

Prediction of Compressive Strength of Geopolymers Using Multi-objective Feature Selection



Lasyamayee Garanayak, Sarat Kumar Das and Ranajeet Mohanty

Abstract To reduce the carbon dioxide emission to the environment, production of geopolymer is one of the effective binding materials to act as a substitute of cement. The strength of the geopolymer depends upon different factors such as chemical constituents, curing temperature, curing time, super plasticizer etc. In this paper, prediction models for compressive strength of geopolymer is presented using recently developed artificial intelligence techniques; multi-objective feature selection (MOFS), functional network (FN), multivariate adaptive regression spline (MARS) and multi gene genetic programming (MGGP). The MOFS is also used to find the subset of influential parameters responsible for the compressive strength of geopolymers. MOFS has been applied with artificial neural network (ANN) and non-dominated sorting genetic algorithm (NSGA II). The parameters considered for development of prediction models are curing time, NaOH concentration, $\text{Ca}(\text{OH})_2$ content, superplasticizer content, types of mold, types of geopolymer and $\text{H}_2\text{O}/\text{Na}_2\text{O}$ molar ratio. The developed AI models were compared in terms of different statistical parameters such as average absolute error, root mean square error correlation coefficient, Nash-Sutcliffe coefficient of efficiency.

Keywords Geopolymer • Compressive strength • Multi-objective feature selection • Artificial neural network • NSGA II • Multivariate adaptive regression spline • Genetic programming • Functional network

L. Garanayak (✉) • R. Mohanty
Civil Engineering Department, National Institute of Technology, Rourkela, India
e-mail: lizaoc@gmail.com

R. Mohanty
e-mail: ranajeetmohanty@gmail.com

S. K. Das
Civil Engineering Department, Indian Institute of Technology (ISM), Dhanbad, India
e-mail: saratdas@rediffmail.com; sarat@nitrrkl.ac.in

1 Introduction

Climatic change, one of the biggest global issues, primarily caused by elevated concentrations of carbon dioxide that increased from 280 to 370 ppm mainly due to industry resources [17]. Every ton of cement consumes 1.5 tons of raw materials i.e. limestone and sand [26] and 0.94 tons of carbon dioxide [22]. The world is attaining an uphill task in terms of sustainable development by using the waste industrial byproducts as an alternate resource of binder material infrastructure development and producing green environment without consumption of natural resources with low-energy, low-CO₂ binders [15]. After lime, ordinary Portland cement and its variants, geopolymer or Alkali-activated material (AAM) in general is considered as third generation cement. In stable geopolymer, source material should be highly amorphous consisting with sufficient reactive glass content and should consist of low water demand, which able to release the aluminum easily [28].

Strength of the geopolymer depends on different factors such as—solid solution ratio, curing temperature, curing time, chemical concentration, molar ratio of alkali solution, type of alkali solution, type of primary materials composed of silica and aluminium, type of admixtures and additives [23, 26]. There is a complex relationship between the compressive strength of the geopolymer with the above-discussed factors, particularly for different types of geopolymer. Hence, in order to achieve a desired compressive strength, it needs a trial and error approach to fix the above parameters, which is cumbersome and time-consuming.

Now a day, artificial intelligence (AI) techniques are found to be more efficient in the development of prediction models compared to traditional statistical methods [16, 30, 31]. Nazari and Torgal [19] used artificial neural network (ANN) for predicting the compressive strength of different types of geopolymers. They used the database of Pacheco-Torgal et al. [23–25] which contained different types of geopolymers obtained from waste materials based on different compositions containing aluminosilicate as an elementary source. First, set of dataset contained 180 data samples where the basic material used was tungsten mine waste [23], thermally treated at a temperature of 950 °C for 24 h. For mortar test authors used crushed sand as fine aggregates having a specific gravity of 2.7, fineness modulus of 2.8 and 0.9% water adsorption for 24 h. Sodium hydroxide (NaOH) solution was used as an activator by dissolving NaOH flakes in distilled water. The solution and extra water were added to the dry mix of sand, tungsten mine waste and calcium hydroxide (Ca(OH)₂) maintaining 4% as water to dry solid ratio. The compressive strength of 50 mm cube cured under ambient condition was per ASTM C109 [1], were determined. Output data collected was the compressive strength of the cube, which was the average of three specimens. The second set of dataset belongs to Pacheco-Torgal et al. [25], which contains total 144 data samples. Its geopolymer is also based on metakaoline, which was subjected to thermal treatment at 650 °C temperature. Fine aggregate of specific gravity 3, 1% water absorption, fineness modulus of 2.8 were used for mortar preparation. Different concentration of NaOH like 10M, 12M, 14M and 16M were developed by mixing of NaOH flakes with

distilled water and then mixed with sodium silicate solution of 1:2.5 as mass ratio. Mortar was added with admixtures like superplasticizer content of 1, 2, and 3%. The $\text{Ca}(\text{OH})_2$ was used as a replacement of metakaoline in the proportion of 5%, 10% while the mass ratio of sand to metakaoline to activator was kept as 2.2:1:1. The samples of $40 \times 40 \times 160 \text{ mm}^3$ prism specimens were obtained according to EN 1015-11, which was cast and cured at room temperature. And finally the rest amount of data was collected from Pacheco-Torgal et al. [24]. It had tungsten mine waste as base material activated by mixing of two alkali solutions like 24M NaOH and sodium silicate keeping 1:2.5 as mass ratio. 10% of $\text{Ca}(\text{OH})_2$ was used as percentage substitution of tungsten mine waste of $50 \times 50 \times 50 \text{ mm}^3$ cube sample was prepared by mixing the solution with a dry mix of sand, mine waste mud, and $\text{Ca}(\text{OH})_2$ with the ratio of mine waste mud to activator as 1:1. Extra water (7, 10%) was added to improve the workability of the mix. In it, water to dry solid binder was 3.6%. Compressive strength was obtained from those three papers which followed the ASTM C109 [1]. Nazari et al. [18] developed ANN models and found to better compare to other prediction models [19]. It may be mentioned here that the developed ANN model had two hidden layers with 12 and 10 number of neurons in hidden layer 1 and 2, respectively, hence number of parameters (weights and biases) were more and the model was not comprehensive. The ANN model also suffers from a lack of a comprehensive procedure for testing the robustness and generalization ability and attains local minima. In the recent decade, AI techniques such as genetic programming (GP), multivariate adaptive regression spline (MARS), functional network (FN) have shown very promising results in overcoming the above-discussed drawbacks.

Usually, in machine learning major portion of the data is used for training and a smaller portion for testing through random sampling of the data to ensure that the testing and training sets are similar for minimising the effects of data discrepancies and to better understand the characteristics of the model and also to limit problems like overfitting and to have an insight on how the model will generalize to an independent dataset [6]. Therefore, the model is trained for a number of times to reduce this effect.

