



OPEN Soft-computing models for predicting plastic viscosity and interface yield stress of fresh concrete

Waleed Bin Inqiad^{1,2}, Muhammad Faisal Javed^{3,4}✉, Deema Mohammed Alsekait⁵✉, Naseer Muhammad Khan^{6,7}, Majid Khan¹✉, Fahid Aslam⁸ & Daa Salama Abd Elminaam^{9,10}

Interface yield stress and plastic viscosity of fresh concrete significantly influences its pumping ability. The accurate determination of these properties needs extensive testing on-site which results in time and resource wastage. Thus, to speed up the process of accurately determining these concrete properties, this study tends to use four machine learning (ML) algorithms including Random Forest Regression (RFR), Gene Expression Programming (GEP), K-nearest Neighbor (KNN), Extreme Gradient Boosting (XGB) and a statistical technique Multi Linear Regression (MLR) to develop predictive models for plastic viscosity and interface yield stress of concrete. Out of all employed algorithms, only GEP expressed its output in the form of an empirical equation. The models were developed using data from published literature having six input parameters including cement, water, time after mixing etc. and two output parameters i.e., plastic viscosity and interface yield stress. The performance of the developed algorithms was assessed using several error metrics, k-fold validation, and residual assessment etc. The comparison of results revealed that XGB is the most accurate algorithm to predict plastic viscosity (training $R^2 = 0.959$, testing $R^2 = 0.947$) and interface yield stress (training $R^2 = 0.925$, testing $R^2 = 0.965$). To get increased insights into the model prediction process, shapely and individual conditional expectation analyses were carried out on the XGB algorithm which highlighted that water, cement, and time after mixing are the most influential parameters to estimate both fresh properties of concrete. In addition, a graphical user interface has been made to efficiently implement the findings of this study in the civil engineering industry.

Keywords Machine learning, Interface yield stress, Gene expression programming, Plastic viscosity, Shapley additive explanatory analysis

Abbreviations

GEP	Gene Expression programming
XGB	Extreme Gradient Boosting
RFR	Random Forest Regression
KNN	K-nearest neighbour
MLR	Multi Linear Regression
AI	Artificial Intelligence
ML	Machine Learning

¹School of Civil and Resources Engineering, University of Science and Technology, 17, Beijing 100083, China.

²Military College of Engineering (MCE), National University of Science and Technology (NUST), Islamabad 44000, Pakistan. ³Department of Civil Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Swabi 23640, Pakistan. ⁴Western Caspian University, Baku, Azerbaijan. ⁵Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. ⁶Department of Sustainable Advanced Geomechanical Engineering, Military College of Engineering (MCE), National University of Science and Technology (NUST), Islamabad 44000, Pakistan. ⁷MEU Research Unit, Middle East University, Amman 11831, Jordan. ⁸Department of Civil Engineering, College of Engineering in Al-Kharj, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia. ⁹Faculty of Computers and Artificial Intelligence, Benha University Egypt, Banha, Egypt. ¹⁰Jadara Research Center, Jadara University, Irbid 21110, Jordan. ✉email: arbabfaisal@giki.edu.pk; Dmalsekait@pnu.edu.sa; majid@ustb.edu.cn

PI	Performance Index
MAE	Mean Absolute Error
P	Pressure
OF	Objective Function
Q	Flowrate
GP	Genetic Programming
SHAP	Shapely Additive Explanatory Analysis
GUI	Graphical user Interface
ICE	Individual Conditional Expectation
R	Coefficient of Correlation
DT	Decision Tree
R ²	Coefficient of Determination
RMSE	Root Mean Squared Error
ANN	Artificial Neural Networks

In the construction industry, the utmost attention is usually given to the hardened properties of concrete specifically compressive strength. However, the procedure of transporting freshly made concrete from ground to the intended location in the construction of dams, high rise buildings, and long-span bridges etc. is very challenging and requires due attention^{1–3}. Although, the concrete is widely transported to great distances and huge vertical heights by means of a concrete pump in a steel pipe, the problem of blockage of pump and pipes and segregation of concrete while transportations remain a serious issue. These issues not only result in degradation of concrete quality, but also cause loss of time and resources on the worksite^{4–6}.

The ability of a concrete to easily pass through a pipe without altering its fresh and hardened properties using pump is known as pumpability of concrete. There are mainly two types of studies related to this, one those who focus on pump and pipe configuration and those who focus on the concrete fresh properties itself. Since, for a given worksite setting, one doesn't have much choice in terms of pump and pipe related properties like pump capacity, pipe diameter, pipe length etc. Thus, the study and optimization of fresh concrete properties plays a key role in achieving the desired outcomes^{7–9}. To quote an example, in the construction of world's tallest building in Dubai, the Burj Khalifa, to ensure a smooth and single stage pumping of concrete over 600 m without blockage, the engineers performed extensive tests to finalize a concrete mixture to achieve the desired properties^{10,11}. Similarly, the hurdles faced while constructing Head Race Tunnel project in India had been described by Hazaree and Mahadevan¹². The project required concrete lining up to a distance of 6.3 km which was successfully achieved by extensive testing and meticulous planning done by the engineers. The crucial processes of material selection, equipment and pipeline selection were done by careful evaluation of priming and pumping methodologies and thorough pre-testing before the actual concrete pumping. The plastic viscosity of concrete is of particular importance in this regard. It is a crucial rheological property of concrete to consider in situations where concrete is subjected to shear forces like pumping, casting, and placement etc. It is an indication of how easily the concrete can be mixed, transported, and placed without loss of quality (bleeding & segregation). It is also important to indicate the ease with which concrete passes through the pump without blockage or clogging. Similarly, interface yield stress of concrete is a crucial parameter which that determines the strength of bond between concrete and the reinforcement bars^{13–15}.

Previous studies regarding estimation of fresh concrete properties

The flow profile of concrete inside a pipe along with three layers named as the lubrication layer at the interface between the pipe and concrete, a shear induced region where concrete is under shear and a plug flow zone where concrete flows under translational motion^{16–18} are shown in Fig. 1. In previous studies, the calculation of concrete pumpability is related to pumping pressure (P) and flow rate (Q) of a concrete mixture. Jang et al.¹⁹ introduced a bi-linear mode to estimate the P-Q relationship. The first assumes that only a boundary layer of small thickness is subjected to shear while the second part considers partial shear during pumping. Scholars have also devised different equations to compute pumping pressure^{19–21} for the purpose of determining concrete pumpability which are given by Eqs. (1) and (2). In below equations, η is viscous constant of the lubricating layer measured by tribometer ($Pa.s/m$), τ_{io} is the interface yield stress (Pa), while τ_o is the yield stress and μ is plastic viscosity ($Pa.s$) of the concrete measured with the help of a viscometer. Also, k represents pipe filling

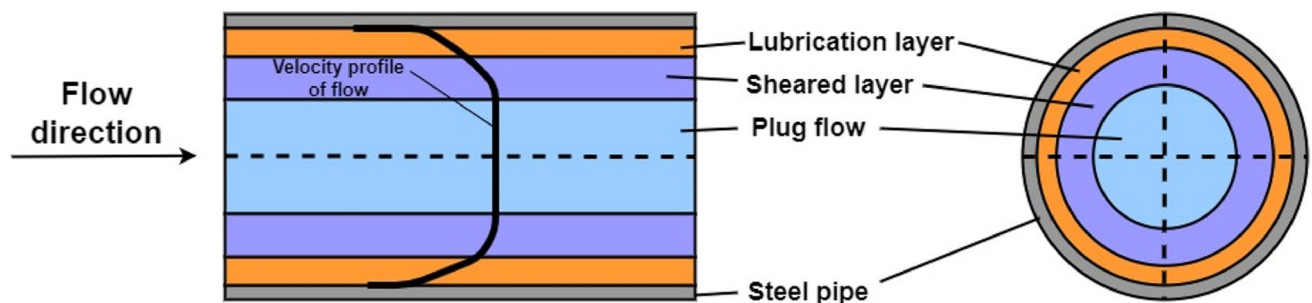


Figure 1. Schematic diagram to show flow of fresh concrete inside a pipe²³.

coefficient while R and L are the radius and length of pipe respectively measured in meters. Equation (1) gives the concrete movement to where only a small boundary layer is subjected to shear and Eq. (2) is applicable when bulk concrete is partially sheared during pumping. The concrete pumping pressure is estimated with the help of its rheological and tribological properties when the pumping configuration, concrete flow rate and pipe dimensions are pre-defined²². Thus, predicting plastic viscosity and interface yield stress is crucial to obtaining an optimal pumping pressure.

$$P = \frac{2L}{R} \left(\tau_{io} + \frac{Q \times \eta}{3600 \pi R^2 k} \right) \quad (1)$$

$$P = \frac{2L}{R} \left[\tau_{io} + \left(\frac{Q}{3600 \pi R^2 k} - \frac{R}{4\mu} \tau_{io} + \frac{R}{3\mu} \tau_o \right) \times \frac{\eta}{(1 + \frac{R}{4\mu} \eta)} \right] \quad (2)$$

In another study conducted by Choi et al.²⁴ the authors calculated flow rate by integration of velocity and radius. The developed models utilized rheological properties of each layer, and the four aforementioned parameters were determined via experimentation. More recently, several scholars suggested using a portable tribometer to determine τ_{io} and η ^{15,25–31}. Similarly, He et al.³² studied the influence of nano titanium oxide on both porous and normal concrete and the authors concluded that the micromechanics approach might be feasible for predicting mechanical concrete properties. However, among these previously developed models, the model developed by Ngo et al.³³ is valid only for normal concrete. In practice, it has been identified that for normal concrete, interfacial properties are crucial to determine pumping pressure^{34–37}. So, the interfacial properties which characterize the lubrication layer properties, have been the subject of interest for many researchers. For example, Ngo et al.³³ conducted experiments in which they estimated variation in pumping parameters with change in concrete compositions including variation in superplasticizer, water cement ratio and paste volume etc. The authors concluded that concrete composition affects the lubrication layer formation and concrete tribological properties. Also, Dong et al.³⁸ documented that yield stress generally decreases with increase in finer content of concrete. Similarly, T. Ngo (2009) concluded that increase in cement paste volume results in a linear decrease in viscosity and a minimal decrease in interface yield stress³⁹. Moreover, the variation in water-cement ratio also affects plastic viscosity and yield stress in the same fashion as cement paste volume. The increase in gravel to sand ratio also causes a decrease in viscosity^{40–43}. According to Choi et al.⁴⁴ both under question parameter also varied greatly with coarse aggregate size, and plastic viscosity exhibited a linear increasing trend with increase in coarse aggregate size.

Concrete plastic viscosity and yield stress are commonly measured by tribometers^{38,45,46}. Based on the findings from these tribometers, researchers have attempted to develop empirical prediction models. Ferraris and Larrard⁴⁷ developed a model to investigate the relationship between properties of finer mixtures and mortar parameters. Ngo et al.³³ studied fresh concrete properties in relation to lubrication layer thickness and ratio of solid concentration to maximum compactness of the granular mixture. But all these models are constructed with respect to specific grading of the aggregates and concrete additives, so their application is greatly limited. In addition, there are very few works in the literature about prediction of plastic viscosity and interface yield stress of concrete. Thus, to ensure smooth concrete pumping, it is important to accurately estimate yield stress and plastic viscosity of concrete mixes as accurately as possible based on the mixture composition.

The optimization of concrete mixes to be used for pumping and transportation through pipes also requires the paste-distance correlations to be investigated. But these tests consume significant time and resources. The rheological properties of concrete mixes depend upon its composition and are inherently complex to understand^{48–50}. Also, the properties of concrete mixes in fresh state are time dependent. Therefore, it is imperative to have a thorough understanding of concrete properties during the fresh state to make the concrete pumping process smooth.

Artificial Intelligence (AI) in civil engineering

Since there is a transition to artificial intelligence-based systems in almost every industry, civil and materials industry is also making its transition towards intelligence-based designs and methodologies driven by data. AI in the realm of civil engineering mainly refers to amalgamation of advanced machine learning (ML) and deep learning techniques to optimize material usage, foster sustainable construction practices and intelligent design of infrastructure to improve the overall efficiency and accuracy of construction industry. ML is a subtype of artificial intelligence that enables computers to recognize patterns from the data and use that learning to make accurate predictions for future unseen data without explicit human intervention⁵¹. Machine learning tools have emerged as an effective way to make decisions based on learnings from the data. The ML techniques understand the complex and hidden relationships between independent and dependent variable(s) in the data obtained from experimental investigations. These learnings from the data are then used by professionals like engineers etc. to make accurate decisions without going into the timely and costly testing procedures. In the technologically advanced world of today, ML techniques are seen as a superior alternative of traditional statistical approaches and are extensively used in modelling complex material properties^{52–64}. Thus, there is a need to apply the robust and widely used ML models to predict fresh concrete properties to solve the aforementioned problems in transportation of concrete over long distances. These ML-based models developed using the generalized data from past studies can be used as reliable tools to predict fresh properties and their use not only results in time and cost savings, but also increases accuracy in estimating rheological concrete properties.

Research significance

From the above discussion, it is evident that a purely experimental approach to assess different rheological properties like plastic viscosity and interface yield stress requires extensive resources and is not the most

feasible option given the diverse site conditions. In past, there have been several attempts to develop empirical relationships to obtain concrete plastic constant using statistical techniques^{13,39,65}. However, these statistical models have limited applicability and are sensitive to the material properties of concrete constituents etc. Also, it is evident from the overview of relevant literature that are few analytical models available for determining plastic viscosity and interface yield stress of concrete but there is a need for exploration on the subject using advanced and accurate ML techniques to predict these properties. The data driven models based on the already available experimental data can help greatly in this regard to accurately predict rheological concrete properties thus eliminating the need for extensive testing and increasing the productivity of the construction industry.

The employment of advance ML techniques will help the professionals to effectively calculate these values without carrying out extensive testing. Thus, this study intends to apply random forest regression (RFR), K-nearest Neighbor (KNN), Extreme gradient Boosting (XGB), Gene Expression Programming (GEP) and also a statistical analysis technique called Multi Linear Regression (MLR) to predict the aforementioned properties of concrete in fresh state and contribute to the literature about predicting plastic viscosity and interface yield stress of concrete mixes using different ML algorithms because the existing literature on the subject is extremely limited. The advantage of using evolutionary technique like GEP alongside other algorithms is that it gives the output in the form of an empirical equation unlike other employed algorithms. This property makes GEP a grey-box model as compared to other black-box model algorithms⁶⁶. GEP provides increased transparency and insights into the prediction process in the form of an empirical equation that relates the predicted parameter to the input parameters. It also offers the advantage of not requiring the pre-optimization of model architecture like Artificial neural networks etc., thus requires less computing power and time to complete the model training and prediction process⁶⁷. Similarly, the tree-based algorithms (RFR and XGB) are known for their accuracy and in giving insights about the importance of different input features. Also, the use of MLR can serve as a way to compare the accuracy of newly developed ML algorithms to predict rheological concrete parameters compared to the conventional statistical approach. The particular significance of this study lies in the fact that it draws a comparison between GEP, tree-based algorithms, KNN, and MLR to highlight their strengths and weaknesses for predicting important rheological properties of concrete. Also, this study intends to utilize a novel explanatory analysis technique (SHAP) analysis on the algorithm of highest accuracy to highlight the importance of different input features in predicting the two outputs. This thing has not been explored in the literature to the best of author's knowledge.

Research methodology

The identification of research problem and collection of data regarding that problem are the most important steps. Thus, the data was collected from literature to be used for training and testing the algorithms. The collected data underwent several statistical analyses to reveal more information about the distribution and correlation of the data. After that, in Step 3, four ML algorithms including GEP, RFR, XGB, KNN and one statistical regression technique i.e., MLR were employed to predict the two concrete properties. After algorithm development, they were validated using several techniques including k-fold cross validation, error assessment, and residual distribution. After the models were validated, they were compared to point out a single accurate model which will be used for further analysis. In step 4, the most accurate model was used for conducting contemporary explanatory analysis techniques including SHAP and individual conditional expectation (ICE) analysis. Finally, a graphical user interface (GUI) has been developed to help implement the findings of this study in the practical settings. The schematic representation of the current study's methodology is given in Fig. 2 while the explanation of the working mechanism of the employed algorithms is given in the subsequent sections.

Gene expression programming (GEP)

GEP is an AI technique which uses evolutionary algorithms to find solution to a problem. It is an extension of Genetic Programming (GP) and based on Charles Darwin's principles of natural selection. GEP finds solution to the problem using genes and performing evolutionary rules on these genes⁶⁷. The genes are build using various mathematical functions called primitive functions. These functions range from simple arithmetic functions like addition, subtraction to more complex like square root, sine, cosine etc⁶⁸. These genes which can be expressed as expression trees represent a small piece of code and are combined in various ways using crossover and mutation to create a full program which represents the solution to the given problem⁶⁹. Crossover involves combining two expressions to develop a new one and mutation means changing an existing expression⁷⁰. These processes in GEP are similar to the genetic mutation and crossover in humans. The pictorial representation of the processes of crossover and mutation is given in Fig. 3.

The prediction process starts with the formation of a random population of computer programs called chromosomes. These chromosomes are evaluated for their ability to solve the given problem using a pre-defined fitness function. The chromosomes with good performance based on the fitness function are selected for the next generation while the worst performing ones are deleted from the population⁷². This process repeats several times until a population of chromosomes with the highest accuracy is reached⁷³. The flowchart of the overall process is given in Fig. 4.

