**Development of Prediction Model for Strength Properties of Concrete using Gene Expression Programming**

**Abstract**

The prediction models developed for the mixture proportion of conventional concrete are mostly established using traditional regression models. However, these traditional modelling techniques showed less accuracy in predicting the strength characteristics of concrete. This lack of accurate predictive empirical models led to the demand for deploying modern machine learning (ML). In the current study, a novel ML technique known as Gene Expression Programming (GEP) is utilized to develop prediction models for dynamic modulus (*DM*), modulus of elasticity (*E*), and modulus of rupture (*MOR*) of conventional concrete. For this purpose, a total of 440 values for *E* and *MOR* and 279 values for *DM* were collected from the literature. Five input parameters such as water-to-cement ratio (w:cm), air entraining admixture (AEA), water reducing admixture (WRA), slump, and age for developing the GEP prediction model. The accuracy and validity of the proposed models were assessed via different statistical metrics and external validation conditions. The proposed models showed higher accuracy with significant levels of R i.e., 0.96, 0.98, and 0.97 for *MOR*, *E*, and *DM*, respectively. Furthermore, SHAP analysis was carried out to get an insight into the input parameter's influence and importance to response concrete strength characteristics. It revealed that age significantly influenced all three considered strength properties and higher age enhances these concrete properties. Moreover, the developed ML model was compared with the regression model and it was observed that the ML model has greater accuracy as compared to traditional regression models. In conclusion, the proposed model will help the designer and builder to avoid costly and time-consuming laboratory testing.

**Keywords:** concrete, strength properties, artificial intelligence, machine learning, SHAP analysis

1. **Introduction**

Concrete is the most often used construction material in the world. Due to frequent growth in population and urbanization, Concrete demand is estimated to reach 18 billion by 2050 [1]–[3]. A deeper understanding of concrete performance is required, to enhance the design of concrete structures, based on the correct measurement of its mechanical and physical properties. To determine the mechanical characteristic of concrete conventional testing have frequently utilizes which are costly, laborious, and time-consuming [4]. Furthermore, Several statistical and mathematical techniques based on empirical formulas have been developed to forecast these properties using the composition of materials as inputs [5]–[7]. With progress in the artificial intelligence field, numerous modelling approaches have been applied to evaluate the mechanical characteristics of concrete using prediction models. However, artificial intelligence (AI) algorithms such as machine learning (ML) models, are more reliable and accurate compared to mathematical and statistical models, for the prediction of concrete properties [8]. ML is focused on the evaluation of computer innovation that learn from past data to make projections for the future. ML techniques have applications everywhere in today's world, specifically in Civil Engineering including anticipation of risk at worksites using image detection, forecasting the rate of change in the cost materials, optimizing mining and building operations, and so on [9]–[11]. Forecasting the characteristics of materials using ML models is one of the most useful applications [12]. The higher capacity of ML models to predict material characteristics considering its physical compositions and input descriptors might reduce the need for physical testing, reducing time and money spent on casting and testing of materials.

Many studies have utilized various ML methods for the prediction of *CS* such as decision trees (DT), artificial neural networks (ANN), GEP, and support vector machines (SVM) [13] [14], [15]. ANN has been used to forecast the characteristics of materials for more than 21 years [16]. however, limited work has been published to forecast mechanical characteristics such as modulus of rupture (*MOR*), modulus of elasticity (*E*), and dynamic modulus (*DM*) using ML models, and the genetic programming (GP) approach has been examined for prediction of concrete properties very infrequently. Calculating the *CS* of concrete is comparatively easy than forecasting these characteristics because the relationship between these characteristics and the composition of concrete materials is significantly non-linear. As a result, more advanced and unusual methods for forecasting concrete characteristics are required [17]. Ensemble techniques are a method for creating complicated algorithms by integrating several distinct models. These are effective ML tools for dealing with extremely complicated situations that cannot be tackled by a single ML algorithm. Limited studies used the ensemble method for predicting characteristics. As a result, this research focuses on using basic to advanced machine learning technique, such as gene expression programming (GEP), for prediction of these mechanical characteristics of concrete.