Also in prediction type modelling identification of the controlling parameters is important, as the inclusion of all the features/parameters increases the complexity of the model with a small increase in predictive capability of the developed model. Thus, researchers are constantly looking for reliable predictive models which are not only low in complexity but also high in its predictive capability. One such algorithm, feature selection (FS) algorithm not only minimises the number of features but also maximises the predictive accuracy (minimisation of error) of the model. The above-described objectives are mutually conflicting in nature, a decrease in one result in an increase in the other. Therefore, multi-objective evolutionary algorithms (MOEA) can be implemented, which simultaneously minimises all the objective functions involved. Feature selection (FS) algorithm is of three types: wrapper, filter and embedded. In wrapper technique, a predictive model is used to evaluate each feature subsets. Each new subset is used to train a model and tested and then ranked based on their accuracy rate or error rate. In filter technique, a proxy measure is used which is fast to compute. Some of the measures

used in filter technique are mutual information [12], pointwise mutual information [34], Pearson product-moment correlation coefficient, inter/intra class distance or the scores of significance tests for each class/feature combinations [9, 34]. Filter selects a feature set, which is not tuned to a particular type of model thus resulting into be more general as compared to wrapper technique. Embedded technique uses a catch-all group method performing feature selection as a part of the modelling process. LASSO algorithm [2, 35] is one such technique where during linear modelling the regression coefficients are penalized with an L1 penalty, shrinking many of them to zero. In terms of computational complexity embedded technique is in between filters and wrappers. Implementation of evolutionary algorithms for FS has been made using differential evolution (DE) [13], genetic algorithms (GA) [36], genetic programming (GP) [21], and particle swarm optimisation (PSO) [5, 32, 33].

In this paper prediction models have been developed using multi-gene genetic programming (MGGP), MARS and FN to predict the compressive strength of geopolymers (alkali activated tungsten waste and metakaoline) based on the database available in the literature. A novel type of algorithm known as multi-objective feature selection (MOFS) is also implemented in this paper. In this proposed MOFS (wrapper type approach), artificial neural network (ANN) is combined with non-dominated sorting genetic algorithm (NSAG II) [8], where ANN acts as the learning algorithm and NSGA II performs the feature subset selection and minimises the errors for the developed AI model at the same time. By using three objectives for minimisation (a subset of features, training error, and testing error), a variant of MOEA (modified non-dominated sorting genetic algorithm or NSGA II) is applied to investigate if a subset of features exists with cent percent correct predictions for both training and testing datasets. The features fed to the MOFS are represented in binary form where 1 indicates selection of the feature and 0 indicates its non-selection. The performance of the AI model is evaluated in terms of mean square error; which NSGA II minimises during the multi-objective optimisation process.

2 Methodologies

In the present study artificial intelligence techniques, FN, MARS, and MGGP have been used for development of prediction models. As these techniques are not very common to professional engineering a brief introduction to the above techniques is presented as follows. The feature selection algorithm MOFS is also presented in this section.

2.1 Functional Network (FN)

Functional network (FN) proposed by Castillo et al. [3, 4] is a recent technique, which is being used as an alternate tool to ANN. In FN, network's preliminary topology is derived, centred, around the modelling properties of the real domain, or in other words, it is related to the problems of the domain knowledge, whereas in ANN, by the use of trial and error approach, the required number of hidden layers and neurons are determined, so that a good fitting model to the dataset can be obtained. After the availability of initial topology, functional equations are utilized to reach a much simpler topology. Therefore, functional networks eliminate the problems of artificial neural networks by utilizing together the data knowledge and the domain knowledge from, which the topology of the problem is derived. By the help of domain knowledge FN determines the network structure and from the data, it estimates the unknown neuron function. Initially, arbitrary neural functions are allocated with an assumption that the functions are of multi-argument type and vector valued in nature.

Functional networks can be classified into two types based on their learning methods. They are:

1. Structural learning: In this stage, the preliminary topology of the network is built on the assets obtainable to the designer and further simplifying is done by the help of functional equations.
2. Parametric learning: In this stage estimation of the neuron function is based on the combination of functional families, which is provided initially and then from the available data the associated parameters are estimated. It is similar to the estimation of the weights of the connections in artificial neural networks.

2.1.1 Working with Functional Networks

The main elements around which a functional network is built can be itemized as:

1. Storing Units
 - The inputs— $x_1, x_2, \text{ and } x_3 \dots$ require 1 input layer of storing unit.
 - The outputs— f_4, f_5, \dots require 1 output layer of storing unit.
 - Processing units containing 1 or several layers, which evaluates the input from the preceding layers to deliver it to the succeeding layer, f_6 .
2. Computing unit's layer, $f1, f2, f3$: In this computing unit, the neuron evaluates the inputs coming from the preceding layer to deliver the outputs to the succeeding layer.

3. Directed links set: Intermediary functions are not random in nature but it depends on the framework of the networks, such as $x_7 = f_4(x_4, x_5, x_6)$.

All the elements described above together form the functional network architecture. The network architecture defines the topology of the functional network and determines the functional capabilities of the network.

The steps for working with the functional network are as follows:

- Step1: Physical relationship of inputs with outputs.
 Step2: Preliminary topology of the functional network depends on the dataset of the problem. Artificial neural network selects the topology by trial and error approach, whereas functional networks select the topology on the properties of the data, which ultimately leads to a solo network structure.
 Step3: Functional equation simplifies the initial network structure of FN. It is done by constantly searching for a simpler network in comparison to the existing one, which will predict the same output from the same set of inputs. Once a simpler network is found the complicated network is replaced with the simpler one.
 Step4: A sole neuron function is selected for the specific topology, which yields a set of outputs.
 Step5: In this step data is collected for the training of the network.
 Step6: On the basis of the data, which is acquired from Step5, and a blend of the functional families, the neuron function is estimated. Learning stage of the network can be linear or non-linear in nature, which directly depends on the linearity of the neuron function.
 Step7: Once a model has been developed it is checked for error rate and also is validated with a different set of data.
 Step8: If the model is found to be satisfactory in the cross-validation process, it is prepared to be used.

In FN the learning method is selected on the basis of the neural function, which depends on the type of data $U = \{I_i, O_i\}$, $\{i = 1, 2, 3, 4, \dots, n\}$. Learning procedure involves minimisation of the Euclidean norm of the error function and it is represented as:

$$E = \frac{1}{2} \sum_{i=1}^m (O_i - F(i))^2 \quad (1)$$

Estimated neural functions $f_i(x)$ can be arranged in the following order:

$$f_i(x) = \sum_{j=1}^m a_{ij} \phi_{ij}(X) \quad (2)$$

where ϕ is the shape function, having algebraic expressions, exponential functions, and/or trigonometry functions. A set of linear or non-linear algebraic equations is obtained by the help of associative optimisation functions.

Previous information about the functional equation is vital for working with the functional network. The functional equation can be defined as a set of functions, unknown in nature, which excludes the integral and differential equations. Cauchy's functional equation is the most common instance for the functional equations and it is as follows:

$$f(x + y) = f(x) + f(y); \quad x, y \in R \quad (3)$$

For more details, readers can refer Das and Suman [7].

2.2 *Multivariate Adaptive Regression Splines (MARS)*

MARS correlates between a set of input variables to an output variable through adaptive regression method. In MARS, a non-linear, non-parametric approach is used to develop a prediction model without any prior assumption of any relationship between the input (independent variables) and the output (dependent variable). MARS algorithm creates these relationships by using sets of coefficient and basis functions from the dataset as discussed above. Due to this, MARS is favourable over other learning algorithms where the numbers of inputs (independent variables) are more in number i.e. high.