K-nearest neighbor (KNN)

KNN is a widely used ML technique which can solve both classification and regression problems⁷⁴. The concept of KNN is based on the principle that if the k closest neighboring samples to a given sample in the space belong to a specific category, then the given sample should also belong to that category. In other words, the KNN algorithm is used to decide the category of an unknown sample by looking at the category of its k closest neighboring samples. For solving a regression problem, the KNN algorithm computes the distance between an input unknown sample and all other samples present in the dataset and selects the first k samples with the

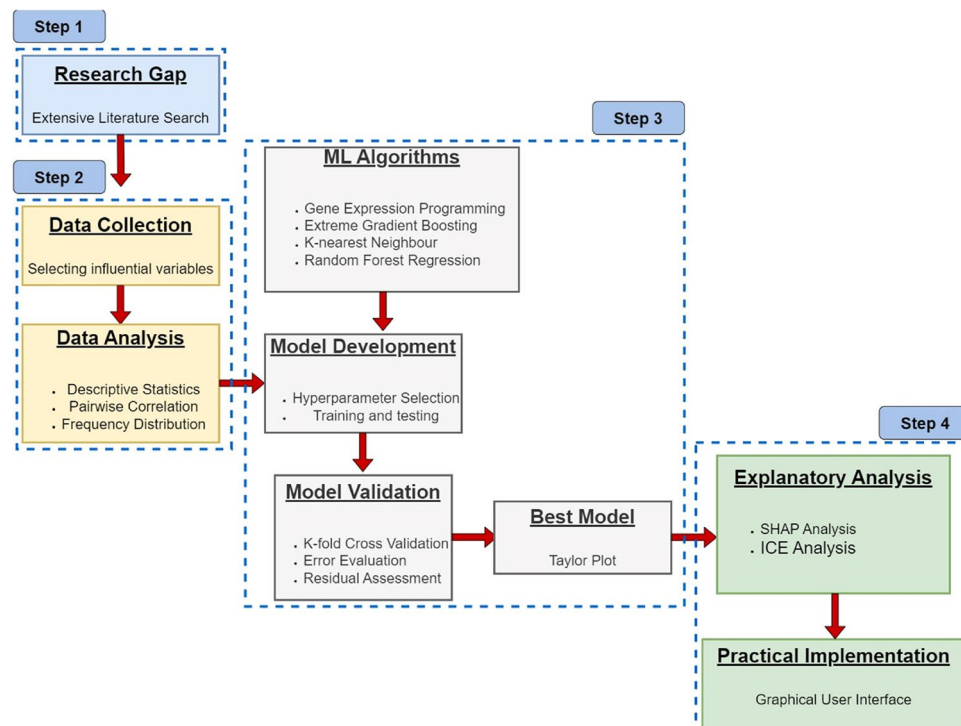


Figure 2. Overall research methodology of the study.

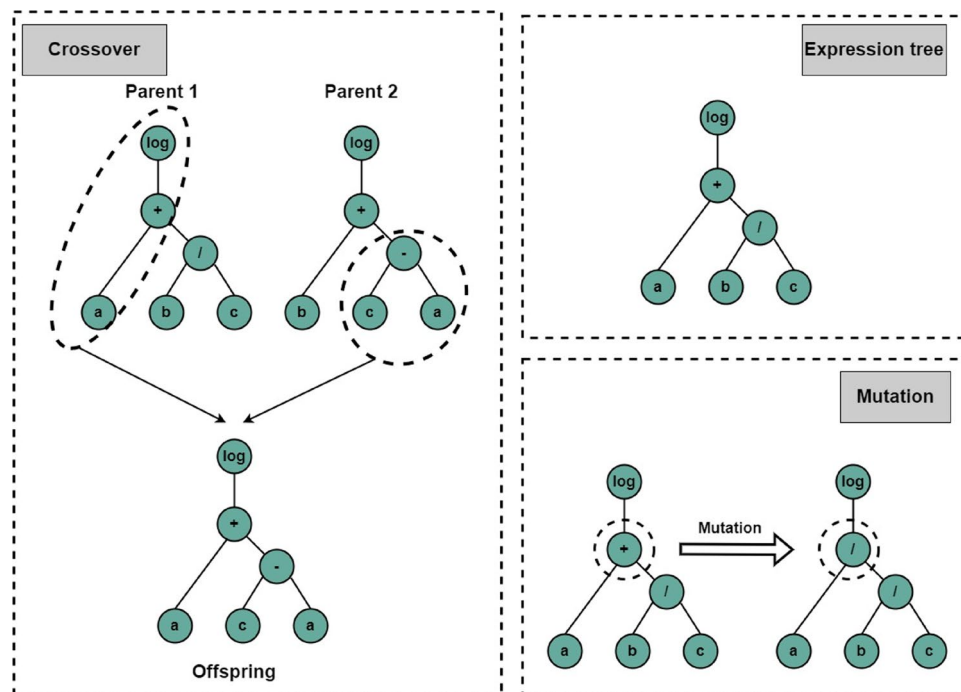


Figure 3. Representation of expression tree and genetic processes⁷¹.

shortest distances. KNN algorithm computes the value of unknown data samples by computing the average of relevant variables and comparing the results with the k -nearest samples in the training set⁷⁵. The algorithm's effectiveness is highly dependent on the k value⁷⁶, which determines the number of samples to be considered in the regression process. A smaller value of k may result in a higher level of noise in the process, while a larger k value may lead to over smoothing and lower accuracy. Therefore, choosing an appropriate k value is crucial for optimization of the KNN performance. The distance between neighboring points used for the regression

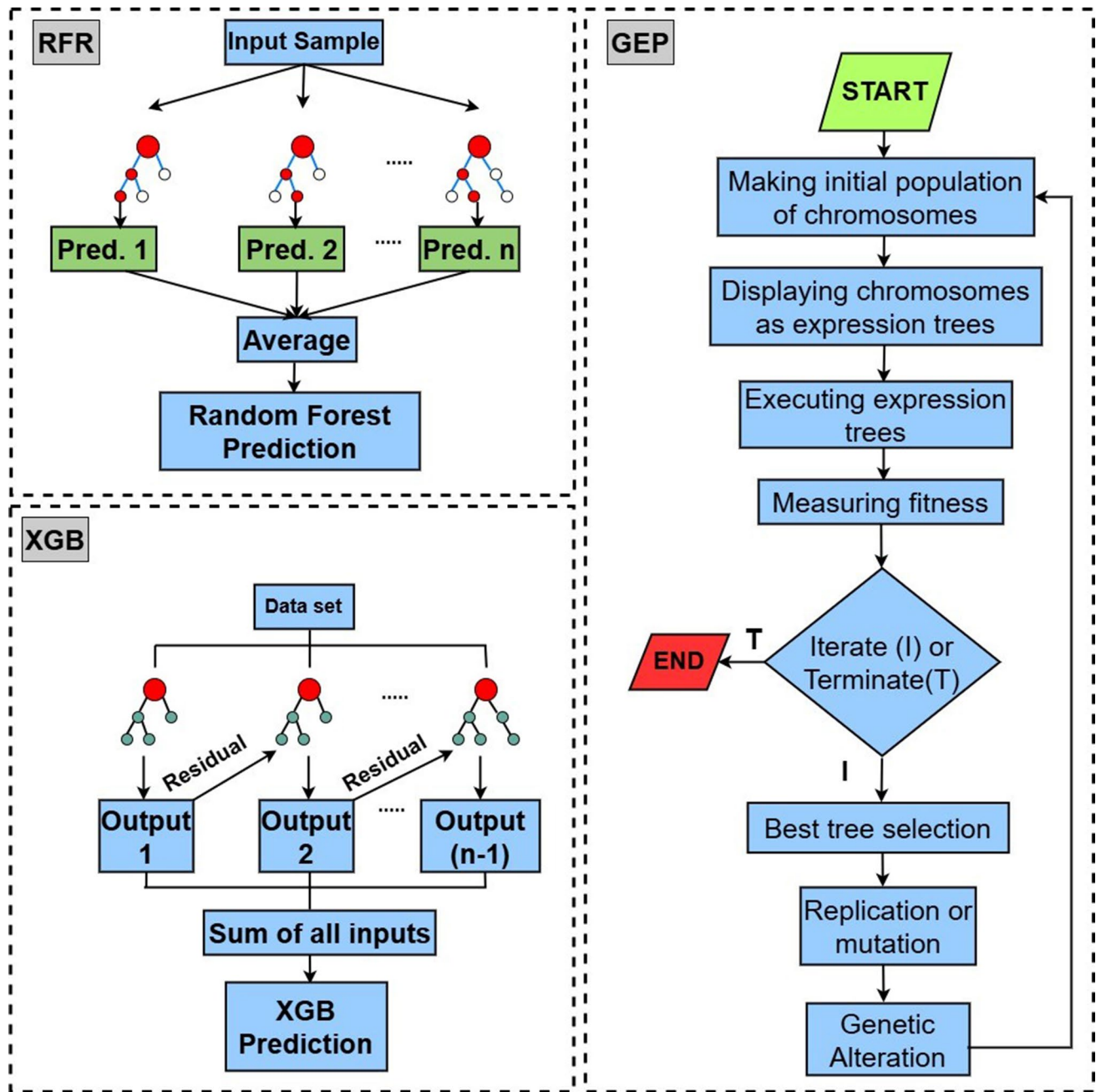


Figure 4. Graphical representation of ML models.

problems is calculated commonly by using Minkowski function as given in Eq. (3). In Eq. (3), x_i and y_i are the i th dimensions, and q indicates the order between the points x and y .

$$F(mi) = \left(\sum_{i=0}^f (|x_i - y_i|^q) \right)^{\frac{1}{q}} \quad (3)$$

The KNN method has several benefits over other ML algorithms. Firstly, it is simple and easy to understand. Secondly, it can easily understand nonlinear decision boundaries of both regression and classification problems. Moreover, the flexibility of the KNN approach is enhanced by the ability to adjust the k value to provide a suitable choice limit.

Random forest regression (RFR)

RFR is an ensemble ML algorithm consisting of two components i.e., decision tree (DT) and bagging algorithm and it can be utilized to solve both regression and classification problems based on the given dataset. It can swiftly relate input and output variables to make accurate predictions and has the ability to simulate complex

and multivariate relationships between different dependent and independent variables. The algorithm runs by iteratively partitioning the feature space into several sub-regions and this partitioning continues until a stopping criterion is met. Although the decision tree algorithm was robust for solving regression problems in civil engineering, Breiman proposed RFR algorithm to overcome the learning limitations of DTs⁷⁷ and since its development, RFR has proven to be useful in solving various civil and geotechnical related problems^{78–81}. RFR treats DT as base learner and uses bagging technique on them to make predictions. It generates a number of DTs that are correlated by using different samples from the bagging technique and the results of all DTs are averaged to make the final predictions to overcome the overfitting issue. It is reported that RFR has several advantages over the traditional DT algorithm including ability to seamlessly handle large data sets, robustness towards outliers and over-fitting⁸². Also, it can simultaneously deal with various variables without skipping any of them and is simpler than other algorithms like ANN etc. The architecture of the RFR algorithm is illustrated in Fig. 4.

Extreme gradient boosting (XGB)

In 2016, Chen and Guestrin developed an ensemble ML algorithm of remarkable accuracy named as XGB⁸³ which uses gradient boosting and DT algorithm to make predictions. Since its development, it has found numerous applications in diverse sectors such as identifying diseases, classifying traffic accidents, and predicting customer behaviours etc^{84–86}. The higher accuracy of XGB as compared to other boosting techniques is by virtue of its ability to learn fast and run several processes in parallel and reduce overfitting⁸⁶. XGB treats DTs as base learners and uses gradient boosting to build more efficient trees. The DTs used in XGB provide useful information about the importance of different variables in predicting the output⁸⁷. The XGB algorithm enhances this ability of DTs by optimizing a loss function to reduce the error of previously generated trees⁸⁸. In simpler words, the process of applying gradient boosting on DTs is simply combining several weak learners to create a strong learner through an iterative process⁸². The XGB algorithm adds new trees in a sequence to reduce the error of previous trees⁸⁶. This feature enables the algorithm to quickly fit the training data, but it can also become overfitted to the training data if this addition of trees is done at a swift rate. Thus, it is advised to reduce the rate of fitting of algorithm to the training data by means of newly added trees⁸⁹. This limitation of learning rate is done by applying a factor called shrinking rate to the predictions of newly constructed trees. The value of this shrinking rate is advised to be between 0.01 and 0.1 in the literature⁸⁷. The selection of a lower value of shrinking rate or learning rate makes sure that new trees are gradually added to the model to reduce the residual of previously trained trees, and the model is not overfitted. The visual representation of XGB prediction process is given in Fig. 4.

Multi-linear regression (MLR)

It is a common statistical practice to model the relationship between two variables using linear regression. Multiple linear regression (MLR) is an advanced version of basic linear regression which can be used to model linear relationship between several variables. In the current study, MLR model is developed to predict both plastic viscosity and interface yield stress of concrete. The general form of linear regression equation is given by Eq. (4). In Eq. (4), c is the intercept, $X_1, X_2, \dots, X_n$ are variables and $b_1, b_2, \dots, b_n$ are coefficients.

$$Y = c + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n \quad (4)$$

Data collection and analysis

A reliable database plays a crucial role in building ML-based models. Thus, a database of 142 instances was collected from published literature (given in Table A) which contained concrete mixture composition and experimental results for plastic viscosity and interface yield stress. The dataset was used from a study in which the authors designed an experiment to investigate the relationship between concrete mixture composition and the two fresh properties of concrete i.e., plastic viscosity and interface yield stress²³. In particular, the aim of the experiment was to examine the interaction between pipe walls of a concrete pump and the outer layer of concrete as (shown in Fig. 1) using a device called tribometer. A tribometer has three parts: (i) an agitator equipped with a speed electronic control device for measuring the torque, a container, and a steel cylinder. A lubricating layer is formed with the help of two co-axial cylinders and the rheological properties of concrete were computed using the Bingham model. The authors varied the concrete mixture composition to study their effect on the rheological properties of concrete. Also, the effect of time was investigated by measuring the rheological properties at different times after mixing the concrete. The data extracted from the tribometer included rotational velocity V measured in rotations per second and torque T measured in newton meter which were used to calculate plastic viscosity and interface yield stress were using a Bingham model. A detailed description of the procedure of extracting the plastic viscosity and interface yield stress from the measured values of torque and rotational velocity along with the detailed material properties of concrete constituents like type of cement, maximum diameter of aggregate, and type of superplasticizer etc. can be found in²³. The dataset used in current study has six input parameters including cement, water, coarse aggregate, fine aggregate, superplasticizer, and time after mixing. These six input parameters were used in the original experimental setup and are categorized as influencing parameters to predict plastic viscosity and interface yield stress of concrete. According to the recommendations from previous studies, the collected dataset has been divided into training (70%) and testing (30%) sets to make sure that the developed models can perform well on the unseen testing data. The algorithms will be trained using training data and testing data will be used to assess the accuracy of developed models afterwards. This type of data splitting makes sure that models are not overfitted to the training data and can perform equally well on the unseen data and^{90,91}. The distribution characteristics of data like minimum and maximum value, average, standard deviation, 25%, 50% and 75% percentile value are given in Table 1.

A statistical analysis tool called correlation matrix is frequently utilized to check the interdependence between various variables used to develop the models. It is used to investigate the effect of explanatory variables on each

	Cement	Water	Fine Aggregate	Coarse Aggregate	Superplasticizer	Time after mixing	Plastic Viscosity	Interface yield stress
Units	kg	kg	kg	kg	Litres/ 100 kg cement	minutes	Pa/s	Pa
Symbol	A	B	C	D	E	F	Y ₁	Y ₂
Maximum	604	275	1064	1175	4.6	90	1324	125.8
Minimum	420	147	553	709	0	0	105.6	15.7
Mean	460.19	190.10	776.29	958.14	2.142	35.660	553.10	41.55
Standard Deviation	29.804	29.461	104.04	81.628	1.938	33.758	230.667	21.145
Median	460	184	792	968	1	30	516	36.42
25%	450	171	731	918	1	0	352.80	27.43
50%	460	184	813	968	1	30	465.68	40.025
75%	460	184	859	1010	3.80	60	603.75	50.29

Table 1. Statistical description of dataset.

other because it quantifies the relationship between different variables in the form of coefficient of correlation (R). The extent of interdependence between variables can be judged by looking at the R values between two variables. Generally, a value above 0.8 indicates the presence of a good correlation between two variables⁹². It is important to check interdependence between variables before developing a ML model. It is due to the fact that if most of the variables are strongly correlated to each other, then it might cause complications in the model, a problem commonly referred to as multi-collinearity⁹³. Notice from Fig. 5 that the correlations between variables are less than 0.8 mostly so they will not give rise to the problem of multi-collinearity.

Violin plots are also constructed to visualize the distribution of parameters used in the study as shown in Fig. 6. In the violin plots, the density of data is shown by the width of the violin plot. Also, the thick black bar represents the 25–75% quartile range, and the small blue dot indicates the median value. Moreover, the extended black line represents the 95% confidence interval. It can be seen from violin plots that both input and output variables are spread across a wide range. The literature recommends having this type of wide data distribution to make widely acceptable and generalized ML models⁹⁴.