1. **Overview of genetic programming (GP) and gene expression programming (GEP)**

Koza [18] presented GP, as an application of the concepts of genetic and natural selection [19], [20]. GP introduces nonlinear structures (parse trees) as a replacement of fixed-size binary strings (used in the genetic algorithm), proving it a flexible programming tool. It is an approach in which a computer program is developed for the solution of an issue based on Darwinian reproduction and naturally occurring genetic operations like mutation, crossover, and reproduction [18], [21]. At the reproduction stage, a method is developed to determine which program should be terminated. At the stage of implementation, a defined tree percentage with the lowest fitness is terminated, and again using the remaining trees the population is filled based on the method proposed by Nazari [21], [22]. The mutation process defends against premature convergence [21]. The technique to develop a program for solving a problem using GP methodology is shown in Figure 1.

Five primary features that must be defined in GP methodology are the terminals set (input constants and variables), domain-specific functions, run regulating factors (crossover, population size, etc.), fitness function, result designation technique, and termination criteria [18], [21]. Although, in the GP approach only three genetic operators have been specified, specifically only tree crossover is used, resulting in a huge size of parse-trees population [18]. An additional disadvantage of GP is lacking an autonomous genome. Because GP's non-linear architecture must operate as both the phenotype and genotype, it is impossible to develop basic and primitive expressions [20].

Ferreira suggested GEP [20], is a variation of GP that is dependent on population genetic theory. Basic chromosomes of constant length (genetic algorithm) and Parse-trees (GP) are combinedly utilized by GEP. The five parameters that must be defined are the same as for the GP. This approach considers a fixed-length character string, while the GP model utilizes fluctuating length parse-tree, during computer algorithm processing. Each individual is encoded into strings of fixed-length, later on, displayed as nonlinear feature of various sizes and different shapes defined as expression trees (ETs), which are branching structures that represent chromosomes [21]. That is comparable to saying that genotype and phenotype are separated in GEP and that programming can gain from all developmental benefits [20]. A notable change in GEP is that just the genome is carried to the coming initiation, eliminating the need for replication and altering the overall structure because all mutations occur in a single linear framework. Another distinguishing aspect is that populations are generated by only one chromosome with several genes that are further divided as head and tail [21]. In GEP, every gene is comprised of a variable of a constant length, terminal sets of constants, and functions set based on arithmetic operations. There is a one-to-one correspondence between the chromosomal symbol and the matching function. At the chromosomal level, the genetic operation simplifies the emergence of genetic alteration in GEP [19]. The coded data in the chromosomes is required to decode into empirical relationships by introducing a newly designed language which is called Karva. If the gene sequence is provided, it is feasible to infer the precise phenotype [20]. The change of Karva to the ETs begins at the top of the ET and continues down the string. By nodes recording from the root to the deeper layer, ET may be turned into the K-expression [23]. The size of ETs fluctuates throughout the GEP method, resulting in a specific number of needless entities that are not used for mapping of genome. As a result, the GEP gene and the length of K-expression lengths may be identical or different. GEP Schematic diagram is provided in Figure 2.

