Min	Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	Maximum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Average	0.38	0.45	0.18	0.21	0.38	0.27	0.67	0.32	0.34	0.26	
	Standard Dev.	0.25	0.22	0.16	0.15	0.25	0.20	0.14	0.24	0.27	0.25	
	Skewness	0.52	0.15	1.93	1.83	0.72	1.03	−0.83	1.44	0.82	1.39	
	kurtosis	−0.70	−0.62	6.06	6.62	−0.40	1.29	4.75	1.60	−0.68	0.87	
Z-Score	Minimum	−1.55	−2.00	−1.13	−1.35	−1.48	−1.30	−4.71	−1.33	−1.25	−1.02	
	Maximum	2.49	2.47	5.08	5.17	2.46	3.59	2.33	2.78	2.44	2.91	
	Average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	Standard Dev.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Skewness	0.52	0.15	1.93	1.83	0.72	1.03	−0.83	1.44	0.82	1.39	
	kurtosis	−0.70	−0.62	6.06	6.62	−0.40	1.29	4.75	1.60	−0.68	0.87	
Robust	Minimum	−0.88	−1.34	−0.87	−1.19	−0.82	−0.89	−5.51	−1.00	−0.32	−1.29	
	Maximum	1.70	1.62	4.58	4.32	1.88	2.76	2.67	3.36	1.59	4.38	
	Average	0.11	−0.01	0.12	−0.05	0.19	0.08	−0.03	0.41	0.33	0.19	
	Standard Dev.	0.64	0.66	0.88	0.84	0.69	0.75	1.16	1.06	0.52	1.44	
	Skewness	0.52	0.15	1.93	1.83	0.72	1.03	−0.83	1.44	0.82	1.39	
	kurtosis	−0.70	−0.62	6.06	6.62	−0.40	1.29	4.75	1.60	−0.68	0.87	
Log	Minimum	2.09	1.67	0.93	0.29	0.40	−2.00	2.84	1.88	0.70	0.00	
	Maximum	2.58	2.25	2.00	1.82	1.33	1.11	3.26	2.42	1.30	2.08	
	Average	2.33	2.01	1.34	1.08	0.92	0.43	3.16	2.12	0.97	1.31	
	Standard Dev.	0.12	0.13	0.23	0.32	0.22	0.35	0.05	0.13	0.16	0.44	
	Skewness	0.05	−0.40	0.09	−0.85	−0.12	−1.08	−2.12	0.73	0.48	−0.22	
	kurtosis	−0.84	−0.60	−0.42	0.35	−0.72	5.41	10.34	0.18	−1.09	−0.47	
Sigmoid	Minimum	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
	Maximum	0.98	0.99	1.00	1.00	0.98	0.99	1.00	0.98	0.98	0.98	
	Average	0.78	0.84	0.72	0.76	0.77	0.75	0.98	0.75	0.73	0.69	
	Standard Dev.	0.13	0.12	0.14	0.13	0.13	0.14	0.05	0.12	0.14	0.14	
	Skewness	−0.35	−0.89	0.02	−0.47	−0.12	−0.23	−7.54	0.13	0.40	0.63	
	kurtosis	−0.82	−0.21	−1.03	−0.75	−1.14	−1.07	62.77	−0.71	−1.33	−0.65	

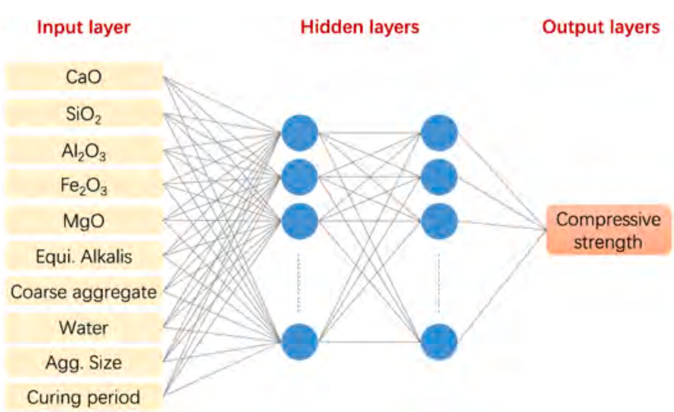

Fig. 4. ANN architecture used for the prediction of compressive strength.

bias for the output layer.

During training, the weights are adjusted using optimization algorithms, such as Gradient Descent, to minimize the loss function.

2.3.2. K-nearest neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, yet effective, machine learning algorithm used for classification and regression tasks. It operates on the principle that similar data points are likely to have similar outcomes. When predicting a target value, KNN identifies the kk training samples that are closest to the query point based on a distance metric, such as Euclidean distance [52]. The algorithm then makes predictions by averaging the outcomes (for regression) or by majority voting (for classification) of these nearest neighbors. The KNN architecture used for predicting compressive strength is shown in Fig. 5.

One of the key advantages of KNN is its simplicity and ease of implementation. It does not require any assumptions about the underlying data distribution, making it a non-parametric method. KNN is particularly effective in scenarios where the decision boundaries are irregular or complex, as it relies on local information from the data [52]. Additionally, because KNN utilizes the entire training dataset during the prediction phase, it can adapt easily to new data without requiring retraining. However, KNN has several limitations. Its performance can significantly degrade with high-dimensional data due to the "curse of dimensionality," where the distance between points becomes less informative as the number of dimensions increases [53].

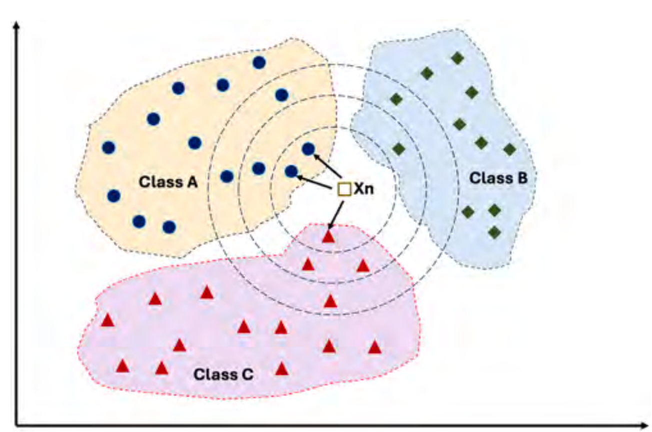

Fig. 5. KNN architecture used for the prediction of compressive strength.

2.3.3. Support vector regression (SVR)

Support Vector Regression (SVR) is a powerful machine learning technique derived from Support Vector Machines (SVM) used for regression tasks. It aims to find a function that deviates from the actual target values by a value no greater than a specified margin, while also being as flat as possible [54]. SVR is particularly effective in capturing complex relationships in data while maintaining robustness against overfitting. One of the primary advantages of SVR is its ability to handle high-dimensional data effectively. It employs a kernel trick in a transformed feature space, where linear relationships can be established even in non-linear datasets. This makes SVR versatile for various types of data distributions [55]. Additionally, it provides a clear margin of tolerance, allowing for a certain degree of error in predictions and making it less sensitive to outliers compared to other regression methods. However, choosing the appropriate kernel and tuning hyperparameters can be challenging and require significant computational resources [56]. Furthermore, while SVR can be effective in many scenarios, it may not perform as well with extensive datasets due to its higher computational complexity. Additionally, it can struggle when the underlying relationship in the data is highly complex or when there is insufficient training data to learn from [57]. The SVR architecture used to predict compressive strength is shown in Fig. 6.

Mathematically, SVR can be formulated as follows. Given a training dataset (x_i,y_i) for i = 1, 2, …, N, where x_i is the input features, and y_i are the corresponding target values, the goal is to find a function $f(x)$ that approximates the target values. This function can be expressed as Eq. (10) [51]:

$$f(x)= \langle w,\phi(x)\rangle+b \tag{10}$$

where w is the weight vector, $\phi(x)$ is the feature mapping, and b is the bias term.

To achieve the goal of minimizing the prediction error within a specified margin ϵ, the optimization problem can be formulated as Eq. (11):

$$\min_{w,b}\frac{1}{2}\|w\|^2 \tag{11}$$

Subject to the constraints:

$$y_i - f(x_i) \leq \epsilon + \xi_i$$

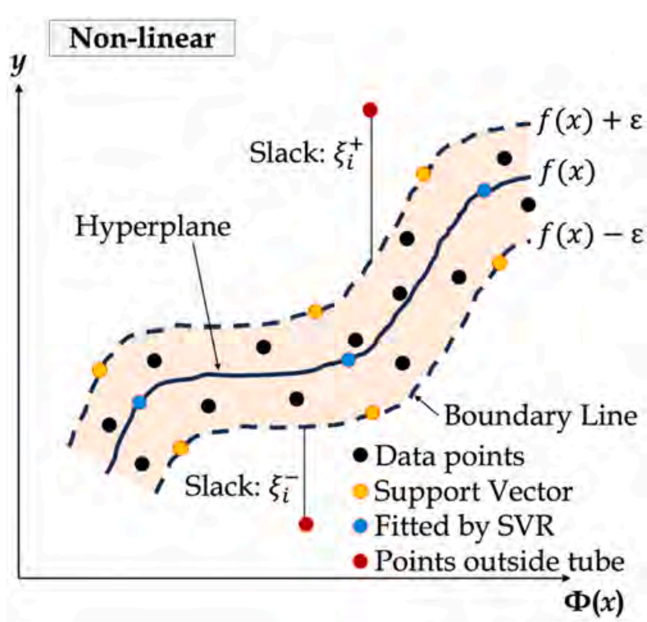

Fig. 6. SVR architecture used for the prediction of compressive strength.

$$f(x_i) - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

Here, ξ_i and ξ_i^* are slack variables allowing deviations greater than ϵ. By solving this optimization problem, SVR finds the best-fitting function that strikes a balance between minimizing error and model complexity.

2.3.4. Extreme gradient boosting (XGB)

Extreme Gradient Boosting (XGB) uses the concept of boosting, where weak learners (typically decision trees) are combined to form a powerful ensemble model. The mathematical formulation underlying XGB involves an objective function that combines a loss function and a regularization term to control model complexity [48]. The model is built additively, where each iteration adds a new tree to improve predictions, using gradient boosting to minimize the loss function iteratively. Overall, XGB is a powerful tool for predicting compressive strength, offering numerous advantages while presenting some challenges [58]. The XGB architecture used for the prediction of compressive strength is shown in Fig. 7. The mathematical formulation is as follows:

The objective function is defined as Eq. (12) [51].

$$L(\theta) = \sum_{i=1}^{n} l(y_i \hat{y}_i) + \sum_{j=1}^{k} \Omega(f_i) \tag{12}$$

where, l is the loss function (error). $\hat{y}_i$ is the predicted value, and Ω is a regularization term that controls the model's complexity.

Each iteration adds a new tree f_t to the model, as shown in Eq. (13).

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{13}$$

The model uses gradients to minimize the loss function iteratively, as shown in Eq. (14)

$$f_t = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)} + f(x_i)\right) + \Omega(f) \right] \tag{14}$$

XGB is a powerful tool for predicting compressive strength, with advantages such as high performance and flexibility. However, the model's complexity can make it harder to interpret compared to simpler models, and it requires careful tuning of hyperparameters, which can be time-consuming. Additionally, XGB can be sensitive to noisy data, potentially leading to poor generalization if not appropriately managed. While it is efficient, significant computational resources may still be required for extensive datasets.

2.3.5. Impact of normalization on specific models

The selection of normalization techniques affects the performance of different models. Understanding these impacts is crucial for selecting the most suitable preprocessing method for each model, ultimately enhancing prediction accuracy and model stability.

- ANN: ANNs, especially when using gradient-based optimization techniques, are sensitive to the scale of input features. Features with large scales can cause the gradient descent to converge more slowly or get stuck in local minima. Normalization helps to ensure faster convergence and improves model performance, particularly in training.
- KNN: Since KNN computes the distance between data points, normalization is crucial. Without normalization, features with large values, such as water content or CaO, could dominate the distance metric, leading to poor model performance. For KNN, normalization helps ensure that each feature contributes equally to the similarity measure, making the algorithm sensitive to all input variables.
- SVR: Similar to KNN, SVR is affected by the scale of input features. SVR models often use kernels that rely on the distance between points, so normalization ensures that the kernel's output is not biased

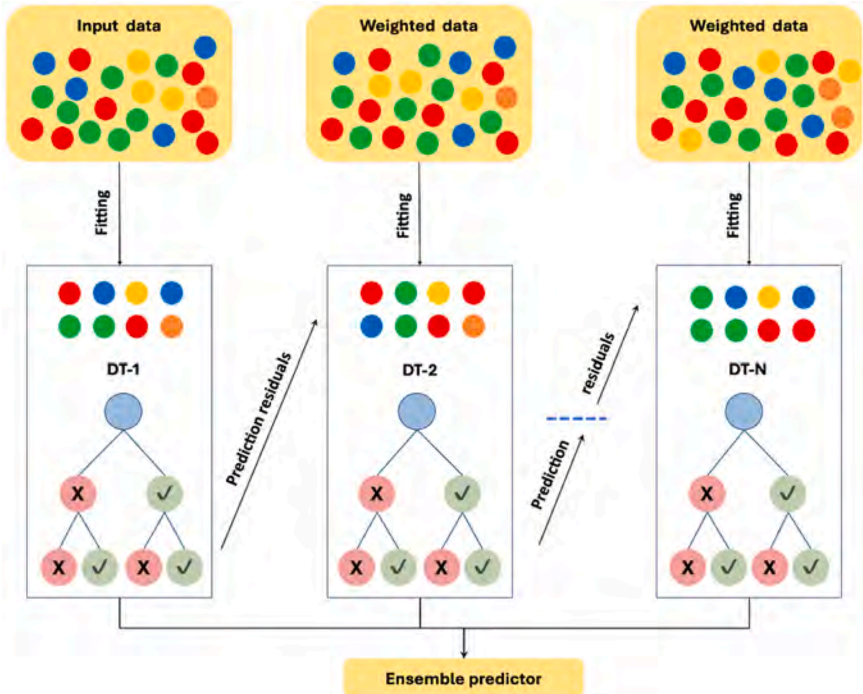

Fig. 7. XGB architecture is used for the prediction of compressive strength.

by features with larger scales. Additionally, Z-score normalization is particularly suited for SVR because it standardizes data to a Gaussian distribution, which SVR models often perform best with.