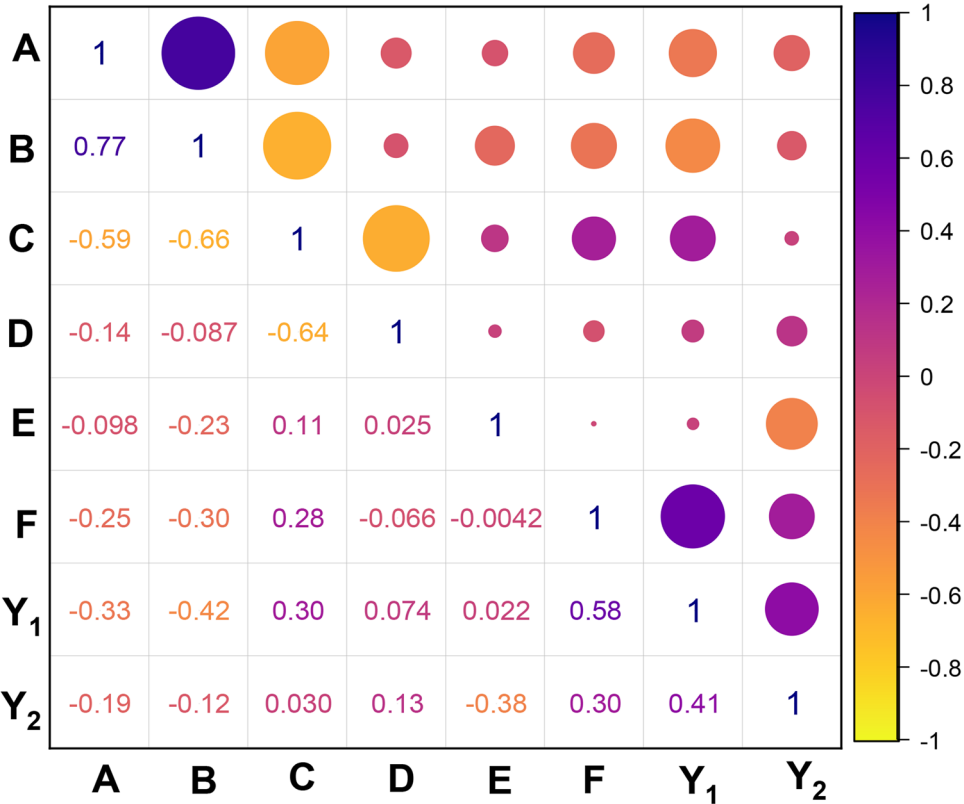


Figure 5. Correlation matrix of variables.

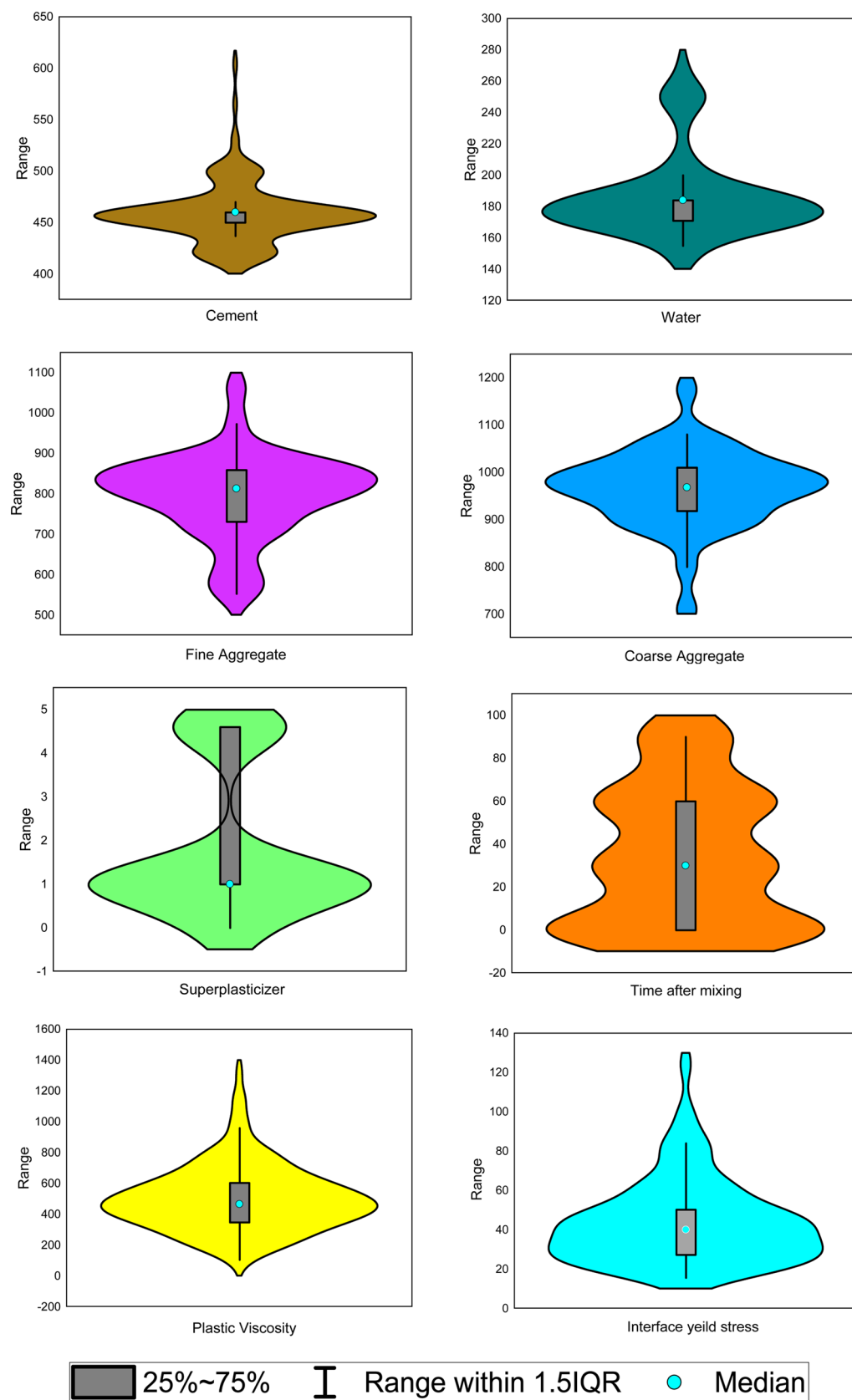


Figure 6. Violin plot distribution of parameters.

Performance assessment and model development

Development of GEP model

A software package was used in current study for efficiently for implementing GEP algorithm known as GeneXpro Tools version 5.0. The dataset fed into the software and split into training and testing data. The training was used for training of GEP and after training, the testing data was used to assess the accuracy of trained model. After splitting of the data, the next step was to specify some GEP hyperparameters because each dataset and each algorithm needs a specific set of hyperparameters. The important hyperparameters to tune for GEP algorithm include head size, number of genes, set of functions, and number of chromosomes etc. Each of these hyperparameters control a critical aspect of the GEP algorithm (as indicated by Table 2) and are crucial for the development of a robust GEP model. For instance, head size and number of chromosomes dictate the convergence of the model towards the solution. Similarly, number of genes specify the number of sub expression trees that will be formed by the algorithm and joined by the linking function to get the final equation. The selection of right set of these parameters is crucial since the model's performance and accuracy is dependent on these hyperparameters. Thus, in current study, these parameters were carefully chosen after consulting previous studies and extensive testing. The initial set of parameter values were chosen using a trial-and-error method and previous literature recommendations^{95,96}. Then these parameters were varied across a range of possible values and the set of values which yielded highest accuracy were used to build the GEP model for predicting both outputs. The model development was started using chromosome number as 10 and it was varied in increments of 10 until the maximum accuracy was achieved at 100 chromosomes for predicting plastic viscosity and at 80 chromosomes for interface yield stress. In the same way, head size was initially selected as 5, and the model presented in current study achieved maximum accuracy at head size equal to 20 and 30 for plastic viscosity and interface yield stress respectively. The accuracy of the model may be improved by increasing these values however increasing these values beyond a certain point may result in overfitting of the model. Selecting higher values of these parameters also increases the run time and computational power required by the algorithm. Thus, the GEP modelling essentially refers to the process of finding the optimal balance between accuracy, model complexity, and computational power required by the algorithm and the model developed using GEP in this study presents the optimal balance between these aspects. Furthermore, notice that the set of functions used to be included in the final equation consist of simple arithmetic functions and trigonometric functions. It was deliberately done in order to keep the resultant equations simple and compact for fast computation and implementation. The set of GEP hyperparameters employed in current study is represented in Table 2.

Development of RFR, XGB, and KNN models

In contrast to GEP algorithm, XGB, RFR, and KNN algorithms were developed with the help of python programming language in Anaconda software. All the prerequisite steps of model development like data loading, data splitting, and actual model development (hyperparameter settings) were done with the help of python codes as shown in Fig. 7. For algorithms like XGB and RFR, the tuning process of hyperparameters holds paramount importance since the accuracy of these algorithms depend directly on these hyperparameter settings⁹⁷. The careful selection of the hyperparameters will enable the algorithms to mitigate overfitting and make them suitable for variable data configurations⁹⁸. There are several approaches available for hyperparameter selection

Parameters	Significance	Search Space	Optimal Values	
			Y ₁	Y ₂
GEP Parameters				
No. of Chromosomes	Specifies the chromosomes in the population	10 to 100	100	80
No. of Genes	Controls the number of sub-expression trees made	2 to 10	5	5
Head Size	Controls size of each sub-expression	10 to 50	20	30
Linking Function	Used to link sub-expressions from all the genes	Addition, Subtraction, Division, Multiplication	Addition	Addition
Functions	Arithmetic functions used to make the equation	+, -, x, ÷, sqrt, sin, cos, tan	+, -, x, ÷, sqrt, sin, cos, tan	+, -, x, ÷, sqrt, sin, cos, tan
KNN Parameters				
k-neighbors	The number of samples averaged by the algorithm to give the prediction	2 to 10	8	5
Distance Function	The metric used to calculate the distance of a sample from its neighbouring samples	Euclidean, Minkowski	Euclidean function	Euclidean function
RFR Parameters				
Max. depth	Maximum size of each decision tree	2 to 10	5	5
n-estimators	Number of base learners present in the algorithm	10 to 50	20	20
XGB Parameters				
Max. depth	Maximum size of each DT	5 to 20	10	10
n-estimators	Number of DTs made by the algorithm	50 to 200	100	100
Learning rate	Shrinking factor applied on the predictions of each new tree	0.01 to 0.1	0.01	0.01

Table 2. Hyperparameter settings of ML models.

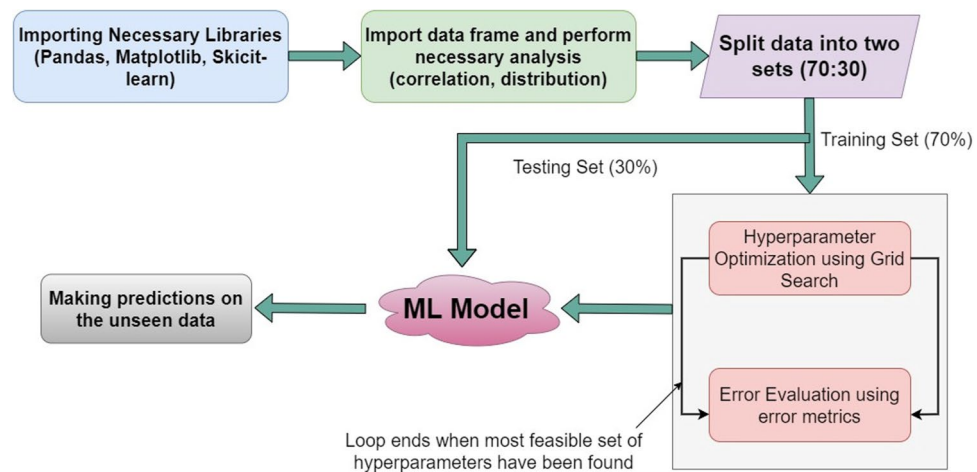


Figure 7. Model development process using python programming language.

to avoid manual selection of parameters. The grid search approach is the most famous and efficient one out of all the available techniques and is widely used for hyperparameter selection⁹⁹. This approach exhaustively tests all possible combinations of hyperparameters in a grid configuration to find the best set of parameters and it has the advantage that it does not further alter the good-performing values¹⁰⁰. For XGB, RFR, and KNN models in current study, the 'GridSearchCV' function in sklearn library of python was used to implement the GridSearch algorithm¹⁰¹. Despite the development of other techniques like random search approach etc. for parameter optimization, GridSearch approach is still the most efficient and widely used technique because:

- (i) it is easy to implement.
- (ii) it is reliable and results in low dimensionality.
- (iii) it results in greater accuracy than other optimization techniques¹⁰².

There are two common hyperparameters which need to be tuned in case of XGB and RFR called n-estimators and maximum depth which specify the number of trees build by the algorithm and depth of each tree respectively. The RFR algorithm builds different trees and takes mean of all the predictions made by trees to reach at the final prediction. The number of trees developed by the algorithm is controlled by the parameter "n-estimators". Thus, the value of this parameter must be carefully chosen so that the algorithm builds a considerable number of trees and takes their mean to make the final prediction. Similarly, the parameter "k-neighbours" of KNN algorithm specifies the number of neighboring samples that the algorithm considers and takes into account in making the decision. The accuracy of KNN algorithm largely depends upon this parameter and selecting an abnormally larger value of this parameter may result in the algorithm being overfitted. The XGB algorithm adds new trees in a sequence to reduce the residual of previously constructed trees and it can fit very quickly to the training data by virtue of this technique, but it can also lead to overfitting of the algorithm to the training data⁸⁹. Thus, it is beneficial to limit the rate at which the algorithm adds new trees to reduce the residual of previous tree. This can be done by applying a shrinking factor to the prediction made by the newly constructed trees. Therefore, a factor called learning rate is applied to the new predictions which limits the rate at which the algorithm fits to the training data. The literature suggests keeping its value between 0.01 and 0.1¹⁰³ so that the residual reduction by each added tree is done slowly. Thus, the value of learning rate has been selected as 0.01 for prediction of both outputs. The values of other hyperparameters used in current study found using grid search approach are given in Table 3 along with the search space in which GridSearch algorithm searched for those values and the significance of each hyperparameter in the development of the ML algorithm.

Performance assessment of models

The developed models will be checked by means of different error metrics because it is necessary to make sure that the models can effectively solve the given problem without abnormally large errors¹⁰⁴. The coefficient of determination R^2 is used as an indicator of model's general accuracy¹⁰⁵. However, due to its insensitivity towards some conditions such as a constant value being multiplied or divided with the output¹⁰⁶, several other metrics are used alongside it. The metrics MAE and RMSE are described as crucial for evaluating a ML based model¹⁰⁴. MAE measures average error between real and predicted values and RMSE gives an indication of presence of larger errors¹⁰⁷. The metrics PI and OF are commonly used as indicators of model's overall performance because they simultaneously consider different metrics and calculates a value based on them¹⁰⁸. Thus, every metric has its own importance. The employed error metrics, their mathematical formulas, their range and suggested criteria for acceptable models are given in Table 3. The criteria given in Table 3 will be used to check the suitability of the resulting ML models.

No.	Metric	Abbreviation	Formula	Range	Suggested value
1.	Mean absolute error	MAE	$\frac{\sum x - y }{N}$	0 to $+\infty$	Close to zero
2.	Root mean square error	RMSE	$\sqrt{\frac{\sum (x - y)^2}{N}}$	0 to $+\infty$	Close to zero
3.	Coefficient of determination	R^2	$1 - \frac{\sum (x - y)^2}{\sum (y - y_{mean})^2}$	0 to 1	$R^2 > 0.8$
4.	Performance index	PI	$\frac{RRMSE}{1 + R}$	0 to $+\infty$	$PI < 0.2$
6.	Objective Function	OF	$\left(\frac{N_{Training} - N_{Testing}}{N}\right) P_{Training} + 2 \left(\frac{N_{Testing}}{N}\right) P_{Testing}$	0 to $+\infty$	$OF < 0.2$

Table 3. Summary of error evaluation criteria.

Results and discussion

GEP results

The empirical equations for predicting plastic viscosity and interface yield stress are given by Eqs. (5) and (6). Since the number of genes were selected as 3 for plastic viscosity modelling, the algorithm yielded three expression trees which were combined using the chosen linking function (addition) to get the final equation. While the number of genes were selected as 5 for interface yield stress prediction thus the algorithm yielded four sub-expression trees which are linked by the addition linking function to get the final result in Eq. (6). The predictive abilities of the GEP algorithm for prediction of both outputs are given in the form of scatter plots for predicting plastic viscosity in Fig. 8. Notice from Fig. 8 (a) that the data points are scattered across the line of linear fit with some points lying at great distances from the line. It depicts that the GEP algorithm failed to accurately predict plastic viscosity at several points. It can also be seen from the training and testing error metrics of GEP given in Table 4. Notice that the R^2 value for training and testing of GEP is 0.74 and 0.77 which is less than the recommended value of 0.8. It means that the general performance of GEP algorithm is not satisfactory. The PI values for training and testing set are 0.11 and 0.07 respectively which is less than the upper threshold of 0.2. Also, the OF value of GEP is 0.086 which is also less than the recommended value of 0.2. Although the PI and OF values suggest that the performance of GEP algorithm is satisfactory, the correlation between the actual and predicted values is less than 0.8 which suggests that its general accuracy is not satisfactory.