Overall, the procedure starts with the generation of constant-length chromosomes for the entire population at random. The expression of chromosomes as ETs followed by evaluation of fitness function for each individual. Individuals who are most suited for the reproduction process are selected. The iteration procedure is repeated with fresh individuals until the optimal solution is found. Furthermore, for the conversion of population, genetic procedures like mutation, crossover, or reproduction are performed.

|  |
| --- |
|  |
| Figure 1. Flowchart of GP algorithm |

1. **Methodology**
   1. **Data collection**

The dataset utilized in this study was compiled as part of a study [24] to evaluate the mechanical characteristics of concrete to be utilized by the AASHTO MEPDG. The effects of changing the mix proportions of concrete on its mechanical characteristics, particularly the *CS, E, MOR*, *DM*, Poisson’s ratio (μ), coefficient of thermal expansion, and splitting tensile strength, (seven properties) were studied. Reference [24] contains the mix proportions, component material characteristics, and procedures for measuring concrete parameters. To explore the effects of concrete composition on its properties, different mixes were prepared by including two different fine aggregates, fifteen different coarse aggregates, two types of slag cement (Grade 100 and 120), two different Type (I) ordinary Portland cement, and three various Class C fly ash. The listed characteristics could be evaluated and used to approximate characteristics of future mixtures in the MEPDG program by changing the mix proportions. The dataset has 440 samples for all properties except *DM*, for which 279 samples were tested. Each mechanical characteristic was measured after 7, 14, 28, and 90 days of curing. For each attribute, each of these curing ages was examined individually, resulting in four predictions for *CS*, *E*, *MOR*, *DM*, μ, and splitting tensile strength. In this study, the three mechanical properties (modulus of rupture, modulus of elasticity, and dynamic modulus) were forecasted using five input variables which are given in Table 1 along with the unit of the variables and database statistical indicators.

|  |
| --- |
|  |
| Figure 2. Schematic diagram of the GEP algorithm |

Table 1. Various statistical parameters of input & output properties of concrete

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Input parameters | | | | |  | Output Properties | | |
| **Statistics** | *w:cm ratio* | *AEA (mL)* | *WRA (mL)* | *Slump (in)* | *Age (days)* |  | *MOR (psi)* | *E (psi)* | *DM (psi)* |
| Mean | 0.38 | 16.43 | 31.37 | 2.41 | 34.75 |  | 759.55 | 4262128 | 5442519 |
| Standard Error | 0.0008 | 0.23 | 1.80 | 0.03 | 1.56 |  | 5.07 | 29917.45 | 36144.45 |
| Mode | 0.4 | 12 | 0 | 3 | 7 |  | 731 | 4295000 |  |
| Median | 0.4 | 16 | 10 | 2.5 | 21 |  | 756 | 4260417 | 5517865 |
| Sample Variance | 0.0003 | 23.26 | 1430.98 | 0.31 | 1077 |  | 11302.01 | 3.94×1011 | 3.641×1011 |
| Standard Deviation | 0.02 | 4.82 | 37.82 | 0.59 | 32.81 |  | 106.31 | 627553.8 | 603731.3 |
| Skewness | -1.50 | 0.59 | 0.88 | -0.66 | 0.98 |  | 0.27 | 0.10 | -0.34 |
| Kurtosis | 1.46 | 0.08 | -0.50 | -0.68 | -0.81 |  | -0.55 | -0.69 | -0.39 |
| Range | 0.07 | 23 | 125 | 2 | 83 |  | 572 | 3205667 | 2968227 |
| Minimum | 0.33 | 8 | 0 | 1 | 7 |  | 508 | 2838333 | 3908690 |
| Maximum | 0.4 | 31 | 125 | 3 | 90 |  | 1080 | 6044000 | 6876917 |
| Sum | 170.56 | 7230 | 13788 | 1062 | 15290 |  | 334204 | 1.88×109 | 1.52×109 |
| Count | 440 | 440 | 440 | 440 | 440 |  | 440 | 440 | 279 |

|  |  |  |
| --- | --- | --- |
| Project Path: D:\COMSATES RESEARCH\origin\paper 1graphs.opju PE Folder: /paper 1graphs/Folder1/ Short Name: wc |  |  |
| (a) | (b) |  |

|  |  |
| --- | --- |
|  |  |
| (c) | (d) |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | (e) | (f) |

|  |  |
| --- | --- |
|  |  |
| (g) | (h) |

Figure 3. Distribution histogram of (a) w:cm; (b) AEA; (c) WRA (d) slump; (e) age; (f) *MOR*; (g) *E*; (h) *DM*