- XGB: XGB is a tree-based method, and trees are generally insensitive to feature scale. However, normalization can still improve performance, particularly in terms of convergence speed and model stability. Although XGB is less sensitive to normalization compared to KNN or SVR, applying it can help the model handle feature variability more efficiently.

2.4. Performance indicators

Coefficient of Determination (R^2): R^2 measures the proportion of variance in the dependent variable that can be explained by the independent variable(s) in a regression model. It ranges from 0 to 1, where a higher value indicates a better fit. The mathematical formulation for R^2 is given by Eq. (15).

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \tag{15}$$

Where: y_i is the actual values, $\hat{y}_i$ is the predicted values and $\bar{y}_i$ is the mean of actual values

Root Mean Square Error (RMSE): RMSE measures the average magnitude of the errors between predicted and actual values, showing how far predictions deviate from actual outcomes. Lower RMSE values indicate better model performance. Ideally, RMSE should be as close to 0 as possible. The mathematical formula for RMSE is given by Eq. (16).

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2} \tag{16}$$

where: n is the number of observations, y_i is the actual values and $\hat{y}_i$ is the predicted values

Mean Absolute Error (MAE): The MAE represents the average absolute difference between the predicted and actual values, providing a straightforward measure of prediction accuracy. Like RMSE, lower MAE values indicate better performance. An ideal MAE is also as close to 0 as possible. The mathematical formula for MAE is given by Eq. (17).

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \tag{17}$$

Scatter Index (SI): The SI quantifies the dispersion of errors relative to the magnitude of the predicted values, helping to assess the reliability of predictions. A lower SI indicates better model performance, with values closer to 0 preferred. The mathematical formula for SI is given by Eq. (18).

$$SI = \frac{RMSE}{\bar{y}} \tag{18}$$

a20 index: a20 measures the proportion of predictions that fall within 20 % of the actual values, serving as a gauge for prediction accuracy. Higher a20 values (closer to 1) are better.

Venn Diagram: In the study, a Venn diagram was employed to systematically analyze the distribution of prediction errors across four different machine learning models. The analysis focused on a significant threshold of 20 % prediction error to categorize the performance of each model. This approach enabled the visualization of overlaps and unique error characteristics among the models, thereby enhancing the understanding of their relative strengths and weaknesses.

Prediction error distribution: The model performance evaluation was conducted using two key metrics: prediction error analysis and area under the curve (AUC) for regression error characteristic (REC) curves. For the prediction error analysis, maximum and minimum error values were calculated for each model on the training and testing datasets. This analysis provided insight into the models' prediction accuracy and highlighted any tendencies to overestimate or underestimate compressive strength. Additionally, skewness was analyzed to understand the distribution of prediction errors, revealing potential biases in the models' predictions. AUC values were also computed to evaluate the models' ability to discriminate between different compressive strength levels. Insights gained from this comprehensive comparison focused on identifying trends in prediction accuracy, examining distribution characteristics, and assessing the overall robustness of the models across different preprocessing methods.

3. Results and discussions

3.1. Hyperparameter tuning

Table 4 summarizes the hyperparameter optimization for various machine learning models used to predict compressive strength, employing different data processing methods. Different activation functions, such as linear, Gaussian, and TanH, are tested, with the Tanh Function frequently emerging as a strong performer for the ANN model. The number of layers in the models ranges from 1 to 16, suggesting that a higher number of layers may lead to improved performance, although this must be balanced against the risk of overfitting. The node configurations, varying from 0.01 to 0.5, directly impact the model's complexity and capacity to learn from data. Learning rates are also crucial for convergence speed and stability, with optimized values revealing the sensitivity of results to this parameter.

For K-Nearest Neighbors (KNN), the optimal value of K tends to be around 2–3, suggesting that a smaller neighborhood can enhance predictions. In Support Vector Machines (SVM), hyperparameters such as kernel type and cost significantly influence model performance, with linear and radial basis function (RBF) kernels demonstrating varying effectiveness based on the dataset. Decision trees and ensemble methods are evaluated through parameters such as tree depth and subsampling rates, highlighting the importance of striking a careful balance between bias and variance for optimal predictive performance. The results indicate that no single model universally outperforms others across all hyperparameter settings; instead, model performance is highly dependent on the specific dataset and chosen hyperparameters. Certain hyperparameters, particularly the learning rate and number of nodes, significantly influence predictive accuracy, underscoring the importance of hyperparameter tuning in machine learning.

3.2. K-fold validation

The K-fold validation results (RMSE) for various machine learning methods and data processing techniques reveal significant insights into their performance, as shown in Fig. 8. Among the methods evaluated, XGB stands out as the best performer on both the training and testing datasets. On the training dataset, XGB achieved a remarkable RMSE of 1.03 MPa, accompanied by a standard deviation of 0.35 MPa, indicating accuracy and consistency in its predictions. This trend continues in the testing dataset, where XGB posted an RMSE of 2.85 MPa, further solidifying its reputation for robustness across different data conditions.

In contrast, although it showed competitive performance with a training RMSE of 1.86 MPa, the ANN method experienced a significant drop in predictive accuracy on the testing dataset, yielding an RMSE of 4.35 MPa. This discrepancy suggests that ANN may be prone to overfitting, as it performs well on the training data but struggles to generalize to new data. Similarly, KNN exhibited a large gap between its training RMSE of 2.39 MPa and testing RMSE of 5.09 MPa, indicating another potential case of overfitting. Other methods, such as SVR and various preprocessing techniques (Max-Min, Z-Score, Robust, and Log), showed varying performance levels. For instance, SVR achieved a training RMSE of 3.28 MPa, which improved to 2.69 MPa with Log transformation in testing, demonstrating its adaptability but still trailing behind XGB. The Max-Min scaling method with KNN also showed promise with a training RMSE of 1.23 MPa, but like ANN, it suffered in testing. These results indicate that XGB emerged as the superior machine learning method across both datasets, benefiting from its ability to minimize error effectively. The analysis highlights concerns regarding overfitting, particularly for ANN and KNN, which exhibited significant discrepancies between training and testing performances.

3.3. Performance of the models and effect of data processing

3.3.1. Artificial neural network

Fig. 9 illustrates the ANN model's predicted vs. measured compressive strength using various data processing techniques. The results in Table 5 indicate the performance of various data processing methods applied to ANN for predicting compressive strength. The raw data processing method achieved the highest R^2 value of 0.96 for the training dataset, indicating a strong correlation between the predicted and actual compressive strength. However, it showed a notable decrease to 0.81 in the testing dataset, suggesting potential overfitting. In contrast, the log transformation method also exhibited strong performance, with an R^2 of 0.97 in training and a relatively high 0.90 in testing, indicating better generalization.

The RMSE values for the training dataset ranged from 1.78 MPa to 2.10 MPa, with the log transformation yielding the lowest RMSE of 1.78 MPa, indicating that it had the smallest average prediction error. The log processing performed best for the testing dataset with an RMSE of 3.14 MPa, closely followed by the raw and Z-score methods. Regarding MAE, the log transformation achieved the lowest values of 1.21 MPa in training and 1.98 MPa in testing, indicating more accurate predictions on average. The SI values were relatively consistent across methods, with log transformation yielding an SI of 0.11 in training, indicating less variability in prediction errors.

When comparing the different processing methods, raw data yielded high training R^2 values but significantly dropped in testing, suggesting poor generalization. Max-Min and Z-score normalization provided

Table 4
Hyperparameter optimization for prediction models.

Model	Hyperparameters	Analyzed Values	Raw	Max-Min	Z-Score	Robust	Log	Sigmoid
ANN	Activation	[Linear, Gaussian,	TanH	TanH	TanH	TanH	TanH	TanH
	Functions	TanH]	2	2	2	2	2	2
	Number of layers	[1,2]	[5,11]	[5,10]	[15]	[9,15]	[14,16]	[9,9]
	Nodes per layer	[1–16]	0.05	0.05	0.05	0.1	0.1	0.1
	Learning rate	[0.01, 0.25, 0.05, 0.1, 0.5]						
KNN	K	[1–10]	3	3	2	3	2	3
SVR	Kernel	[linear, Radial basis	Radial basis	Radial basis	Radial basis	Radial basis	Radial basis	Radial basis
	Cost	function]	function	function	function	function	function	function
	Gamma	[0.01–5]	4.73902	4.85500	4.70928	4.93898	3.32549	2.54962
		[0.001–0.5]	0.10715	0.13050	0.09520	0.17673	0.32676	0.10027
XGB	Max_depth	[1–7]	7	5	7	7	6	7
	Subsample	[0.5–1]	0.8778	0.9675	0.9835	0.7527	0.9354	0.8840
	Colsample_by tree	[0.5–1]	0.7036	0.8262	0.9629	09503	0.8239	0.7577
	Min_child_weight	[1–3]	1.4922	1.0964	1.9100	1.0148	1.2301	1.5933
	Alpha	[0–0.5]	0.1300	0.2030	0.4294	0.2787	0.1699	0.3489
	Lambda	[0–2]	1.3667	0.3379	0.0466	1.1174	0.3164	0.3088
	Learning rate	[0.05–0.2]	0.3880	0.3483	0.17278	0.3746	0.3231	0.3464
	Iterations	[20–50]	49	42	32	50	45	32

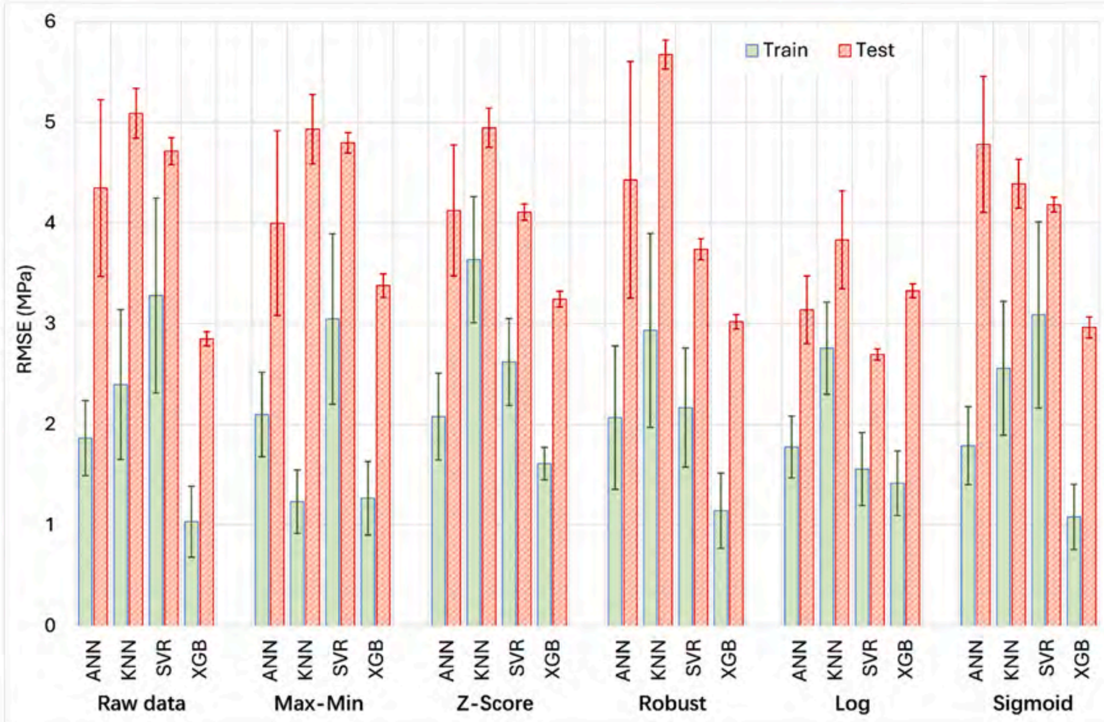

Fig. 8. RMSE variation for K-fold validation for different machine learning and data processing techniques.