Regarding the prediction of interface yield stress, the scatter plots in Fig. 8 (b) provide a useful way to visualize the predictive capabilities of the algorithm. The line of linear fit represents the points where actual and predicted values are the same. It can be seen that some data points are lying close to the line while some are scattered in both training and testing plots. It indicates that there are some points at which GEP didn't accurately predict the interface yield stress. The error evaluation summary of GEP model is also given in Table 5. The R^2 value for both training and testing sets lie in the range of 0.6 which is not up to the recommended value for a good model (0.8). Also, the average error between actual and predicted yield stress is 9.17 and 12.80 for training and testing data respectively. The PI value for training is 0.161 and 0.20 for testing while it is recommended to have PI value less than 0.2 for an acceptable model. Thus, despite the advantage that GEP algorithm yielded an empirical Eq.

$$Y_1 = T_1 + T_2 + T_3 \quad (5)$$

Where

$$\begin{aligned}
 T_1 &= \left\{ \left(\frac{F}{2B} - \tan \left(\sqrt{A \cdot E^5} \right) \right) \left((\sqrt{B} - \sin A)^5 - \tan(\sqrt{C} + \cos D) \right) \right\} + \sqrt{C} \\
 T_2 &= \left\{ \left(\frac{(C(E \cdot C) + D)}{(\sqrt{B})^4} - \tan(E + C) \right) + \sqrt{\left(\frac{F}{C} \right) (A - B)} \right\} + F \\
 T_3 &= (A - B) + \left\{ \left(\cos((F^6 + B^6) + \left(\frac{\sqrt{\frac{2B \cdot D}{(\tan(E - A))^4}}}{B} \right)^3)^{10} \right) \right\} \\
 Y_2 &= T_4 + T_5 + T_6 + T_7 + T_8 \quad (6) \\
 T_4 &= \tan \left[\tan[D \cdot \left(\frac{\sin \{ (-C) ((D \cdot E) - E^2) + E \} - \{ (E^2(A + 9.192)C)(D \cdot E - (B \cdot E)) \}}{\cos(-9.192)} \right) C] \right] \\
 T_5 &= \tan(B + D^4) \\
 T_6 &= \frac{\cos \left[\left\{ \left((0.5827 \cdot (E + D)) + \frac{\cos D}{B + 0.285} \right) + C \right\} + \left\{ \frac{\frac{6.686 - A}{2.5373}}{\tan \left(\frac{-9.812}{D} + 0.285 \right)} \right\} \right]}{\cos \left\{ \left(\frac{4.767 \cdot C}{0.285} \right)^4 - C \right\}}
 \end{aligned}$$

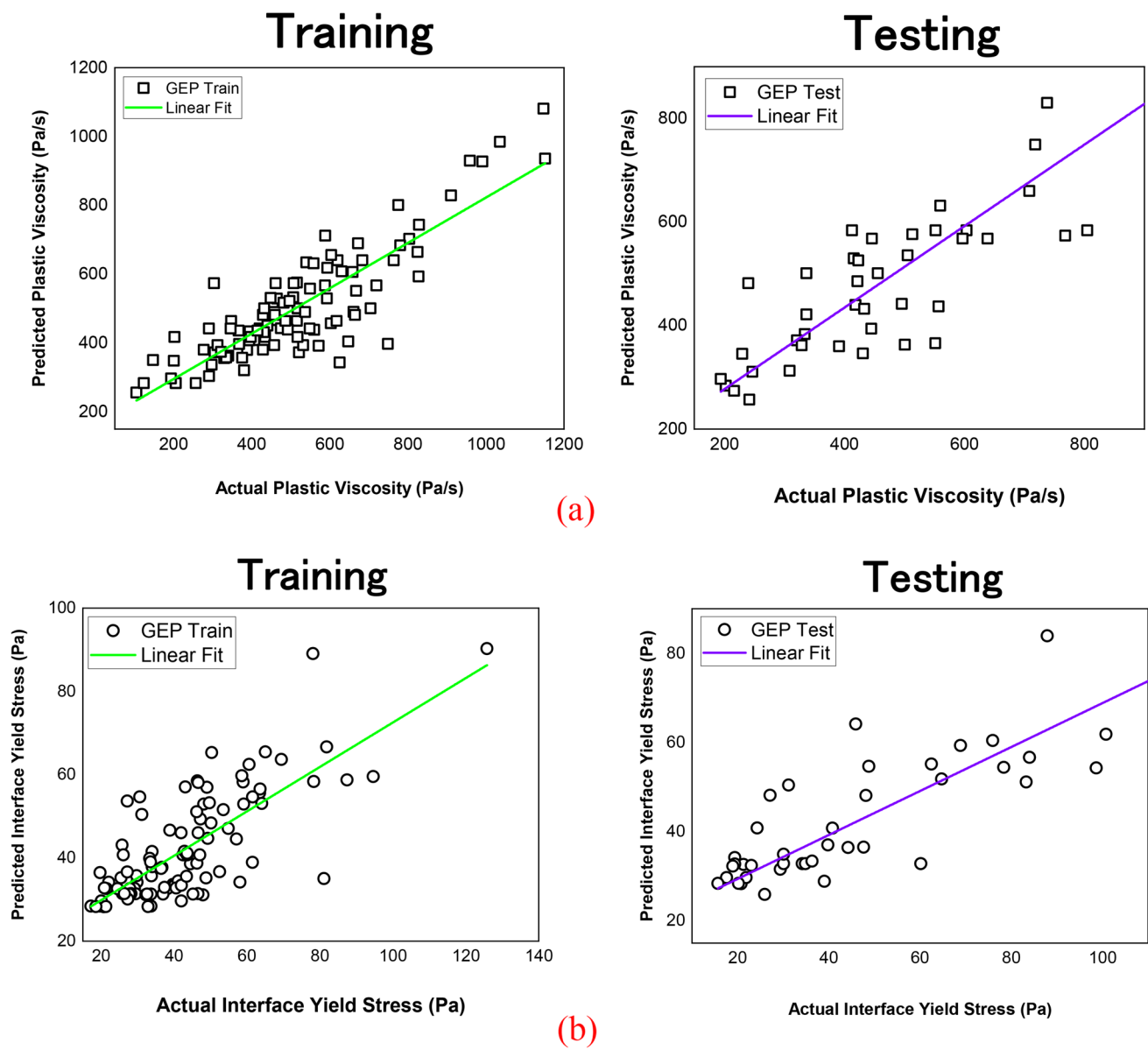


Figure 8. Scatter plots to GEP training and testing: (a) for plastic viscosity; (b) for interface yield stress.

	GEP		KNN		XGB		MLR		RFR	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	83.38	83.28	91.15	98.77	22.39	29.28	106.66	160.92	46.55	90
RMSE	109.22	66.06	119.96	138.4	57.10	73.32	144.43	196.1	61.02	128.28
R ²	0.7419	0.7765	0.6608	0.7219	0.959	0.947	0.5093	0.2073	0.9149	0.6429
PI	0.11	0.07	0.137	0.142	0.059	0.071	0.184	0.222	0.065	0.12
OF	0.086		0.139		0.066		0.206		0.097	

Table 4. Error metrics of developed models for plastic viscosity prediction.

	GEP		KNN		XGB		MLR		RFR	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	9.17	12.80	8.8	12.98	1.966	2.45	13.81	15.64	4.02	12.53
RMSE	12.18	16.4	11.9	16.2	4.63	5.24	17.5	21.22	6.188	18.15
R^2	0.5655	0.6825	0.5209	0.6788	0.9259	0.965	0.3489	0.2278	0.8806	0.695
PI	0.161	0.20	0.168	0.17	0.057	0.053	0.25	0.321	0.078	0.20
OF	0.165		0.184		0.054		0.292		0.150	

Table 5. Error metrics of developed models for interface yield stress prediction.

$$T_7 = \left(\left[\cos \left\{ (E^5 + (8.45.A)) + \left(\frac{76.73.C}{E + 8.35} \right) \right\} + \left\{ \sqrt{\tan \left(\frac{F}{A} \right)} \right\} \right] + 6.55 \right)^2$$

$$T_8 = \tan \{ \tan (B) \}$$

KNN results

The KNN algorithm didn't yielded an equation for prediction of interface yield stress and plastic viscosity. However, the KNN results are shown as scatter plots in Fig. 9 for the prediction of both outputs and it can be seen that the points are scattered and not lie much closer to the line of linear fit. The error metrics of KNN for plastic viscosity prediction given in Table 4 also suggest the same. The R^2 value is 0.66 for training which shows that there is only 60% accuracy in the predictions made by KNN. The accuracy improved a little in testing set as shown by 0.729 R^2 value but it is still less than the recommended value of 0.8. The average errors of KNN are also greater than GEP which shows that KNN will give larger errors than GEP when used to make predictions. The PI values are 0.137 and 0.142 for training and testing which are close to the upper value of 0.2. The OF value is also 0.139 which suggests that the overall performance of KNN algorithm is also marginally satisfactory according to the performance indicators. In the same way, the result for interface yield stress prediction is shown as scatter plots in Fig. 9 (b) and the error evaluation is given in Table 5. It can be seen from Table 5 that the R^2 value is 0.520 for training while it is 0.678 for testing. Both of these values are greater than GEP but lower than the recommended literature value of 0.8. The trend is same for other error metrics of KNN like MAE and RMSE. This shows that the performance of KNN isn't satisfactory to accurately predict the interface yield stress. This is also visible from the scatter plots in Fig. 9 where the data points are very much scattered and not lie close to the linear fit line. The PI values are 0.168 and 0.17 for training and testing respectively which is less than the recommended value of 0.2. However, despite the PI and OF values lying in the recommended range, the KNN-based predictive model has a difference of 8 to 12 Pa between actual and predicted values.

MLR results

The MLR result for plastic viscosity prediction is given as Eq. (7) while the equation for yield stress is given as Eq. (8). The performance of MLR is also given as scatter plots in Fig. 10 (a) for plastic viscosity and performance evaluation is given in Table 4. Firstly, the R^2 value is 0.50 for training and 0.20 for testing which suggests that MLR accuracy is very less than the other algorithms. Similarly, the testing PI and OF value of MLR model are greater than 0.2 which suggests that MLR failed to exhibit overall satisfactory performance to predict plastic viscosity. The scatter plots for interface yield stress prediction are shown in Fig. 10 (b) for MLR training and testing sets. The inability of MLR to accurately predict output and lack of correlation between actual and predicted values is evident from the scatter plots. Also from Table 5, notice that the MAE and RMSE values of MLR are significantly higher than other algorithms while the R^2 value is at the lowest of about 0.2. Notably, both the training and testing PI values of MLR (0.25 and 0.321 respectively) are considerably higher than 0.2 which suggests the overall performance of MLR model is acceptable to be used for prediction. Moreover, the OF value of MLR being 0.292 also suggests that it can't be used for accurate prediction of the output. It further necessitates the utilization of ML models for predicting properties of concrete.

$$Y_1 = -3529.81 + 3.04A - 0.49B + 1.37C + 1.55D - 10.99E + 4.34F \quad (7)$$

$$Y_2 = 37.89 - 0.1A + 0.03B + 0.0007C + 0.045D - 5.90E + 0.25F \quad (8)$$

RFR results

The error metrics calculated for RFR-based predictive model for plastic viscosity are given in Table 4 and also shown as scatter plots in Fig. 11 (a). Notice that the R^2 value for RFR training is 0.91 which indicates that the RFR predicted values are very close to the actual values. It can also be depicted by the very less training MAE value of 46.55. However, the performance of RFR algorithm drops significantly when tested against unseen data of the training set. The R^2 drops to a lowest value of 0.64 and MAE increases to 90 Pa/s from 46 Pa/s which is a significant loss in accuracy. Similarly, the training PI suggests that the performance of algorithm is satisfactory in training phase having a value of 0.065 but it increases to 0.12 in testing phase which shows that the algorithm could not exhibit the same level of accuracy when used for unseen data. It shows that the algorithm is overfitted to the training set. Similarly, Fig. 11 (b) shows the RFR result for interface yield stress prediction and the summary of error evaluation of RFR model is also given in Table 5. First of all, the R^2 value

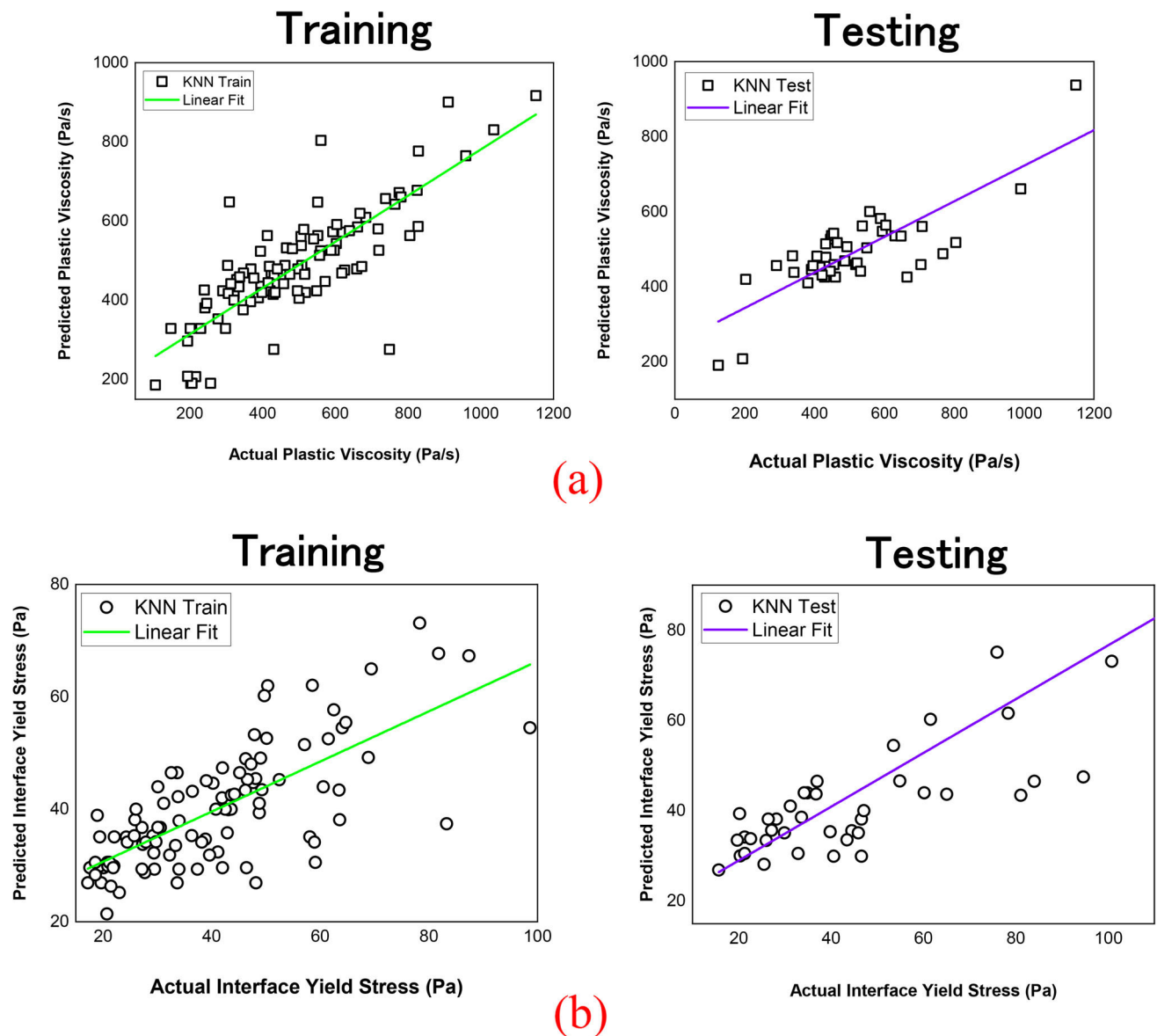


Figure 9. Scatter plots to KNN training and testing; (a) for plastic viscosity; (b) for interface yield stress.

for RFR training is 0.88 which is greater than the recommended value of 0.8. The good performance of algorithm in training phase can also be depicted by the MAE value of only 4.02 Pa. However, the performance of drops significantly in the testing phase with R^2 reaching the value of 0.695. Same trend was observed for RMSE and MAE. Also, the training PI suggests that the performance of algorithm is satisfactory having a value of 0.078 but it increases significantly to 0.2 in testing phase which is equal to the upper threshold for PI values. It depicts that the algorithm could not exhibit the same level of accuracy when used for testing data and is overfitted to the training set. It can also be seen from Fig. 11 that the data points in training plot lie close to the linear fit line while they are scattered in the testing data plot.