The backbone of MARS algorithm is founded on divide and conquer strategy in, which the dataset is split into a number of groups of piecewise linear segments known as splines, which varies in gradient. MARS is comprised of knots, which are basically the end points of splines and the functions (piece-wise linear function/ piece-wise cubic function) between these knots are called as basis function (BF). In this paper for the case of simplicity of the model, only piece-wise linear basis functions are used.

MARS algorithm proposed by Friedman [10] is a 2-step process to fit data's and is explained below:

- i. Forward stepwise algorithm: Basis functions are added in this step. First, the model is developed only by the help of an initial intercept known as β_o . Then in each successive step, a basis function, which shows the greatest decrease in the training error, is annexed. Like this, the whole operation is continued until the number of basis functions reaches its maximum value which has been predetermined beforehand. As a result, an over-fitted model is obtained. Searching of knots among all the variables are done by the adaptive regression algorithm
- ii. Backward pruning algorithm: Elimination of the over-fitting of data is done in this phase. The terms in the model are snipped (one by one removal of the terms) in this operation. The best viable sub model is obtained by removing the

least effective term. Then the subset of models is equated among themselves by means of the generalized cross-validation (GCV) process.

For a better understanding of MARS algorithm (refer [7]), examine a dataset, which contains an output y for a set of inputs $X = \{X_1, X_2, X_3, \dots, X_p\}$, which consists of p input variables. MARS generates a model of the form:

$$y = f(X_1, X_2, X_3, \dots, X_p) + e = f(X) + e \tag{4}$$

where, e = distribution of error; $f(x) = a$ function which is approximated by BFs (piece-wise linear function/piece-wise cubic function).

For the case of simplicity, only piece-wise linear functions have been discussed in this paper for its easy interpretability. The piece-wise linear function is represented as $\max(0, x - t)$ where t is the location of the knot. Its mathematical form is,

$$\max(0, x - t) = \{x - t, \text{ if } x > t \text{ or } 0, \text{ otherwise}\} \tag{5}$$

And finally, $f(x)$ = linear combination of BFs, and interactions between them is defined as,

$$f(X) = \beta_0 + \sum_{i=1}^M \beta_m \lambda_m(X) \tag{6}$$

where, λ_m = basis function, which is a single spline or product of 2 or more than 2 splines; β = coefficients of constant values calculated by least square method.

An illustration containing 22 data samples as inputs with an output is taken. Random numbers between one and twelve comprised the input matrix $\{X\}$ with a single output $\{Y\}$, which is obtained as per the equation, given below:

$$Y_i = \frac{1}{\sin(X_i)} - \frac{1}{\cos(X_i)} \tag{7}$$

Also, the data samples are normalized in the range of zero to one and MARS analysis is conducted. The MARS model developed for this dataset is represented as:

$$\hat{Y} = -0.143 + 4.066BF1 - 5.336BF2 + 1.852BF3 \tag{8}$$

where \hat{Y} = predicted values;

$$BF1 = \max(0, X_i - 0.40) \tag{9}$$

$$BF2 = \max(0, X_i - 0.65) \tag{10}$$

$$BF3 = \max(0, 0.65 - X_i) \tag{11}$$

In this MARS model the knots are situated at, $x = 0.65$ and $x = 0.40$. The R value for this model is 0.805. Proper care should be taken to use normalized values of X_i (Eqs. 9–11) and the denormalized values of the predicted Y_i can be obtained as per Eq. 12.

$$\hat{Y}_{denorm} = \hat{Y}_{norm} (X_{i(max)} - X_{i(min)}) + X_{i(min)} \tag{12}$$

Therefore, models developed using MARS algorithm has not only better efficiency but also simplifies the complex equations just like Eq. 7 to a simple linear equation.

2.3 Multi-gene Genetic Programming (MGGP)

Multi-gene genetic programming (MGGP) is a variation of GP where a model is built from the combination of several GP trees. Each tree is composed of genes, which represents a lower non-linear transformation of input variables. The output is created from a weighted linear combination of these genes and is termed as ‘multi-gene’. For an MGGP model, the model complexity and accuracy can be controlled by controlling the maximum depth of GP tree (d_{max}) and the maximum allowable number of genes (G_{max}). With the decrease in G_{max} and d_{max} values, the complexity of the MGGP model decreases, whereas its accuracy is hampered. Thus, there exist optimum values of G_{max} and d_{max} which gives fairly accurate results with the relatively compact model [27] for a given problem. The linear coefficients (c_1 and c_2) termed as weights of the gene and bias (c_0) of the model are got by ordinary least square method on the training data.

First, population initialization is done by creating a number of randomly evolved genes with lengths varying from 1 to G_{max} . Then, for each generation, a new population is chosen from the initial population as per their merit and then implementation of reproduction, followed by crossover, followed by mutation operations are performed on the function and terminal sets of the selected GP trees. In subsequent runs, the population is generated by addition and deletion of genes using traditional crossover mechanisms from GP and special MGGP crossover mechanisms. Few distinctive MGGP crossover mechanisms [27] are briefly described below.

2-Point High Level Crossover

The process of mating between two individual parents to swap genes between them is called as a 2-point high level crossover. Suppose there are 2 trees having four genes and three genes respectively marked by G_i to G_n . Assume that the G_{max} value

for the model is five. A crossover point, represented by $\{\dots\}$ is selected for each individual.

$$[G_1, \{G_2, G_3, G_4\}], [G_5, G_6, \{G_7\}]$$

Genes enclosing the crossover points are interchanged and thus, 2 new offspring are formed as shown below.

$$[G_1, \{G_7\}], [G_5, G_6, \{G_2, G_3, G_4\}]$$

The number of genes in any individual is not allowed to be more than G_{max} . But if it exceeds then, randomly genes are selected and eliminated till each individual has G_{max} genes. This process leads to the creation of fresh genes for both the individuals, as well as the deletion of some genes.

2-Point Low Level Crossover

Standard crossover of GP sub-trees in MGGP algorithm is known as 2-point low level crossover. First, a gene is arbitrarily chosen from each of the individuals and then exchanging of the sub-trees under the selected nodes is done. The newly created trees swap the parent trees in an otherwise unchanged individual in the subsequent generation. There are 6 types of mutations, which can be performed on this stage [11]. For achievement of best MGGP model probability of reproduction, crossover and mutation have to be given, such that the sum of the probability of these operations should not exceed 1.

2.4 Multi-objective Feature Selection (MOFS)

2.4.1 Non-dominated Sorting Genetic Algorithms (NSGA II)

NSGA II [8] is an elitist non-dominated sorting genetic algorithm and is very popular in the application of multi-objective optimisation. Not only does it adopt an elite preservation strategy but also uses the explicit diversity preservation technique. In this first the parent population is initialized, from which the offspring population is created and then both the population are combined and finally classified based on non-dominated sorting. After the completion of non-dominated sorting, filling of the new population starts with the best non-dominated front with the assignment of rank as 1 and this continues for successive fronts and assignment of ranks simultaneously. Along with the non-dominated sorting, another niching strategy adopted is the crowding distance sorting in which the distance reflects the closeness of a solution to its neighbours, greater the distance better is the diversity of the Pareto front. Offspring population is created from parent population by using crowded tournament selection, crossover and mutation operators and this whole operation continues until a termination criterion is met. More details of the algorithm can be found in Deb et al. [8].