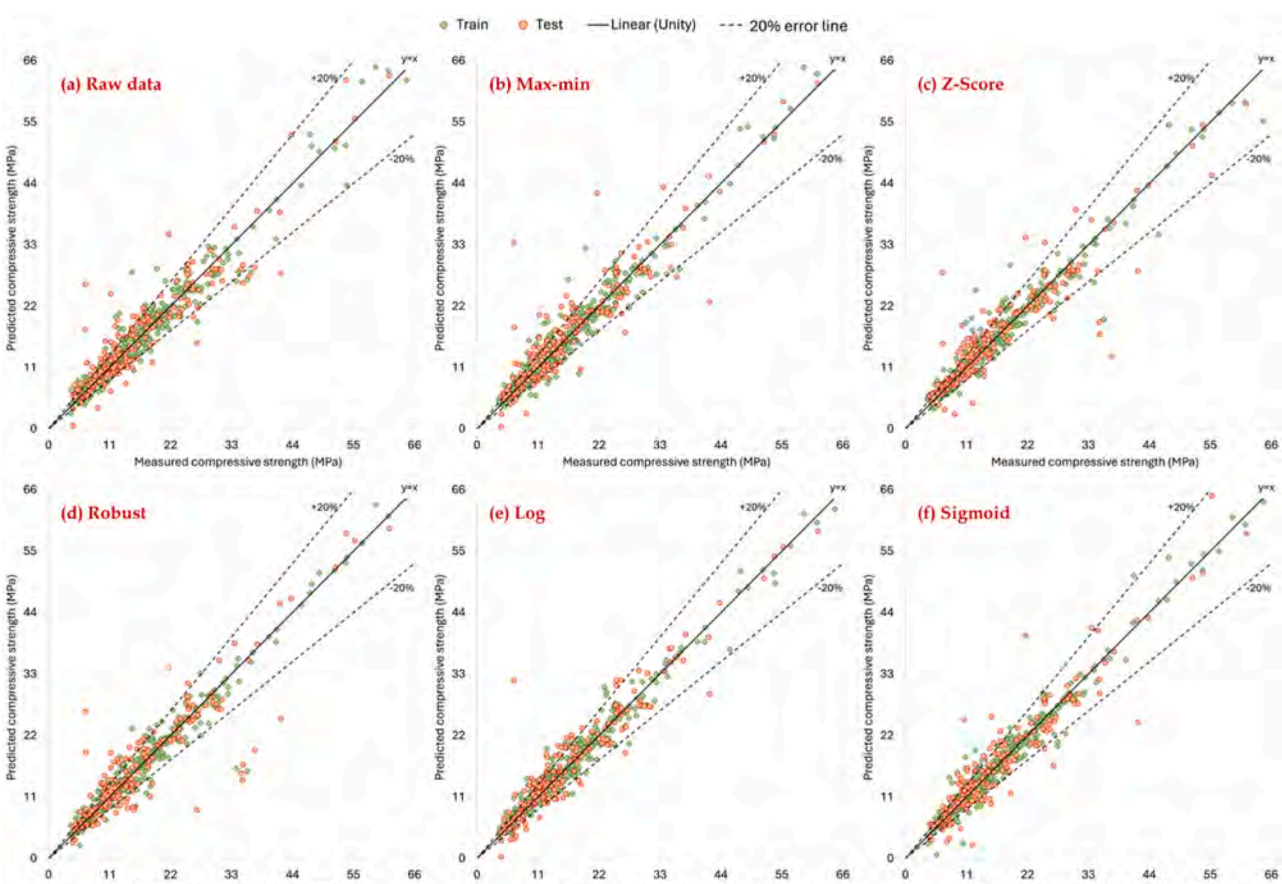

Fig. 9. Predicted vs. actual compressive strength for the dataset after different data processing techniques for the ANN model.

Table 5
Performance indicators for the various machine learning models with different data processing techniques.

		Train					Test				
ML	Data process	R^2	RMSE (MPa)	MAE (MPa)	SI	a20	R^2	RMSE (MPa)	MAE (MPa)	SI	a20
ANN	Raw	0.96	1.86	1.25	0.12	0.90	0.81	4.35	2.40	0.26	0.78
	Max-Min	0.95	2.10	1.43	0.13	0.87	0.84	4.00	2.42	0.24	0.71
	Z-score	0.95	2.08	1.23	0.13	0.89	0.83	4.13	2.34	0.25	0.76
	Robust	0.95	2.07	1.16	0.13	0.90	0.81	4.43	2.45	0.27	0.72
	Log	0.97	1.78	1.21	0.11	0.91	0.90	3.14	1.98	0.19	0.73
	Sigmoid	0.96	1.79	1.18	0.11	0.91	0.78	4.78	2.51	0.29	0.73
KNN	Raw	0.94	2.40	1.32	0.15	0.91	0.75	5.09	2.68	0.31	0.77
	Max-Min	0.98	1.23	0.42	0.08	0.95	0.76	4.93	2.76	0.30	0.73
	Z-score	0.85	3.64	1.78	0.23	0.85	0.76	4.95	2.95	0.30	0.68
	Robust	0.91	2.93	1.63	0.18	0.87	0.68	5.67	3.12	0.34	0.73
	Log	0.92	2.76	1.31	0.17	0.90	0.86	3.83	2.46	0.23	0.76
	Sigmoid	0.93	2.56	1.49	0.16	0.88	0.81	4.39	2.53	0.27	0.76
SVR	Raw	0.88	3.28	1.80	0.21	0.86	0.78	4.72	2.66	0.29	0.74
	Max-Min	0.90	3.05	1.72	0.19	0.87	0.77	4.80	2.62	0.29	0.72
	Z-score	0.92	2.62	1.58	0.16	0.89	0.83	4.11	2.43	0.25	0.74
	Robust	0.95	2.17	1.39	0.14	0.91	0.86	3.74	2.21	0.23	0.79
	Log	0.97	1.56	1.16	0.10	0.93	0.93	2.69	1.84	0.16	0.78
	Sigmoid	0.89	3.09	1.80	0.19	0.86	0.83	4.18	2.51	0.25	0.73
XGB	Raw	0.99	1.04	0.67	0.06	0.98	0.92	2.85	1.74	0.17	0.86
	Max-Min	0.98	1.27	0.87	0.08	0.97	0.89	3.38	2.06	0.21	0.80
	Z-score	0.97	1.61	1.23	0.10	0.93	0.90	3.24	1.97	0.20	0.82
	Robust	0.99	1.14	0.78	0.07	0.98	0.91	3.02	1.85	0.18	0.86
	Log	0.98	1.42	1.03	0.09	0.95	0.89	3.33	2.11	0.20	0.80
	Sigmoid	0.99	1.08	0.73	0.07	0.98	0.91	2.96	1.87	0.18	0.83

consistent results but did not outperform the log transformation. The robust scaling method demonstrated consistent performance but did not excel in any metrics, suggesting that while it may mitigate outlier effects, it does not significantly enhance predictive performance. The log transformation consistently delivered the best results across all indicators, demonstrating its effectiveness in stabilizing variance and improving ANN's predictive capability. While the sigmoid transformation performed adequately, its lower testing performance

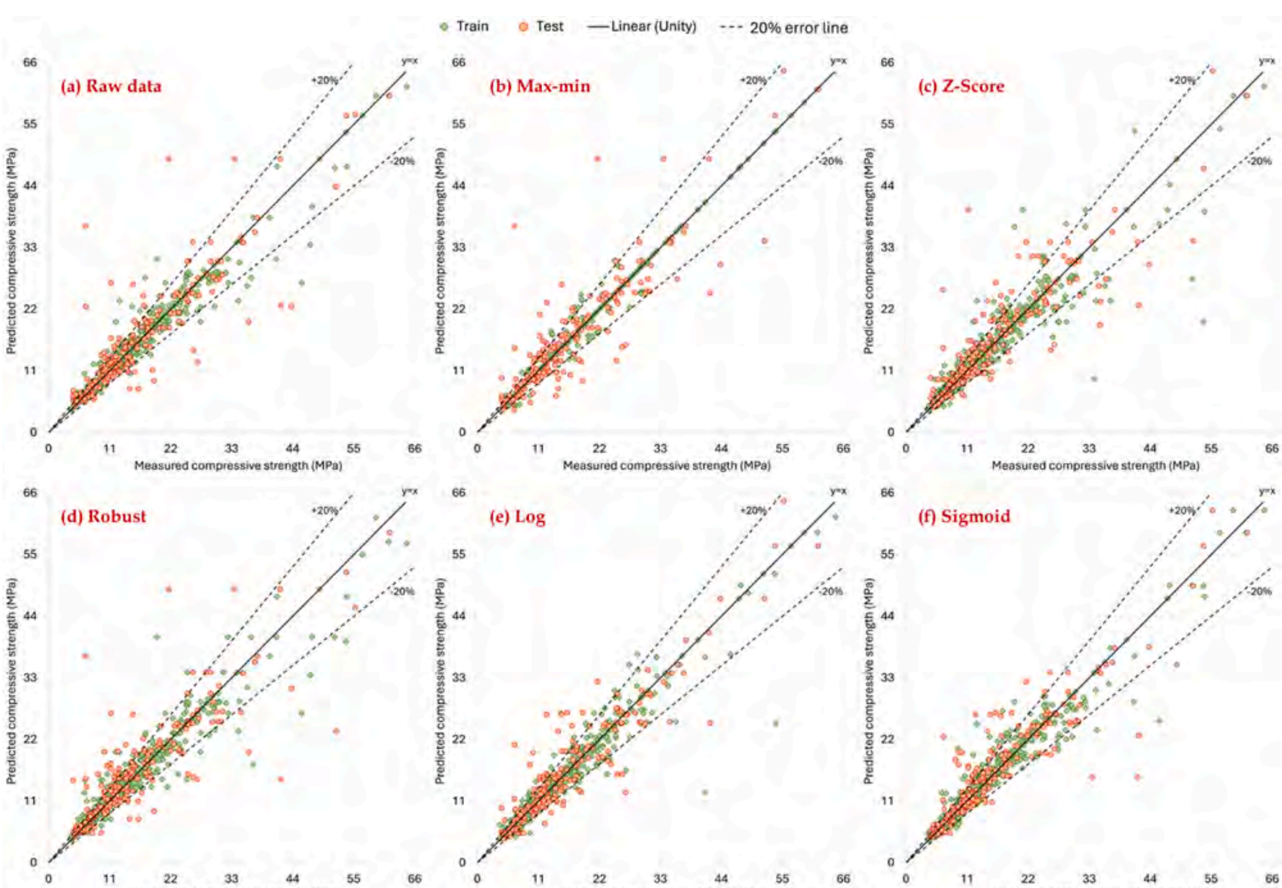

Fig. 10. Predicted vs. actual compressive strength for the dataset after different data processing techniques for the KNN model.

indicated a failure to capture underlying data patterns as effectively as other methods.

3.3.2. K-nearest neighbors

The results of predicting compressive strength using the KNN algorithm with various data processing methods are illustrated in Fig. 10. The raw data yielded strong training performance ($R^2 = 0.94$, RMSE = 2.40 MPa), but there was a notable drop in testing performance ($R^2 = 0.75$, RMSE = 5.09 MPa), suggesting potential overfitting. In contrast, Max-Min normalization achieved the highest training R^2 (0.98) but exhibited generalization issues in testing ($R^2 = 0.76$, RMSE = 4.93 MPa). This indicates that while this method effectively captures training data, it may not perform well on unseen data. The Z-score normalization method demonstrated lower training performance ($R^2 = 0.85$, RMSE = 3.64 MPa), indicating it may not adequately represent the data distribution.

Robust scaling provided decent training results ($R^2 = 0.91$) but struggled with generalization, similar to the raw data. Conversely, the log transformation produced balanced results, with training R^2 of 0.92 and testing R^2 of 0.86, suggesting that it stabilizes variance while effectively capturing data structure. The sigmoid transformation showcased solid training performance ($R^2 = 0.93$) yet still exhibited some overfitting, as reflected in its testing metrics ($R^2 = 0.81$).

The SI values indicate the relative scaling of the error metrics, with lower values suggesting better scaling performance. In this study, the Max-Min normalization method achieved the best SI score of 0.08 during training, indicating highly effective scaling, while the robust scaling method had an SI of 0.18, suggesting moderate scaling performance. The log transformation and sigmoid methods also displayed competitive SI values, reinforcing their effectiveness in error scaling.

When examining the a20 metric, which reflects the proportion of predictions that fall within a 20 % error margin, the Max-Min normalized data achieved an a20 score of 0.73 during testing, indicating that 73 % of the predictions were within 20 % of the actual values, which highlights its strong predictive capacity despite the drop in R^2. The log transformation also yielded a noteworthy a20 of 0.76, suggesting it maintained a robust predictive performance. In contrast, the Z-score normalization method had a lower a20 score of 0.68, indicating that it struggled to keep a significant portion of predictions within an acceptable error margin.