XGB results

The XGB algorithm failed to give an empirical equation like GEP, however the result of XGB to predict plastic viscosity is given in the form of scatter plots in Fig. 12 (a). It can be seen from the plots that most of the data points lie close to the linear fit line which shows the excellent accuracy of XGB. Also, the error metrics for XGB is given in Table 4. Notice that the R^2 value of XGB is 0.959 and 0.947 for training and testing respectively and the MAE value is also considerably less than the other algorithms. It is an indicator of XGB's excellent accuracy to predict the plastic viscosity values. The RMSE values are 57.10 for training and 73.32 for testing which suggests that there will not be any abnormally larger errors in the predictions made by the XGB algorithm. Moreover, the OF value of XGB is considerably less than the other algorithms (0.066) which again suggests that its performance is better than all other algorithms. Also, the error metrics calculated for XGB-based predictive model for interface yield stress are given in Table 5 and also shown as scatter plots in Fig. 12 (b). Notice that the

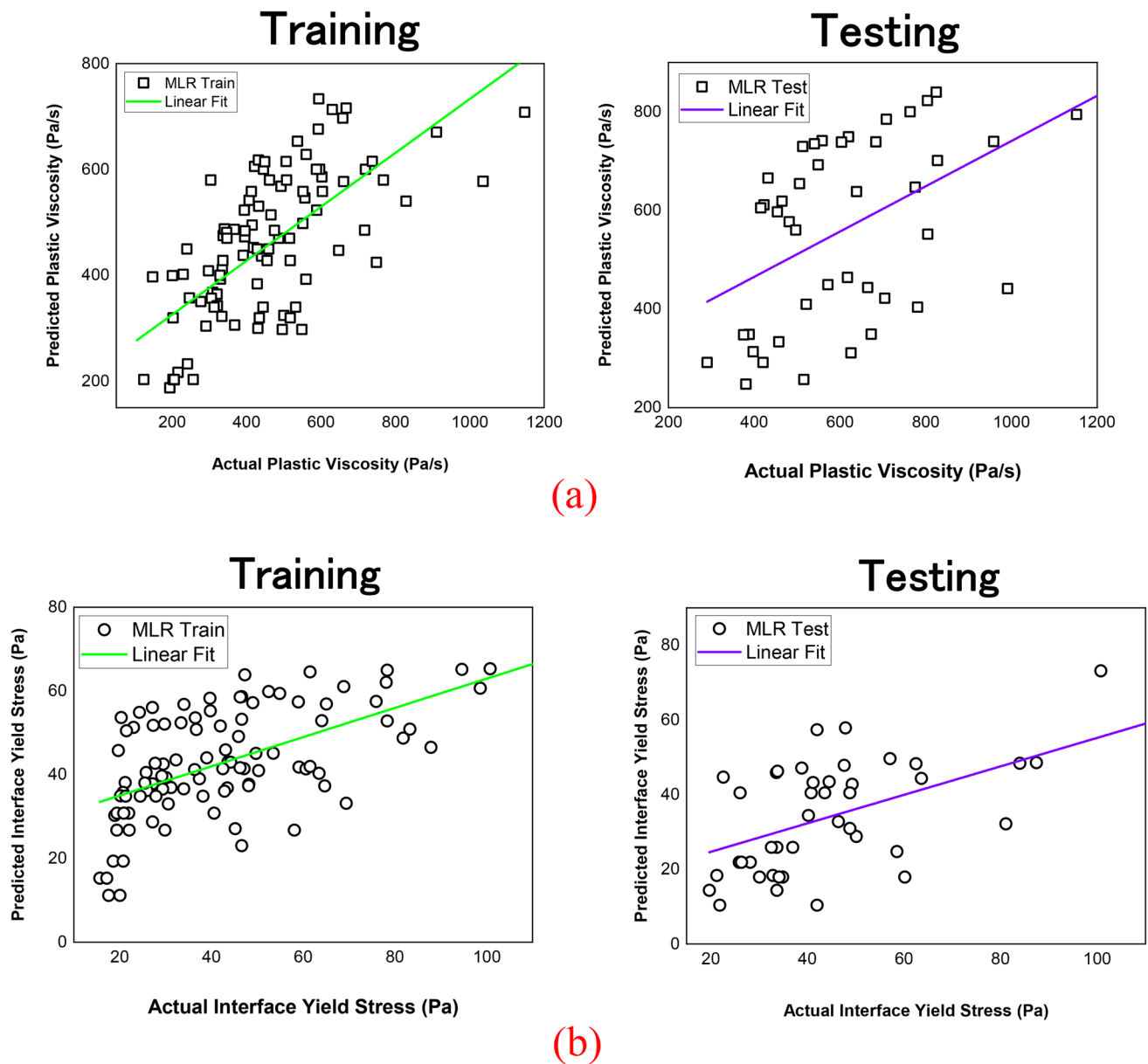


Figure 10. Scatter plots to MLR training and testing; (a) for plastic viscosity; (b) for interface yield stress.

R^2 value for XGB training is 0.92 which indicates that the XGB predicted values are very close to the actual values. The performance of XGB algorithm doesn't drop when tested against unseen data of the training set. Similarly, the training PI suggests that the performance of algorithm is satisfactory in training phase having the lowest value of 0.057 and it remained almost same in testing phase (0.053) which shows that the algorithm exhibited the same level of accuracy when used for unseen data. It shows that the algorithm is not overfitted to the training set and can be effectively used for predictions on the new data. This can also be visualized from scatter plots in Fig. 12 where the data points in training and testing plots lie very close to the linear fit line. Thus, from Table 5 it can be concluded that XGB proved to be the most accurate algorithm for prediction of interface yield stress of concrete.

K-fold Validation

K-fold cross-validation is a statistical method for evaluating the generalization abilities of ML models and helps to confirm that the models are resistant to overfitting or underfitting. Overfitting occurs when a model performs exceptionally well on the training data but fails to generalize to unseen data. It can result from the model capturing irrelevant patterns in the training set. K-fold validation mitigates overfitting by ensuring that the model is trained and validated on different subsets of the data, promoting more robust learning patterns. This technique works by dividing the whole dataset into 'k' equal-sized subsets, known as folds. During every iteration, one fold is used for testing, while all other folds are cumulatively used for testing. This process is

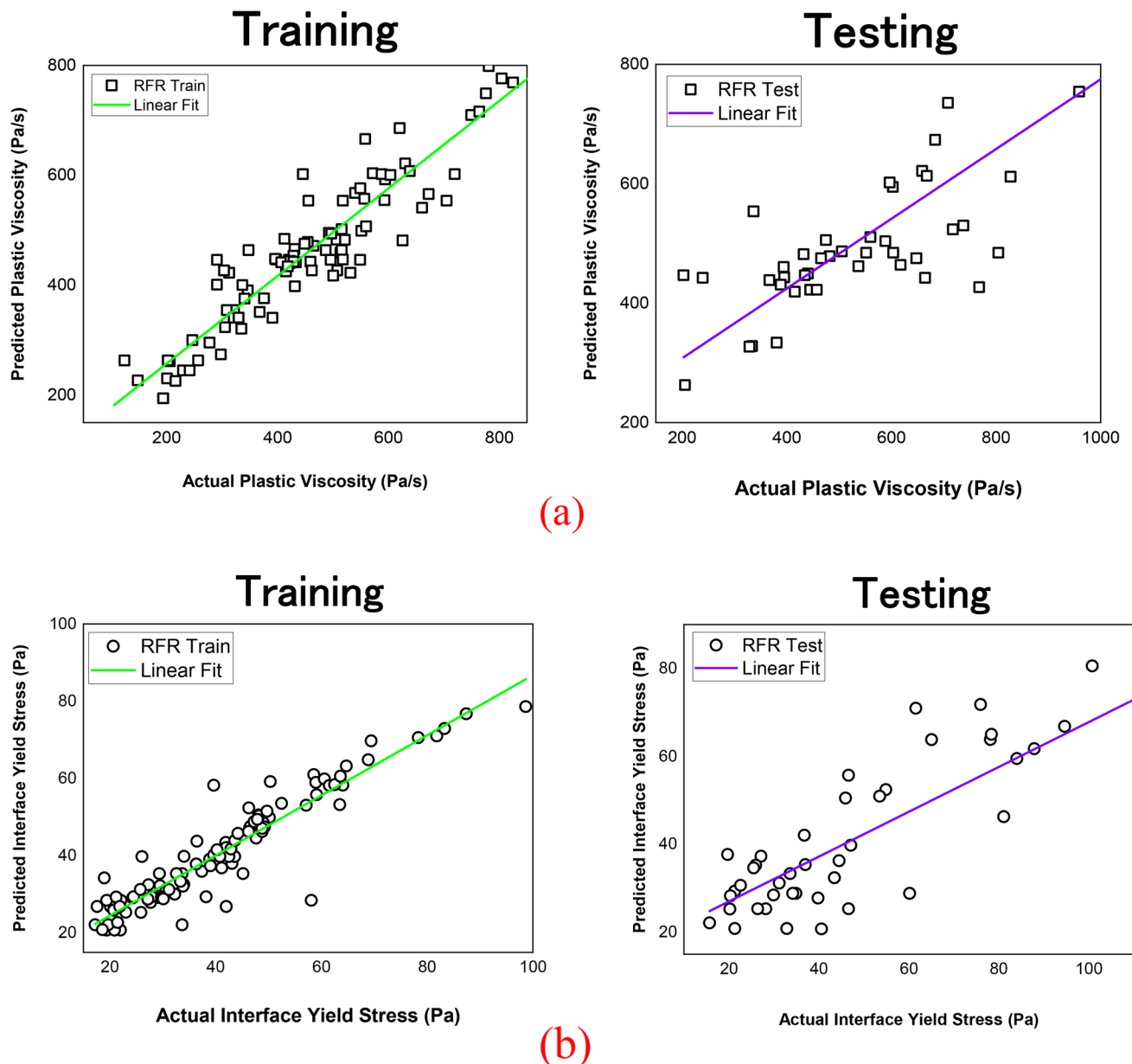


Figure 11. Scatter plots to RFR training and testing; (a) for plastic viscosity; (b) for interface yield stress.

repeated 'k' times, enabling each fold to act as testing data exactly one time, as shown by Fig. 13. This ensures that the developed models are generalizable to unseen data, thereby eliminating overfitting. This study utilizes 8-fold cross-validation for all three algorithms, with the accuracy of each algorithm in each fold measured by RMSE, as shown in Figs. 14 and 15.

Figure 14 shows the k-fold results of algorithms for prediction of plastic viscosity. Plot (a) in Fig. 14 corresponds to the results of GEP algorithm. It can be seen that the RMSE values corresponding to the 8 folds of GEP have some difference in the values. Although it is closer to the RMSE values of GEP given in Table 4, the disparity between the RMSE values of different folds suggests that the performance of algorithm is marginally accurate, and the algorithm may be subject to slight overfitting. Next, plot (b) of Fig. 14 shows the KNN results of k-fold validation. In contrast to GEP, the RMSE values of different folds are very close to each other suggesting that although KNN has less accuracy (as indicated in Table 4), it is not overfitted. Moving on, plot (c) shows MLR results in which some folds yield slightly higher values of RMSE indicating slight overfitting. In contrast, the RFR results in plot (d) show a substantial difference in RMSE values of different folds conforming with the previous findings that RFR model is overfitted to the training set. In particular, RFR yields RMSE values as high as 130 Pa/s in some folds while this value is only 60 Pa/s in other folds. It suggests that the performance of RFR algorithm can significantly change based on the distribution of the data. In other words, the RFR model developed in current study is not well generalized for making predictions. It may be due to the fact that RFR algorithm picked up some unnecessary correlations from the data which affected its accuracy. Finally, plot (e) of

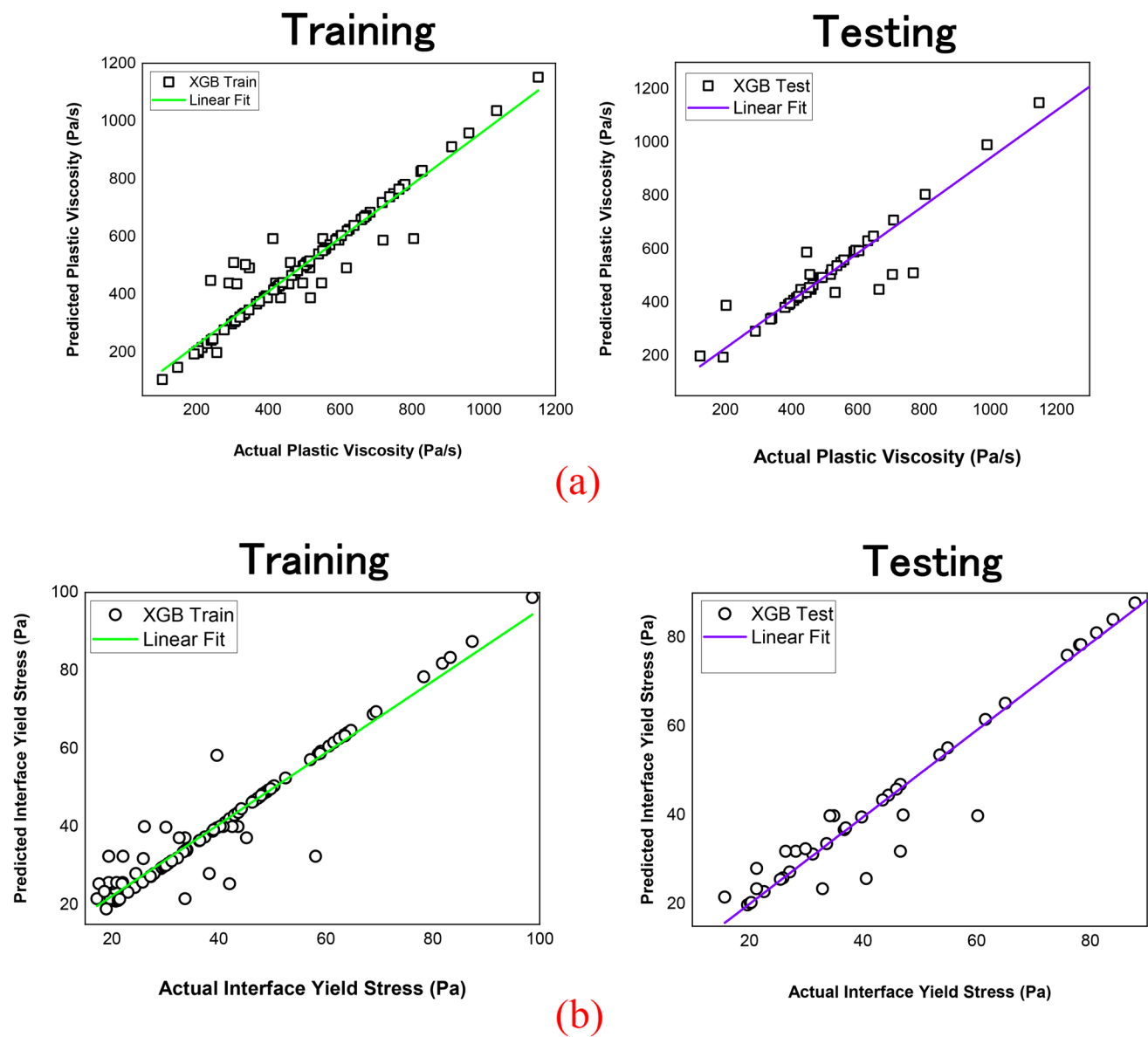


Figure 12. Scatter plots to XGB training and testing; (a) for plastic viscosity; (b) for interface yield stress.

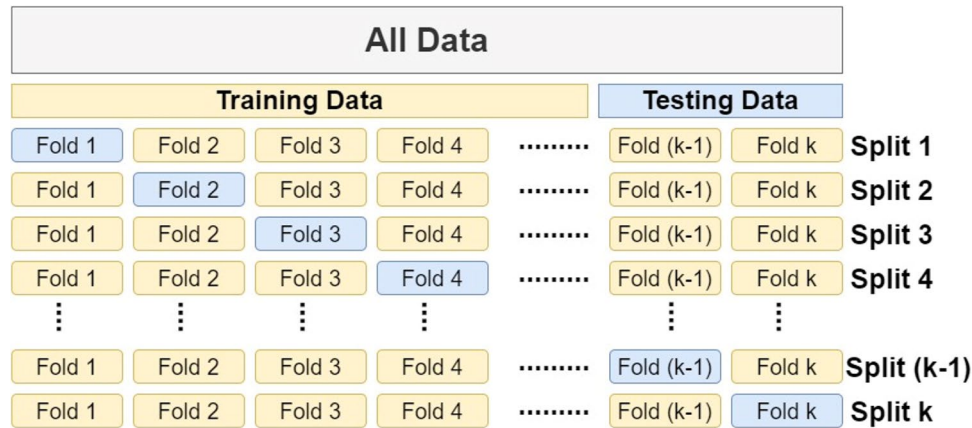


Figure 13. Schematic representation of k-fold cross validation.

