Resources, Conservation & Recycling 190 (2023) 106870

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

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

Full length article

# Deep learning from physicochemical information of concrete with an artificial language for property prediction and reaction discovery

Soroush Mahjoubi, Rojyar Barhemat, Weina Meng, Yi Bao *

*Department of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA*

ARTICLE INFO

*Keywords:*
Deep learning
Physicochemical information
Concrete properties
Artificial language
Reaction discovery

ABSTRACT

Existing machine learning-based approaches to investigate and design concrete mainly use the mixture design variables to predict concrete properties and do not consider the physicochemical properties of ingredients such as the particle size distribution and chemical composition of various binders and aggregates. This paper presents an approach to discover the intrinsic relationships between the physicochemical properties of the ingredients and mechanical properties of concrete. Specifically, this research creates an artificial language to represent concrete mixtures and the physicochemical information of their ingredients, develops a feature extraction method based on character-level N-grams, and proposes a method to configure deep learning models automatically. The proposed approach has been implemented to predict the compressive strength of complex concrete mixtures, assess the importance of variables, and discover chemical reactions, showing high accuracy and high generalizability. This research advances the capabilities of understanding the underlying reactions for complex concrete mixtures and designing low-carbon cost-effective concrete.

## 1. Introduction

Concrete is one of the most widely used construction materials worldwide, with an annual consumption of 30 billion tons in 2017 (Monteiro et al., 2017). The production of cement was responsible for more than 8% of the total carbon emissions in 2018 (Lehne and Preston, 2018), which drives the development of low-carbon concrete, aiming to mitigate climate change. Currently, the dominant approach to develop low-carbon concrete is to use low-carbon ingredients such as recycled concrete aggregates and/or fines (Long et al., 2022; Wang et al., 2020) as well as green binders such as supplementary cementitious materials (SCMs). Many types of SCMs have been utilized to produce concrete, and the representative examples include fly ash (Aubert et al., 2004) and slag (Shi et al., 2008). SCMs have been utilized to replace cement in producing ultra-high-performance concrete (de Larrard and Sedran, 1994) and strain-hardening cementitious composites (Li, 2003). It was found that appropriate use of SCMs improved the workability (Hunger, 2010), mechanical properties (Borosnyói, 2016), and durability (Elahi et al., 2021) due to their physical and chemical properties such as the small particle sizes and chemical composition. Finely tuning particle size distribution is able to maximize the particle packing density, and tailoring the chemical composition is able to maximize the degree of hydration of cement and refine the microstructures of concrete. Recent research has shown that off-specification fly ash, which was known to degrade concrete properties, can be utilized to produce ultra-high-performance concrete with high mechanical strengths, superior durability, and the self-consolidating property (Du et al., 2022).

Existing design approaches of concrete incorporating SCMs are mainly based on experiments. Usually, a large number of experiments are conducted to investigate the effects of SCMs on the fresh and hardened properties and optimize the mixture proportions. The experiments of concrete are costly and time-consuming and involve additional carbon emissions. To reduce experiments, approaches guided by artificial intelligence (AI) have been proposed based on machine learning. Various machine learning models were developed to correlate concrete mixtures with properties, such as the mechanical properties (Liang et al., 2022; Mahjoubi et al., 2021; S. Mahjoubi et al., 2022; Asteris et al., 2021; Zhang et al., 2022), interfacial properties (S. Mahjoubi et al., 2022), workability (S. Mahjoubi et al., 2022), and porosity (S. Mahjoubi et al., 2022). Those models were trained using experimental data and applied to predict concrete properties by inputting mixture proportions of ingredients, such as the water-to-binder ratio, the sand-to-binder ratio, and the fiber content. Those models were able to consider different types of ingredients such as cement, fly ash, slag, sand, and

* Corresponding author.
*E-mail address:* yi.bao@stevens.edu (Y. Bao).

https://doi.org/10.1016/j.resconrec.2023.106870
Received 24 September 2022; Received in revised form 12 December 2022; Accepted 6 January 2023
Available online 10 January 2023

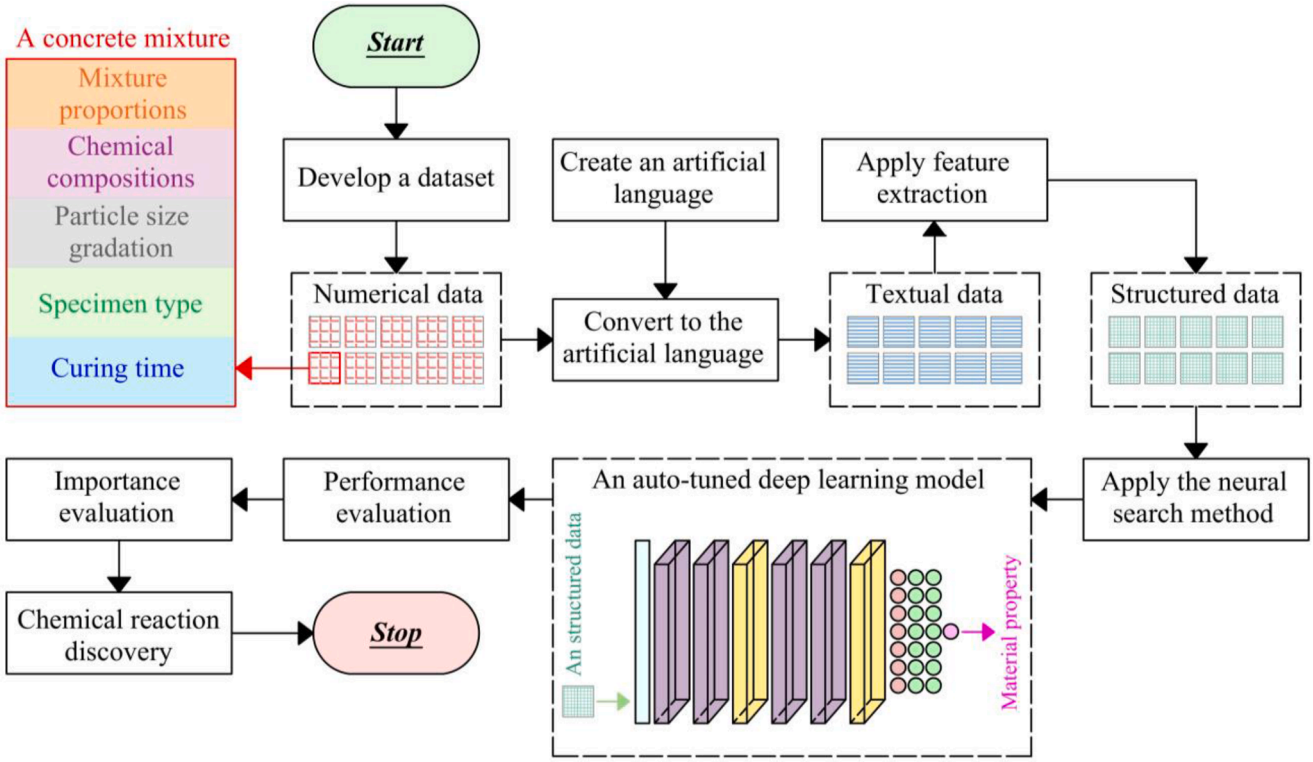


**Fig. 1.** Flowchart of the proposed data-driven approach to predict the mechanical properties and discover chemical reactions of cementitious composites.

fibers. Thermodynamics were coupled with machine learning methods to develop alkali-activated mixtures (Ke and Duan, 2021). A trained model predicted the properties of mixtures as the mixture proportions were changed (S. Mahjoubi et al., 2022). The state-of-the-art models achieved high accuracy in predicting the fresh and hardened concrete properties when particular ingredients were used. Advanced techniques have been employed to improve the datasets (S. Mahjoubi et al., 2022), enrich the datasets by generating artificial data (Guo et al., 2021), and optimize the architectures of machine learning models as well as hyperparameters (Mahjoubi et al., 2021; S. Mahjoubi et al., 2022). The predictive models have been integrated with multi-objective optimization techniques for efficient design (Mahjoubi et al., 2021).

Despite the exciting advances, it remains challenging to predict concrete properties when the ingredients and experimental conditions are changed. For example, a predictive model trained using a dataset with binder systems composed of cement and fly ash is invalid to predict the concrete properties when the fly ash is replaced by slag or other SCMs. In real practices, the problem is exacerbated by the large variations in the physicochemical properties of various SCMs. Two batches of fly ash produced from the same power plant often have different particle sizes and chemical compositions. Even for the same category of fly ash such as the Class F fly ash, the chemical composition involves large variations, thus disabling the use of the trained models (del Visoet al., 2008). The above problems originated from the neglection of the physicochemical natures of concrete ingredients, such as the change of the particle sizes and chemical composition of fly ash. The applicability of the existing machine learning models is limited to specific concrete ingredients with specific physicochemical properties. Previous research showed that the shape and size of specimens largely influence the test results of concrete properties (del Visoet al., 2008; Zhang et al., 2021). Different standards have been issued and utilized to regulate the shape and size of specimens (ACI Committee 318 2022; China Academy of Building Research 2010). Therefore, the data collected from various publications involve different shapes and sizes of specimens. It is essential to consider the shape and size of specimens as variables in predicting concrete properties. Approaches that are robust to the changes of concrete ingredients, physicochemical information, and specimen geometry are highly desired.

This study intends to address these limitations by establishing a new paradigm for AI-guided design of concrete. The idea is that the key physicochemical information of concrete ingredients as well as curing time and specimen type are considered in the representation of concrete mixtures. Based on this idea, this paper creates an artificial language to represent the particle size distribution and chemical composition of concrete ingredients as well as concrete mixtures and experimental conditions. With the novel representation of concrete, textual data is converted into structured data through feature extraction (Shannon, 1948). Although this study presents feature extraction for textural data of concrete mixtures to predict concrete properties and chemical reaction discovery for the first time, feature extraction from biochemical textual data has been proposed earlier for drug discovery (Öztürk et al., 2020). Next, a high-performance deep learning pipeline is automatically configured and trained to link concrete mixture design variables, including the physicochemical information of ingredients, and concrete property (Jin et al., 2019). The trained deep learning model is then utilized to evaluate the importance of physicochemical variables and discover reactions between chemical compounds. Different types of concrete are considered in this research to test the applicability of the proposed method. The considered types of concrete include conventional concrete, self-compacting concrete, high-strength concrete, and ultra-high performance concrete. These types of concrete use various types of ingredients, which are considered as input features. For example, steel fibers are often used in high-strength concrete and ultra-high performance concrete to enhance the tensile properties.

This research will advance the concrete design capability, offer a new approach to explore the reactions in complex mixtures, and greatly facilitate the adoption of different SCMs into concrete for efficient design and manufacturing of high-performance low-carbon cost-effective concrete, offering an efficient avenue for the development of sustainable concrete. The remainder of the paper is organized as follows:

Section 2 presents the methods. Section 3 introduces the collected dataset. Section 4 elaborates the results and discussion. Section 5 summarizes the conclusions.