3.3.3. Support vector regression

The results of predicting compressive strength using SVR with various data processing methods are illustrated in Fig. 11. The raw data yielded a training R^2 of 0.88 and a testing R^2 of 0.78, indicating solid training performance but potential overfitting. Max-Min normalization improved the training R^2 to 0.90; however, the testing performance dropped to 0.77, highlighting generalization issues. Z-score normalization showed an intense training R^2 of 0.92 but a testing R^2 of 0.83, suggesting it may not adequately capture the underlying data distribution. Robust scaling achieved an impressive training R^2 of 0.95 and a testing R^2 of 0.86, indicating effective modelling. Notably, the log transformation yielded the highest training R^2 of 0.97 and a testing R^2 of 0.93, indicating that it effectively stabilizes variance and captures the data structure. The sigmoid transformation displayed decent training results ($R^2 = 0.89$) but struggled to maintain performance during testing ($R^2 = 0.83$).

The raw data yielded a training RMSE of 3.28 MPa and a testing RMSE of 4.72 MPa, indicating a decline in performance and suggesting potential overfitting. With Max-Min normalization, the training RMSE improved to 3.05 MPa; however, the testing RMSE rose slightly to 4.80 MPa, indicating ongoing challenges with generalization. Z-score

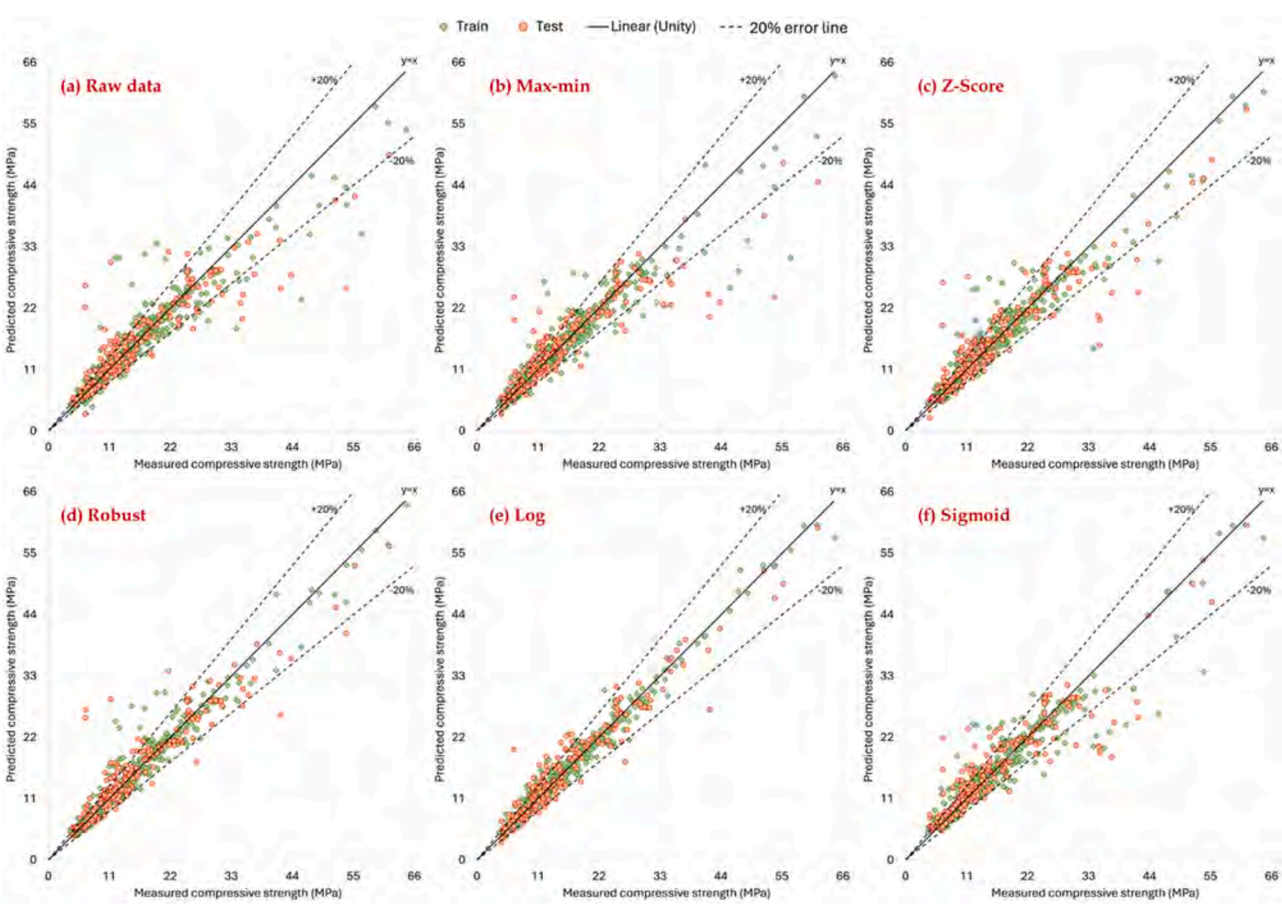

Fig. 11. Predicted vs. actual compressive strength for the dataset after different data processing techniques for the SVR model.

normalization yielded a training RMSE of 2.62 MPa and a testing RMSE of 4.11 MPa, indicating effective training but suboptimal performance on unseen data. Robust scaling achieved an impressive training RMSE of 2.17 MPa, with a testing RMSE of 3.74 MPa, indicating effective model generalization. In comparison, the log transformation yielded the lowest RMSE values of 1.56 MPa for training and 2.69 MPa for testing, indicating its ability to capture data patterns effectively. The sigmoid transformation yielded a training RMSE of 3.09 MPa and a testing RMSE of 4.18 MPa, suggesting that this method struggles to maintain performance across data splits.

The log transformation achieved the lowest SI of 0.10, indicating superior error scaling, while robust scaling followed closely with an SI of 0.14. The a20 metric, crucial for assessing practical predictive accuracy, revealed that the log transformation had an a20 score of 0.93 during testing, indicating that 93 % of predictions fell within acceptable error limits. Robust scaling also performed well with an a20 of 0.79, whereas Max-Min normalization and sigmoid transformations exhibited lower a20 scores of 0.72 and 0.73, respectively. In conclusion, the choice of data processing method has a significant impact on SVR performance in predicting compressive strength. The log transformation is the most effective approach, offering high training and testing performance, superior SI, and a20 metrics.

3.3.4. Extreme gradient boosting

The results of predicting compressive strength using the XGB algorithm with various data processing methods are illustrated in Fig. 12. The raw data yielded a training R^2 of 0.99 and a testing R^2 of 0.92, indicating exceptional model performance during training. However, the testing performance suggests a slight decline in generalization capability. The RMSE for the raw data was 1.04 MPa for training and 2.85 MPa for testing, indicating a low prediction error but also highlighting potential overfitting.

Max-Min normalization resulted in a training R^2 of 0.98 and a testing R^2 of 0.89, with RMSE values of 1.27 MPa for training and 3.38 MPa for testing. While this method maintained high training performance, the increase in testing RMSE indicates some challenges in generalization. Z-score normalization showed a training R^2 of 0.97 and a testing R^2 of 0.90, with RMSE values of 1.61 MPa for training and 3.24 MPa for testing, suggesting that this method may not fully capture the data distribution. Robust scaling achieved a training R^2 of 0.99 and a testing R^2 of 0.91, with RMSE values of 1.14 MPa for training and 3.02 MPa for testing, indicating effective modelling with balanced performance across training and testing. The log transformation provided a training R^2 of 0.98 and a testing R^2 of 0.89, with RMSE values of 1.42 MPa for training and 3.33 MPa for testing. Finally, the sigmoid transformation exhibited a training R^2 of 0.99 and a testing R^2 of 0.91, with RMSE values of 1.08 MPa for training and 2.96 MPa for testing, indicating strong predictive capability.

The raw data had an SI of 0.06 during training, indicating excellent error scaling, while the Max-Min normalization and Z-score normalization methods had SIs of 0.08 and 0.10, respectively. The robust scaling method achieved an SI of 0.07, indicating effective error management, similar to the sigmoid transformation, which also had an SI of 0.07. The a20 metric revealed that the raw data achieved an impressive score of 0.86, demonstrating that 86 % of its predictions fell within the 20 % error margin. In contrast, Max-Min normalization and log transformation resulted in an a20 score of 0.80. The robust scaling and sigmoid methods achieved a20 scores of 0.86 and 0.83, respectively, indicating reliable predictive accuracy.

Overall, among the models, XGB consistently demonstrated superior predictive capability, achieving the highest training and testing R^2 values alongside the lowest RMSE, indicating its robustness in capturing

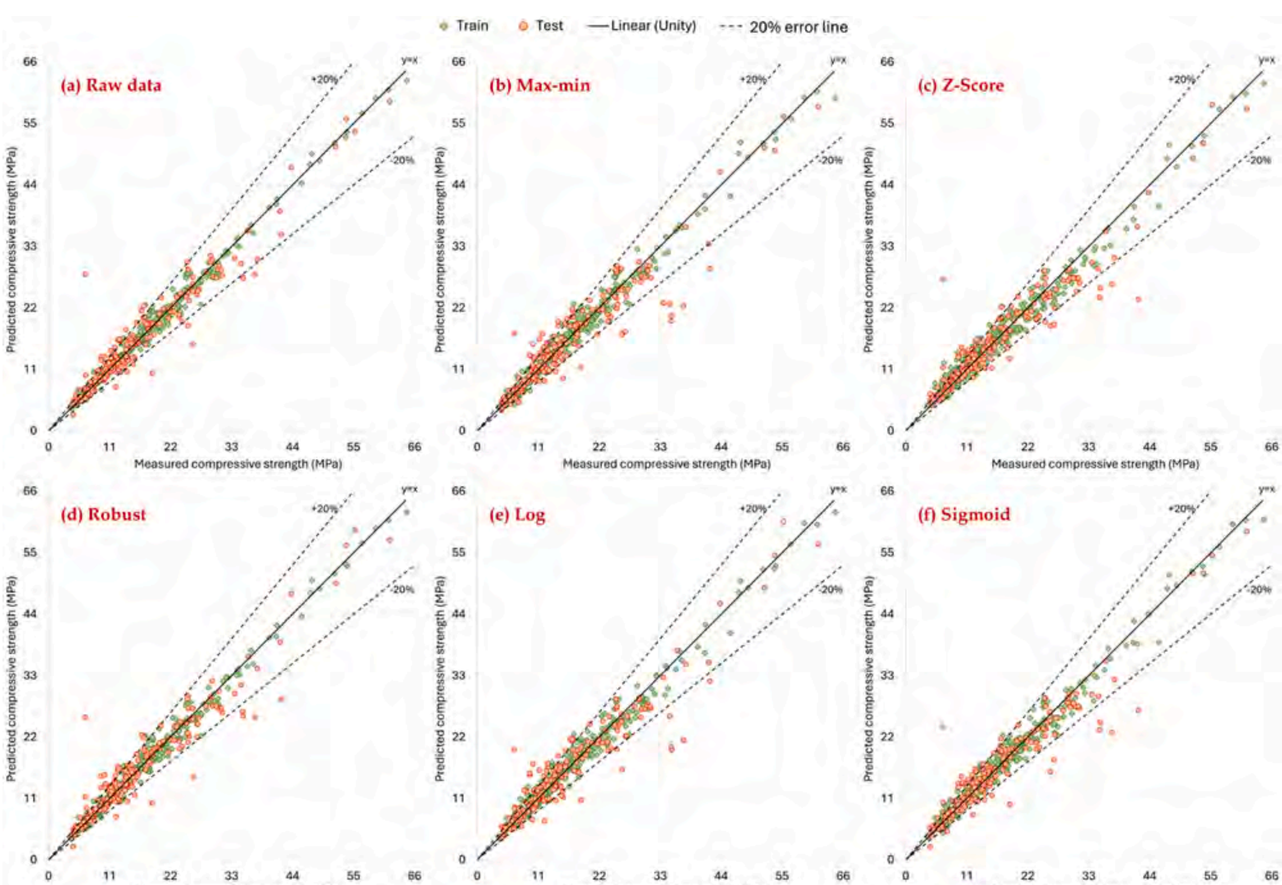

Fig. 12. Predicted vs. actual compressive strength for the dataset after different data processing techniques for the XGB model.

the underlying patterns in the data. This suggests that XGB is the most effective model for this task, as it benefits from its ability to handle complex relationships and interactions within the dataset.

3.3.5. Data processing methods

The log transformation and robust scaling techniques proved particularly effective in data processing methods, yielding high performance across most models, especially with XGB. The log transformation stabilized variance and improved the model's ability to generalize to unseen data. Robust scaling provided a balanced approach, minimizing the influence of outliers while maintaining predictive accuracy. While Max-Min normalization also showed competitive results, it did not consistently outperform the log transformation or robust scaling across all models. Overall, combining the XGB model with raw

Fig. 13. Venn diagram for the number of data points with prediction error within 20 % of actual values. The areas where the ellipses overlap indicate the datapoints shared between the machine learning models, while the non-overlapping areas represent elements unique to each set.

data and log transformation appears to be the most promising strategy for accurately predicting compressive strength, underscoring the importance of selecting the correct algorithm and effective preprocessing techniques to enhance model performance and generalization capabilities.