2.4.2 NSGA II with ANN for Feature Selection

In this study to solve the feature selection problem wrapper type approach is implemented where binary chromosomes are used to represent the features with a value of 0 and 1, 0 indicating that the required feature is not selected and 1

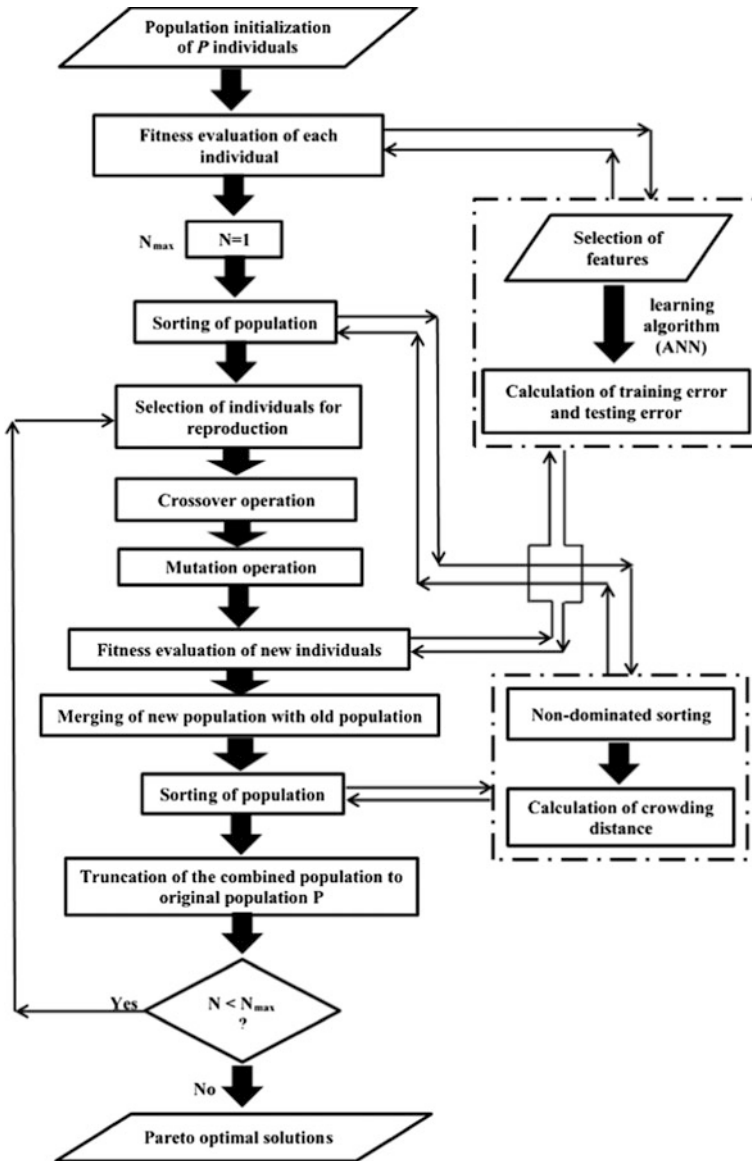


Fig. 1 Flowchart of MOFS algorithm

indicating that the required feature is selected. Three objectives are defined in the NSGA II algorithm, first being the minimisation of the number of selected features, second being the minimisation of training error rate and third being the minimisation of testing error rate in the learning algorithm. The training error and testing error are calculated based on mean square error. Learning algorithm used is feed-forward artificial neural networks (ANNs). Basic flowchart of the MOFS algorithm is presented in Fig. 1.

3 Database and Pre-processing

In this study, 384 data samples were taken with eight parameters from the literature Pacheco-Torgal et al. [23–25]. In those papers tungsten mine waste and metakaoline were used to develop geopolymer activating by alkali solution and extra admixture like superplasticizers as well as calcium hydroxide contents were added with different percentages. Seven variables i.e. curing time in days (T), percentage content of calcium hydroxide by weight (C), superplasticizer percentage by weight (S), NaOH concentration (N), mould type (M), type of geopolymer (G) and H_2O/Na_2O ratio (H) are taken as input parameters and compressive strength (Q_m) is the output. Table 1 presents the statistical values of the dataset used and Fig. 2 shows the variation of input and output parameters of the dataset.

It can be observed (Fig. 2) that when the curing time of the geopolymer is maximum, its compressive strength is minimum and vice versa. Also, with the addition of superplasticizers compressive strength increases but up to a certain extent. Also, when the molar ratio (H) and $Ca(OH)_2$ content is high, compressive strength is less.

The training (288 data samples) and testing (96 data samples) dataset were normalized between 0 and 1 for its implementation in FN and MGGP. For MGGP 500 was taken as the population size and 200 as maximum number of generation keeping 15 as tournament size. Crossover and mutation probability were considered as 0.84 and 0.14 respectively. For MARS modelling 70% data were used for

Table 1 Statistical values of the dataset

Variables	Range	Mean	Std. dev (σ)
Curing time (days) (T)	1.0–90.0	32.67	31.26
$Ca(OH)_2$ content (wt%) (C)	0.0–22.5	12.86	7.61
Superplasticizer (wt%) (S)	0.0–3.0	1.50	1.12
NaOH concentration (N)	6.0–24.0	12.86	5.54
Mold type (M)	1.0–2.0	1.50	0.50
Geopolymer type (G)	1.0–3.0	2.00	0.82
H_2O/Na_2O molar ratio (H)	8.9–19.1	14.63	2.86
Compressive strength (MPa) (Q_m)	1.5–79.0	30.81	16.85

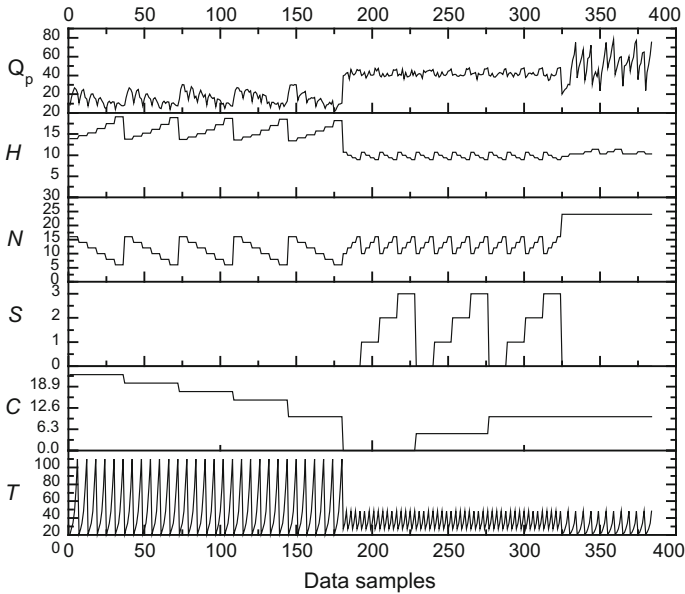


Fig. 2 Variation of input and output parameters of the dataset

training and rest 30% data were used for testing for the normalized values of the dataset in the range of 0–1.