## 2. Methods

### 2.1. Overview

Fig. 1 shows the flowchart of the proposed approach to predict the properties and discover the chemical reactions of cementitious composites. A deep learning framework is proposed through integrating five main steps: (1) A dataset is established to relate the concrete property, such as the compressive strength, to the key concrete design variables, such as the curing time, specimen type, mixture proportion, chemical composition, and particle size distribution (see Section 2.2). (2) The numerical data in the dataset are converted to textual data using an artificial language created in this research (see Section 2.3). (3) Feature extraction is performed using N-grams to convert the textual data into structured data through text mining (see Section 2.4). (4) A high-performance deep learning model is developed through token embedding, neural search, and model retraining (see Section 2.5). (5) The performance of the trained deep learning model is evaluated using three performance metrics (see Section 2.6). With the predictive model, the importance of chemical compositions is evaluated (see Section 2.7), and chemical reactions of ingredients are discovered (see Section 2.8).

### 2.2. Dataset development

A dataset is developed to represent the mixture proportion, processing, and testing of concrete. The dataset covers five types of numerical variables: (1) Mixture proportion. In this study, the mass ratios of ingredients are included in the dataset. (2) Particle size distribution. Three statistical parameters are used to represent the particle size distribution of granular materials, which are $D_{10}$, $D_{50}$, and $D_{90}$, representing the 10th, 50th, and 90th percentiles of mass, respectively. (3) Chemical composition. The chemical composition of each ingredient is represented by the mass percentage of each compound in the ingredient. (4) Curing time. The curing time is presented by the number of days. (5) Specimen type. The shape and dimensions of specimens are included. To develop the approach, this research focuses on the compressive strength of concrete under the standard curing condition, but the approach will be applicable to other properties such as the workability and the tensile properties, as long as the training dataset of compressive strengths is replaced by a dataset of the other properties or other curing conditions, as elaborated in reference (Mahjoubi et al., 2021).

The generalizability of the model is evaluated based on a test dataset unseen to the model, meaning that the dataset is not used to train the model. In this study, the developed dataset is split into three datasets: training (65%), validation (15%), and test (20%) sets, according to references (Xu et al., 2021; Malinin et al., 2020). The training dataset is utilized to train the deep learning model; the validation dataset is used to estimate prediction performance for model selection; and the test dataset is used to evaluate the generalizability.

### 2.3. Artificial language

This study creates an artificial language to provide an explanatory essay for each cementitious composite. Particularly, the artificial language converts the numerical data of the curing time, mixture proportion, chemical composition, and particle size distribution of ingredients into a unified textual form. The following linguistic rules are established:

- Notations and numbers are the words of this language. The notations are as follows: “d”, “c”, and “cyl” represent the days of curing, cubic specimens, and cylinder specimens, respectively. “D10”, “D50”, and “D90” are used to describe particle size distribution. “SP” and “SF” represent the superplasticizer and steel fibers, respectively. Chemical formulas are used to represent chemical compounds. For example, $H_2O$, CaO, and $SiO_2$ are used to represent water, calcium oxide, and silica dioxide, respectively.
- Each textual data instance comprises multiple sentence-like elements, as shown in Eq. (1):

$$\mathbf{W} = [\mathbf{A}][\mathbf{B}][\mathbf{S_1}][\mathbf{S_2}]\ldots[\mathbf{S}_N] \tag{1}$$

where **A** is the sentence related to the curing time and test specimen; sentence **B** describes the mixture proportions of admixtures and fibers; and $\mathbf{S_i}$ describes the mass proportion, particle size distribution, and chemical composition of the $i$ th ingredient, where $i = 1, 2, \ldots, N$, and $N$ is the number of raw materials. **A** is expressed as shown in Eq. (2):

$$\mathbf{A} = aged, sw \times l \tag{2}$$

where *age* is the curing time of cementitious composite in days; $s$ is the shape of specimen which is either cube (“c”) or cylinder (“cyl”); $w$ and $l$ are the width and length of specimen.

- The second sentence (**B**) describes the mixture proportions of superplasticizer and fibers:

$$\mathbf{B} = SP : sp,\ SF = sf \tag{3}$$

where $sp$ and $sf$ are the mass of superplasticizer and fibers in a unit volume of the mixture.

- The third to the last sentences describe the mixture proportions, particle size distribution, and chemical composition of constituents. The $i$ th ingredient is defined as shown in Eq. (4):

$$\mathbf{S}_i = W_i,\ D10 : d_{10,i}, D50 : d_{50,i}, D90 : d_{90,i}, CC_{1,i} : cc_{1,i}, CC_{2,i} : cc_{2,i}, \ldots, CC_{n,i} : cc_{n,I} \tag{4}$$

where $W_i$ is the mass of the $i$ th ingredient in a unit volume of the cementitious composite; $d_{10,i}$, $d_{50,i}$, $d_{90,i}$ are the $D_{10}$, $D_{50}$, and $D_{90}$ of the particle size distribution of the $i$ th ingredient; $cc_{j,i}$ is the percentage of the $j$-th chemical compound in the $i$ th ingredient.

To clarify the artificial language, an example is provided in Appendix A based on a concrete mixture in reference (Horsakulthai, 2021). The numerical data related to the mixture proportions, physicochemical properties of raw ingredients, curing time, and specimen dimensions, as well as the compressive strengths are given in Appendix A. Next, the textual data obtained from the artificial language from the numerical data is reported.

### 2.4. N-grams characterization of cementitious composites

Feature extraction is utilized to convert the textual data into structured data based on character-level N-grams (Shannon, 1948) and tokenization in two sequential steps. First, tokenization is performed to segment a text into pieces known as tokens or N-grams. N-grams are the sequence of either words or characters. A word is a sequence of characters delimited by two delimiters, while a delimiter is a character that specifies the boundaries of words, such as space and punctuation marks (e.g., period and bracket) (Pibiri and Venturini, 2019). N-grams have been extensively used in natural language processing, such as machine translation, auto-completion in search engines, spelling correction, and automatic speech recognition (Pibiri and Venturini, 2019; Reshamwala

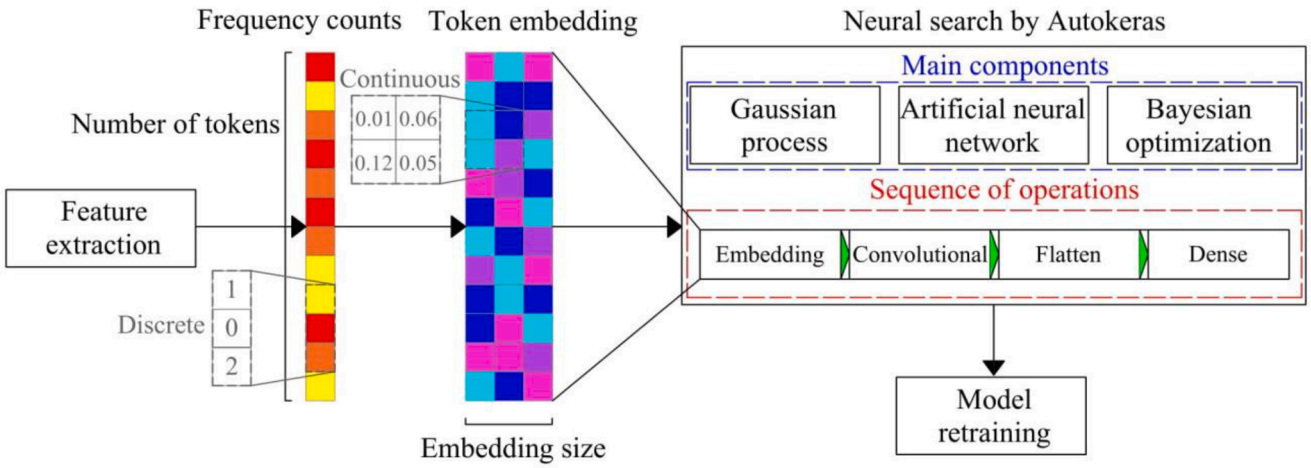


**Fig. 2.** Development of a high-fidelity deep learning model using token embedding, neural search, and model training.

et al., 2013).

Then, the occurrences of N-grams are counted to convert the text into the occurrence number (Masud et al., 2008). The method based on character-level N-grams is proposed since the sequence of characters of a corpus in the artificial language is more informative than the sequence of words, as shown in Appendix B. This is because the textual data representing the mixture proportions, particle size distribution, and chemical composition have both digits and letters, as indicated in Section 2.3. To clarify the character-level feature extraction method, an example is shown in Appendix B, where the proposed feature extraction method is applied to a textual data given in Appendix A.

The parameter N influences the extracted features. When N is equal to 1, N-grams is known as unigram with limited semantic information because the sequence of characters is not considered. When N exceeds a certain value, the number of occurrences of N-grams will be small, known as rare N-grams, with limited information as well. This study proposes an approach to determine N according to the length of words in a given text. First, the distribution of the number of characters in the words is determined. Second, the lower boundary of N is set to 1 while the upper boundary of N is set to the length of a word that is longer than 90% of the other words in the given text. In this way, the extracted N-grams will cover most of the words while eliminating long and infrequent sequences of words. A parametric study is conducted to show the effectiveness of the proposed method, and the results are reported in Appendix B.

It is essential to ensure that the prediction accuracy of the deep learning model is independent of the sequence of words. For example, the semantic meaning of [..., $a$ = x, $b$ = y, ...] is the same as that of [..., $b$ = y, $a$ = x, ...]; and the semantic meaning of $[\mathbf{A}][\mathbf{B}][\mathbf{S_1}]\ldots[\mathbf{S_j}][\mathbf{S_{j+1}}]\ldots[\mathbf{S_N}]$ is the same as that of $[\mathbf{A}][\mathbf{B}][\mathbf{S_1}]\ldots[\mathbf{S_{j+1}}][\mathbf{S_j}]\ldots[\mathbf{S_N}]$. The investigation results are given in Section 4.3.

### 2.5. Deep learning pipeline

As shown in Fig. 2, the deep learning model is developed by sequentially performing token embedding, model configuration, and model retraining. Token embedding is applied to extract semantic information. Automatic model configuration is realized using a neural search method based on Auto-Keras (Jin et al., 2019). Finally, the deep learning model is retrained to improve accuracy. The three steps are elaborated in Sections 2.5.1 to 2.5.3.

#### 2.5.1. Token embedding

There are two main issues regarding the feature representations obtained by converting a given text into a feature vector for feature extraction: (1) The feature representation of a textual data reflects the number of occurrences of tokens rather than their semantic meanings (Tsai et al., 2020). Although a feature representation provides the frequency of tokens, it has no semantic meaning of the words such as the number expression and the chemical formula. (2) A component of a feature vector is a non-negative integer representing the number of occurrences of specific N-grams. In mathematics, a continuous variable is a variable that can be any value within a range, while a discrete variable is a variable that is an integer. Discrete representations of the extracted features are more difficult than continuous representations for neural networks to learn (Zhou et al., 2019). To address the above issues, an embedding approach is proposed to map textual data to a meaningful continuous space, where the distance between textual data quantifies semantic similarity, as shown in Eq. (5):

$$\chi = \rho \mathrm{D} \tag{5}$$

where $\rho$ is the embedding matrix; $\mathrm{D} \in \mathbb{Z}^m$ is a feature vector with the length of $m$ containing the number of occurrences of N-grams; and $\chi \in \mathbb{R}^{m \times L}$ is the projection of D in the embedding space with dimension $L$. The components of the matrix $\rho$ and the parameter $L$ are the trainable parameters of the embedding layer. In this study, the embedding layer is a part of the deep learning model, meaning that the parameters are tuned during the learning process of the deep learning model.