These results show that data normalization is often less critical for XGB than other machine learning models. First, XGB is a gradient-boosting algorithm that builds decision trees, which are not sensitive to the scale of input features. Decision trees split data based on feature thresholds, making the range of the data less relevant. Additionally, XGB handles features with different scales and distributions effectively, as the algorithm inherently manages feature importance and can cope with varying magnitudes. Furthermore, tree-based methods, including XGB, are generally robust to outliers and do not require normalization to mitigate their effects, unlike linear models, which extreme values can heavily influence. XGB also incorporates regularization techniques (L1 and L2) to control overfitting without normalized data. Empirical studies show that XGB often performs well without normalization, and in many cases, it does not lead to significant improvements in model performance. In contrast, models KNN and SVR can significantly benefit from normalization, as they are sensitive to the scale of input features, ensuring that each feature contributes equally to distance calculations and optimization processes.

Fig. 13 illustrates the efficacy of various machine learning models and data processing techniques in predicting outcomes with a 20 % error margin, based on actual data. The highest number of predictions falling within the 20 % error margin by all the machine learning models was observed with the Log transformation (474 predictions), followed closely by the Robust technique (467 predictions). This suggests that the Log transformation effectively normalizes the data, reducing skewness and stabilizing variance, which enhances model performance. Among the ML models, XGB demonstrated a robust performance across various data processing techniques. Although it did not achieve the highest number of predictions, it maintained a consistent performance, indicating its reliability in handling complex relationships within the data.

Raw Data yielded 462 predictions, demonstrating that even unprocessed data can provide a reasonable baseline for model performance. The results highlight the importance of effective outlier management in predictive modelling. The Robust technique, designed to mitigate the influence of outliers, showcased significant predictive capability, evidenced by 467 predictions within the 20 % error margin. Based on the analysis, the Log transformation paired with XGB appears to be the most suitable approach for predictions, yielding the highest accuracy. This combination not only maximizes the number of predictions within the acceptable error range but also leverages the strengths of XGB in managing complex datasets.

3.4. Prediction error

Fig. 14 shows the analysis of error distribution characteristics for predicting compressive strength using four machine learning techniques across six data processing techniques. In the raw data analysis of the training dataset, the ANN model demonstrates the least negative minimum error (-10.12 MPa) compared to KNN (-18.94 MPa) and SVR (-22.11 MPa), indicating its superior predictive capability in the positive range. The maximum errors for ANN (13.03 MPa) and KNN (11.42 MPa) suggest that these models occasionally predict values higher than actual results, while SVR and XGB maintain lower maximum errors, indicating more consistent predictions. The skewness values show that ANN and XGB are more symmetrically distributed, whereas KNN and SVR exhibit left skewness, highlighting potential biases in their predictions.

When examining the testing dataset, KNN's maximum error (30.21 MPa) is notably higher than the other models, indicating its tendency to overestimate compressive strength in specific scenarios. The skewness for KNN (1.18) and XGB (1.34) indicates a positive skew, suggesting these models tend to predict higher values more frequently. This trend raises concerns about the reliability of KNN, particularly when assessing its performance across various data distributions. In the context of max-min data processing, the training dataset reveals that SVR has a notably low minimum error (-25.51 MPa), indicating challenges with under-predictions. The skewness for SVR (-2.45) reinforces concerns about its performance in specific data ranges. In the testing dataset, KNN maintains a maximum error of 30.21 MPa, while ANN displays balanced performance with a maximum error of 12.84 MPa, indicating robustness across different distributions.

With Z-score data processing, skewness values suggest that ANN and KNN exhibit relatively balanced distributions in the training dataset. However, KNN shows substantial left skewness (-2.70) in the testing dataset, indicating increased prediction errors in lower ranges. The robust data processing technique highlights a significant negative skewness for ANN (-3.19) in the training dataset, implying frequent under-predictions. In contrast, KNN's nearly symmetrical distribution in the testing dataset suggests improved reliability. Log scale data processing results indicate that KNN again presents a high negative skewness (-4.83), suggesting it may be less reliable in predicting compressive strength. The maximum error for KNN (11.64 MPa) is significantly reduced compared to raw data processing, indicating that data transformation techniques can enhance model performance.

Finally, the sigmoid data processing technique reveals that ANN maintains a balanced performance, while KNN's high negative skewness (-1.22) in the testing dataset underscores its challenges with prediction accuracy. Overall, the performance of machine learning techniques in predicting compressive strength varies significantly based on the applied data processing technique.

Fig. 15 presents two sets of curves for the training and testing datasets, illustrating the performance of various data processing techniques across different machine learning algorithms in terms of prediction accuracy. Analyzing the AUC (Area Under the Curve) values in predicting compressive strength using diverse data processing techniques and machine learning algorithms reveals nuanced insights into their performance dynamics, and is summarized in Table 6. Among the data processing methods employed for training datasets, max-min normalization stands out, particularly in enhancing the performance of the KNN algorithm. With an impressive training AUC of 0.987, this underscores the technique's efficacy in scaling features to a bounded range, thereby allowing KNN to capture the dataset's intricacies more effectively. Similarly, the ANN benefits from robust scaling, achieving a training AUC of 0.960, indicating its capability to model complex relationships while maintaining stability against outliers. Conversely, z-score normalization presents a mixed-performance landscape. Although it improves the training AUC for several algorithms, including SVR, at 0.950, its effectiveness diminishes during testing, particularly for KNN, which drops to 0.916. This discrepancy highlights potential overfitting issues, where the model performs well on training data but struggles to generalize to unseen data. The log transformation also proves beneficial, especially for ANN, increasing its testing AUC to 0.930, which suggests its role in addressing skewed data distributions and enhancing model robustness.

When examining the testing AUC values, it becomes evident that while the performance across models decreases, the relative rankings remain consistent. The testing AUC for ANN and XGB is noteworthy, with values of 0.947 and 0.938, respectively, indicating that these models, when paired with effective preprocessing techniques, can maintain strong predictive capabilities. Overall, the results highlight the crucial relationship between data preprocessing methods and machine learning algorithms, demonstrating that tailored preprocessing can substantially enhance model performance. As such, this study advocates for a meticulous approach in selecting data processing techniques that align with the specific characteristics of the chosen machine learning models, ultimately aiming to enhance predictive accuracy in regression tasks.

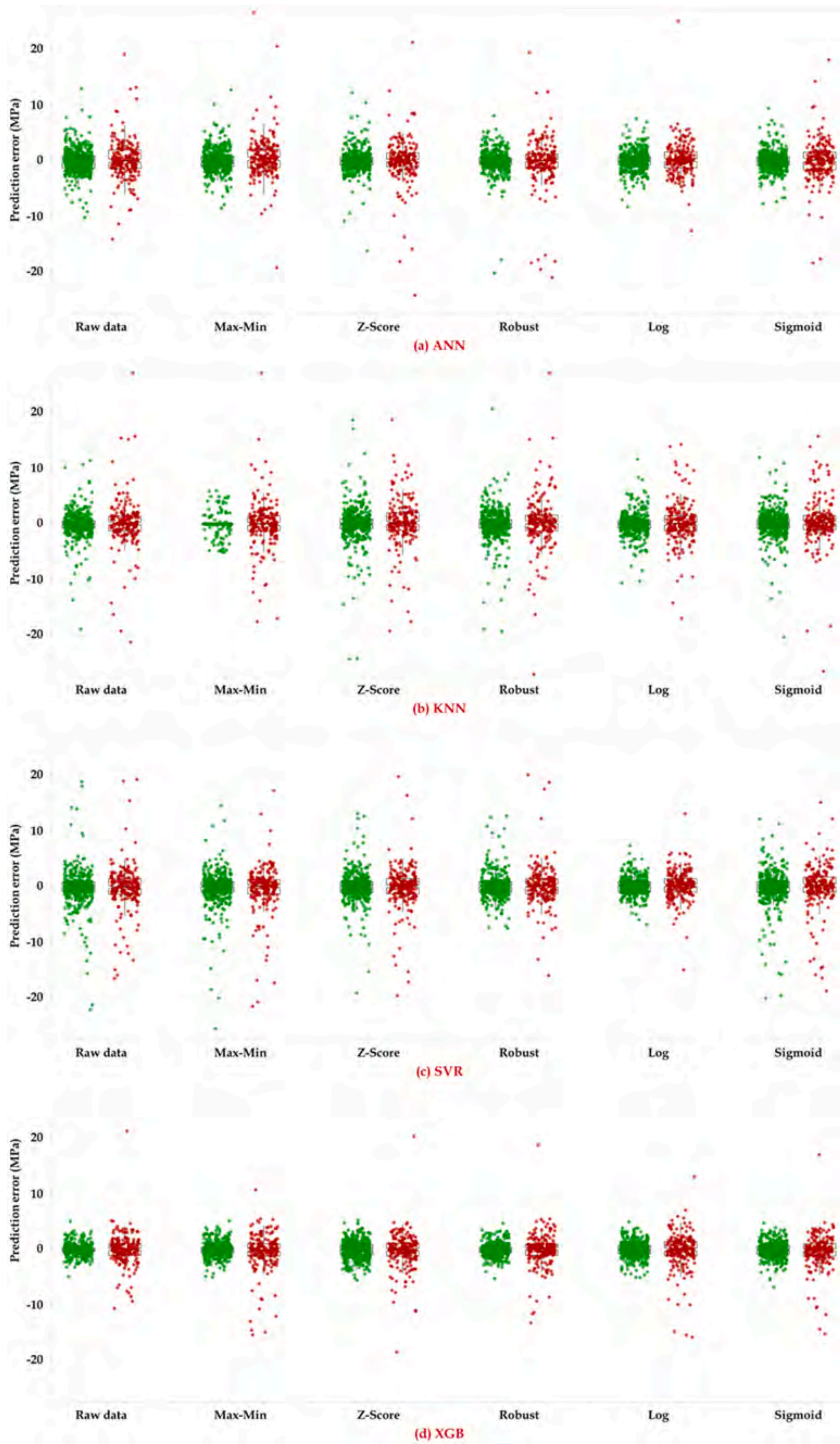

Fig. 14. Prediction error distribution for different machine learning models with various data processing techniques (color code: Green: training dataset and Red: testing dataset).

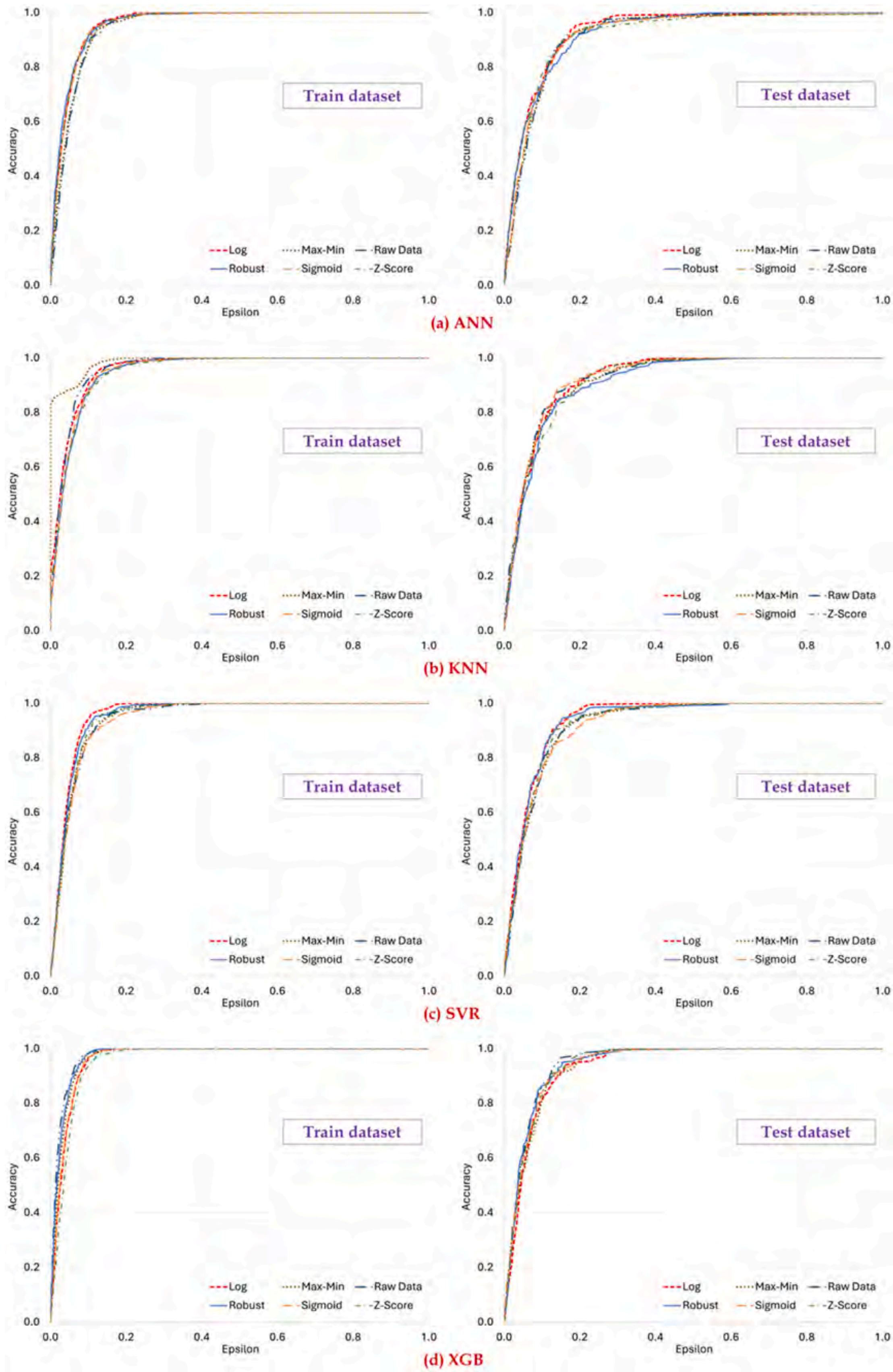

Fig. 15. Regression error characteristic (REC) curves for the prediction of compressive strength using different machine learning models with various data processing techniques.