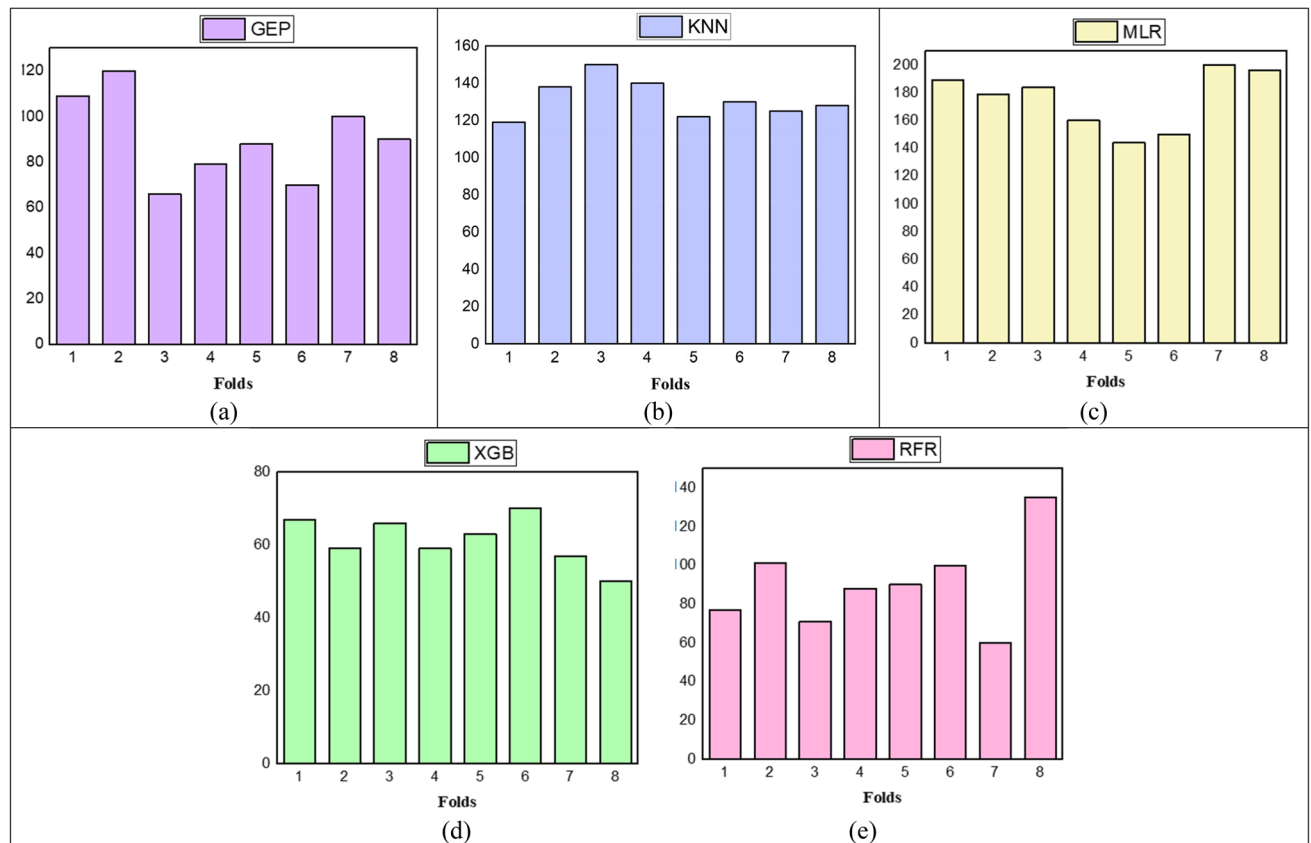


Figure 14. K-fold Validation results of algorithms for prediction of plastic viscosity.

Fig. 14 show the k-fold validation results of XGB algorithm. It can be seen from the plot that the RMSE values from each fold are lying sufficiently closer to one another. Moreover, these values are also closer to the training and testing RMSE values of XGB given earlier in Table 4. It confirms that XGB is resistant to overfitting and can be used effectively on unseen dataset.

Similarly, the k-fold results of algorithms for prediction of interface yield stress are shown in Fig. 15. The GEP results for yield stress prediction are shown by plot (a) of Fig. 15 from which it can be seen that there is some disparity between the RMSE values of different folds. Previously, in Table 5, GEP algorithm also showed a difference in accuracies for training and testing data. It can be attributed to the fact that GEP is an evolutionary technique which solves the problem by means of chromosomes (mathematical expressions). These expressions are randomly generated and later modified to fit the data. Due to this randomness in the structure of GEP, it is unable to give a generalized and robust model for prediction of interface yield stress. In next plot (b), which has been made with k-fold values for KNN algorithm, the difference between RMSE values for various folds is less prominent compared to GEP. It suggests that KNN is not overfitted to the training data and has good generalization capabilities although with a lower accuracy. It is also true for the MLR results. Its accuracy is less than that of KNN with RMSE values approaching 24 Pa but it is generalized, nevertheless. However, the RFR model again shows prominent signs of overfitting with substantial differences between RMSE values of different folds indicating that the model's performance varies for diverse data distributions. It highlights that RFR is not generalized well and thus should not be used practically for the prediction of subject output. At last, XGB again stands out having the least difference among different RMSE values. Also, the RMSE values for XGB are very less compared to its counterparts. This suggests that XGB, in addition to accurately predicting plastic viscosity on unseen data, can be effectively leveraged to predict interface yield stress of unseen data too.

Residual assessment

In addition to error evaluation and k-fold validation, it is imperative to check the residuals of individual algorithms to confirm the maximum error which can be encountered while using the algorithm. Also, the residual assessment of algorithm results will help us to understand the range in which the errors lie most of the time. Thus, Fig. 16 shows the residual assessment of developed algorithms for plastic viscosity prediction. Notice from plot (a) (which shows the GEP residuals) that most of the residuals are concentrated in the range -100 to 200 . Although there are some observations for which the residual is more than 300 Pa/s, but mostly the error remains within the range of -150 to 150 . It is a fairly larger range considering the importance of accurate predictions for plastic viscosity of concrete. In the next plot, the KNN residuals are depicted which shows a similar range of residuals with some residuals going as high as 500 Pa/s. However, in this plot too, the range

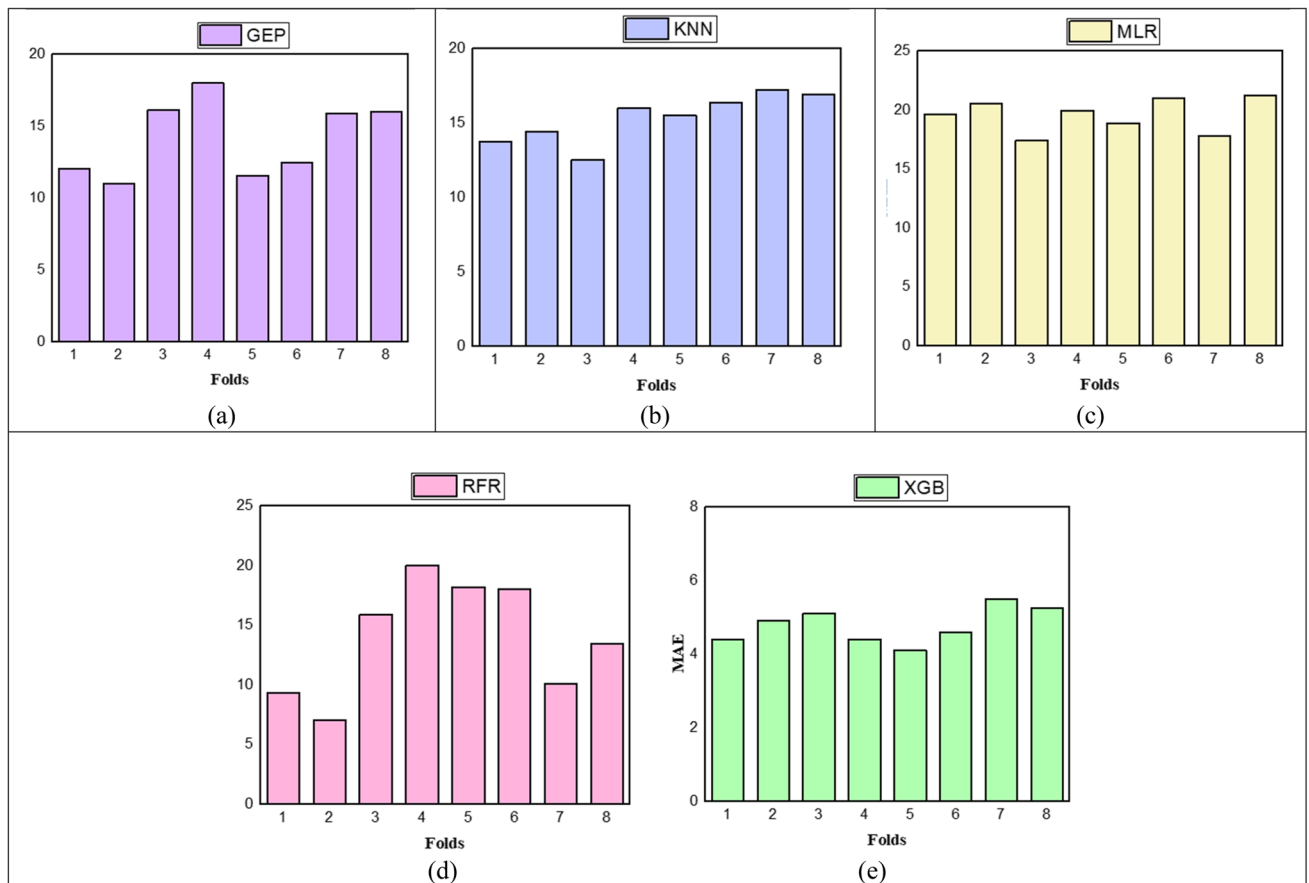


Figure 15. K-fold Validation results of algorithms for prediction of interface yield stress.

for existence of majority of the residual is -200 to 300. It is again a larger range, and it suggests that KNN fails to predict accurate values at various observations. Moving on, plot (3) for MLR shows a similar picture with residuals as high as 600 Pa/s. In contrast to GEP and KNN, larger number of residuals are lying beyond the range -400 to 400 suggesting that MLR has higher error rates. It can be attributed to the fact that MLR is a simple statistical technique which is unable to capture the complexity of relationships that exist between plastic viscosity and concrete mixture composition. Plot (d) developed using RFR residuals shows significant improvement. There are no errors going beyond -200 to 300 in the plot while it can be seen that most residuals are concentrated in the range of -100 to 100. It underlines that the accuracy of RFR is better than KNN, MLR, and GEP algorithms. However, it is overfitted to the training data (discussed previously) implying that it may give unexpected errors when tested on unseen data. At last, plot (e) shows the XGB residuals with an impressive improvement in the prediction accuracy. More than 95% of the XGB residuals lie in the narrowest range of -100 to 100. It shows that almost 96% of the times, the XGB error between actual and predicted plastic viscosity will not exceed 100 Pa/s. In addition, there is not a single residual going beyond the range -200 to 300 further highlighting XGB's superior performance as compared to other algorithms.

Figure 17 highlights the residual assessment of models developed for the purpose of prediction of interface yield stress of concrete. Starting from plot (a) of Fig. 17 for GEP algorithm, it can be seen that the range for all residuals is -30 to 50 while most residuals lie in the range -20 to 20 Pa. This suggests that GEP can give error up to 50 Pa when used to predict interface yield stress values. The KNN residuals are shown next in plot (b) which paints a similar picture to that of plot (a). Most of the residuals are concentrated between -20 Pa and 30 Pa while the maximum residual goes as high as 48 Pa. In plot (c), the range is increased for MLR with residuals lying in the widest range of -40 to 80. The maximum observed residual in case of MLR is almost 70 Pa which suggests a significant drop in accuracy compared to GEP and KNN algorithms. The RFR residuals in plot (d) of Fig. 17 highlights an interesting phenomenon. Although there is an improvement in accuracy because most of the residuals are lying from -10 Pa to 20 Pa. However, there are some residuals which go beyond this range to as large as 60 Pa. It shows that while 95% of the times the error will be around 10 Pa, there can be situations where RFR can yield larger than 50 Pa errors. This abnormal distribution of residuals can be attributed to the fact that RFR showed significant overfitting tendencies in the previous sections and this distribution is a result of that. Finally, the XGB residual assessment is given in plot (e) which highlights two important developments: (i) the range for errors is significantly reduced compared to previous algorithms and now the range is -7 to 7 compared to -30 to 50 for GEP, (ii) most of the errors lie really close to zero and only some errors go beyond the 10 Pa

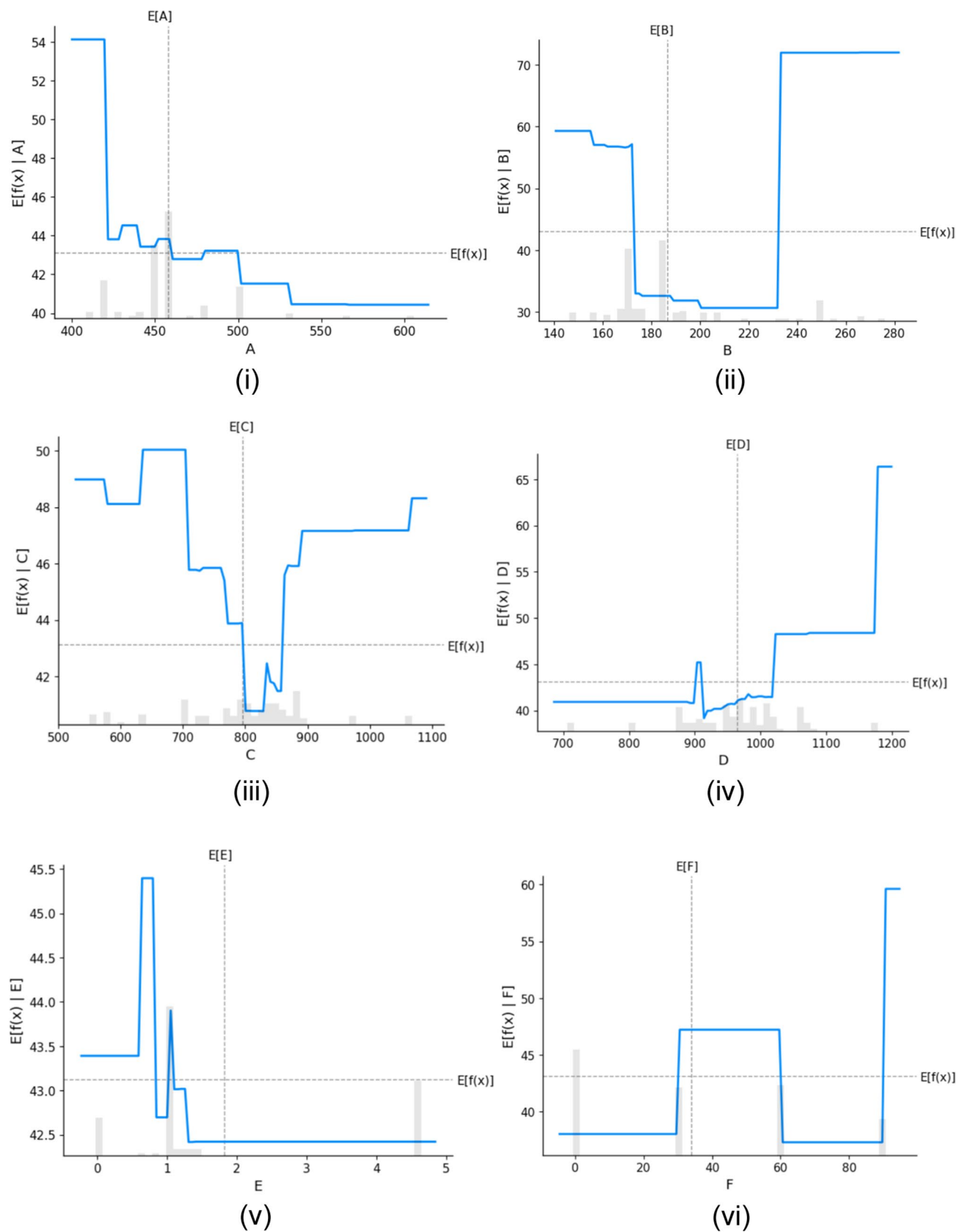


Figure 24. ICE analysis for investigating relationships between inputs and interface yield stress.

been identified by Ngo et al.³³. The amount of water in the concrete directly impacts the paste volume which in turn affects the plastic viscosity of concrete³⁹. Similarly, the influence of finer content (sand) and additives like superplasticizer on rheological concrete properties have been documented by various studies^{17,38,125}. Thus, it can be concluded that XGB is the most robust algorithm to predict fresh concrete properties not only because it depicted higher accuracy on training and testing datasets but also because the insights into the XGB prediction process provided by the SHAP, and ICE analyses are also the same as observed in experimental investigations.