* 1. **Development of model**

In order to develop an ML model, the most important step is to pick input variables that can affect the mechanical characteristics of concrete. To develop a generalized relationship between concrete properties and input variables, the impact of all parameters available in the database was examined and the most influential input parameters were selected. Therefore, modulus of rupture, elastic modulus, and dynamic modulus of concrete are considered to be a function of [water-to-cement ratio (w:cm), water reducing admixture (WRA), slump, air entraining admixture (AEA), and Age] as expressed in equation (1).

(1)

It is important to note that the robustness and generalization capability of the model can significantly be impacted by fitting parameters. The fitting parameters of the GEP algorithm were calculated Based on literature recommendations and several initial runs [19]. The population size controls the program's running time (number of chromosomes). The population size was 50, 100, or 150, depending on the number and complexity of the model. Furthermore, the head size and the number of genes are utilized to evaluate the architecture of the models, with the former defining each term's complexity and the latter governing the number of sub-ETs in the program. In this study, three different head sizes 7, 8, or 10 are considered and the number of genes was fixed at 3 or 4.

Table 2. The setting of parameters, numerical constants, & genetic operations for the GEP algorithm.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Settings** | | | | | |
| Parameters | Chromosomes | Head size | Genes | Linkage function | Function set |
| MOR | 150 | 10 | 4 | Addition | +, -, ×, ÷, log. |
| E | 150 | 10 | 4 | Addition | +, -, ×, ÷, exp. |
| DM | 100 | 8 | 3 | Addition | +, -, ×, ÷, exp. |
| **Numerical constants** | | | | | |
| Constant / gene | | 10 | | Lower-bound | -10 |
| Type of data | | Floating numbers | | Upper-bound | 10 |
| **Genetic operations** | | | | | |
| Rate of mutation | | 0.00138 | | Inversion rate | 0.00546 |
| IS transportation rate | | 0.00546 | | RIS transportation rate | 0.00546 |
| Recombination rate (one-point) | | 0.00227 | | Recombination rate (two-point) | 0.00227 |
| Recombination rate of gene | | 0.00227 | | Transportation rate of gene | 0.00227 |

**3.3. Performance assessment and external validation**

In a training, testing, or validation set, the accuracy of the model can be examined via statistical errors such as mean absolute error (MAE), root mean square error (RMSE), relative standard error (RSE), root mean squared logarithmic error (RMSLE), and correlation coefficient (R). The value of R is the best for the assessment of the model. The models are assessed using different statistical indicators and the calculation metrics of the error. These measurements can give some information about your model's faults. The validation of the model is determined by its R, between actual and predicted results. R levels of 0.8 to 0.87 show satisfactory results, whereas values less than 0.70 imply poor results. The relative accuracy between the actual and forecasted results is taken into consideration by the RMSLE algorithm. It is described as the difference between predicted and real logarithms. This is useful when dealing with data which is right-skewed, as the log converts the data range to how it was before. The MAE is the gap between the predicted and measured results when all input data points have the same weightage. Errors absolute values are utilized to ignore the minus sign. It computes the exact magnitude of the occurred errors and reports the results in similar units. Similarly, RMSE is the average of the squared differences between predicted and real observations. It evaluates the mean square value of the error. It is the standard deviation of the estimated mistake. In this method, big deviations, such as outliers, are given more weight, resulting in larger and smaller squared differences. The RMSE measures the mean estimation error of the algorithm when predicting the results for a particular input. The model is more precise when the RSME is low, Equations (1-5) can be used to determine the mentioned error.

= (1)

= (2)

= (3)

= (4)

= (5)

where ,and represent the ith actual, actual mean, model predicted and model predicted mean values, respectively, and n represents the total count of data points used for the development of the model.