And for the MOFS algorithm ANN training function used was Levenberg-Marquardt type consisting of 3 hidden neurons and performance of the neural network was based on MSE. 70% of the data samples were used for training and the remaining 30% for testing. Data were normalized in the range [0, 1]. In NSGA II uniform crossover technique was applied where replacement of the genetic material of the two selected parents takes place uniformly at several points. Conventional mutation operator was used on each bit separately and changing randomly its value. Parameters used in NSGA II were population size = 50, crossover probability = 0.95, mutation probability = 0.1 and mutation rate = 0.1.

4 Results and Discussion

Statistical comparisons of all the AI models developed in this study was done in terms of average absolute error (AAE), root mean square error (RMSE), correlation coefficient (R) and Nash-Sutcliffe coefficient of efficiency (E) and are presented in Table 5. Also, the overfitting ratio, which is the ratio between the RMSE of testing and training was found out and presented in Table 6. Overfitting ratio indicates the generalization of the prediction models. Cumulative probability of the developed models can be expressed as the ratio between the predicted compressive strength

(Q_p) to the measured compressive strength (Q_m) of the geopolymer. The ratio Q_p/Q_m are sorted in ascending order and its respective cumulative probability is found out from Eq. 13.

$$P = \frac{i}{n + 1} \tag{13}$$

where; i = order number for the respective Q_p/Q_m and n = total number of data samples. From the cumulative probability distribution (Fig. 8) P_{50} , the ratio of Q_p/Q_m corresponding to 50% probability and P_{90} corresponding to 90% probability are found out. For P_{50} less than one, under prediction is inferred and for greater than one over prediction is implied, with the best model being exactly equal to one. P_{50} and P_{90} values for all the four AI models are given in Table 6. Also, residual plots (residual error between the measured and the predicted values) of all the 4 AI models developed in this research has been presented in Figs. 4, 5, 6 and 7 for the testing dataset (performance on the testing dataset indicates the robustness and generalization capability of the prediction model). If the residuals appear to behave randomly (equally distributed on both sides of the zero line), it suggests that the model fits the data well otherwise it is a poorly fitted model.

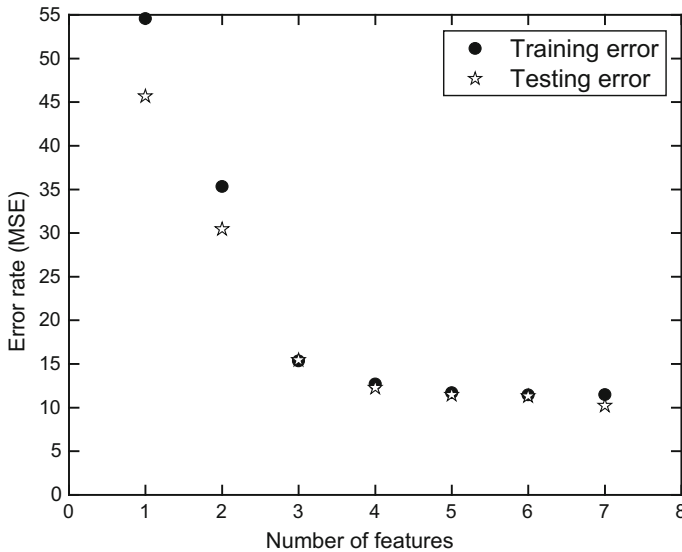


Fig. 3 Pareto optimal solutions

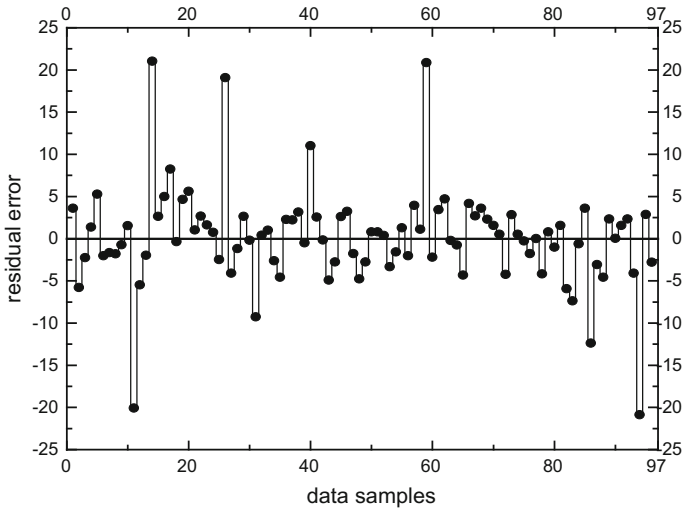


Fig. 4 Residual error of FN model (testing)

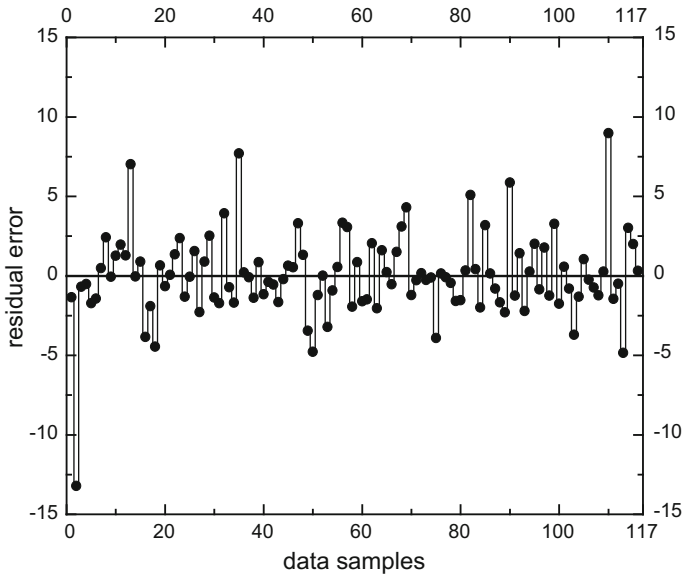


Fig. 5 Residual error of MARS model (testing)

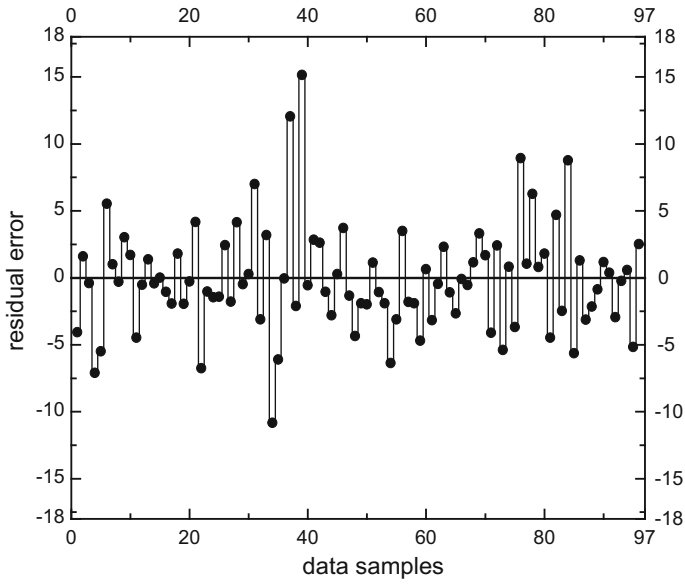


Fig. 6 Residual error of MGGP model (testing)