#### 2.5.2. A neural search method for semi-automated deep learning

This study implements Auto-Keras (Jin et al., 2019) to construct a high-performance deep learning model for predicting the compressive strength of concrete. The method configures a deep convolutional neural network architecture customized for the specific dataset. The hyperparameters are tuned via neural search, aiming to maximize the prediction accuracy.

Auto-Keras involves four main components: (i) A regression technique, called Gaussian process, is used to estimate the accuracy of a given neural network with a specific architecture and hyperparameters. (ii) A neural network is used to map the variables describing the architecture and hyperparameters to a latent space. It is impractical to directly vectorize a neural network architecture due to the uncertain numbers of layers and parameters of a network (Jin et al., 2019). Therefore, a neural network is used as a kernel to convert the vector of each neural network architecture to a unified form. (iii) Bayesian optimization is used to automatically search the architecture and the hyperparameters of a neural network. (iv) An acquisition function is defined to estimate the potential utility of a given architecture. The acquisition function is minimized using the Bayesian optimization to select the most promising architecture to test (Jin et al., 2019).

In this study, a search space is defined for Auto-Keras to search for a high-performance deep learning model with a specific sequence of

operations. The deep learning models selected by Auto-Keras have one embedding layer, followed by one or multiple convolutional, max pooling, flatten, and dense layers:

- Embedding layer: The layer is responsible for token embedding, as discussed in Section 2.5.1.
- Convolutional block: Individual or multiple convolution layers with dropouts. A dropout layer randomly sets a set of neurons to zero, aiming to mitigate overfitting (Srivastava et al., 2014). Overfitting is a modeling error that occurs when a predictive model learns the training dataset so well that it performs poorly on unseen data. Max pooling is a down-sampling convolution operation where the feature map delivered by the convolutional layers is categorized into a set of regions. The maximum values of the regions are selected, and other values are discarded. This operation has two major benefits: (I) It reduces the dimensionality of the feature map, and thus lowers the computational burden for training. (II) It reduces the chance of overfitting: The overfitting phenomenon may occur when the dataset has many features. The model may learn spurious correlations in the training data that are not reflected in the unseen data. Max pooling layer reduces the dimensionality of the outputs obtained by convolutional layers while effectively preserving feature information. Thus, the max pooling layer is able to avoid overfitting.
- Flatten layer: Flattening converts the multidimensional feature vector from the convolution block into a 1-dimensional array. The layer makes the feature vector linear and helps the vector pass a dense layer.
- Dense block: Dense layers are used to generate the output of the deep models. The neurons of a dense layer are connected to every neuron of its preceding layer. Dense block involves multiple dense and activation layers to perform predictions based on the 1-dimensional array of the previous flatten layer. The head of the network is a dense layer with one neuron, which is responsible for generating the output of the network.

The parameters of the neural search method are set as follows: The search for a high-performance model is stopped when the number of tried neural networks reaches 200. To balance the computational cost and accuracy, the number of data points in each batch is set to 50. The number of epochs to train each model during the search is set to 20. The parameters were set based on trial-and-error. Early stopping is applied to avoid overfitting. The training process is terminated if the loss function on the validation set is not improved for 5 consecutive epochs. Finally, root mean square error (RMSE) is utilized to evaluate the accuracy of the deep learning models:

$$\mathrm{RMSE}(P,A)=\sqrt{\frac{\sum_{i=1}^{n}(p_i-a_i)^2}{n}} \tag{6}$$

where $P=[p_1, p_2, \ldots, p_n]$ and $A=[a_1, a_2, \ldots, a_n]$ are vectors containing the predicted and the actual values; and $n$ represents the number of predictions.

An automated learning approach, Microsoft Azure automated machine learning (Fusi et al., 2018; Copeland et al., 2015), is used to validate the proposed neural search method. Azure automated machine learning creates a high-performance machine learning pipeline via Bayesian optimization of ensemble learning and various machine learning methods, such as extreme gradient boosting (Chen and Guestrin, 2016), light gradient boosting machine (Ke et al., 2017), and random forest (Breiman, 2001).

### 2.5.3. Model retraining

The neural search method compares the performance of models trained with 20 epochs to save computational cost. It is speculated that the models suffer from undertraining, meaning the number of epochs used to train the neural network models in the neural search is insufficient. It is promising to improve the generalizability by reducing the validation loss through increasing the number of training epochs. Therefore, the best performing trained deep learning model is retrained with 100 more epochs, aiming to maximize the generalization performance.

## 2.6. Performance metrics

Root mean square error (RMSE) indicates how far predictions fall from actual values based on Euclidean distance (Barhemat et al., 2022; Mahjoubi et al., 2023), as shown in Eq. (6). Apart from RMSE, three other performance metrics were used to evaluate the performance of the trained predictive models:

(1) Mean absolute error (MAE) measures the average significance of absolute errors (Mahjoubi et al., 2021; Mahjoubi et al., 2023):

$$\mathrm{MAE}(\mathrm{P},\mathrm{A})=\sum_{i=1}^{n}|\mathrm{p_i}-\mathrm{a_i}|^2 \tag{7}$$

(2) Median absolute deviation (MAD) measures the variability of errors (Mahjoubi et al., 2021; Mahjoubi et al., 2023):

$$\mathrm{MAD}(\mathrm{P},\mathrm{A})=\mathrm{median}(|\mathrm{p_1}-\mathrm{a_1}|,|\mathrm{p_2}-\mathrm{a_2}|,\ \ldots,|\mathrm{p_n}-\mathrm{a_n}|) \tag{8}$$

(3) Coefficient of determination ($R^2$) shows the extent of uncertainty (Mahjoubi et al., 2021; Barhemat et al., 2022; Mahjoubi et al., 2023):

$$\mathrm{R}^2(\mathrm{P},\mathrm{A})=1-\frac{\sum_{i=1}^{n}(\mathrm{p_i}-\mathrm{a_i})^2}{\sum_{i=1}^{n}[\mathrm{a_i}-\mathrm{mean}(\mathrm{a_i})]^2} \tag{9}$$

The variables used in the performance metrics are defined in Section 2.5.2. In general, $R^2 > 0.8$ indicates a very strong correlation between predictions and actual values, and the performance of the predictive model is satisfying, according to reference (Miot, 2018).

## 2.7. Variable importance

A novel importance metric is presented to evaluate the importance of the chemical compounds and the percentile values of particle size distribution on the concrete properties. The concept of the variable importance metric is that the importance of a variable is reflected by the increase of the error when the variable is removed from the training set. The variable importance is defined as:

$$VI_\tau=\delta(R,\tau)-\delta(E) \tag{10}$$

where $VI_\tau$ is the variable importance of variable $\tau$; $\delta(R,\tau)$ is the RMSE of the model with the test dataset when the information about variable $\tau$ is removed from the test dataset; $\delta(E)$ is the RMSE for the test set when all information is provided in the test dataset. A high variable importance means that the variable greatly impacts the concrete property.

## 2.8. Variable interaction

The chemical interactions between the chemical compounds are identified using variable interaction. The interaction between two variables, which are two chemical compounds, implies chemical reactions between the two compounds (Wang et al., 2015). This study proposes a measurement to quantify the interaction between variables, inspired by mutual information (Kraskov et al., 2004). Variable interaction refers to a two-way interaction measure that shows whether two variables interact with each other and the extent of the interaction between the

**Table 1**
Statistics of the chemical compounds of cement.

| Variable | Unit | Range | Skewness | Kurtosis |
|---|---|---|---|---|
| $Al_2O_3$ | % | 2.76 – 7.31 | 0.83 | 1.21 |
| CaO | % | 59.52 – 68.05 | 0.24 | -0.49 |
| $Fe_2O_3$ | % | 2.24 – 5.1 | 1.18 | 1.69 |
| $K_2O$ | % | 0.11 – 1.027 | 0.69 | 0.51 |
| MgO | % | 0.6 – 4.7 | 0.38 | -0.47 |
| $Na_2O$ | % | 0 – 0.85 | 1.17 | 1.30 |
| $P_2O_5$ | % | 0.06 – 0.74 | 2.29 | 5.48 |
| $SiO_2$ | % | 19.03 – 23.25 | 0.11 | -0.26 |
| $SO_3$ | % | 0.55 – 3.49 | -1.55 | 5.39 |

**Table 2**
Statistics of the chemical compounds of SCMs.

| Variable | Unit | Range | Skewness | Kurtosis |
|---|---|---|---|---|
| $Al_2O_3$ | % | 0 – 43 | 1.39 | 1.68 |
| CaO | % | 0 – 66.06 | 1.40 | 0.69 |
| $Fe_2O_3$ | % | 0 – 21.93 | 2.22 | 4.53 |
| $K_2O$ | % | 0 – 7.5 | 2.51 | 8.30 |
| MgO | % | 0 – 10.8 | 1.83 | 2.20 |
| $Na_2O$ | % | 0 – 7.5 | 4.81 | 25.70 |
| $P_2O_5$ | % | 0 – 2.83 | 2.68 | 7.43 |
| $SiO_2$ | % | 17.33 – 99.8 | -0.31 | -1.61 |
| $SO_3$ | % | 0 – 4.85 | 2.58 | 7.68 |

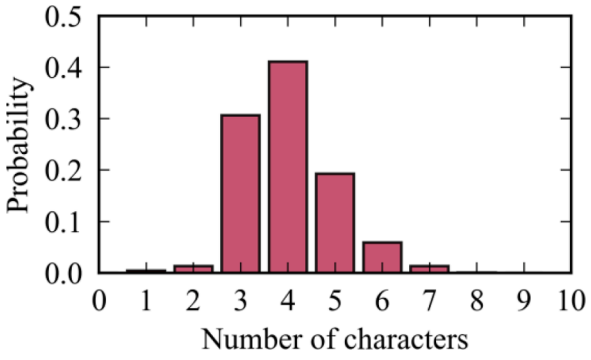


**Fig. 3.** Distribution of the length of characters in words. The vertical axis shows the density of probability, while the horizontal axis shows the number of characters.

two variables. The variable interaction between $x$ and $y$ is defined as:

$$\mathrm{VI}(x,y) = |f(x,y) - f(x) - f(y)| \tag{11}$$

where $x$ and $y$ are two variables; $f(x)$ is the range of average of predictions associated with the variability of $x$; the range of a variable is the difference between the lowest and highest values; and $f(x,y)$ is the range of average of predictions associated with the variability of both variables. To determine $f(x)$, variable $x$ is randomly assigned for each data points in the dataset, while all the other variables are kept constant. The random value is within the range of the values in the dataset. This process is repeated 50 times to obtain $f(x)$ and $f(x,y)$. Next, the outputs of the machine learning model are determined for the manipulated data points. Finally, $f(x)$ is determined as the average of the outputs. The normalized variable interaction can be expressed as:

$$\mathrm{NVI}(x,y) = \frac{\mathrm{VI}(x,y)}{\max(\mathrm{VI})} \tag{12}$$

where $\mathrm{NVI}(x,y)$ is the normalized variable interaction between $x$ and $y$; and $\max(\mathrm{VI})$ is the maximum of variable interactions for all possible pairs of variables.