Table 6
AUC (Area Under the Curve) values in predicting compressive strength using diverse data processing techniques and machine learning algorithms.

	Training				Testing			
	ANN	KNN	SVR	XGB	ANN	KNN	SVR	XGB
Raw data	0.948	0.961	0.946	0.978	0.920	0.923	0.924	0.947
Max-min	0.950	0.987	0.947	0.971	0.919	0.920	0.925	0.938
Z-Score	0.958	0.950	0.950	0.958	0.920	0.916	0.928	0.942
Robust	0.960	0.951	0.954	0.974	0.922	0.911	0.935	0.944
Log	0.959	0.960	0.959	0.965	0.930	0.923	0.939	0.936
Sigmoid	0.957	0.953	0.943	0.966	0.917	0.926	0.924	0.942

3.5. Sensitivity analysis

The SHAP (SHapley Additive exPlanations) analysis provides a robust framework for understanding how various factors influence the compressive strength of concrete. By assigning SHAP values to each parameter and summary plot in Figs. 16 and 17, it can discern their contributions and the underlying trends in their effects on strength.

Al_2O_3 content stands out with the highest positive SHAP value, indicating that increases in Al_2O_3 concentration led to significant improvements in compressive strength. The trend suggests that as Al_2O_3 levels rise, the concrete's ability to form stable bonds is enhanced, likely due to its role in the pozzolanic reactions that occur during hydration. This finding aligns with existing literature that underscores the importance of alumina in promoting the formation of calcium-aluminate hydrates, which are crucial for strength development. The curing period exhibits a positive correlation with compressive strength, with higher values corresponding to increased strength. The trend suggests that extending the curing time enables more complete hydration of cement particles, resulting in a denser microstructure. Conversely, insufficient curing can result in negative SHAP values, highlighting the detrimental effects of premature drying, which can inhibit the hydration process. This highlights the crucial role of curing practices in concrete technology, where optimal curing can substantially improve mechanical properties.

The water content exhibits a complex trend, with higher water levels yielding positive SHAP values, while lower levels result in negative values. The positive association is attributed to the essential role of water in facilitating the hydration process, which is critical for strength development. However, excessive water can dilute the cement paste, lowering density and reducing strength. Thus, achieving an optimal water-to-cement ratio is paramount. This dual trend highlights the need for precise control over water content during mixing to achieve maximum compressive strength. MgO content has a positive SHAP value, suggesting moderate levels can enhance compressive strength. The trend suggests that MgO can contribute to the formation of additional hydration products, thereby improving the concrete microstructure. However, the potential for unsoundness with excessive MgO is critical; thus, its incorporation must be carefully managed to prevent adverse outcomes.

The size of aggregates also exhibits a favorable trend, with larger aggregates typically resulting in increased compressive strength. Larger aggregates tend to occupy more volume and reduce the amount of voids in the concrete mix, improving load-bearing capacity. However, if aggregate sizes are substantial, they may negatively impact the workability and bonding of the concrete mix, leading to a decline in strength. This highlights the importance of selecting an appropriate aggregate size that strikes a balance between strength and workability. Aggregate content exhibits the least influence among the parameters assessed, indicating that while aggregates are essential, their contribution to strength is overshadowed by other factors. The trend suggests that optimizing the blend of aggregates with other components is crucial for achieving desired strength characteristics.

Both equivalent Na_2O and Fe_2O_3 exhibit nuanced trends. Higher levels of Eq. Na_2O can improve workability but may negatively influence long-term strength if overused, as indicated by the slight negative SHAP values at extreme concentrations. Fe_2O_3, on the other hand, generally

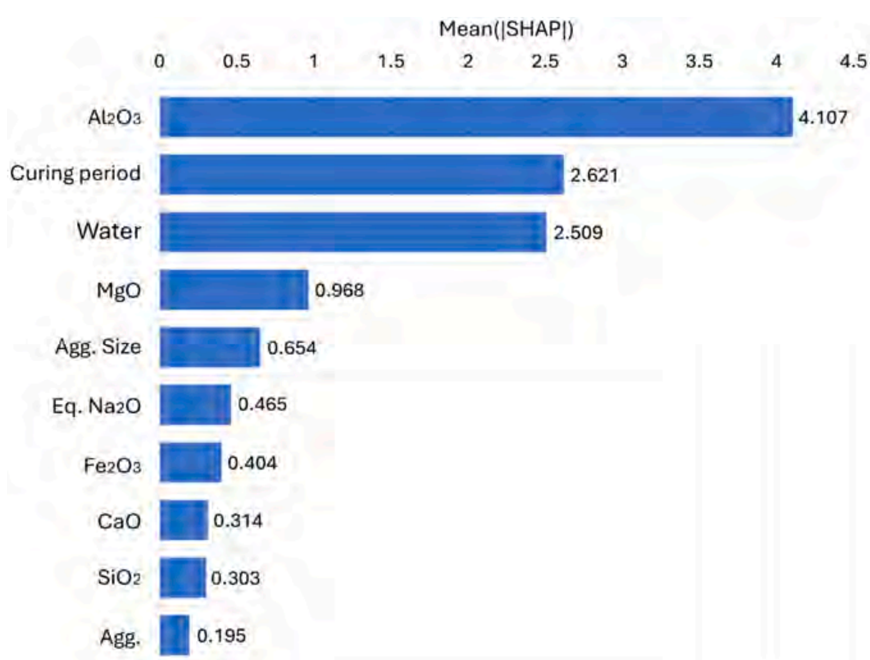

Fig. 16. Mean SHAP plot for output of XGB model with raw data.

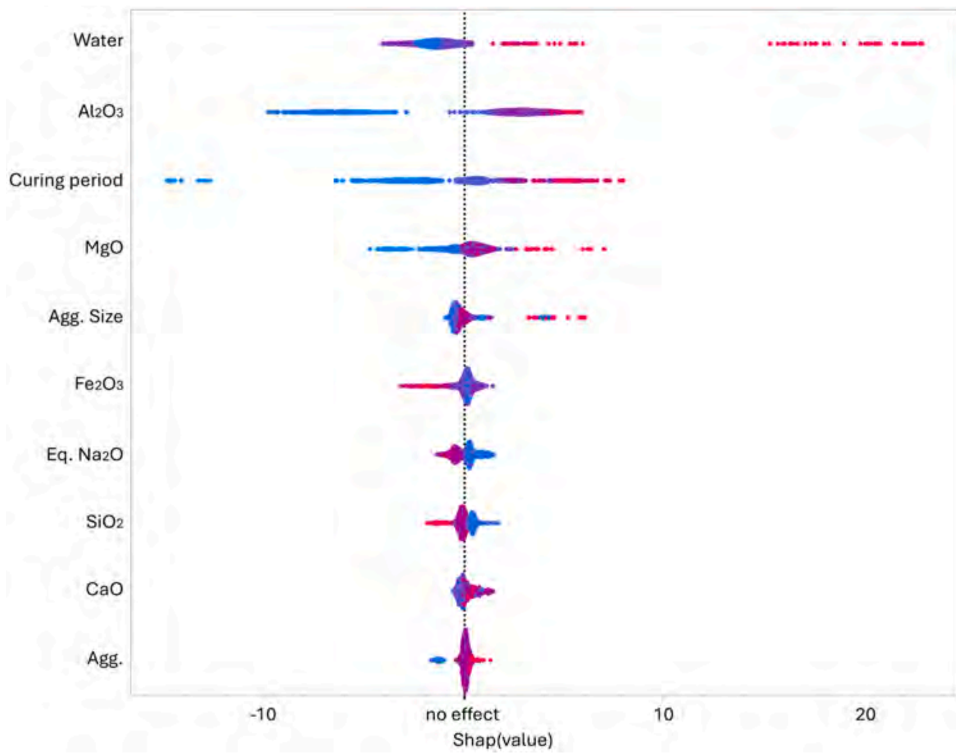

Fig. 17. Summary SHAP plot for the output of the XGB model with raw data.

contributes negatively to compressive strength at higher levels, likely due to the formation of weaker phases within the concrete matrix. This trend highlights the importance of careful formulation to mitigate potential adverse effects. SiO_2 and CaO have complex relationships with compressive strength. The SHAP analysis suggests that while SiO_2 contributes positively at lower concentrations, its impact diminishes as levels rise, potentially due to reduced reactivity in hydration. Conversely, CaO shows a consistent positive trend, reinforcing its role as a key component in enhancing the cementitious properties of the mix.

The SHAP analysis reveals intricate trends in how various factors interact to influence the compressive strength of concrete. Key parameters, such as Al2O3, curing period, and water content, are pivotal in driving strength performance. Additionally, MgO, aggregate size, and the chemical composition of aggregates also play significant roles. Understanding these trends enables the formulation of more adequate concrete mixes, ultimately enhancing structural performance.

3.6. Feature importance

The feature importance study for predicting compressive strength reveals critical insights into the impact of various parameters on model performance, as shown in Table 7. Given that the XGB model accurately predicted compressive strength using the raw dataset, this combination was utilized for future feature importance analysis. Initially, when all ten parameters are included, the model demonstrates an impressive training R^2 value of 0.99, indicating a near-perfect fit to the training data. The testing R^2, while slightly lower at 0.92, still suggests robust predictive capability. Additionally, the low RMSE of 1.04 MPa and MAE of 0.67 MPa for the training set, as well as similar metrics for testing, highlight the model's accuracy.

As parameters are removed one by one, significant changes in performance metrics are observed. For instance, removing CaO leads to a noticeable decline in training and testing R^2 values, dropping to 0.97 and 0.89, respectively. This change is accompanied by increases in RMSE and MAE, indicating that CaO plays a vital role in the model's predictive accuracy. Similarly, removing SiO_2 results in a drop in R^2 and an increase in error metrics, further underscoring its importance in the predictive framework. The parameter Al_2O_3 exhibits a slightly less severe impact than CaO and SiO_2, but its removal still results in a decline in performance, indicating that it contributes meaningfully to the model. Conversely, the removal of Fe_2O_3 results in a minimal reduction in training R^2 but a more significant decrease in testing R^2, indicating its

Table 7
Performance indicators for the XGB model with one input variable left out.

Removed parameter	Train R^2	RMSE (MPa)	MAE (MPa)	SI	a20	Test R^2	RMSE (MPa)	MAE (MPa)	SI	a20	Rank
None	0.99	1.04	0.67	0.06	0.98	0.92	2.85	1.74	0.17	0.86	-
CaO	0.97	1.68	1.25	0.11	0.92	0.89	3.30	2.05	0.20	0.83	3
SiO_2	0.98	1.37	1.01	0.09	0.95	0.90	3.16	2.06	0.19	0.80	4
Al_2O_3	0.98	1.33	0.97	0.08	0.96	0.91	3.07	1.99	0.19	0.80	5
Fe_2O_3	0.99	1.13	0.77	0.07	0.98	0.91	3.11	1.77	0.19	0.86	8
MgO	0.98	1.25	0.89	0.08	0.96	0.91	3.02	1.79	0.18	0.85	9
Eq.NaO	0.97	1.51	1.07	0.09	0.94	0.91	3.10	1.87	0.19	0.83	6
Agg.	0.99	1.02	0.64	0.06	0.98	0.91	3.09	1.86	0.19	0.83	7
Water	0.95	2.14	1.45	0.13	0.91	0.85	3.92	2.53	0.24	0.76	2
Agg.Size	0.98	1.28	0.88	0.08	0.97	0.92	2.83	1.77	0.17	0.86	10
Curing period	0.72	5.01	3.10	0.31	0.64	0.55	6.80	4.42	0.41	0.56	1

role in improving the model's generalization to unseen data. Parameters such as MgO and Eq.NaO also exhibits performance metric declines upon removal, indicating its moderate importance in the model. The aggregates parameter (Agg.) shows a slight decrease in performance, highlighting its relevance, albeit not as critical as the previously mentioned parameters.

The water content is particularly significant, as its removal substantially increases both RMSE and MAE, indicating its essential role in accurate predictions. In contrast, the Aggregate Size parameter appears to have the least influence on the model's predictive capability, suggesting it may be less critical than others. Lastly, the curing period has the most pronounced effect on the model's performance; its removal leads to drastic declines in all metrics, solidifying its status as the most crucial parameter for predicting compressive strength. In summary, the analysis highlights the importance of feature selection in predictive modeling. It highlights that while several parameters contribute to the accuracy of compressive strength predictions, specific key parameters such as CaO, SiO_2, and curing period are particularly vital.