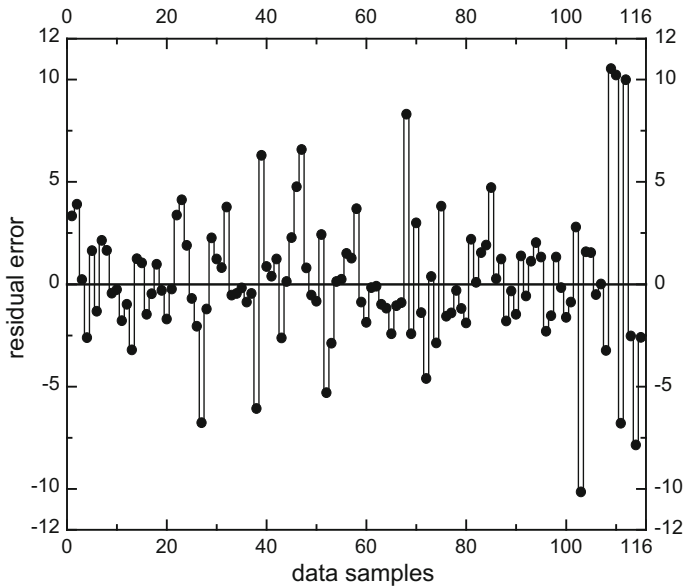


Fig. 7 Residual error of MOFS (ANN) model (testing)

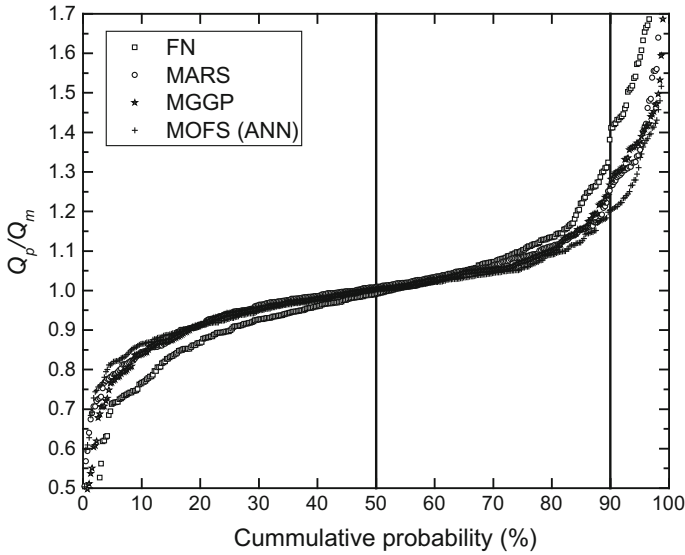


Fig. 8 Cumulative probability distribution of training and testing data

4.1 FN Model

FN models were developed from randomly selected 288 data samples, which were normalized in between 0 and 1. Its prediction value was obtained from the following equation.

$$y = a_0 + \sum_{i=1}^m \sum_{j=1}^m f_i(x_j) \tag{14}$$

where, n = no. of variables and m = degree of variables. The best model was found to be of associative type with 25 numbers of degree and tanh BF. As the degree of the model was very high, therefore, it was found to be unsuitable for developing a comprehensive model equation. Figure 4 shows the residual error plot between the measured and the predicted compressive strength of geopolymer for the testing data. It can be seen from Fig. 4 that the model fits well along with a maximum deviation of 20 MPa on both sides of the zero line. It can be seen from Table 5 that the values of R in training and testing are same i.e. 0.941, which indicates a strong correlation between predicted and observed values according to Smith [29]. Generally, R is a biased estimate for the prediction models [6]. So another indicator for the goodness of the model can be presented by the help of E. The values of E (Table 5) for training and testing are 0.885 and 0.885 respectively. RMSE and AAE for the FN model as shown in Table 5 are 5.669 MPa, 5.841 MPa and 3.867 MPa and 3.726 MPa for training and testing respectively. The overfitting ratio (Table 6)

for the FN model is 1.030, which indicates that the FN model developed is well generalized. Also, P_{50} and P_{90} (Fig. 8) as indicated in Table 6 are 0.992 and 1.411, implying that the model is slightly under predictive.

4.2 MARS Model

In MARS modelling, the best model was obtained corresponding to 11 basis functions and the equivalent model equation is given below.

$$\begin{aligned}
 Q_{p(n)} = & 0.49 + 0.5 \times BF1 - 0.61 \times BF2 - 3.09 \times BF3 - 0.33 \times BF4 \\
 & - 10.1 \times BF5 - 0.05 \times BF6 - 4.06 \times BF7 + 0.36 \times BF8 \\
 & - 2.45 \times BF9 + 0.74 \times BF10 + 2.22 \times BF11
 \end{aligned}
 \tag{15}$$

Details of the respective BF are presented in Table 2. De-normalized value of $Q_{p(n)}$ can be obtained from the following equation:

$$Q_p = Q_{p(n)}(78.38 - 1.55) + 1.55
 \tag{16}$$

The residual error plot of the MARS model for testing is shown in Fig. 5. It can be seen that the scatter of the error around the zero line is random with a maximum error of approx. -13 MPa from the measured value. The values of R and E in training and testing for the MARS model are 0.963; 0.926 and 0.988 and 0.975 respectively as indicated in Table 5. RMSE and AAE for training and testing (Table 5) are 4.602 MPa; 3.155 MPa and 2.639 MPa; 1.794 MPa respectively. From Table 6 it can be inferred that the MARS model is under-fitted (overfitting ratio = 0.573) and the developed MARS model is good for prediction as its P_{50} value is 1.009 which is nearly same as one.

4.3 MGGP Model

In MGGP model the modelling equation can be developed as follows:

Table 2 Details of the BFs for the MARS model

BF1	$\max(0, G - 0.5)$	BF7	$\max(0, 0.02 - T)$
BF2	$\max(0, 0.5 - G)$	BF8	$BF1 \times \max(0, 0.7 - H)$
BF3	$BF1 \times \max(0.15 - T)$	BF9	$\max(0, N - 0.22) \times \max(0, C - 0.89)$
BF4	$\max(0, 0.22 - N)$	BF10	$\max(0, N - 0.22) \times \max(0, H - 0.08)$
BF5	$BF1 \times \max(0.14 - H)$	BF11	$\max(0, N - 0.22) \times \max(0, 0.08 - H)$
BF6	$\max(0, T - 0.02)$		

$$\begin{aligned}
 Q_p(n) = & 18.8 \times N - 0.142 * T + 0.826 \times H - 19.4 * \exp(N \times H) \\
 & + 2.16 \times \exp(M) + 30.5 \times \exp(H) - 7.42 \times N \\
 & + 0.215 \times (T)^{1/4} - 41.7 \times H + 0.826 \times (H)^{1/2} \\
 & - 0.142 \times (H)^4 + 0.533 \times (T)^{1/2} \times (G)^4 - 25.9
 \end{aligned}
 \tag{17}$$

where the predicted value of Q_p can be obtained using Eq. 16. Figure 6 shows that the maximum error in prediction for the MGGP model is around 15 MPa on either side of the zero error line for the testing dataset. R between measured and predicted values of compressive strength for the MGGP model as per Table 5 is 0.979 and 0.976 respectively for training and testing. Also from Table 5, E, RMSE and AAE are given as 0.958, 3.405 MPa, 2.449 MPa and 0.950, 3.934 MPa, 2.868 MPa respectively for training and testing. The values of overfitting ratio, P₅₀ and P₉₀ for the MGGP model as indicated in Table 6 are 1.155, 1.004 and 1.281 respectively. MGGP model is also a good model for prediction as its P₅₀ value is close to unity.