A performance index (ρ) suggested by Gandomi and Roke [25], a function of both R and RRMSE can be used to determine the model’s efficiency. Due to extensive data training overfitting of a model is a concern in ML modelling. Training errors may keep decreasing but testing errors are rapidly growing [23]. To minimize overfitting, the model is selected by reducing the value of the objective function (OF) [25]. The validation and training sets are represented by the subscripts V and T, respectively. Furthermore, external validation of the developed model was evaluated using criteria suggested by other authors, as presented in Table 3.

= (6)

= ) + 2() (7)

Table 3. Statistical indicators of the models for external validation

|  |  |  |  |
| --- | --- | --- | --- |
| SNo. | Mathematical expression | conditions | Recommended by |
| 1 |  |  | [26] |
| 2 |  |  | [26] |
| 3 | Where, |  | [27] |

1. **Results and discussion**

For the forecasting of mechanical characteristics of concrete, GEP method is utilized. The output of the algorithm is presented as an ET as given in Figure (4-6) from which empirical relationships are derived. The ETs for *MOR*, *E* and *DM* comprises basic operations (+, -, ×, ÷), with the addition of natural logarithm (ln) in *MOR*, and with additional exponential (e) in *E* and *DM*.

|  |
| --- |
|  |
| Figure 4. Expression tree for the GEP model of *MOR* |

|  |
| --- |
|  |
| Figure 5. Expression tree for GEP model of *E* |

|  |
| --- |
|  |
| Figure 6. Expression tree for the GEP model of *DM* |

* 1. **Formulation of *MOR*, *E*, and *DM* of concrete using GEP**

The model for *MOR* is developed by setting the head size and genes number as 10 and 4, respectively. Equation (8) is a simplified expression for the prediction of *MOR* of concrete proposed by the GEP model.

(8)

Where,

(9)

(10)

) (11)

(12)

The comparison between actual and predicted values of *MOR* is shown in Figure (7). Testing, training and validation data points are concentrated on the ideal fit line (slope equal to 1) which shows an excellent correlation between actual and forecasted values. Furthermore, the fitting nature of data sets to the ideal fit line illustrates the well-trained behavior and strong generalization capacity of the proposed model, so it will work similarly well on unseen data. From the statistical analysis of error, it is noticed that about 50% of results having error values less than 20 psi, 93% are below 50 psi, and only 7% of results have more error values than 50 psi. The average error is 23 psi, which is only 3.28% of the mean experimental result.

|  |
| --- |
| Project Path: UNTITLED.opju PE Folder: /UNTITLED/Folder1/ Short Name: Graph1 |
| Figure 7. Comparison of model-predicted MOR with actual results |

Equation (13) can be used for the prediction of *E* of concrete, developed by GEP. It is notable from these equations that the impact of all five input variables to predict the *E* of concrete has been accurately accounted by the proposed model.

(13)

Where,

(14)

(15)

(16)

(17)

The actual results of the modulus of elasticity are compared with GEP-predicted values, as shown in Figure 8. The generated model seems to have been well-trained on the input data since it produces reliable predictions of the experimental *E*. The possibility of the model getting overfitted has been reduced. The more data points included in the modelling procedure, the more accurate and generalizable the results [95]. Therefore, in the collected database, the maximum number of points available (440) were picked for *E*, giving in a high degree of accuracy with low statistical errors.

|  |
| --- |
| Project Path: D:\COMSATES RESEARCH\E.opju PE Folder: /E/Folder1/ Short Name: Graph1 |
| Figure 8. Comparison of model predicted E with actual results |

Similarly, empirical equations (18-22) proposed by GEP can be used to predict the *DM* of concrete. Figure 9 illustrates a graphical comparison of predicated and experimental outcomes for *DM.*  A significant correlation is shown between experimental outputs and GEP-predicted values of dynamic modulus in this figure, as all the data points are lies close to the ideal fit (45 degrees) line.