Practical implications

It is important to highlight the practical implications of such empirical studies to further highlight their importance. It has already been discussed in [introduction](#) section that accurate determination of plastic viscosity and interface yield stress are crucial for determining rheology and pumpability of concrete. However, the accurate determination of these properties calls for significant time and resources on-site. Thus, this study was conducted to develop empirical prediction models for determining these two properties of concrete in fresh state. Also, with the recent surge in use of AI and ML in every aspect of the world, it is important to provide these types of frameworks through which the construction industry as a whole and the concrete industry in particular can make its transition towards informed decision-making techniques using the transformative power of ML. This study employed XGB, RFR, GEP, and KNN algorithms for the purpose of fresh properties prediction of concrete. Out of all these algorithms, only GEP expressed its output as an empirical equation for the prediction of subject outputs. However, the accuracy of GEP wasn't satisfactory compared to the other algorithms. Also, the developed equations are lengthy and computationally extensive which increases the risk of human error while calculating the values. In contrast, the accuracy of XGB was exceptionally good for the prediction of both outputs. However, its black-box nature prevented it from yielding an equation. Thus, to overcome these limitations, a GUI has been developed using python programming language to help compute the plastic viscosity and interface yield stress of concrete automatically based on the set of input variables considered in this as shown in [Fig. 25](#). The user can enter desired input values and upon pressing the predict button, result will be obtained instantly in the output box specified in the GUI. This GUI can serve as an invaluable tool for professionals in the civil engineering industry and different stakeholders to effectively predict concrete properties by entering the mixture composition without requiring extensive technical expertise. Also, with the increasing shift of civil engineering industry towards informed decision making, this GUI is an important tool which reflects the practical application of ML utilization in civil engineering industry. The accurate determination of these properties without going into extensive testing will help the workers to make informed decisions quickly on the site without waiting for the test results. Also, a lot of resources which would have otherwise wasted by conducting the tests on concrete can be saved by employing these ML models. In situations where concrete is used in bulk as batches, there can be several time, resource, and apparatus limitations which can be prevented by using these ML models for predicting concrete properties. The two outputs considered in this study i.e., plastic

GUI for Prediction of Plastic Viscosity and Interface Yield Stress of Concrete

Inputs

Cement (Kg)

410

Water (Kg)

176

Fine Aggregate (Kg)

865

Coarse Aggregate (Kg)

1005

Superplasticizer (Liters)

1

Time (Minutes)

0

Predict

Outputs

Plastic Viscosity (Pa/s)

308.2

Interface Yield Stress (Pa)

43.6

Figure 25. GUI developed for prediction of outputs.

viscosity and interface yield stress are of particular importance in situations where concrete is subjected to shear forces such as during pumping, placement, and casting etc. Thus, this study holds paramount importance in this regard that it provides a framework to accurately determine these properties using the transformative power of ML. The predictive models in this study can be effectively utilized to determine these properties on site which will save time and help in making swift decisions.

Conclusions

This study was conducted to develop ML-based prediction models for yield stress and plastic viscosity of concrete since efficient pumpability of concrete through pumps for high-rise construction requires accurate estimation of these concrete properties. Four ML algorithms including GEP, XGB, RFR and KNN along with a statistical technique i.e., MLR were used for this purpose using a database of 142 instances and the conclusions of this study were:

- Among all the algorithms employed in this study, only GEP expressed its output in the form of an empirical equation to predict plastic viscosity and interface yield stress.
- The deployed algorithms were validated using error metrics, k-fold validation, and residual assessment which proved that XGB is the most accurate algorithm to predict the subject outputs while GEP and KNN had marginal accuracies. The algorithm comparison by means of Taylor plots also confirmed this.
- The results of error evaluation showed that XGB is the most accurate algorithm to predict both outputs having R^2 value greater than 0.92 and minimal values of other error metrics like objective function and MAE etc.
- SHAP and ICE analysis were conducted on the XGB algorithm which revealed that time after mixing, water, and cement etc. are the most important features to predict both fresh concrete properties.

Recommendations

Despite the fact that this study is a significant contribution to the literature and a crucial step towards prediction of critical rheological properties of concrete using diverse ML techniques, it is important to highlight its limitations and provide suggestions for future studies:

- The ML models presented in current study were developed using a limited dataset of 142 experimental collected from literature. For the development of more reliable, robust, and widely applicable models, it is recommended to consider much larger datasets in future studies.
- The output variables in this study i.e., plastic viscosity and interface yield stress can be affected by different material properties like nominal aggregate diameter, type of superplasticizer and variable site conditions like mixing and pumping temperature etc. This study doesn't exclusively incorporate the effect of these variables but recommends considering them in future studies.
- The models presented in this study were developed using data whose range is given in Table 1. For the development of more widely applicable models, it is necessary to consider a wider range of data.
- Other widely used interpretable techniques like partial dependence plots (PDP) and permutation importance analysis etc. must be considered in future studies to reveal more about the nature of black-box models.

Data availability

Data is available in the supplementary materials.

Received: 30 January 2024; Accepted: 22 October 2024

Published online: 28 March 2025

References

1. Yücel, K. T. Examination of behavior of fresh concrete under pressure. *Int. J. Thermophys.* **33**(5), 885–894. <https://doi.org/10.1007/S10765-012-1180-6> (2012).
2. Wu, Y. et al. A Study on the Ultimate Span of a Concrete-Filled Steel Tube Arch Bridge. *Buildings*. **14**(4), 896. <https://doi.org/10.3390/BUILDINGS14040896> (2024).
3. Chen, R. S. et al. Experimental study on ultimate bearing capacity of short thin-walled steel tubes reinforced with high-ductility concrete. *Structures* **68**, <https://doi.org/10.1016/j.istruc.2024.107109> (2024).
4. Liu, C. et al. The role of TBM asymmetric tail-grouting on surface settlement in coarse-grained soils of urban area: Field tests and FEA modelling. *Tunn. Undergr. Space Technol.* **111**, <https://doi.org/10.1016/j.tust.2021.103857> (2021).
5. He, H. et al. Exploring green and efficient zero-dimensional carbon-based inhibitors for carbon steel: From performance to mechanism. *Constr. Build. Mater.* **411**, <https://doi.org/10.1016/j.conbuildmat.2023.134334> (2024).
6. Liu, X., Liu, X. & Zhang, Z. Application of red mud in carbon capture, utilization and storage (CCUS) technology, Sep. 01, Elsevier Ltd. doi: (2024). <https://doi.org/10.1016/j.rser.2024.114683>
7. Yao, X. et al. AI-based performance prediction for 3D-printed concrete considering anisotropy and steam curing condition. *Constr. Build. Mater.* **375**, <https://doi.org/10.1016/j.conbuildmat.2023.130898> (2023).
8. He, H., Qiao, H., Sun, T., Yang, H. & He, C. Research progress in mechanisms, influence factors and improvement routes of chloride binding for cement composites. *J. Building Eng.* **86**, <https://doi.org/10.1016/j.jobe.2024.108978> (2024).
9. Fan, S. et al. Machine learning-based classification of quality grades for concrete vibration behaviour. *Autom. Constr.* **167**, <https://doi.org/10.1016/j.autcon.2024.105694> (2024).
10. Aldred, J. Burj Khalifa - A new high for high- Performance concrete. *Proc Inst Civil Eng Civil Eng* **163**(2), 66–73. <https://doi.org/10.1680/cien.2010.163.2.66> (2010).
11. Hu, W., Qin, K., Lu, F., Li, D. & Cohen, J. B. Merging TROPOMI and eddy covariance observations to quantify 5-years of daily CH₄ emissions over coal-mine dominated region. *Int J Coal Sci Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00700-1> (2024).
12. Hazaree, C. & Mahadevan, V. Single stage concrete pumping through 2.432 km (1.51 miles): Weather and execution challenges. *Case Stud. Constr. Mater.* **3**, 56–69. <https://doi.org/10.1016/j.cscm.2015.05.001> (2015).

13. Wei, J. et al. Seismic performance of concrete-filled steel tubular composite columns with ultra high performance concrete plates. *Eng. Struct.* **278**, <https://doi.org/10.1016/J.ENGSTRUCT.2022.115500> (2023).
14. Tan, L. et al. Moisture penetration and distribution characterization of hard coal: a μ -CT study. *Int. J. Coal Sci. Technol.* **11** (1), <https://doi.org/10.1007/S40789-024-00713-W> (2024).
15. Modiga, A., Eterigho-Ikelegbe, O. & Bada, S. Extractability and mineralogical evaluation of rare earth elements from Waterberg Coalfield run-of-mine and discard coal. *Int. J. Coal Sci. Technol.* **11**(1), 65. <https://doi.org/10.1007/S40789-024-00702-Z> (2024).
16. Secrieru, E., Khodor, J., Schröfl, C. & Mechtcherine, V. Formation of lubricating layer and flow type during pumping of cement-based materials. *Constr. Build. Mater.* **178**, 507–517. <https://doi.org/10.1016/j.conbuildmat.2018.05.118> (2018).
17. Feys, D., Khayat, K. H., Perez-Schell, A. & Khatib, R. Prediction of pumping pressure by means of new tribometer for highly-workable concrete. *Cem. Concr Compos.* **57**, 102–115. <https://doi.org/10.1016/j.cemconcomp.2014.12.007> (2015).
18. Zhao, R., Li, C. & Guan, X. Advances in Modeling Surface Chloride Concentrations in Concrete Serving in the Marine Environment: A Mini Review. *Multidiscip Digital Publishing Inst (MDPI)* <https://doi.org/10.3390/buildings14061879> (2024).
19. Jang, K. P. et al. Experimental Observation on Variation of Rheological Properties during Concrete Pumping. *Int J Concr Struct Mater.* **12**(1), <https://doi.org/10.1186/s40069-018-0310-3> (2018).
20. Li, X. Coupling mechanism of dissipated energy–infrared radiation energy of the deformation and fracture of composite coal-rock under load, *ACS PublicationsX Li, H Li, Z Yang, H Zuo, W Sun, H Li, Y LiACS omega, 2022 ACS Publications*, 7 (9), 8060–8076 (2022). <https://doi.org/10.1021/acsomega.1c07289>
21. Soleimani, F., Si, G., Roshan, H. & Zhang, J. Numerical modelling of gas outburst from coal: a review from control parameters to the initiation process. *Int. J. Coal Sci. Technol.* **10**(1), <https://doi.org/10.1007/S40789-023-00657-7> (2023).
22. Secrieru, E., Fataei, S., Schröfl, C. & Mechtcherine, V. Study on concrete pumpability combining different laboratory tools and linkage to rheology. *Constr. Build. Mater.* **144**, 451–461. <https://doi.org/10.1016/j.conbuildmat.2017.03.199> (2017).
23. Nguyen, T. D., Tran, T. H. & Hoang, N. D. Prediction of interface yield stress and plastic viscosity of fresh concrete using a hybrid machine learning approach. *Adv. Eng. Inform.* **44**, <https://doi.org/10.1016/j.aei.2020.101057> (2020).
24. Choi, M. S., Kim, Y. J. & Kwon, S. H. Prediction on pipe flow of pumped concrete based on shear-induced particle migration. *Cem. Concr Res.* **52**, 216–224. <https://doi.org/10.1016/J.CEMCONRES.2013.07.004> (2013).
25. Kwon, S. H., Park, C. K., Jeong, J. H., Jo, S. D. & Lee, S. H. Prediction of concrete pumping: Part i-development of new tribometer for analysis of lubricating layer. *ACI Mater J.* **110**(6), 647–655. <https://doi.org/10.14359/51686332> (2013).
26. Zhang, W., Zhang, S., Wei, J. & Huang, Y. Flexural behavior of SFRC-NC composite beams: An experimental and numerical analytical study. *Structures* **60**, <https://doi.org/10.1016/J.ISTJUC.2023.105823> (2024).
27. Lu, H. et al. Study of synergistic behavior during bituminous coal-cow manure co-gasification: The role of intrinsic AAEM and organic matter. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00694-W> (2024).
28. Zhang, Y. et al. Research on coal-rock identification method and data augmentation algorithm of comprehensive working face based on FL-Segformer. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00704-X> (2024).
29. Sun, M. et al. The release and migration mechanism of arsenic during pyrolysis process of Chinese coals. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00715-8> (2024).
30. Amoi, F. et al. Comparative study on different coals from the Lorraine basin (France) by sorption isotherms, thermogravimetric analysis and breakthrough curves for CO₂-ECBM recovery. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00696-8> (2024).
31. Li, W., Wang, C., Yu, Z. & Song, Y. Catalytic coal gasification: mechanism, kinetics, and reactor model. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00712-X> (2024).
32. He, R., Huang, X., Zhang, J., Geng, Y. & Guo, H. Preparation and evaluation of exhaust-purifying cement concrete employing titanium dioxide. *Materials*. **12**(13), <https://doi.org/10.3390/ma12132182> (2019).
33. Ngo, T. T., Kadri, E. H., Bennacer, R. & Cussigh, F. Use of tribometer to estimate interface friction and concrete boundary layer composition during the fluid concrete pumping. *Constr. Build. Mater.* **24**(7), 1253–1261 (2010).
34. Choi, M., Roussel, N., Kim, Y. & Kim, J. Lubrication layer properties during concrete pumping. *Cem Concr Res.* **45**(1), 69–78. <https://doi.org/10.1016/J.CEMCONRES.2012.11.001> (2013).
35. Mai, C. T., Kadri, E. H., Ngo, T. T., Kaci, A. & Riche, M. Estimation of the pumping pressure from concrete composition based on the identified tribological parameters. *Adv. Mater. Sci. Eng.* **2014**(1), 503850 (2014).
36. Sun, L. et al. Experimental investigation on the bond performance of sea sand coral concrete with FRP bar reinforcement for marine environments. *Adv Struct Eng.* **26**(3), 533–546. <https://doi.org/10.1177/13694332221131153> (2023).
37. Huang, H., Yao, Y., Liang, C. & Ye, Y. Experimental study on cyclic performance of steel-hollow core partially encased composite spliced frame beam. *Soil Dyn. Earthq. Eng.* **163**, <https://doi.org/10.1016/J.SOILDYN.2022.107499> (2022).
38. Dong, J. F., Wang, Q. Y. & Guan, Z. W. Structural behaviour of recycled aggregateconcrete filled steel tube columns strengthened by CFRP. *Eng. Struct.* **48**, 532–542 (2013).
39. Ngo, T. *Influence de la composition des bétons sur les paramètres de pompage et validation d'un modèle de prévision de la constante visqueuse* (2009) Accessed: 01 Jan 2024. Available: <https://www.theses.fr/2009CERG0404>.
40. Song, X. et al. Data-driven modeling for residual velocity of projectile penetrating reinforced concrete slabs. *Eng. Struct.* **306**, <https://doi.org/10.1016/j.engstruct.2024.117761> (2024).
41. Zhou, C. et al. Study on signal characteristics of burst tendency coal under different loading rates. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00724-7> (2024).
42. Wu, L. & Wang, J. Infrared radiation features of coal and rocks under loading. *Int. J. Rock Mech. Min. Sci.* **35**(7), 969–976 (1998).
43. Thiruchittampalam, S., Singh, S. K., Banerjee, B. P., Glenn, N. F. & Raval, S. Spoil characterisation using UAV-based optical remote sensing in coal mine dumps. *Int. J. Coal Sci. Technol.* **10**(1), <https://doi.org/10.1007/S40789-023-00622-4> (2023).
44. Choi, M. S., Kim, Y. J., Jang, K. P. & Kwon, S. H. Effect of the coarse aggregate size on pipe flow of pumped concrete. *Constr Build Mater.* **66**, 723–730. <https://doi.org/10.1016/J.CONBUILDMAT.2014.06.027> (2014).
45. Huang, H., Huang, M., Zhang, W. & Yang, S. Experimental study of predamaged columns strengthened by HPFL and BSP under combined load cases. *Struct. Infrastruct. Eng.* **17** (9), 1210–1227. <https://doi.org/10.1080/15732479.2020.1801768> (2021).
46. Lu, S. et al. A review of coal permeability models including the internal swelling coefficient of matrix. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00701-0> (2024).
47. Ferraris, C. F. & de Larrard, F. Testing and modelling of fresh concrete rheology. (1998). <https://doi.org/10.6028/NIST.IR.6094>
48. Secrieru, E., Mohamed, W., Fataei, S. & Mechtcherine, V. Assessment and prediction of concrete flow and pumping pressure in pipeline. *Cem. Concr Compos.* **107**, <https://doi.org/10.1016/J.CEMCONCOMP.2019.103495> (2020).
49. Liu, W., Zhou, H., Zhang, S. & Zhao, C. Variable Parameter Creep Model Based on the Separation of Viscoelastic and Viscoplastic Deformations. *Rock Mech Rock Eng.* **56**(6), 4629–4645. <https://doi.org/10.1007/S00603-023-03266-7> (2023).
50. He, L. et al. A quasi-exponential distribution of interfacial voids and its effect on the interlayer strength of 3D printed concrete. *Addit. Manuf.* **89**, <https://doi.org/10.1016/j.addma.2024.104296> (2024).
51. Mahmood, M. S. et al. Enhancing compressive strength prediction in self-compacting concrete using machine learning and deep learning techniques with incorporation of rice husk ash and marble powder. *Case Stud. Constr. Mater.* **19**, <https://doi.org/10.1016/j.cscm.2023.e02557> (2023).
52. Wasim, M., Duc Ngo, T. & Law, D. Durability performance of reinforced waste-based geopolymer foam concrete under exposure to various corrosive environments. *Case Stud. Constr. Mater.* **15**, <https://doi.org/10.1016/j.cscm.2021.e00703> (2021).