## 3. Collected dataset

A total of 760 cementitious composites with their corresponding compressive strengths were collected from references (Du et al., 2022; Horsakulthai, 2021; Yu et al., 2015; Liu and Wei, 2021; Pezeshkian et al., 2021; Huang et al., 2017; Kang et al., 2019; Jaturapitakkul et al., 2004; Shi et al., 2021; Zhan et al., 2021; Li and Zhang, 2022; Hasnat and Ghafoori, 2021; Wang et al., 2022; Van et al., 2014; Jamil et al., 2016). The unit of compressive strength is MPa. In addition, the collected compressive strengths are in the range of 0.04 MPa to 204.9 MPa. As described in Section 2.2, the information on the mixture proportions, chemical composition, percentile values of the particle size distribution of ingredients, shape size and type of specimen, and curing time were considered for each data instance.

This dataset includes 15 types of cement, 37 types of SCMs, 8 types of fillers, and 14 types of aggregates, which had different particle sizes and chemical compositions. The included SCMs are different fly ash (Yu et al., 2015; Jaturapitakkul et al., 2004; Shi et al., 2021; Li and Zhang, 2022; Hasnat and Ghafoori, 2021), silica fume (Liu and Wei, 2021; Pezeshkian et al., 2021; Huang et al., 2017; Kang et al., 2019; Jaturapitakkul et al., 2004; Shi et al., 2021; Hasnat and Ghafoori, 2021; Wang et al., 2022; Van et al., 2014), nano silica (Yu et al., 2015; Shi et al., 2021), natural zeolite (Pezeshkian et al., 2021), natural pozzolan (Hasnat and Ghafoori, 2021), slag (Yu et al., 2015; Shi et al., 2021; Zhan et al., 2021; Li and Zhang, 2022; Hasnat and Ghafoori, 2021; Van et al., 2014), metakaolin (Zhan et al., 2021), glass powder (Zhan et al., 2021), rice husk ash (Van et al., 2014; Jamil et al., 2016), and recycled concrete powder (Horsakulthai, 2021). The fillers were limestone powder (Yu et al., 2015; Huang et al., 2017; Kang et al., 2019; Wang et al., 2022), red mud (Li and Zhang, 2022), and quartz powder (Liu and Wei, 2021; Kang et al., 2019; Van et al., 2014). The aggregates were calcined bauxite (Liu and Wei, 2021), river sand (Liu and Wei, 2021; Jaturapitakkul et al., 2004; Hasnat and Ghafoori, 2021; Wang et al., 2022; Jamil et al., 2016), glass sand (Pezeshkian et al., 2021), quartz sand (Pezeshkian et al., 2021; Huang et al., 2017; Shi et al., 2021; Li and Zhang, 2022), masonry sand (Du et al., 2022), dolomite (Li and Zhang, 2022), and gold mine tailings (Wang et al., 2022). Straight steel fibers measuring 13 mm in length and 0.2 mm in nominal diameter (Du et al., 2022; Liu and Wei, 2021; Shi et al., 2021; Hasnat and Ghafoori, 2021; Van et al., 2014) were adopted. Different types of concrete were included in the dataset: self-compacting mortar (Horsakulthai, 2021; Jamil et al., 2016), ultra-high performance concrete (Du et al., 2022; Yu et al., 2015; Liu and Wei, 2021; Pezeshkian et al., 2021; Huang et al., 2017; Kang et al., 2019; Shi et al., 2021; Zhan et al., 2021; Hasnat and Ghafoori, 2021; Wang et al., 2022; Van et al., 2014), high-strength concrete (Jaturapitakkul et al., 2004), and conventional concrete (Li and Zhang, 2022).

Table 1 and Table 2 respectively list the statistics of the chemical compounds of cement and SCMs. Skewness and kurtosis are used to measure the non-normality. Skewness reflects the asymmetry of distribution, and kurtosis indicates the outlier-prone extent of distribution.

## 4. Results and discussions

This section presents the results including feature extraction of textual data (Section 4.1), development, configuration, and performance evaluation of the deep learning model (Section 4.2), effect of the order of information on the prediction accuracy (Section 4.3), importance of chemical composition (Section 4.4), and chemical discovery in cementitious composites (Section 4.5).

### 4.1. Artificial language processing

Fig. 3 shows the distribution of the number of characters for words in the textual data obtained from the developed dataset. The figure indicates that 90% of words have five or fewer characters. Based on the discussions in Section 2.4, the upper bound of N in the feature extractor

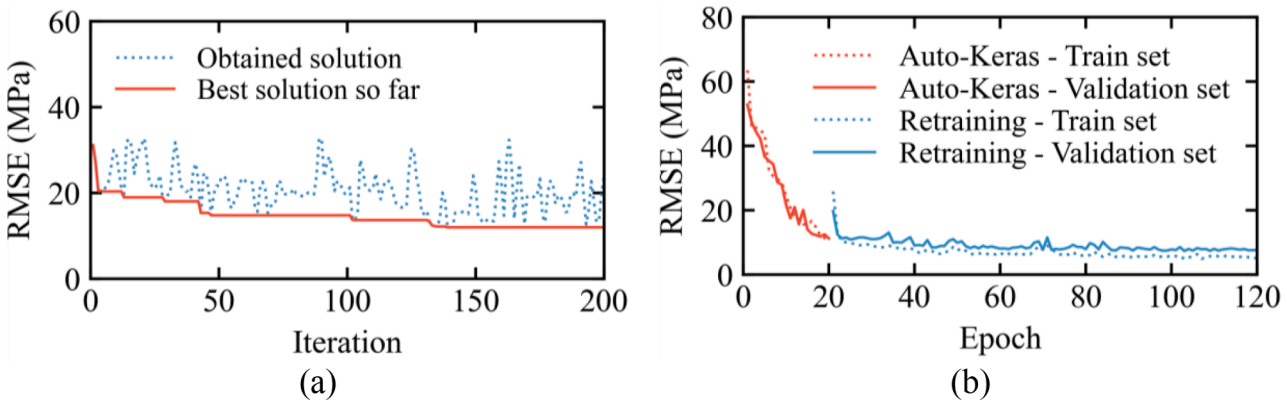


**Fig. 4.** Exploration for a high-performance model: (a) RMSE of the models iteratively selected by Auto-Keras, and (b) learning curves of the final model in neural search and model retraining.

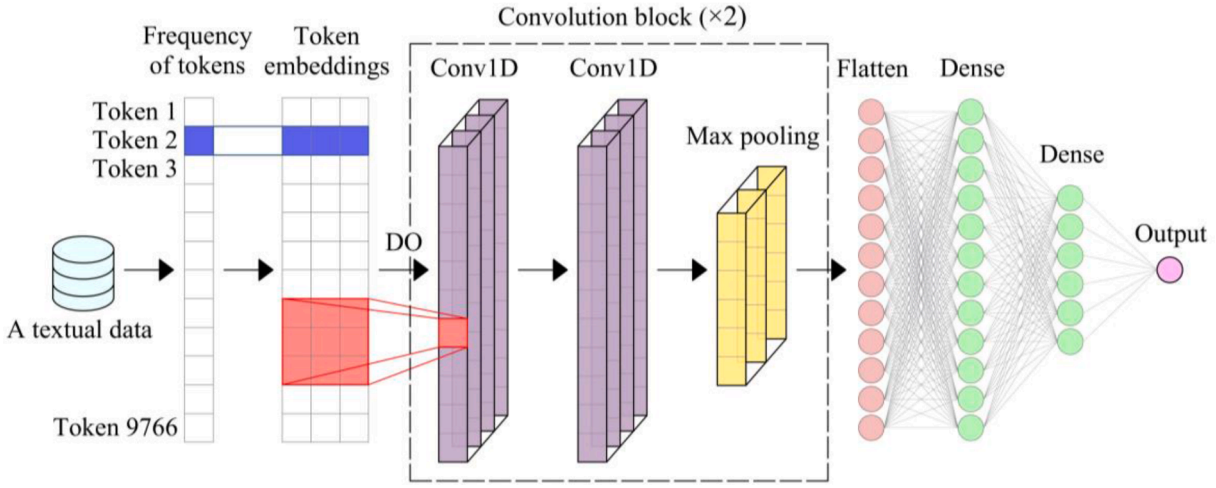


**Fig. 5.** Architecture of the deep neural network developed to predict the concrete properties. DO stands for dropout, and Conv1d stands for 1D convolution layer.

**Table 3**
The configuration of the final deep learning model.

| Layer | Output shape | Kernel size | Stride length | Activation function | Number of parameters |
|---|---|---|---|---|---|
| Input | 9766 | N.A. | N.A. | N.A. | 0 |
| Embedding | 9766×256 | N.A. | N.A. | N.A. | 5120,256 |
| Dropout | 9766×256 | N.A. | N.A. | N.A. | 0 |
| Conv1D-1 | 9764×32 | 3 | 1 | ReLu | 24,608 |
| Conv1D-2 | 9762×512 | 3 | 1 | ReLu | 49,664 |
| Max pooling-1 | 4881×512 | 2 | 2 | N.A. | 0 |
| Conv1D-3 | 4879×32 | 3 | 1 | ReLu | 49,184 |
| Conv1D-4 | 4877×32 | 3 | 1 | ReLu | 3104 |
| Max pooling-2 | 2438×32 | 2 | 2 | N.A. | 0 |
| Flatten | 78,016 | N.A. | N.A. | N.A. | 0 |
| Dense-1 | 128 | N.A. | N.A. | ReLu | 9986,176 |
| Dense-2 | 32 | N.A. | N.A. | ReLu | 4128 |
| Dense-3 | 1 | N.A. | N.A. | ReLu | 33 |
| | | | | Total parameters: | 15,237,153 |

* "N.A." stands for "not applicable".

is $N = 5$. The N-grams with five or fewer unique characters are extracted from textural data for tokenization. Next, the frequencies of the extracted tokens are fed to the token embedding layer. A total of 9766 N-grams were identified in the training set by the feature extractor. Each data was converted to a vector with 9766 components, meaning that the number of input variables of the deep learning model is 9766, and each variable represents the frequency of a unique N-grams.