4.4 MOFS (ANN) Model

Pareto optimal solutions given by MOFS algorithm are presented below with the number of input parameters used for modelling and the error rate of training and testing in terms of MSE. Results of the multi-objective optimisation are presented in Fig. 3 and the details of the Pareto front are given in Table 3. From Table 3 it can be inferred that the most influential features responsible for the compressive strength of the geopolymer are curing time (T) and molar ratio (H), as these two parameters are selected for a maximum number of times.

Table 3 Details of the Pareto front obtained from the MOFS (ANN) model

Selected features							MSE	
						H	Training	54.563
							Testing	45.658
T						H	Training	35.341
							Testing	30.439
T					G	H	Training	15.350
							Testing	15.431
T		S	N			H	Training	12.697
							Testing	12.260
T		S	N		G	H	Training	11.704
							Testing	11.441
T		S	N	M	G	H	Training	11.469
							Testing	11.303
T	C	S	N	M	G	H	Training	11.472
							Testing	10.205

Table 4 Connection weights and biases of the MOFS (ANN) model

Neuron (hidden)	Weights (w_{jk})								Biases	
	Input							Output	b_{hk}	b_0
	T	C	S	N	M	G	H	Q_p		
k_1	-0.318	-0.031	-19.694	-5.261	22.643	1.404	-9.094	736.668	5.164	-169.285
k_2	-0.320	-0.037	0.005	-5.273	1.447	1.410	-9.101	-736.554	5.173	-
k_3	14.642	-0.781	17.010	-0.775	0.966	0.164	0.557	169.499	4.507	-

Figure 3 clearly shows that MSE for training and testing decreases with increase in the number of input variables/features. Also, the difference between the training error and testing error which indicates the generalization of a model (small difference means more generalized is the model) is almost negligible when number of input features is 3 followed closely by when number of features selected is 6; ANN model developed in this paper is for 7 input features (details given in Table 3).

It is evident from Table 5 that model is well generalized as the R values for both training and testing are nearly same (0.981 and 0.979). Thus, the model developed has a good generalized fit between the independent (input parameters) and dependent variables (output). E, RMSE, and AAE of the MOFS (ANN) model are 0.962, 3.387 MPa, 2.395 MPa and 0.958, 3.195 MPa, 2.221 MPa for training and testing dataset respectively (Table 5). From Fig. 7 it can be observed that the MOFS (ANN) model is a good fit model with a maximum residual error of 10 MPa. The input weights, layer weights, and biases of the selected MOFS (ANN) model are given in Table 4. Based on the connection weights and biases (Table 4) of the MOFS (ANN) model, equation is formulated as follows:

$$A_1 = 5.164 - 0.318T - 0.031C - 19.694S - 5.261N + 22.643M + 1.404G - 9.094H \tag{18}$$

$$A_2 = 5.173 - 0.32T - 0.037C + 0.005S - 5.273N + 1.447M + 1.41G - 9.101H \tag{19}$$

Table 5 Statistical values of AI models used in this study

		FN	MARS	MGGP	MOFS (ANN)
R	Training	0.941	0.963	0.979	0.981
	Testing	0.941	0.988	0.976	0.979
RMSE (MPa)	Training	5.669	4.602	3.405	3.387
	Testing	5.841	2.639	3.934	3.195
E	Training	0.885	0.926	0.958	0.962
	Testing	0.885	0.975	0.950	0.958
AAE (MPa)	Training	3.867	3.155	2.449	2.395
	Testing	3.726	1.794	2.868	2.221

Table 6 Overfitting ratio and cumulative probability of the AI models

	Overfitting ratio	P ₅₀	P ₉₀
FN	1.030	0.992	1.411
MARS	0.573	1.009	1.264
MGGP	1.155	1.004	1.281
MOFS (ANN)	0.943	1.003	1.204

$$A_3 = 4.507 + 14.642T - 0.781C + 17.01S - 0.775N + 0.966M + 0.164G + 0.557H \tag{20}$$

$$Q_p = 76.83 \left[\frac{-169.285 + 736.668 \tanh(A_1) - 736.554 \tanh(A_2) + 169.499 \tanh(A_3)}{736.554 \tanh(A_2) + 169.499 \tanh(A_3)} \right] + 1.55 \tag{21}$$

The input values of the variables used in Eqs. 18–20 are normalized values in the range [0, 1]. Table 6 shows that the MOFS (ANN) model is slightly under fitted (overfitting ratio = 0.943) and the model is good in prediction as the P50 value is 1.003 (close to one).

Hence, it can be easily concluded that out of the 4 AI models developed, MOFS (ANN) model is best followed closely by MGGP model as indicated in the statistical comparison (Table 5). However, the model equation developed by MOFS (ANN) model is quite complex (not comprehensive), so for practical on field use MGGP model equation can be utilized, but again in the MGGP model not all the parameters of the geopolymer are used (Ca(OH)₂ and superplasticizer content are absent). Thus, it really depends on the user on the choice of AI model to be used. Also from Table 6 all the AI models are good in prediction except the FN model (graphical representation is given in Fig. 8).

5 Conclusion

The present study deals with the compressive strength of geopolymers based on the experimental database available in the literature using different AI methods. Identification of the subset of features responsible for the predictive capacity of the model is addressed here by considering it as a multi-objective optimization problem. Based on different statistical parameters like R, E, RMSE and AAE values, MOFS (ANN) algorithm is found to be more efficient as compared to other AI techniques. The R, E, RMSE and AAE values of the present ANN model, are 0.981, 0.962, 3.387 MPa and 2.395 MPa, respectively, for training and 0.979, 0.958, 3.195 MPa and 2.221 MPa, respectively, for testing data. The model equations are also presented, which can be used by quality control professional engineers to identify the proper proportion of different constituent and the condition of different curing etc. for a desired compressive strength. It was observed that though, the model equation as per the MGGP model is comprehensive, but out of seven parameters of the geopolymer, two important parameters (Ca(OH)₂ and

superplasticizer content) are not part of the model equation. But, the MOFS (ANN) model is best and though the model equation is not comprehensive, but the model equation is presented in a tabular form. The model equation will help the professional engineers particularly at the initial level to predict the compressive strength of geopolymer, which is a very complex phenomenon.