3.7. Comparison with previous studies

This study employed machine learning models to predict the compressive strength of pervious concrete incorporating supplementary cementitious materials (SCMs), such as fly ash. Among these models, XGB demonstrated the best performance, with a testing R^2 of 0.92 and RMSE of 2.85 MPa. These results align with the findings from previous studies but demonstrate further advancements, particularly in the inclusion of a broader range of input variables, including the chemical compositions of SCMs, which are crucial in determining the strength development of concrete.

Several prior studies have applied machine learning models to predict compressive strength in concrete, including pervious concrete, using similar methodologies. Wadhawan et al. [17] conducted a comparative study using ANN, Decision Tree (DT), SVR, and XGB for fly ash-based concrete. They found that XGB outperformed other models, achieving an R^2 of 0.95 and an RMSE of 4.33 MPa. These findings are comparable to those of the current study, where XGB also demonstrated high performance, albeit with a lower RMSE value, indicating improved prediction accuracy in our model. In addition to these standard parameters, the current study uniquely integrates chemical compositions such as the SiO_2/CaO ratio, which was not fully explored in this work. Jef et al. [19] also employed ANN and SVR models for predicting compressive strength, with a primary focus on fly ash content and the water-to-binder ratio. Their results showed that SVR performed better than ANN. In contrast, the current study goes further by incorporating a wider set of chemical variables, such as the SiO_2/CaO ratio, alongside traditional mix design parameters, offering a more nuanced and robust prediction of compressive strength. Jaf et al. [18] investigated the compressive strength of self-compacting concrete using ANN and SVR, focusing on parameters like fly ash content, water-to-binder ratio, and aggregate size. Although the study showed promising results with ANN and SVR, it did not incorporate as wide a range of chemical composition variables as would be desirable. The current study builds upon this by incorporating the chemical compositions of CaO, SiO_2, and other oxides into the mix, resulting in a more comprehensive and effective model.

Sathiparan [13], similar to the current study, examined the prediction of compressive strength in pervious concrete using ANN, KNN, SVR, and XGB. Their results showed that XGB and ANN were the best performers, which is consistent with the findings of the present study. However, the present research introduces a key innovation by integrating feature importance analysis and sensitivity analysis using SHAP values, allowing for deeper insights into which specific features, such as fly ash content and curing conditions, most influence the predictions. Sathiparan et al. [8] focused on the impact of fly ash and utilized several machine learning models, including ANN, XGB, and Random Forest, for predicting the compressive strength of pervious concrete. They found XGB to be the most effective, achieving an R^2 of 0.98 and an RMSE of 1.62 MPa. Similarly, their study on silica fume-blended pervious concrete confirmed the superiority of XGB, achieving an R^2 of 0.99 and a significantly lower RMSE of 0.28 MPa.

While previous studies have made important contributions to predicting compressive strength in pervious concrete using machine learning, the present study advances the field by incorporating a broader range of variables. The inclusion of chemical composition data, particularly the SiO_2/CaO ratio, and the use of advanced data preprocessing techniques, such as log transformation and Max-Min scaling, enhance the accuracy and applicability of the model. These advancements not only improve the model's predictive performance but also provide a deeper understanding of how SCMs influence the compressive strength of pervious concrete. The findings of this study contribute to the optimization of mix designs and the promotion of more sustainable construction practices.

4. Conclusions

This study demonstrates an approach for predicting the compressive strength of pervious concrete by combining machine learning techniques and chemical composition analysis. The novelty of this study lies in the integration of chemical composition analysis with advanced machine learning techniques to predict the compressive strength of pervious concrete. By considering a broader range of SCMs and incorporating diverse data processing methods, this approach offers a more robust and adaptable solution compared to traditional empirical models. The findings make a significant contribution to the field of sustainable construction and the optimization of concrete mix designs. The key findings from this research are as follows:

- Machine Learning Integration: Advanced machine learning models, specifically Extreme Gradient Boosting (XGB), outperformed traditional methods, achieving high predictive accuracy with a testing R^2 of 0.92 and an RMSE of 2.85 MPa.
- Data Processing Effectiveness: Various data processing techniques, such as log transformation and robust scaling, have been shown to significantly improve model performance, particularly for complex models like ANN and KNN. The log transformation, in particular, stabilized variance and reduced prediction error across models.
- Chemical Composition Impact: The chemical components, particularly calcium oxide (CaO) and silicon dioxide (SiO_2), as well as the curing period, played a significant role in predicting compressive strength, highlighting the importance of chemical composition in mix design optimization.
- Practical Application: The study underscores the importance of incorporating comprehensive chemical composition data into predictive models, offering valuable insights for optimizing pervious concrete formulations. This work contributes to the development of more sustainable construction practices by enhancing the accuracy of compressive strength predictions, which is critical for designing resilient urban infrastructure.

Limitations of the study

This study presents valuable insights into the predictive capabilities of various machine learning techniques for estimating the compressive strength of pervious concrete. However, several limitations should be acknowledged that may affect the interpretation and generalizability of the findings:

- Dataset constraints: The dataset used in this study was limited to specific studies that reported on pervious concrete with supplementary cementitious materials (SCMs). While it includes a variety of compositions, it may not capture the full spectrum of materials and conditions encountered in real-world applications. This could limit the applicability of the models to broader contexts.

- Model overfitting: Certain models, particularly artificial neural Networks (ANNs), exhibited signs of overfitting, as evidenced by significant discrepancies between training and testing performance. This raises concerns about the robustness of the predictions, particularly in practical applications where unseen data may differ from the training set.
- Input variable selection: The selection of input variables was based on the available literature, which may overlook other potentially influential factors affecting compressive strength. For instance, additional parameters such as environmental conditions during curing or the presence of specific admixtures could be critical but were not included in the analysis.
- Normalization techniques: The effectiveness of various normalization techniques was assessed, but the choice of method can significantly impact model performance. While this study examined several approaches, the optimal method may vary depending on the specific characteristics of the dataset, and further exploration of normalization effects could be beneficial.
- Generalizability of results: The findings from this study may not be universally applicable to all types of pervious concrete or other concrete mixes. Variations in local materials, environmental conditions, and construction practices may influence the compressive strength outcomes differently from those observed in the study.
- Lack of experimental validation: While the study relies on predictive modelling, it lacks direct experimental validation of the machine learning predictions. Future studies should aim to correlate model outcomes with laboratory tests to ensure that predictions are statistically significant and practically relevant.
- Complex interactions: The study may not fully account for the complex interactions between different input variables. Advanced techniques, such as feature interaction modelling, could provide deeper insights into how combinations of factors influence compressive strength.
- Computational resources: The computational intensity required for training some machine learning models, particularly ensemble methods like XGB, may pose a barrier to practical implementation in resource-limited settings. This raises questions about the feasibility of deploying such models in routine practice.

While this study provides important contributions to understanding the predictive power of machine learning in assessing compressive strength, acknowledging these limitations is crucial for guiding future research and practical applications in sustainable construction practices.

Future research directions

This study has laid the groundwork for understanding the predictive capabilities of machine learning techniques in estimating the compressive strength of pervious concrete. However, several key areas warrant further investigation to enhance the robustness and applicability of these findings:

- Expanded dataset development: Future research should focus on developing a more comprehensive dataset that includes a broader variety of pervious concrete compositions and environmental conditions. Incorporating data from different geographical regions and construction practices can improve the generalizability of predictive models.
- Exploring additional machine learning techniques: Investigating a broader range of machine learning algorithms, including deep learning models and hybrid approaches, could provide valuable insights into their effectiveness compared to traditional methods.
- Integration of experimental validation: Conducting experimental studies to validate the predictions made by machine learning models is essential. Future research should aim to correlate model outcomes with laboratory tests, providing empirical evidence of model reliability and enhancing confidence in practical applications.
- Feature importance and interaction analysis: Further studies should investigate the importance of additional input variables and their interactions. Advanced techniques such as Shapley values or permutation feature importance can help identify critical factors influencing compressive strength beyond those considered in this study.
- Long-term performance predictions: Research could explore the long-term behaviour of pervious concrete, assessing how factors such as ageing, environmental exposure, and mechanical loading influence compressive strength over time. This would provide valuable insights into the durability and sustainability of pervious concrete in real-world applications.
- Economic and environmental impact assessments: Research should also focus on conducting economic analyses of implementing machine learning models in construction practices, evaluating cost-effectiveness, and potential savings in material usage. Additionally, assessing the environmental impacts of optimized pervious concrete mixtures could further support sustainable construction practices.
- User-friendly tools and software development: Developing accessible software tools that integrate these predictive models into design practices can facilitate their adoption by engineers and practitioners. Future research should aim to create user-friendly interfaces that allow for easy input of variables and provide immediate predictions of compressive strength.

By addressing these areas, future research can significantly enhance the understanding and application of machine learning techniques in predicting the compressive strength of pervious concrete, ultimately contributing to more sustainable and resilient construction practices.

Author contribution

Study conception and design: N.S and D.N.S; Data acquisition: N.S; analysis and interpretation of results: N.S and P.J; draft manuscript preparation: N.S and P.J; manuscript review & editing: D.N.S.

Code availability

Not applicable.

Consent for publication

Not applicable.

Consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Ethics approval

Not applicable.

Funding

This work was supported by the National Research Council of Sri Lanka: (Grant Number 24–031)

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.nxmate.2025.100947.

Data availability

Data can be made available on request by interested parties.

References

[1] T.S. Tie, K.H. Mo, A. Putra, S.C. Loo, U.J. Alengaram, T.-C. Ling, Sound absorption performance of modified concrete: A review, J. Build. Eng. 30 (2020) 101219, https://doi.org/10.1016/j.jobe.2020.101219.

[2] O. Addo-Bankas, T. Wei, Y. Zhao, X. Bai, A.E. Núñez, A. Stefanakis, Revisiting the concept, urban practices, current advances, and future prospects of green infrastructure, Sci. Total Environ. 954 (2024) 176473, https://doi.org/10.1016/j.scitotenv.2024.176473.

[3] N. Sathiparan, D.H.H.P. Dassanayake, D.N. Subramaniam, Utilization of supplementary cementitious materials in pervious concrete: a review, Int. J. Environ. Sci. Technol. 21 (6) (2024) 5883–5918, https://doi.org/10.1007/s13762-023-05440-4.

[4] S.R. Islam, S.N. Majumder, R. Mutsuddy, Life cycle assessment and mechanical strength of cement composites with conventional, and recycled fine aggregate, Sustain. Struct. 4 (2) (2024) 000052, https://doi.org/10.54113/j.sust.2024.000052.

[5] F. Althoey, W.S. Ansari, M. Sufian, A.F. Deifalla, Advancements in low-carbon concrete as a construction material for the sustainable built environment, Dev. Built Environ. 16 (2023) 100284, https://doi.org/10.1016/j.dibe.2023.100284.

[6] G.O. Claudino, G.G.O. Rodrigues, A.B. Rohden, E.F.T. Mesquita, M.R. Garcez, Mix design for pervious concrete based on the optimization of cement paste and granular skeleton to balance mechanical strength and permeability, Constr. Build. Mater. 347 (2022) 128620, https://doi.org/10.1016/j.conbuildmat.2022.128620.

[7] K. Ćosić, L. Korat, V. Ducman, I. Netinger, Influence of aggregate type and size on properties of pervious concrete, Constr. Build. Mater. 78 (2015) 69–76, https://doi.org/10.1016/j.conbuildmat.2014.12.073.

[8] N. Sathiparan, P. Jeyananthan, D.N. Subramaniam, Quantifying the impact of chemical composition on pervious concrete strength: a comparative analysis using full quadratic model and artificial neural network, Int. J. Pavement Eng. 25 (1) (2024) 2424381, https://doi.org/10.1080/10298436.2024.2424381.

[9] S.H.B. Wijekoon, A. Janarth, J. Dharmar, P. Vinojan, N. Sathiparan, D. N. Subramaniam, Optimizing compressive strength prediction of pervious concrete using artificial neural network, Eng. Res. Express 7 (1) (2025) 015418, https://doi.org/10.1088/2631-8695/adb129.

[10] M.A. DeRousseau, E. Laftchiev, J.R. Kasprzyk, B. Rajagopalan, W.V. Srubar, A comparison of machine learning methods for predicting the compressive strength of field-placed concrete, Constr. Build. Mater. 228 (2019) 116661, https://doi.org/10.1016/j.conbuildmat.2019.08.042.

[11] N. Sathiparan, Prediction model for compressive strength of rice husk ash blended sandcrete blocks using a machine learning models, Asian J. Civ. Eng. 25 (6) (2024) 4745–4758, https://doi.org/10.1007/s42107-024-01077-x.

[12] A. Kordbagheri, M. Kordbagheri, N. Tayim, A. Fakhrou, M. Davoudi, Using advanced machine learning algorithms to predict academic major completion: a cross-sectional study, Comput. Biol. Med. 184 (2025) 109372, https://doi.org/10.1016/j.compbiomed.2024.109372.