**Table 4**
Performance metrics of the deep learning model.

| Dataset | MAE (MPa) | MAD (MPa) | RMSE (MPa) | $R^2$ (1) |
|---|---|---|---|---|
| Training | 3.06 | 2.19 | 5.08 | 0.99 |
| Validation | 5.79 | 4.58 | 7.46 | 0.97 |
| Test | 6.31 | 4.15 | 9.15 | 0.96 |

### 4.2. Deep learning model

This section describes the development, configuration, and performance of the deep learning model. Section 4.2.1 elaborates the model development. Section 4.2.2 describes the architecture and hyperparameters of the model. Section 4.2.3 evaluates the performance of the trained model.

#### 4.2.1. Development process

Fig. 4(a) shows the performance of Auto-Keras on the feature vectors extracted by N-grams feature extractor. The RMSE decreased from 31.14 MPa to 12.02 MPa as the number of iterations increased from 1 to 200, indicating that the method iteratively obtained better solutions. Fig. 4(b) shows the learning curves of the final model obtained from neural search and model retraining. The loss function increases when

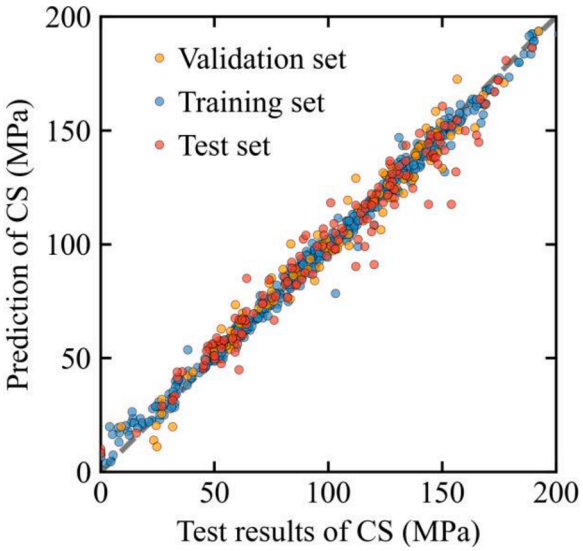


**Fig. 6.** The predictions of compressive strength obtained by the deep learning model versus the actual values obtained by experimental tests.

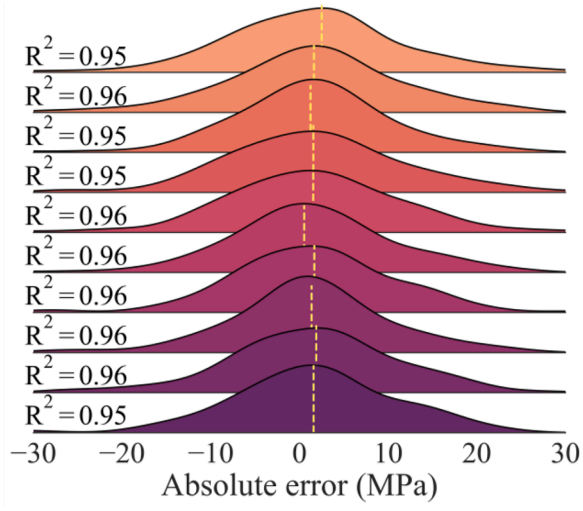


**Fig. 7.** Effect of the order of information on the accuracy. Dashed lines show the median.

the retraining phase is started, as artificial neural networks are prone to abruptly forget previously learned information (Kirkpatrick et al., 2017). Although the loss function increases at the beginning of the retraining phase, model retraining improved the validation loss by about 38%. The lowest validation loss obtained in the neural search phase was 12.02 MPa, and the lowest validation loss obtained in the model retraining phase was 7.46 MPa. Fig. 4(b) indicates that 20 epochs were sufficient to obtain a high-performance model. Further increasing the number of epochs increased the computational burden.

#### 4.2.2. *Deep learning architecture and hyperparameters*

Fig. 5 shows the architecture of the deep learning model determined by the neural search method. The features obtained by the character-level extractor are fed to the embedding layer, followed by a dropout layer with a rate of 0.25. Next, there are two convolution blocks. Each block includes two 1D convolution layers and one max pooling layer, followed by a flatten layer and finally three dense layers.

The output shapes, hyperparameters, and number of trainable parameters of the layers of the final model are provided in Table 3. The model has more than 15 million trainable parameters to be tuned in the learning process. Nearly two-thirds of the parameters belong to the dense layers, and about one-third of parameters belong to the embedding layer.

#### 4.2.3. *Performance evaluation*

Table 4 lists the performance metrics of the trained deep learning model in predicting the compressive strength of cementitious composites. The $R^2$ value of the model for the test set is 0.96, indicating high accuracy and high generalizability of the trained model.

Fig. 6 compares the predictions made using deep learning against the experimental data for the compressive strength. The figure indicates that the predictions agree well with the experimental data, demonstrating the high precision and high accuracy of the proposed method.

The effectiveness of feature extraction is evaluated by comparing the prediction performance of the deep learning model trained on the features extracted from the characters, words, and varied hyperparameter settings. In Appendix B, the proposed character-level feature extraction leads to the highest accuracy. It is concluded that the feature extraction method provides representative features that reflect the characteristics of cementitious composites, and the deep learning model is able to learn from the extracted features.

The ensemble machine learning model automatically designed by Azure automated machine learning is described in Appendix C. The performance metrics indicate that the developed deep learning model achieves the higher accuracy and certainty. The RMSE and $R^2$ of the deep learning model for the test dataset are 9.15 MPa and 0.97, respectively. The RMSE and $R^2$ of the ensemble learning model are 10.84 MPa and 0.92, respectively. The RMSE of the deep learning model

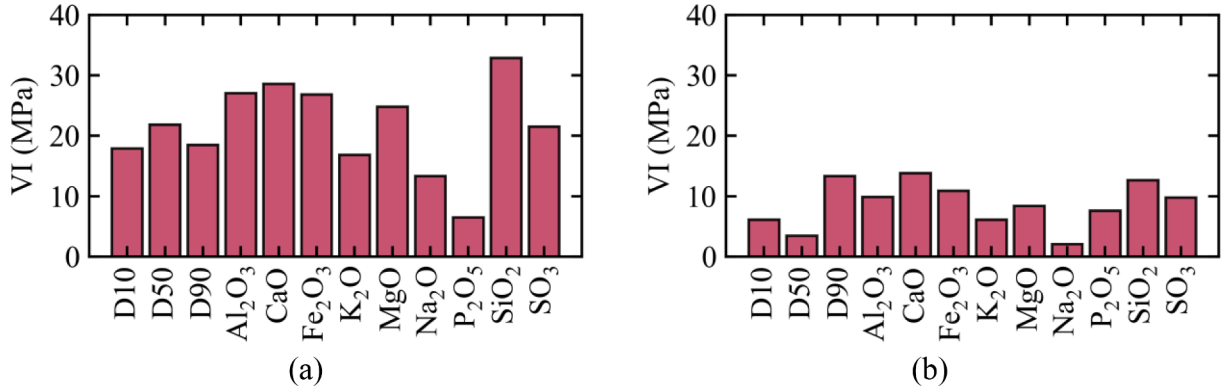


**Fig. 8.** The various importance of chemical compounds and three properties regarding particle gradation for (a) SCMs and (b) cement. Vertical axes show the variable importance of materials.

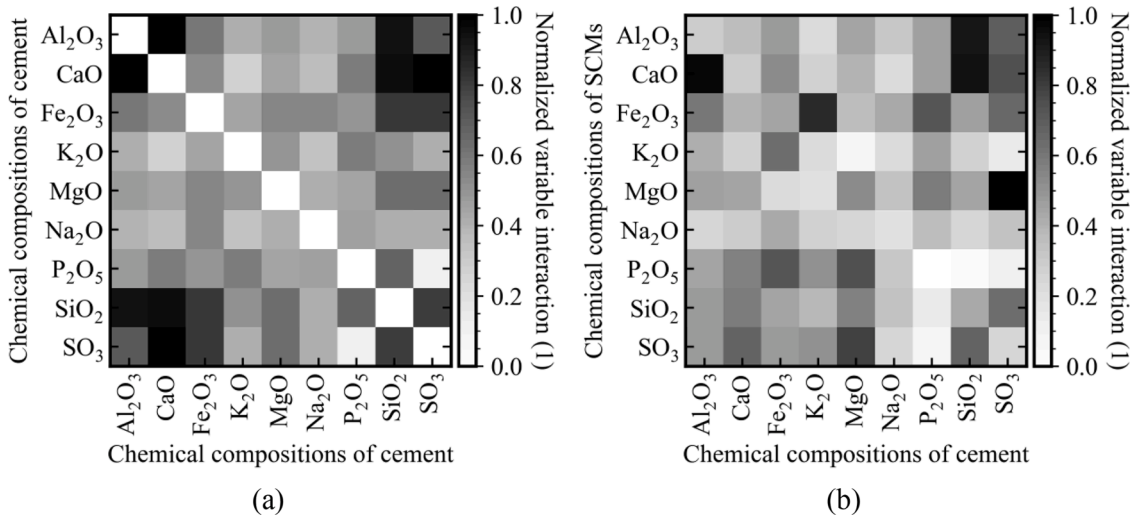


**Fig. 9.** Normalized variable interaction of chemical compounds in cement and SCMs: (a) interactions between chemical compounds of cement and (b) interactions between the chemical compounds of cement and the chemical compounds of SCMs.

**Table A1**
Mixture design and information of ingredients (Horsakulthai, 2021).

| | Portland cement | Recycled concrete powder | Sand | Water | Superplasticizer |
|---|---|---|---|---|---|
| Proportion ($kg/m^3$) | 400 | 100 | 875 | 175 | 16.5 |
| Chemical composition | | | | | |
| $SiO_2$ (%) | 19.95 | 55.19 | 97.7 | 0 | N/A |
| $Al_2O_3$ (%) | 5.18 | 2.18 | 0.5 | 0 | |
| $Fe_2O_3$ (%) | 3.25 | 4.85 | 0.1 | 0 | |
| CaO (%) | 67.84 | 35.02 | 1.4 | 0 | |
| MgO (%) | 0.79 | 0.29 | 0 | 0 | |
| $Na_2O$ (%) | 0.01 | 0.22 | 0 | 0 | |
| $K_2O$ (%) | 0.31 | 0.7 | 0 | 0 | |
| $P_2O_5$ (%) | 0.07 | 0.38 | 0 | 0 | |
| $SO_3$ (%) | 2.36 | 0.51 | 0 | 0 | |
| LOI (%) | 0.29 | 1.42 | 0 | 0 | |
| $H_2O$ (%) | 0 | 0 | 0 | 100 | |
| Particle size distribution | | | | | |
| $D_{10}$ (mm) | 3.10 | 6.80 | 9.93 | N/A | N/A |
| $D_{50}$ (mm) | 13.11 | 8.20 | 50.32 | | |
| $D_{90}$ (mm) | 25.38 | 28.94 | 90.70 | | |

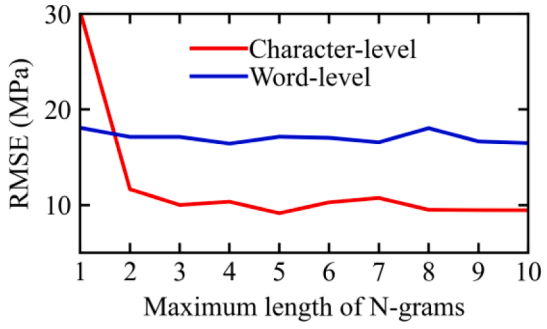


**Fig. B1.** The effects of parameters of feature extraction method on the prediction accuracy of deep learning; the horizontal axis shows the maximum length of N-grams, while the vertical axis shows the RMSE of deep learning models for the test set.

is 19% lower than that of the ensemble model, which indicates that the deep learning model has superior performance over the ensemble model.