References

1. ASTM. (2013). International Standard Test Method for Compressive Strength of Hydraulic Cement Mortars (Using 2-in. or [50-mm] Cube Specimens). (ASTM C109/C109M) West Conshohocken, PA 19428-2959. United States.
2. Bach, F. R. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In A. McCallum & S. T. Roweis (Eds.), *Proceedings of 25th International Conference on Machine Learning, (ICML2008)*, Helsinki, Finland (pp. 33–400).
3. Castillo, E., Cobo, A., Gutierrez, J. M., & Pruneda, E. (1998). *An introduction to functional networks with applications*. Boston: Kluwer.
4. Castillo, E., Cobo, A., Manuel, J., Gutierrez, J. M., & Pruneda, E. (2000). Functional networks: A new network-based methodology. *Computer-Aided Civil and Infrastructure Engineering*, 15, 90–106.
5. Cervante, L., Xue, B., Zhang, M., & Shang, L. (2012). Binary particle swarm optimisation for feature selection: A filter based approach. In *Proceedings of Evolutionary Computation (CEC), 2012 IEEE Congress*, Brisbane, QLD (art. no. 6256452, pp. 881–888).
6. Das, S. K. (2013). Artificial neural networks in geotechnical engineering: Modeling and application issues, Chapter 10. In X. Yang, A. H. Gandomi, S. Talatahari & A. H. Alavi (Eds.), *Metaheuristics in water, geotechnical and transport engineering* (pp. 231–270). London: Elsevier.
7. Das, S. K., & Suman, S. (2015). Prediction of lateral load capacity of pile in clay using multivariate adaptive regression spline and functional network. *The Arabian Journal for Science and Engineering*, 40(6), 1565–1578.
8. Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197.
9. Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289–1305.
10. Friedman, J. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 19, 1–141.
11. Gandomi, A. H., & Alavi, A. H. (2012). A new multi-gene genetic programming approach to nonlinear system modeling. Part II: Geotechnical and Earthquake Engineering Problems. *Neural Computing and Applications*, 21(1), 189–201.
12. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3, 1157–1182.
13. He, X., Zhang, Q., Sun, N., & Dong, Y. (2009). Feature selection with discrete binary differential evolution. In *Proceedings of International Conference on Artificial Intelligence and Computational Intelligence, AICI 2009*, Shanghai (Vol. 4, art. no. 5376334, pp. 327–330).
14. <http://www.geopolymer.org/faq/alkali-activated-materials-geopolymers/>.
15. Juenger, M. C. G., Winnefeld, F., Provis, J. L., & Ideker, J. H. (2011). Advances in alternative cementitious binders. *Cement and Concrete Research*, 41, 1232–1243.

16. Kutylowska, M. (2016). Comparison of two types of artificial neural networks for predicting failure frequency of water conduits. *Periodica Polytechnica Civil Engineering*, <https://doi.org/10.3311/ppci.8737>.
17. Mehta, P. K. (2004). High-performance, high-volume fly ash concrete for sustainable development. In *Proceedings of the International Workshop on Sustainable Development and Concrete Technology*, Beijing, China (pp. 3–14).
18. Nazari, A., Hajjiallahyari, H., Rahimi, A., Khanmohammadi, H., & Amini, M. (2012). Prediction compressive strength of Portland cement-based geopolymers by artificial neural networks. *Neural Computing and Applications*, 1–9.
19. Nazari, A., & Pacheco-Torgal, F. (2013). Predicting compressive strength of different geopolymers by artificial neural networks. *Ceramics International*, 39, 2247–2257.
20. Nazari, A., & Riahi, S. (2012). Prediction of the effects of nanoparticles on early-age compressive strength of ash-based geopolymers by fuzzy logic. *International Journal of Damage Mechanics*, 22(2), 247–267.
21. Neshatian, K., & Zhang, M. (2009). Pareto front feature selection: Using genetic programming to explore feature space. In *Proceedings of 11th Annual conference on Genetic and Evolutionary Computation, GECCO'09* (pp. 1027–1034). New York, NY, USA: ACM.
22. Pacheco-Torgal, F., Abdollahnejad, Z., Camões, A. F., Jamshidi, M., & Ding, Y. (2012). Durability of alkali-activated binders: A clear advantage over Portland cement or an unproven issue? *Construction and Building Materials*, 30, 400–405.
23. Pacheco-Torgal, F., Castro-Gomes, J., & Jalali, S. (2008). Alkali-activated binders: A review. Part 2. About materials and binders manufacture. *Construction and Building Materials*, 22(7), 1315–1322.
24. Pacheco-Torgal, F., Castro-Gomes, J., & Jalali, S. (2007). Investigations about the effect of aggregates on strength and microstructure of geopolymeric mine waste mud binders. *Cement and Concrete Research*, 37, 933–941.
25. Pacheco-Torgal, F., Moura, D., Ding, Y., & Jalali, S. (2011). Composition, strength and workability of alkali-activated metakaolin based mortars. *Construction and Building Materials*, 25, 3732–3745.
26. Rashad, A. M. (2014). A comprehensive overview about the influence of different admixtures and additives on the properties of alkali-activated fly ash. *Materials and Design*, 53, 1005–1025.
27. Searson, D. P., Leahy, D. E., & Willis, M. J. (2010). GPTIPS: An open source genetic programming toolbox from multi-gene symbolic regression. In *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Hong Kong (Vol. 1, no. 3, pp. 77–80).
28. Singh, B., Ishwarya, G., Gupta, M., & Bhattacharyya, S. K. (2015). Geopolymer concrete: A review of some recent developments. *Construction and Building Materials*, 85, 78–90.
29. Smith, G. N. (1986). *Probability and statistics in civil engineering: An introduction*. London: Collins.
30. Tarawneh, B., & Nazzal, M. D. (2014). Optimization of resilient modulus prediction from FWD results using artificial neural network. *Periodica Polytechnica Civil Engineering*, 58(2), 143–154. <https://doi.org/10.3311/ppci.2201>.
31. Ünès, F., Demirci, M., & Kisi, Ö. (2015). Prediction of Millers Ferry Dam reservoir level in USA using artificial neural network. *Periodica Polytechnica Civil Engineering*, 59(3), 309–318. <https://doi.org/10.3311/ppci.7379>.
32. Xue, B., Cervante, L., Shang, L., Browne, W. N., & Zhang, M. (2012). A multi-objective particle swarm optimisation for filter based feature selection in classification problems. *Connection Science*, 24(2–3), 91–116.
33. Xue, B., Cervante, L., Shang, L., Browne, W. N., & Zhang, M. (2014). Binary PSO and rough set theory for feature selection: A multi-objective filter based approach. *International Journal of Computational Intelligence and Applications*, 13(2), art. no. 1450009.

34. Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of Fourteenth International Conference on Machine Learning (ICML'97)* (Vol. 97, pp. 412–420), Nashville, Tennessee, USA.
35. Zare, H., Haffari, G., Gupta, A., & Brinkman, R. R. (2013). Scoring relevancy of features based on combinatorial analysis of Lasso with application to lymphoma diagnosis. *BMC Genomics*, 14, art. no. S14.
36. Zhu, Z., Ong, Y. S., & Dash, M. (2007). Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, 37(1), 70–76.