[13] N. Sathiparan, Predicting compressive strength in cement mortar: the impact of fly ash composition through machine learning, Sustain. Chem. Pharm. 43 (2025) 101915, https://doi.org/10.1016/j.scp.2025.101915.

[14] M. Basil Ishaq, A. Salih Mohammed, Optimizing compressive strength prediction in glass powder-modified concrete: a comprehensive study on silicon dioxide and calcium oxide influence across varied sample dimensions and strength ranges 64 (1) (2025), https://doi.org/10.1515/rams-2025-0095.

[15] M.B. Ishaq, A.S. Mohammed, A.A. Mohammed, Novel exploration of chemical compounds in mono-variate, Bi-variate, and multi-variate scenarios of waste glass powder as a cement replacement in concrete, Sustain. Chem. Pharm. 43 (2025) 101894, https://doi.org/10.1016/j.scp.2024.101894.

[16] M.B. Ishaq, A.S. Mohammed, A.A. Mohammed, Chemical characterization and performance optimization of waste glass-modified concrete for sustainable construction, Sustain. Chem. Pharm. 44 (2025) 101968, https://doi.org/10.1016/j.scp.2025.101968.

[17] S. Wadhawan, A. Bassi, R. Singh, M. Patel, Prediction of compressive strength for fly ash-based concrete: critical comparison of machine learning algorithms, J. Soft Comput. Civ. Eng. 7 (3) (2023) 68–110, https://doi.org/10.22115/scce.2023.353183.1493.

[18] D. Kakasor Ismael Jaf, P. Ismael Abdulrahman, A. Salih Mohammed, R. Kurda, S.M. A. Qaidi, P.G. Asteris, Machine learning techniques and multi-scale models to evaluate the impact of silicon dioxide (SiO2) and calcium oxide (CaO) in fly ash on the compressive strength of green concrete, Constr. Build. Mater. 400 (2023) 132604, https://doi.org/10.1016/j.conbuildmat.2023.132604.

[19] D. Kakasor Ismael Jaf, A.S. Abdulrahman, P.I. Abdulrahman, A. Salih Mohammed, R. Kurda, H.U. Ahmed, R.H. Faraj, Effitioned soft computing models to evaluate the impact of silicon dioxide (SiO2) to calcium oxide (CaO) ratio in fly ash on the compressive strength of concrete, J. Build. Eng. 74 (2023) 106820, https://doi.org/10.1016/j.jobe.2023.106820.

[20] N. Sathiparan, P. Jeyananthan, D.N. Subramaniam, Predicting compressive strength of pervious concrete with fly ash: a machine learning approach and analysis of fly ash compositional influence, Multiscale Multidiscip. Model. Exp. Des. 7 (6) (2024) 5651–5671, https://doi.org/10.1007/s41939-024-00551-y.

[21] N. Sathiparan, P. Jeyananthan, D.N. Subramaniam, Machine learning techniques to evaluate the impact of calcium oxide (CaO) and silicon dioxide (SiO2) in supplementary cement materials on the compressive strength of sustainable pervious concrete, Sustain. Chem. Pharm. 42 (2024) 101818, https://doi.org/10.1016/j.scp.2024.101818.

[22] K.H. Tan, H. Du, Sandless concrete with fly ash as supplementary cementing material, J. Sustain. Cem. Based Mater. 2 (3-4) (2013) 238–249, https://doi.org/10.1080/21650373.2013.827994.

[23] J.R. Prajapati, J.R. Pitroda, AD Raval, Experimental study on pervious concrete with fly ash and metakaolin, J. Emerg. Technol. Innov. Res. 6 (5) (2019) 294–304.

[24] T. Singh, R. Siddique, S. Sharma, Effectiveness of using Metakaolin and fly ash as supplementary cementitious materials in pervious concrete, Eur. J. Environ. Civ. Eng. (2021) 1–24, https://doi.org/10.1080/19648189.2021.1988715.

[25] S. Hwang, J.H. Yeon, Fly ash-added, seawater-mixed pervious concrete: compressive strength, permeability, and phosphorus removal, Materials (2022).

[26] S.K. Boddu, Investigating the Effect of Aggregate Size and Binder Material Proportion on Strength and Permeability of Pervious Concrete by Statistical Modeling, in: M.V.L.R. Anjaneyulu, M. Harikrishna, S.S. Arkatkar, A. Veeraragavan (Eds.), Recent Advances in Transportation Systems Engineering and Management, Springer Nature Singapore, Singapore, 2023, pp. 235–251.

[27] B.S. Kumar, K. Srikanth, A study on properties of pervious concrete with high-volume usage of supplementary cementitious materials as substitutes for cement, Asian J. Civ. Eng. 24 (7) (2023) 1997–2009, https://doi.org/10.1007/s42107-023-00619-z.

[28] M. Nazeer, K. Kapoor, S.P. Singh, Strength, durability and microstructural investigations on pervious concrete made with fly ash and silica fume as supplementary cementitious materials, J. Build. Eng. 69 (2023) 106275, https://doi.org/10.1016/j.jobe.2023.106275.

[29] N. DN Subramaniam, Sathiparan, comparative study of fly ash and rice husk ash as cement replacement in pervious concrete: mechanical characteristics and sustainability analysis, Int. J. Pavement Eng. 24 (2) (2023) 2075867.

[30] M. Rama, V. Shanthi, Study on strength, permeability and microstructure of pervious concrete blended with metakaolin, Jordan J. Od. Civ. Eng. 17 (1) (2023) 10–22.

[31] S. Talsania, J. Pitroda, C. Vyas, Effect of rice husk ash on properties of pervious concrete, Int. J. Adv. Eng. Res. Stud. 4 (2) (2015) 296–299.

[32] H. El-Hassan, P. Kianmehr, Sustainability assessment and physical characterization of pervious concrete pavement made with GGBS, MATEC Web Conf. 120 (2017).

[33] B. Sudhir Kumar, K. Srikanth, T. Eeshwar, Implementation of soft computing techniques in forecasting compressive strength and permeability of pervious concrete blended with ground granulated blast-furnace slag, J. Soft Comput. Civ. Eng. 8 (2) (2024) 19–45, https://doi.org/10.22115/scce.2023.359399.1517.

[34] E. Khankhaje, M. Rafieizonooz, M.R. Salim, R. Khan, J. Mirza, H.C. Siong, Salmiati, Sustainable clean pervious concrete pavement production incorporating palm oil fuel ash as cement replacement, J. Clean. Prod. 172 (2018) 1476–1485, https://doi.org/10.1016/j.jclepro.2017.10.159.

[35] M.A. Tijani, W.O. Ajagbe, A.A. Ganiyu, O.A. Agbede, Sustainable pervious concrete incorporating sorghum husk ash as cement replacement, IOP Conf. Ser. Mater. Sci. Eng. 640 (1) (2019) 012051, https://doi.org/10.1088/1757-899x/640/1/012051.

[36] M.A. Tijani, W.O. Ajagbe, O.A. Agbede, Combined reusing of sorghum husk ash and recycled concrete aggregate for sustainable pervious concrete production, J. Clean. Prod. 343 (2022) 131015, https://doi.org/10.1016/j.jclepro.2022.131015.

[37] A. Li, H. Qiao, Q. Li, T. Hakuzweyezu, B. Chen, Study on the performance of pervious concrete mixed with waste glass powder, Constr. Build. Mater. 300 (2021) 123997, https://doi.org/10.1016/j.conbuildmat.2021.123997.

[38] S. Talsania, J. Pitroda, C. Vyas, A review of pervious concrete by using various Industrial waste materials, J. Int. Acad. Res. Multidiscip. 4 (2) (2015) 296–299.

[39] A. Joshaghani, A.A. Ramezanianpour, O. Ataei, A. Golroo, Optimizing pervious concrete pavement mixture design by using the Taguchi method, Constr. Build. Mater. 101 (2015) 317–325, https://doi.org/10.1016/j.conbuildmat.2015.10.094.

[40] M. Adamu, K.O. Ayeni, S.I. Haruna, Y.E.-H. Ibrahim Mansour, S. Haruna, Durability performance of pervious concrete containing rice husk ash and calcium carbide: A response surface methodology approach, Case Stud. Constr. Mater. 14 (2021) e00547, https://doi.org/10.1016/j.cscm.2021.e00547.

[41] P. Mehrabi, M. Shariati, K. Kabirifar, M. Jarrah, H. Rasekh, N.T. Trung, A. Shariati, S. Jahandari, Effect of pumice powder and nano-clay on the strength and permeability of fiber-reinforced pervious concrete incorporating recycled concrete aggregate, Constr. Build. Mater. 287 (2021) 122652, https://doi.org/10.1016/j.conbuildmat.2021.122652.

[42] H. Benhar, A. Idri, J.L. Fernández-Alemán, Data preprocessing for heart disease classification: a systematic literature review, Comput. Methods Prog. Biomed. 195 (2020) 105635, https://doi.org/10.1016/j.cmpb.2020.105635.

[43] X.H. Cao, I. Stojkovic, Z. Obradovic, A robust data scaling algorithm to improve classification accuracies in biomedical data, BMC Bioinforma. 17 (1) (2016) 359, https://doi.org/10.1186/s12859-016-1236-x.

[44] C.O. Alves, C. Ji, Multiple Normalized Solutions to a Logarithmic Schrödinger Equation via Lusternik–Schnirelmann Category, J. Geom. Anal. 34 (7) (2024) 198, https://doi.org/10.1007/s12220-024-01649-y.

[45] U. Sharma, N. Gupta, M. Verma, Prediction of the compressive strength of Flyash and GGBS incorporated geopolymer concrete using artificial neural network, Asian J. Civ. Eng. 24 (8) (2023) 2837–2850, https://doi.org/10.1007/s42107-023-00678-2.
[46] Y.-S. Kim, M.K. Kim, N. Fu, J. Liu, J. Wang, J. Srebric, Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models, Sustain. Cities Soc. 118 (2025) 105570, https://doi.org/10.1016/j.scs.2024.105570.
[47] J. Wu, Y. Zhao, Machine learning technology in the application of genome analysis: a systematic review, Gene 705 (2019) 149–156, https://doi.org/10.1016/j.gene.2019.04.062.
[48] N. Sathiparan, Predicting compressive strength of grouted masonry using machine learning models with feature importance analysis, Mater. Today Commun. 41 (2024) 110487, https://doi.org/10.1016/j.mtcomm.2024.110487.
[49] N. Sathiparan, P. Jeyananthan, D.N. Subramaniam, Surface response regression and machine learning techniques to predict the characteristics of pervious concrete using non-destructive measurement: ultrasonic pulse velocity and electrical resistivity, Meas. J. Int. Meas. Confed. 225 (2024) 114006.
[50] S.F. Ahmed, M.S.B. Alam, M. Hassan, M.R. Rozbu, T. Ishtiak, N. Rafa, M. Mofijur, A.B.M. Shawkat Ali, A.H. Gandomi, Deep learning modelling techniques: current progress, applications, advantages, and challenges, Artif. Intell. Rev. 56 (11) (2023) 13521–13617, https://doi.org/10.1007/s10462-023-10466-8.
[51] N. Sathiparan, P. Jeyananthan, D.N. Subramaniam, A machine learning approach to predicting pervious concrete properties: a review, Innov. Infrastruct. Solut. 10 (2) (2025) 55, https://doi.org/10.1007/s41062-024-01829-3.
[52] R.K. Halder, M.N. Uddin, M.A. Uddin, S. Aryal, A. Khraisat, Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications, J. Big Data 11 (1) (2024) 113, https://doi.org/10.1186/s40537-024-00973-y.
[53] N. Ukey, G. Zhang, Z. Yang, B. Li, W. Li, W. Zhang, Efficient continuous kNN join over dynamic high-dimensional data, World Wide Web 26 (6) (2023) 3759–3794, https://doi.org/10.1007/s11280-023-01204-9.
[54] C. Comito, C. Pizzuti, Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: a focused review, Artif. Intell. Med. 128 (2022) 102286, https://doi.org/10.1016/j.artmed.2022.102286.
[55] S. Xu, B. Lu, M. Baldea, T.F. Edgar, M. Nixon, An improved variable selection method for support vector regression in NIR spectral modeling, J. Process Control 67 (2018) 83–93, https://doi.org/10.1016/j.jprocont.2017.06.001.
[56] Y. Khaledian, B.A. Miller, Selecting appropriate machine learning methods for digital soil mapping, Appl. Math. Model. 81 (2020) 401–418, https://doi.org/10.1016/j.apm.2019.12.016.
[57] A. Sheikhmohammadi, P. Khakzad, T. Rasolevandi, H. Azarpira, Leveraging artificial intelligence models (GBR, SVR, and GA) for efficient chromium reduction via UV/trichlorophenol/sulfite reaction, Results Eng. 26 (2025) 104599, https://doi.org/10.1016/j.rineng.2025.104599.
[58] D.P.P. Meddage, I. Fonseka, D. Mohotti, K. Wijesooriya, C.K. Lee, An explainable machine learning approach to predict the compressive strength of graphene oxide-based concrete, Constr. Build. Mater. 449 (2024) 138346, https://doi.org/10.1016/j.conbuildmat.2024.138346.