### 4.3. Order of information

Fig. 7 shows the distribution of absolute errors of the data instances when the sentences and information in those sentences are randomly shuffled. Each of the 10 error distributions is related to all the textual data in the test set with randomly shuffled information. The $R^2$ values of all the ten trials ranged from 0.95 to 0.96, indicating that the order of information does not significantly impact the predictions made by deep learning.

### 4.4. Influencing chemical compounds

The variable importance is defined as the range of average compressive strength associated with the variability of each variable. The measurement quantifies how strong is the effect of each variable on the compressive strength. Fig. 8 shows the variable importance of the chemical compounds and the three percentile values that describes particle size distribution. The summation of importance of the three percentile values of cement and SCMs is significant. Therefore, the results confirm that particle size distribution has significant effects on the compressive strength. For both cement and SCMs, $SiO_2$, CaO, $Al_2O_3$, and $Fe_2O_3$ are the most important chemical compounds, consistent with previous research (Kolovos et al., 2002).

### 4.5. Discovery of chemical reactions

Fig. 9(a) shows the variable interactions between the chemical compounds of cement in the prediction of compressive strength. Significant interactions between the chemical compounds of cement are revealed. The strongest interaction effects with their corresponding normalized variable importance (NVI) are listed as follows:

(1) $Al_2O_3$ – CaO interaction (NVI = 1): The identified chemical compounds participate in multiple chemical reactions during hydration of cement (Gu et al., 1997; Taylor, 1997):

$$C_3A + 3C\overline{S}H_2 + 26H \Rightarrow C_6A\overline{S}_3H_{32} \quad (13)$$

$$C_3A + CH + 21H \Rightarrow C_4AH_{22} \quad (14)$$

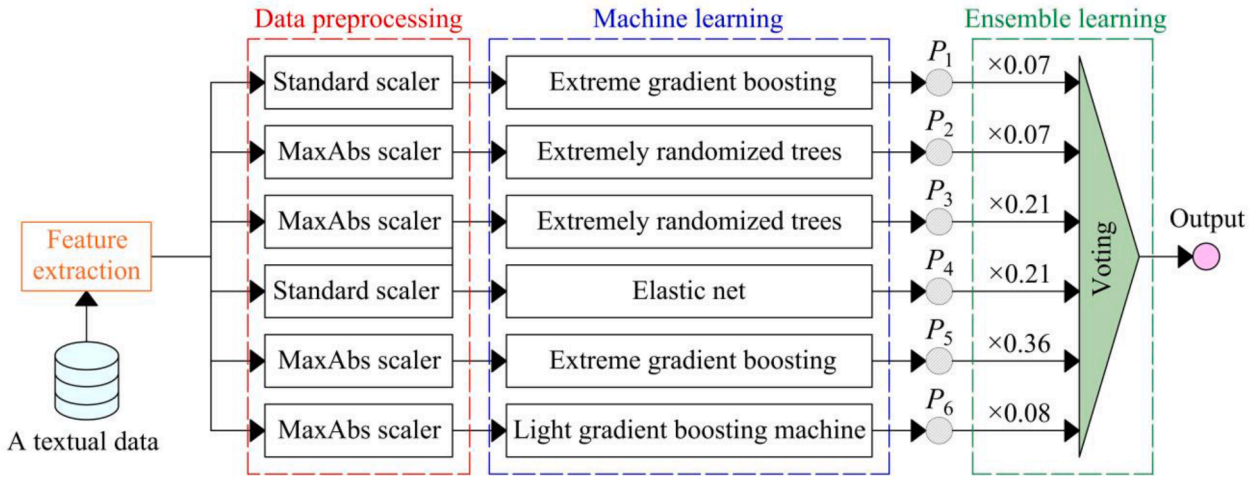


**Fig. C1.** The machine learning pipeline design automatically by Azure automated machine learning to predict the compressive strength of cementitious composites. P*i* is the prediction made by the *i* th machine learning algorithm. The predictions from the machine learning models are multiplied by ensemble weights to obtain the final prediction.

**Table C1**
Comparison of performance metrics between the deep learning and ensemble models.

| Model | Dataset | MAE (MPa) | MAD (MPa) | RMSE (MPa) | $R^2$ (1) |
|---|---|---|---|---|---|
| Deep learning | Training | 3.06 | 2.19 | 5.08 | 0.99 |
| Ensemble learning | | 3.24 | 2.42 | 5.39 | 0.97 |
| Deep learning | Validation | 5.79 | 4.58 | 7.46 | 0.97 |
| Ensemble learning | | 7.21 | 5.67 | 9.23 | 0.93 |
| Deep learning | Test | 6.31 | 4.15 | 9.15 | 0.96 |
| Ensemble learning | | 7.48 | 5.05 | 10.84 | 0.92 |

$$C_3AF + 4CH + 22H \Rightarrow C_4AH_{13} + C_4FH_{13} \quad (15)$$

$$CA + 3C\overline{S}H_2 + 26H \Rightarrow C_6A\overline{S}_3H_{32} \quad (16)$$

where C, A, $\overline{S}$, F, and H represent CaO, $Al_2O_3$, $SO_3$, $Fe_2O_3$, and $H_2O$, respectively; $C_3A$ is Celite (tricalcium aluminate); and $C_3AF$ is Felite (calcium aluminoferrite). Previous studies showed that $C_3A$ has large effect on the mechanical strengths and $C_3AF$ has little effect on the mechanical strengths (Hewlett and Liska, 2019). $C_6A\overline{S}_3H_{32}$ is ettringite which is a primary constituent of hydration of cement and contributes to the early strength of cement (Gu et al., 1997).
(2) CaO – $SO_3$ interaction with variable interaction of 0.99: The two compounds participate in two chemical reactions given in Eq. (13) and Eq. (16).
(3) CaO – $SiO_2$ interaction (NVI = 0.99): It is well known that the two compounds participate in the reactions of calcium silicate phases (Brouwers, 2003):

$$C_3S + (3 - x + y)\ H \Rightarrow C_xSH_y + (3 - x)\ Ca(OH)_2 \quad (17)$$

$$C_2S + (2 - x + y)\ H \Rightarrow C_xSH_y + (2 - x)\ Ca(OH)_2 \quad (18)$$

where C and S represent CaO and $SiO_2$. $Ca(OH)_2$ is calcium hydroxide. $C_3S$ and $C_2S$ represent Alite (tricalcium silicate) and Belite (dicalcium silicate). $C_xSH_y$ represents calcium silicate hydrate (also known as C-S-H) which is the main hydration product of cement, and directly contributes to the strength of concrete (Hewlett and Liska, 2019). The stoichiometry of C-S-H in cement is varied; therefore, x and y are varied (Goñi et al., 2010).
(4) $Al_2O_3$ – $SiO_2$ interaction (NVI = 0.98): The two compounds react with CaO and $H_2O$, according to Eqs. (13) – (18). Therefore, the percentage of $Al_2O_3$ may indirectly affect the reactivity of $SiO_2$ with CaO and formation of C-S-H gel.
(5) $Fe_2O_3$ – $SiO_2$ interaction (NVI = 0.81): $Fe_2O_3$ and $SiO_2$ separately react with CaO and $H_2O$, according to Eq. (15), (17), and (18). Consequently, the percentage of $Fe_2O_3$ indirectly influences the reactivity of $SiO_2$ with CaO and the formation of C-S-H gel.

Fig. 9(b) shows the variable interactions between chemical compounds of SCMs and cement in the prediction of compressive strength. The strongest interaction effects are listed as follows:

- SiO2 of cement – CaO of SCMs interaction (NVI = 1): The chemical compounds participate in hydration reactions given in Eqs. (17) and (18). The interaction between CaO of cement and SiO2 of SCMs is not significant. Perhaps, the reason is the difference between the reactivity of chemical compounds in cement and SCMs.
- Al2O3 of cement – CaO of SCMs interaction (NVI = 0.97): There is a strong interaction between CaO of cement and Al2O3 of SCMs. Similar to SiO2 of cement – CaO of SCMs interaction, there is no significant reaction between Al2O3 of SCMs and CaO of cement.
- SiO2 of cement – Al2O3 of SCMs interaction (NVI = 0.94): The two chemical compounds react with CaO and H2O, according to the chemical reactions given in Eqs. (13) – (18). The percentage of Al2O3 in SCMs indirectly affects the reactivity of SiO2 with CaO.

## 5. Conclusions

This paper presents an approach to link the physicochemical properties of concrete ingredients and the mechanical properties, aiming at overcoming the major challenges of considering the physicochemical properties of ingredients in existing machine learning approaches. This research created an artificial language to represent concrete mixtures and the physicochemical information of their ingredients, developed a feature extraction method based on character-level N-grams, and proposed a method to automatically configure deep learning models. The proposed approach was implemented to predict the compressive strength of complex concrete mixtures, assess the importance of variables, and discover the chemical reactions. Based on the above investigations, the following conclusions are drawn:

- The presented approach is able to consider the physicochemical information such as the chemical composition and particle size distribution of the ingredients as well as the other variables such as the mixture proportion, specimen type, and curing time of concrete in predicting concrete properties. The prediction results agree well with the experimental results, with the coefficient of determination being 0.96. The capability of considering the physicochemical information makes the approach applicable to different types of concrete ingredients with different chemical compositions and/or particle size distributions. The machine learning model was trained using a dataset encompassing various SCMs such as fly ash, slag, recycled concrete, and waste glass, indicating that the model is applicable to various SCMs. The developed capability facilitates the valorization of solid wastes in the design and applications of high-performance low-carbon cost-effective concrete.
- The presented approach is able to handle datasets with nonuniform physicochemical information of concrete ingredients. In this study, a deep learning model was trained using a dataset with the physicochemical information of cement, SCMs, and aggregate, while the physicochemical information of superplasticizer and fibers were missing.
- The proposed approach is able to evaluate the effects of variables representing the particle size distribution and chemical composition on the compressive strength of concrete and rank the importance of those variables. The results of the importance of the oxides and the particle sizes of cement and SCMs were consistent with established knowledge.
- The proposed approach is capable of discovering and assessing the chemical reactions in concrete. The interactions of the oxides of cement differ from the interactions of the oxides of cement and SCMs. Specifically, $Al_2O_3$ and CaO in cement showed the highest interaction, while the interaction between CaO in SCMs and $Al_2O_3$ in cement was low, because the oxides in cement and SCMs exist in different phases.