53. Chen, Z. et al. Strength evaluation of eco-friendly waste-derived self-compacting concrete via interpretable genetic-based machine learning models. *Mater. Today Commun.* **37**, <https://doi.org/10.1016/j.mtcomm.2023.107356> (2023).
54. Iqbal, M. F. et al. Sustainable utilization of foundry waste: Forecasting mechanical properties of foundry sand based concrete using multi-expression programming. *Sci. Total Environ.* **780**, <https://doi.org/10.1016/j.scitotenv.2021.146524> (2021).
55. Choudhary, R., Gupta, R. & Nagar, R. Impact on fresh, mechanical, and microstructural properties of high strength self-compacting concrete by marble cutting slurry waste, fly ash, and silica fume. *Constr. Build. Mater.* **239**, <https://doi.org/10.1016/j.conbuildmat.2019.117888> (2020).
56. Liu, P. et al. Image-based quantitative probing of 3D heterogeneous pore structure in CBM reservoir and permeability estimation with pore network modeling. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00722-9> (2024).
57. Han, Z., Liu, K., Ma, J. & Li, D. Numerical simulation on the dynamic mechanical response and fracture mechanism of rocks containing a single hole. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00718-5> (2024).
58. He, J., Serati, M., Veidt, M. & De Alwis, A. Determining rock crack stress thresholds using ultrasonic through-transmission measurements. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-024-00669-X> (2024).
59. Qin, S., Lin, H., Yang, S. & Wei, Z. A mathematical model for parameter setting in discrete element numerical simulation. *Int. J. Coal Sci. Technol.* **10**(1), <https://doi.org/10.1007/S40789-023-00644-Y> (2023).
60. Yin, J. J., Lei, J., Fan, K. & Wang, S. Integrating image processing and deep learning for effective analysis and classification of dust pollution in mining processes. *Int. J. Coal Sci. Technol.* **10**(1), <https://doi.org/10.1007/S40789-023-00653-X> (2023).
61. Yang, Y. et al. Simulation study of hydrogen sulfide removal in underground gas storage converted from the multilayered sour gas field. *Int. J. Coal Sci. Technol.* **10**(1), <https://doi.org/10.1007/S40789-023-00631-3> (2023).
62. He, S., Qin, M., Qiu, L., Song, D. & Zhang, X. Early warning of coal dynamic disaster by precursor of AE and EMR 'quiet period'. *Int. J. Coal Sci. Technol.* **9**(1), <https://doi.org/10.1007/S40789-022-00514-Z> (2022).
63. Zhang, B., Yang, K., Zhang, K., Wang, Q. & Wu, N. Migration transformation, prevention, and control of typical heavy metal lead in coal gangue: a review. *Int. J. Coal Sci. Technol.* **10**(1), <https://doi.org/10.1007/S40789-023-00656-8> (2023).
64. Gao, X. et al. Three dimensional discrete element modelling of the conventional compression behavior of gas hydrate bearing coal. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/S40789-023-00639-9> (2024).
65. He, L. et al. Development of novel concave and convex trowels for higher interlayer strength of 3D printed cement paste. *Case Stud. Constr. Mater.* **21**, <https://doi.org/10.1016/j.cscm.2024.e03745> (2024).
66. Shahin, M., Jaksa, M. B. & Maier, H. R. Physical Modeling of Rolling Dynamic Compaction View project Artificial neural networks-pile capacity prediction View project. (2008) <https://www.researchgate.net/publication/228364758>.
67. Khan, M. A., Zafar, A., Akbar, A., Javed, M. F. & Mosavi, A. Application of gene expression programming (GEP) for the prediction of compressive strength of geopolymer concrete. *Materials* **14**(5), 1–23. <https://doi.org/10.3390/ma14051106> (2021).
68. Gandomi, A. H., Alavi, A. H., Mirzahosseini, M. R. & Nejad, F. M. Nonlinear Genetic-Based Models for Prediction of Flow Number of Asphalt Mixtures. *J. Mater. Civil Eng.* **23**(3), 248–263. [https://doi.org/10.1061/\(asce\)mt.1943-5533.0000154](https://doi.org/10.1061/(asce)mt.1943-5533.0000154) (2011).
69. Koza, J. R. & Jacks Hall, M. Survey of genetic algorithms and genetic programming. <http://www-cs-faculty.stanford.edu/~koza/>.
70. Huang, H., Huang, M., Zhang, W., Guo, M. & Liu, B. Progressive collapse of multistory 3D reinforced concrete frame structures after the loss of an edge column. *Struct. Infrastruct. Eng.* **18** (2), 249–265. <https://doi.org/10.1080/15732479.2020.1841245> (2022).
71. Nazar, S. et al. Machine learning interpretable-prediction models to evaluate the slump and strength of fly ash-based geopolymer. *J. Mater. Res. Technol.* **24**, 100–124. <https://doi.org/10.1016/j.jmrt.2023.02.180> (2023).
72. Gholampour, A., Gandomi, A. H. & Ozbakkaloglu, T. New formulations for mechanical properties of recycled aggregate concrete using gene expression programming. *Constr. Build. Mater.* **130**, 122–145. <https://doi.org/10.1016/j.conbuildmat.2016.10.114> (2017).
73. Huang, H. et al. Numerical investigation on the bearing capacity of RC columns strengthened by HPFL-BSP under combined loadings. *J. Building Eng.* **39**, <https://doi.org/10.1016/J.JOBE.2021.102266> (2021).
74. Wu, X. et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14** (1), 1–37. <https://doi.org/10.1007/s10115-007-0114-2> (2008).
75. Qian, Y., Zhou, W., Yan, J., Li, W. & Han, L. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sens. (Basel)* **7** (1), 153–168. <https://doi.org/10.3390/rs70100153> (2015).
76. Akbulut, Y., Sengur, A., Guo, Y. & Smarandache, F. NS-k-NN: Neutrosophic set-based k-nearest neighbors classifier. *Symmetry (Basel)* **9**(9), <https://doi.org/10.3390/sym9090179> (2017).
77. Huang, H., Yuan, Y., Zhang, W. & Li, M. Seismic behavior of a replaceable artificial controllable plastic hinge for precast concrete beam-column joint. *Eng. Struct.* **245**, <https://doi.org/10.1016/J.ENGSTRUCT.2021.112848> (2021).
78. Huang, F. et al. Slope stability prediction based on a long short-term memory neural network: comparisons with convolutional neural networks, support vector machines and random forest models. *Int. J. Coal Sci. Technol.* **10**(1), 1–14. <https://doi.org/10.1007/S40789-023-00579-4/FIGURES/5> (2023).
79. Bin Inqiad, W. et al. Comparative analysis of various machine learning algorithms to predict 28-day compressive strength of Self-compacting concrete. *Heliyon* **9**(11), e22036. <https://doi.org/10.1016/j.heliyon.2023.e22036> (2023).
80. Shanmugasundar, G. et al. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. *Processes* **9**(11), <https://doi.org/10.3390/pr9112015> (2021).
81. Li, H., Yang, Y., Wang, X. & Tang, H. Effects of the position and chloride-induced corrosion of strand on bonding behavior between the steel strand and concrete. *Structures* **58**, <https://doi.org/10.1016/J.ISTRUC.2023.105500> (2023).
82. Jin, X. et al. Estimation of Wheat Agronomic Parameters using New Spectral Indices. *PLoS One* **8**(8), <https://doi.org/10.1371/journal.pone.0072736> (2013).
83. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery* 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>.
84. Qu, Y., Lin, Z., Li, H. & Zhang, X. Feature Recognition of Urban Road Traffic accidents based on GA-XGBoost in the context of Big Data. *IEEE Access* **7**, 170106–170115. <https://doi.org/10.1109/ACCESS.2019.2952655> (2019).
85. Xu, H. et al. Identifying diseases that cause psychological trauma and social avoidance by GCN-Xgboost. *BMC Bioinform.* **21**, <https://doi.org/10.1186/s12859-020-03847-1> (2020).
86. Huang, H., Li, M., Zhang, W. & Yuan, Y. Seismic behavior of a friction-type artificial plastic hinge for the precast beam-column connection. *Archives Civil Mech. Eng.* **22**(4), <https://doi.org/10.1007/S43452-022-00526-1> (2022).
87. Cui, L., Chen, P., Wang, L., Li, J. & Ling, H. Application of Extreme Gradient Boosting Based on Grey Relation Analysis for Prediction of Compressive Strength of Concrete. *Adv Civil Eng* **2021**, <https://doi.org/10.1155/2021/8878396> (2021).
88. Wang, C., Xu, S. & Yang, J. Adaboost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement. *Sensors* **21**(17), <https://doi.org/10.3390/s21175682> (2021).
89. Al-Taai, S. R. et al. XGBoost Prediction Model Optimized with Bayesian for the Compressive Strength of Eco-Friendly Concrete Containing Ground Granulated Blast Furnace Slag and Recycled Coarse Aggregate. *Appl Sci (Switzerland)* **13**(15), <https://doi.org/10.3390/app13158889> (2023).
90. Ma, X., Dong, Z., Quan, W., Dong, Y. & Tan, Y. Real-time assessment of asphalt pavement moduli and traffic loads using monitoring data from Built-in Sensors: Optimal sensor placement and identification algorithm. *Mech Syst Signal Process.* **187**, 109930. <https://doi.org/10.1016/J.YMSSP.2022.109930> (2023).

91. Huang, H., Li, M., Yuan, Y. & Bai, H. Experimental Research on the Seismic Performance of Precast Concrete Frame with Replaceable Artificial Controllable Plastic Hinges. *J Struct Eng* **149**(1), 04022222. <https://doi.org/10.1061/JSENDH.STENG-11648> (2022).
92. Sarveghadi, M., Gandomi, A. H., Bolandi, H. & Alavi, A. H. *Development of prediction models for shear strength of SFRCB using a machine learning approach* (Springer, 2019). <https://doi.org/10.1007/s00521-015-1997-6>
93. Iftikhar, B. et al. Predictive modeling of compressive strength of sustainable rice husk ash concrete: Ensemble learner optimization and comparison. *J. Clean. Prod.* **348**, <https://doi.org/10.1016/j.jclepro.2022.131285> (2022).
94. Zhang, J. & Zhang, C. Using viscoelastic materials to mitigate earthquake-induced pounding between adjacent frames with unequal height considering soil-structure interactions. *Soil Dyn. Earthq. Eng.* **172**, <https://doi.org/10.1016/j.soildyn.2023.107988> (2023).
95. Inqiad, W. B., Dumitrascu, E. V., & Dobre, R. A. Forecasting residual mechanical properties of hybrid fibre-reinforced Self-compacting concrete (HFR-SCC) exposed to elevated temperatures. *Heliyon* **10**(12) (2024).
96. Inqiad, W. B., Dumitrascu, E. V., Dobre, R. A., Khan, N. M., Hammood, A. H., Henedy, S. N., & Khan, R. M. A. Predicting compressive strength of hollow concrete prisms using machine learning techniques and explainable artificial intelligence (XAI). *Heliyon* **10**(17) (2024).
97. Inqiad, W. B. et al. Predicting natural vibration period of concrete frame structures having masonry infill using machine learning techniques. *J. Build. Eng.* **96**, 110417 (2024).
98. Bardenet, R., Brendel, M., Kégl, B., Sebag, M. & Fr, S. *Collaborative hyperparameter tuning*. (2013).
99. Belete, D. M. & Huchaiah, M. D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **44** (9), 875–886. <https://doi.org/10.1080/1206212X.2021.1974663> (2022).
100. Inqiad, W. B., Siddique, M. S., Ali, M. & Najeh, T. Predicting 28-day compressive strength of fibre-reinforced self-compacting concrete (FR-SCC) using MEP and GEP. *Sci. Rep.* **14**(1), 17293 (2024).
101. Pedregosa, F. et al. FABIANPEDREGOSA Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. (2011). <http://scikit-learn.sourceforge.net>
102. Ghawi, R. & Pfeffer, J. Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity. *Open Computer Sci.* **9**(1), 160–180. <https://doi.org/10.1515/comp-2019-0011> (2019).
103. Inqiad, W. B., Javed, M. F., Siddique, M. S., Alarifi, S. S. & Alabduljabbar, H. A comparative analysis of boosting and genetic programming techniques for predicting mechanical properties of soilcrete materials. *Mater. Today Commun.* **40**, 109920 (2024).
104. Rostami, A. et al. Rigorous framework determining residual gas saturations during spontaneous and forced imbibition using gene expression programmin. *J Nat Gas Sci Eng.* **84**, <https://doi.org/10.1016/j.jngse.2020.103644> (2020).
105. Alavi, A. H., Gandomi, A. H., Sahab, M. G. & Gandomi, M. Multi expression programming: A new approach to formulation of soil classification. *Eng Comput.* **26**(2), 111–118. <https://doi.org/10.1007/s00366-009-0140-7> (2010).
106. Jalal, F. E., Xu, Y., Iqbal, M., Javed, M. F. & Jamhiri, B. Predictive modeling of swell-strength of expansive soils using artificial intelligence approaches: ANN, ANFIS and GEP. *J. Environ. Manage.* **289**, <https://doi.org/10.1016/j.jenvman.2021.112420> (2021).
107. Despotovic, M., Nedic, V., Despotovic, D. & Cvetanovic, S. Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. *Renew. Sustain. Energy Rev.* **56**, 246–260. <https://doi.org/10.1016/j.rser.2015.11.058> (2016).
108. Inqiad, W. B. et al. Comparison of boosting and genetic programming techniques for prediction of tensile strain capacity of Engineered Cementitious Composites (ECC). *Mater. Today Commun.* **39**, 109222 (2024).
109. Armaghani, D. J. & Asteris, P. G. A comparative study of ANN and ANFIS models for the prediction of cement-based mortar materials compressive strength. *Neural Comput. Appl.* **33**(9), 4501–4532. <https://doi.org/10.1007/s00521-020-05244-4> (2021).
110. Dong, J. et al. Mechanical behavior and impact resistance of rubberized concreteenhanced by basalt fiber-epoxy resin composite. *Constr. Build. Mater.* **435**, 136836 (2024).
111. Gomah, M. E. et al. Prediction of Strength Parameters of Thermally Treated Egyptian Granodiorite Using Multivariate Statistics and Machine Learning Techniques. *Mathematics* **10**(23), <https://doi.org/10.3390/math10234523> (2022).
112. Bin Inqiad, W. et al. Soft computing models for prediction of bentonite plastic concrete strength. *Nature. Com* **14**, 18145–123AD. <https://doi.org/10.1038/s41598-024-69271-0>.
113. Taylor, K. E. Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Atmospheres* **106**(D7), 7183–7192. <https://doi.org/10.1029/2000JD900719> (2001).
114. Sagi, O. & Rokach, L. Explainable decision forest: transforming a decision forest into an interpretable tree. *Inform. Fusion.* **61**, 124–138. <https://doi.org/10.1016/j.inffus.2020.03.013> (2020).
115. Gao, Q. et al. Method for rock fracture prediction and early warning: Insight from fusion of multi-physics field information. *Heliyon* **10**(10), <https://doi.org/10.1016/j.heliyon.2024.e30660> (2024).
116. Thisovithan, P., Aththanayake, H., Meddage, D. P. P., Ekanayake, I. U. & Rathnayake, U. A novel explainable AI-based approach to estimate the natural period of vibration of masonry infill reinforced concrete frame structures using different machine learning techniques. *Results Eng.* **19**, <https://doi.org/10.1016/j.rineng.2023.101388> (2023).
117. Cakiroglu, C., Aydın, Y., Bekdaş, G. & Geem, Z. W. Interpretable Predictive Modelling of Basalt Fiber Reinforced Concrete Splitting Tensile Strength Using Ensemble Machine Learning Methods and SHAP Approach. *Materials* **16**(13), <https://doi.org/10.3390/ma16134578> (2023).
118. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation (2013). Available: <http://arxiv.org/abs/1309.6392>.
119. Alyami, M. et al. Application of metaheuristic optimization algorithms in predicting the compressive strength of 3D-printed fiber-reinforced concrete. *Developments Built Environ.* **17**, 100307. <https://doi.org/10.1016/j.dibe.2023.100307> (2024).
120. Bin Inqiad, W. et al. Comparative analysis of various machine learning algorithms to predict 28-day compressive strength of Self-compacting concrete. *cell.com* **9**, e22036. <https://doi.org/10.1016/j.heliyon.2023.e22036> (2023).
121. Banerjee, R., Chakladar, S., Kumar, A., Chattopadhyay, S. K. & Chakravarty, S. Petrology and association of rare earth elements in magmatically altered high-ash coal of Indian origin. *Int. J. Coal Sci. Technol.* **11**(1), <https://doi.org/10.1007/s40789-024-00709-6> (2024).
122. Ekanayake, I. U., Meddage, D. P. P. & Rathnayake, U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud Construct Mater* **16**, <https://doi.org/10.1016/j.cscm.2022.e01059> (2022).
123. Soualhi, H. et al. Design of portable rheometer with new vane geometry to estimate concrete rheological parameters. *J Civil Eng Manag.* **23**(3), 347–355. <https://doi.org/10.3846/13923730.2015.1128481> (2017).
124. Feys, D. et al. Measuring rheological properties of cement pastes: Most common techniques, procedures and challenges. *RILEM Tech Lett* **2**, 129–135. <https://doi.org/10.21809/rilemtechlett.2017.43> (2017).
125. Chapdelaine, F. & Étude fondamentale et pratique sur le pompage du béton. Available: www.GenieCivilPDF.com.

Acknowledgements

The authors would like to acknowledge the support of Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

All authors contributed equally.

Funding

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-77490-8>.

Correspondence and requests for materials should be addressed to M.F.J., D.M.A. or M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024