Based on the above investigations, future research opportunities are identified:

- The relationships between the deep learning architecture, feature extraction method, and prediction performance are still unknown. It is interesting to investigate the accuracy of various deep learning models with respect to architecture variables, such as the number of layers. In addition, it is important to investigate the performance of deep learning models developed by using other feature extraction and text mining techniques.
- The importance and interactions of other physiochemical properties are still unknown. It is significant to investigate the variable importance of mixture proportions of cement, water, various SCMs, and inert materials in future studies. Variable interaction between the percentile values representing the particle size distributions should be studied too.
- It is unknown whether the deep learning model provides reasonable predictions when the number of N-grams is out of range of the input variables of the training dataset. The prediction accuracy of various deep learning models with respect to the range of N-grams can be investigated in future research.
- It is promising to apply the proposed framework to design low-carbon concrete containing multiple types of SCMs. The proposed paradigm opens a new avenue for the design of low-carbon concrete with multiple types of by-products and/or waste materials because the deep learning method is not limited to specific ingredients.

## CRediT authorship contribution statement

**Soroush Mahjoubi:** Data curation, Formal analysis, Investigation, Software, Visualization, Writing – original draft. **Rojyar Barhemat:** Data curation, Software, Validation, Writing – review & editing. **Weina Meng:** Conceptualization, Funding acquisition, Methodology, Resources, Writing – review & editing. **Yi Bao:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Weina Meng received funding from the National Science Foundation of the United States.

## Data Availability

Data will be made available on request.

## Acknowledgement

This research was funded by the National Science Foundation [grant No. CMMI-2046407].

## Appendix A: Textual characterization

Appendix A shows the conversion of the numerical data representing a concrete mixture into textual data based on the developed artificial language. Table A1 presents the numerical data about physicochemical properties of raw ingredients of a mixture investigated in reference (Horsakulthai, 2021). Cubic specimens with a side length of 50 mm were used. The 7-day compressive strength was 66.3 MPa.

Based on the numerical data in Table A1 and linguistic rules given in Section 2.3, the sentence-like elements are built in Eqs. (A.1) and (A.2):

$$[\mathbf{A}] = [aged, sw \times l] = [7d, c50 \times 50] \tag{A.1}$$

$$[\mathbf{B}] = [SP : sp, SF = sf] = [SP : 16.5] \tag{A.2}$$

where **A** and **B** are sentence-like elements that contain information about the curing time, specimen dimensions, and mixture proportions of superplasticizer and fibers. Since the mixture did not have steel fiber, the term for steel fiber was removed from sentence **B**. The remaining sentences describe the physicochemical properties of raw ingredients. For example, $\mathbf{S_1}$ describes the information about the first ingredient, which is the Portland cement, as shown in Eq. (A.3):

$$[\mathbf{S}1] = [W1,\ \mathrm{D10}:\ d10,1,\ \mathrm{D50}:\ d50,1,\ \mathrm{D90}:\ d90,1,\ \mathrm{CC}1,1:\ cc1,1,\ \mathrm{CC}2,1:\ cc2,1,\ \ldots,\ \mathrm{CC}11,1:\ cc11,1] \tag{A.3}$$

where $\mathbf{S_1}$ is a sentence describing the physicochemical properties of Portland cement. With the values of the variables such as $W_1$ and $d_{10,1}$, Eq. (A.3) is

rewritten as Eq. (A.4):

$$[\mathbf{S1}] = [400.00 : D10 : 3.10, D50 : 13.11, D90 : 25.38, SiO2 : 19.95, Al2O3 : 5.18, Fe2O3 : 3.25, CaO : 67.84, MgO : 0.79, Na2O : 0.01, K2O : 0.31, P2O5 : 0.07, SO3 : 2.36, LOI : 0.29] \quad (A.4)$$

Similarly, the other sentences are written to describe the properties of the other ingredients. All the sentences are gathered to generate the textual data of the mixture, as shown in Eqs. (A.5) and (A.6):

$$\mathbf{W} = [\mathbf{A}][\mathbf{B}][\mathbf{S_1}][\mathbf{S_2}][\mathbf{S_3}][\mathbf{S_4}] \quad (A.5)$$

$$\begin{aligned}\mathbf{W} = {} & [7d,\ c\ 50{\times}50][SP{:}16.5][400{:}\ D10{:}3.10,\ D50{:}13.11,\ D90{:}25.38,\ SiO2{:}19.95,\\ & Al2O3{:}5.18,\ Fe2O3{:}3.25,\ CaO{:}67.84,\ MgO{:}0.79,\ Na2O{:}0.01,\ K2O{:}0.31,\\ & P2O5{:}0.07,\ SO3{:}2.36,\ LOI{:}0.29][100{:}\ D10{:}6.80,\ D50{:}8.20,\ D90{:}28.94,\ SiO2{:}55.19,\\ & Al2O3{:}2.18,\ Fe2O3{:}4.85,\ CaO{:}35.02,\ MgO{:}0.29,\ Na2O{:}0.22,\ K2O{:}0.70,\\ & P2O5{:}0.38,\ SO3{:}0.51,\ LOI{:}1.42][875{:}\ D10{:}9.93,\ D50{:}50.32,\ D90{:}90.70,\\ & SiO2{:}97.70,\ Al2O3{:}0.50,\ Fe2O3{:}0.10,\ CaO{:}1.40][175{:}\ H2O{:}100.00]\end{aligned} \quad (A.6)$$

## Appendix B: Feature extraction

Appendix B shows the effects of the hyperparameters of feature extraction method. There are two main hyperparameters: (1) Level of extraction: The N-grams are extracted based on either the sequence of characters or words. As discussed in Section 2.4, a word is a sequence of characters separated by two delimiters, such as space and punctuation marks. (2) The length of N-grams: The length of N-grams has an impact on the extracted features, as discussed in Section 2.4.

Fig. B1 shows the effect of the two parameters on the prediction accuracy of the deep learning model. Noted that a feature extraction method and a high-performance deep learning model is developed for each trial. The reason is that the size of feature vector changes by changing the two parameters. The deep learning models were developed automatically by the proposed neural search method. The figure shows the superiority of character-level feature extraction: The minimum RMSE of character-level feature extractors is 9.17 MPa, while the minimum RMSE obtained by the word-level feature extractors is 16.42 MPa. Fig. B1 indicates the effectiveness of the proposed parameter setting method elaborated in Section 2.4: The minimum RMSE of the character-level feature extractors corresponds to the maximum length of N-grams is 5. By setting this parameter to 5, 90% of the sequence of characters and words are covered by the N-grams.

According to Fig. B1, the sequence of characters is more informative than words in a piece of text. The reason is that textual data in the artificial language contain numerical values representing the mixture proportions, particle size distribution, and chemical composition. Each number is a sequence of digits. A chemical formula also contains digits and letters. It is concluded that the sequence of characters of a corpus in the artificial language is more informative than words.

The textual data in Eq. (A.6) is used as an example to demonstrate the proposed character-level feature extraction method. Among 9906 unique N-grams, five different N-grams with varied lengths are selected: (1) unigram: "5″, (2) bigrams: "O3", (3) trigrams: "2O3", (4) 4-grams: " D50:", and (5) 5-grams: "AL2O3". Assuming the first five components of the feature vector obtained by the feature extraction method corresponds to the five selected N-grams, the feature vector of the textual data is derived as:

$$FV = [a_1,\ a_2, a_3, a_4, a_5 \ldots,\ a_h] \quad (B.1)$$

where $FV$ is the feature vector of the textual data, given in Eq. (A.6), determined by the feature extraction method; $h$ is the number of unique N-grams, and is equal to 9906; $a_i$ denotes the number of occurrences of the $i$ th N-grams; $a_1,\ a_2, a_3, a_4,$ and $a_5$ are the numbers of occurrences of the five selected N-grams, respectively; $a_1$ is 20 because the unigram "5″ is repeated in the textual data for 20 times; $a_2$ is 8 because the bigrams "O3" is repeated for 8 times; $a_3$ is 6 because the trigrams "2O3" is repeated for 6 times; $a_4$ and $a_5$ are 3 because the 4-grams "D50:" and 5-grams "AL2O3" are repeated for 3 times.

## Appendix C: Ensemble machine learning

Fig. C1 illustrates the machine learning pipeline automatically designed by Azure automated machine learning. The pipeline consists of three steps: data preprocessing, machine learning algorithms, and ensemble learning. In data preprocessing, the numerical data extracted by the feature extractor

is transformed based on two different methods, namely standard scaler, and maximum absolute scaler. The methods transform the features according to the following equations:

$$ss = \frac{x - m}{\sigma} \tag{C.1}$$

$$\text{mas} = \frac{x - m}{\max(x)} \tag{C.2}$$

where $ss$ and mas are the transformed values of $x$, the input variable, according to standard scaler and maximum absolute scaler; $m$ and $\sigma$ represent the average and standard deviation of $x$; and $\max(x)$ is the maximum value of $x$.

As shown in Fig. C1, the pipeline includes six machine learning models based on four different machine learning algorithms: extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016), extremely randomized trees (Geurts et al., 2006), elastic net (Zou and Hastie, 2005), and light gradient boosting machine (LightGBM) (Ke et al., 2017). Finally, weighted voting ensemble combines the predictions from the machine learning models: Predictions from the six models were combined with assigned weights to determine the final prediction; voting weights were tuned by the automated machine learning method, aiming to maximize prediction performance.

Table C1 compares the performance metrics between deep learning and ensemble learning models. It indicates that RMSE and $R^2$ of the ensemble machine learning model for the test dataset are 10.84 MPa and 0.92, respectively, while the RMSE and $R^2$ of the deep learning model are 9.15 MPa and 0.96, respectively. It can be concluded that the prediction accuracy of the deep learning model is higher than that of the ensemble machine learning model.

## References


ACI Committee 318, 2022. ACI CODE-318-19(22): Building code Requirements For Structural Concrete and Commentary (reapproved 2022). American Concrete Institute.

Asteris, P.G., Skentou, A.D., Bardhan, A., Samui, P., Pilakoutas, K., 2021. Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. Cement Concrete Res. 145, 106449.

Aubert, J.-E., Husson, B., Vaquier, A., 2004. Use of Municipal Solid Waste Incineration Fly Ash in Concrete. Cement Concrete Res. 34, 957–963.

Barhemat, R., Mahjoubi, S., Li, V.C., Bao, Y., 2022. Lego-inspired reconfigurable modular blocks for automated construction of engineering structures. Autom. Construct. 139, 104323.

Borosnyói, A., 2016. Long term durability performance and mechanical properties of high performance concretes with combined use of supplementary cementing materials. Construct. Build. Mater. 112, 307–324.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brouwers, H.J.H., 2003. Chemical reactions in hydrated ordinary Portland cement based on the work by Powers and Brownyard. In: Proceedings 15th Ibausil (Internationale Baustofftagung). Weimar, 1, pp. 553–566.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

China Academy of Building Research, 2010. Standard For Evaluation of Concrete Compressive Strength (GB/T 50107-2010). Ministry of Housing and Urban-Rural Development and General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China.

Copeland, M., Soh, J., Puca, A., Manning, M., Gollob, D., 2015. Microsoft Azure and Cloud computing, Microsoft Azure. Springer, pp. 3–26.

de Larrard, F., Sedran, T., 1994. Optimization of ultra-high-performance concrete by the use of a packing model. Cement Concrete Res. 24, 997–1009.

del Viso, J.R., Carmona, J.R., Ruiz, G., 2008. Shape and size effects on the compressive strength of high-strength concrete. Cement Concr. Res. 38, 386–395.

Du, J., Liu, Z., Christodoulatos, C., Conway, M., Bao, Y., Meng, W., 2022. Utilization of off-specification fly ash in preparing ultra-high-performance concrete (UHPC): mixture design, characterization, and life-cycle assessment. Resour. Conserv. Recycl. 180, 106136.

Elahi, M.M.A., Shearer, C.R., Reza, A.N.R., Saha, A.K., Khan, M.N.N., Hossain, M.M., Sarker, P.K., 2021. Improving the sulfate attack resistance of concrete by using supplementary cementitious materials (SCMs): a review. Construct. Build. Mater. 281, 122628.

Fusi, N., Sheth, R., Elibol, M., 2018. Probabilistic matrix factorization for automated machine learning. Adv. Neural Inf. Process. Syst. 31, 3348–3357.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach Learn 63, 3–42.

Goñi, S., Puertas, F., Hernández, M.S., Palacios, M., Guerrero, A., Dolado, J.S., Zanga, B., Baroni, F., 2010. Quantitative study of hydration of C-3-S and C-2-S by thermal analysis: evolution and composition of C–S–H gels formed. J. Therm. Anal. Calorim. 102, 965–973.

Gu, P., Beaudoin, J.J., Quinn, E.G., Myers, R.E., 1997. Early strength development and hydration of ordinary Portland cement/calcium aluminate cement pastes. Adv. Cement Based Mater. 6, 53–58.

Guo, P., Meng, W., Xu, M., Li, V.C., Bao, Y., 2021. Predicting mechanical properties of high-performance fiber-reinforced cementitious composites by integrating micromechanics and machine learning. Materials (Basel) 14, 3143.

Hasnat, A., Ghafoori, N., 2021. Properties of ultra-high performance concrete using optimization of traditional aggregates and pozzolans. Construct. Build. Mater. 299, 123907.

Hewlett, P., Liska, M., 2019. Lea's Chemistry of Cement and Concrete. Butterworth-Heinemann.

Horsakulthai, V., 2021. Effect of recycled concrete powder on strength, electrical resistivity, and water absorption of self-compacting mortars. Case Stud. Construct. Mater. 15, e00725.

Huang, W., Kazemi-Kamyab, H., Sun, W., Scrivener, K., 2017. Effect of cement substitution by limestone on the hydration and microstructural development of ultra-high performance concrete (UHPC). Cement Concrete Compos. 77, 86–101.

Hunger, M., 2010. An Integral Design Concept For Ecological Self-Compacting Concrete. Technische Universiteit Eindhoven (PhD Dissertation).

Jamil, M., Khan, M.N.N., Karim, M.R., Kaish, A.B.M.A., Zain, M.F.M., 2016. Physical and chemical contributions of rice husk ash on the properties of mortar. Construct. Build. Mater. 128, 185–198.

Jaturapitakkul, C., Kiattikomol, K., Sata, V., Leekeeratikul, T., 2004. Use of ground coarse fly ash as a replacement of condensed silica fume in producing high-strength concrete. Cement Concrete Res. 34, 549–555.

Jin, H., Song, Q., Hu, X., 2019. Auto-keras: an efficient neural architecture search system. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1946–1956.

Kang, S.-H., Jeong, Y., Tan, K.H., Moon, J., 2019. High-volume use of limestone in ultra-high performance fiber-reinforced concrete for reducing cement content and autogenous shrinkage. Construct. Build. Mater. 213, 292–305.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, pp. 3146–3154.

Ke, X., Duan, Y., 2021. Coupling machine learning with thermodynamic modelling to develop a composition-property model for alkali-activated materials. Compos. Part B: Eng. 216, 108801.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl Acad. Sci. 114, 3521–3526.

Kolovos, K., Tsivilis, S., Kakali, G., 2002. The effect of foreign ions on the reactivity of the CaO–SiO2–Al2O3–Fe2O3 system: part II: cations. Cement Concrete Res. 32, 463–469.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69, 066138.

Lehne, J., Preston, F., 2018. Making concrete change: innovation in low-carbon cement and concrete, Chatham House. Royal Insti. Int. Affairs.

Li, V.C., 2003. On engineered cementitious composites (ECC) a review of the material and its applications. J. Adv. Concr. Technol. 1, 215–230.

Li, X., Zhang, Q., 2022. Influence behavior of phosphorus slag and fly ash on the interface transition zone in concrete prepared by cement-red mud. J. Build. Eng. 49, 104017.

Liang, M., Chang, Z., Wan, Z., Gan, Y., Schlangen, E., Šavija, B., 2022. Interpretable ensemble-machine-learning models for predicting creep behavior of concrete. Cement Concr. Compos. 125, 104295.

Liu, Y., Wei, Y., 2021. Internal curing efficiency and key properties of UHPC influenced by dry or prewetted calcined bauxite aggregate with different particle size. Construct. Build. Mater. 312, 125406.

Long, W.-J., Wu, Z., Khayat, K.H., Wei, J., Dong, B., Xing, F., Zhang, J., 2022. Design, dynamic performance and ecological efficiency of fiber-reinforced mortars with different binder systems: ordinary Portland cement, limestone calcined clay cement and alkali-activated slag. J. Clean. Prod. 337, 130478.

Mahjoubi, S., Barhemat, R., Guo, P., Meng, W., Bao, Y., 2021. Prediction and multi-objective optimization of mechanical, economical, and environmental properties for strain-hardening cementitious composites (SHCC) based on automated machine learning and metaheuristic algorithms. J. Clean. Prod. 329, 129665.

Mahjoubi, S., Barhemat, R., Meng, W., Bao, Y., 2023. AI-guided auto-discovery of low-carbon cost-effective ultra-high performance concrete (UHPC). Resour. Conserv. Recycl. 189, 106741.
Mahjoubi, S., Meng, W., Bao, Y., 2022b. Logic-guided neural network for predicting steel-concrete interfacial behaviors. Expert Syst.Appl. 198, 116820.
Mahjoubi, S., Meng, W., Bao, Y., 2022a. Auto-tune learning framework for prediction of flowability, mechanical properties, and porosity of ultra-high-performance concrete (UHPC). Appl. Soft. Comput. 115, 108182.
A. Malinin, L. Prokhorenkova, A. Ustimenko, Uncertainty in gradient boosting via ensembles, arXiv preprint arXiv:2006.10562, (2020).
Masud, M.M., Khan, L., Thuraisingham, B., 2008. A scalable multi-level feature extraction technique to detect malicious executables. Inform. Syst. Front. 10, 33–45.
Miot, H.A., 2018. Correlation analysis in clinical and experimental studies. J. Vascular Brasileiro 17, 275–279.
Monteiro, P.J.M., Miller, S.A., Horvath, A., 2017. Towards sustainable concrete. Nat. Mater. 16, 698–699.
Öztürk, H., Özgür, A., Schwaller, P., Laino, T., Ozkirimli, E., 2020. Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discov. Today 25, 689–705.
Pezeshkian, M., Delnavaz, A., Delnavaz, M., 2021. Development of UHPC mixtures using natural zeolite and glass sand as replacements of silica fume and quartz sand. Eur. J. Environ. Civil Eng. 25, 2023–2038.
Pibiri, G.E., Venturini, R., 2019. Handling massive n-gram datasets efficiently. ACM Transa. Inform. Syst. (TOIS) 37, 1–41.
Reshamwala, A., Mishra, D., Pawar, P., 2013. Review on natural language processing. IRACST Eng. Sci. Technol. Int. J. (ESTIJ) 3, 113–116.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Techn. J. 27, 379–423.
Shi, C., Meyer, C., Behnood, A., 2008. Utilization of copper slag in cement and concrete. Resour. Conserv. Recycl. 52, 1115–1120.
Shi, Y., Long, G., Zeng, X., Xie, Y., Wang, H., 2021. Green ultra-high performance concrete with very low cement content. Construct. Build. Mater. 303, 124482.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Taylor, H.F.W., 1997. Cement Chemistry. Thomas Telford.
Tsai, S.-.T., Kuo, E.-.J., Tiwary, P., 2020. Learning molecular dynamics with simple language model built upon long short-term memory neural network. Nat. Commun. 11, 1–11.
Van, V.-T.-A., Rößler, C., Bui, D.-.D., Ludwig, H.-.M., 2014. Rice husk ash as both pozzolanic admixture and internal curing agent in ultra-high performance concrete. Cement Concrete Compos. 53, 270–278.
Wang, C., Liu, J., Luo, F., Deng, Z., Hu, Q.-.N., 2015. Predicting target-ligand interactions using protein ligand-binding site and ligand substructures. BMC Syst. Biol. 9, S2.
Wang, J.N., Yu, R., Ji, D.D., Tang, L.W., Yang, S.C., Fan, D.Q., Shui, Z.H., Leng, Y., Liu, K. N., 2022. Effect of distribution modulus (q) on the properties and microstructure development of a sustainable ultra-high performance concrete (UHPC). Cement Concrete Compos. 125, 104335.
Wang, T., Wu, K., Wu, M., 2020. Development of green binder systems based on flue gas desulfurization gypsum and fly ash incorporating slag or steel slag powders. Constr. Build. Materials 265, 120275.
Xu, X., Elgamal, M., Doddamani, M., Gupta, N., 2021. Tailoring composite materials for nonlinear viscoelastic properties using artificial neural networks. J. Compos. Mater. 55, 1547–1560.
Yu, R., Spiesz, P., Brouwers, H.J.H., 2015. Development of an eco-friendly Ultra-high performance concrete (UHPC) with efficient cement and mineral admixtures uses. Cement Concrete Compos. 55, 383–394.
Zhan, P., Xu, J., Wang, J., Jiang, C., 2021. Multi-scale study on synergistic effect of cement replacement by metakaolin and typical supplementary cementitious materials on properties of ultra-high performance concrete. Construct. Build. Mater. 307, 125082.
Zhang, L.V., Marani, A., Nehdi, M.L., 2022. Chemistry-informed machine learning prediction of compressive strength for alkali-activated materials. Construct. Build. Mater. 316, 126103.
Zhang, Y., Li, H., Abdelhady, A., Yang, J., Wang, H., 2021. Effects of specimen shape and size on the permeability and mechanical properties of porous concrete. Construct. Build. Mater. 266, 121074.
Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H., 2019. On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5745–5753.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Royal Statist. Soc. Series B (Statistical Methodology) 67, 301–320.