(18)

Where,

(19)

(20)

(21)

(22)

|  |
| --- |
| Project Path: D:\COMSATES RESEARCH\origin\dm.opju PE Folder: /dm/Folder1/ Short Name: Graph1 |
| Figure 9. Comparison of model-predicted *DM* with actual results |

* 1. **Performance assessment of GEP models**

Literature has proposed that the data points to input parameters ratio should be at least 3 for acceptable models, and ratio of more than 5 is preferrable [28], [29]. This ratio in the study is substantially greater i.e., 88 for *MOR* and *E*, and 55.8 for *DM*. As previously stated, the model's performance is assessed using MAE, RSE, RMSE, RRMSE, R, ρ, and OF. These parameters for the training, testing, and validation stages for GEP models are listed in Table 4. The MAE, RSE and RMSE, values for the *MOR* model are 21.98, 0.096, and 27.17 for the training set and 22.59, 0.0107, and 31.19 for the testing set, respectively. The calculated outcomes of MAE, RSE, and RMSE for the *E* model are 102014.9, 0.07, and 130323.7 for the training phase and 72713.5, 9.08, and 99048.6 for the validation phase, respectively. Moreover, the measured values for these three indicators for *DM* are 108143.5, 0.16, and 130402.6 for the testing set and 103094.1, 0.15, and 126808.3 for the validation set, respectively. These findings show that the models are trained well, and there is a strong relation between actual and forecasted outputs with small errors. Furthermore, the statistical indicators are nearly similar for training, testing, and validation models which indicate good precision and generalization ability of GEP models.

Table 4. Various statistical parameters of GEP models

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| property |  | *MOR* | | |  | *E* | | |  | *DM* | | |
| Model |  | Training | Testing | Validation |  | Training | Testing | Validation |  | Training | Testing | Validation |
| MAE |  | 21.98 | 22.59 | 28.40 |  | 102014.9 | 117856.2 | 72713.5 |  | 111653.7 | 108143.5 | 103094.1 |
| RMSE |  | 27.17 | 31.19 | 35.28 |  | 130323.7 | 157128.1 | 99048.6 |  | 150071.7 | 130402.6 | 126808.3 |
| RSE |  | 0.096 | 0.107 | 0.249 |  | 0.06 | 0.099 | 0.081 |  | 0.08 | 0.16 | 0.15 |
| RRMSE |  | 0.037 | 0.037 | 0.040 |  | 0.032 | 0.033 | 0.020 |  | 0.028 | 0.022 | 0.021 |
| R |  | 0.96 | 0.95 | 0.92 |  | 0.98 | 0.96 | 0.97 |  | 0.97 | 0.94 | 0.93 |
| ρ |  | 0.019 | 0.019 | 0.02 |  | 0.016 | 0.017 | 0.010 |  | 0.014 | 0.011 | 0.010 |
| OBF |  | 4.08 | | |  | 22188.55 | | |  | 22679.26 | | |

The entire dataset of actual and predicted results along with absolute errors is plotted in Figure (7) to explain the statistics of absolute errors. The average and maximum error in the experimental and forecasted values of *MOR* is 23.03 psi and 87.15 psi, respectively. The values of average and maximum errors are 99995 psi 484157 psi for *E*, and 109836 psi and 604224 psi for *DM*, respectively. It is notable that, the ratio of average error to mean experimental value is only 3.02%, 2.34%, and 2.02% for *MOR*, *E*, and *DM* respectively.

|  |
| --- |
| Project Path: D:\COMSATES RESEARCH\origin\error.opju PE Folder: /error/Folder1/ Short Name: Graph1 |
| (a) |

|  |
| --- |
| Project Path: D:\COMSATES RESEARCH\origin\error.opju PE Folder: /error/Folder1/ Short Name: Graph2 |
| (b) |

|  |
| --- |
| Project Path: D:\COMSATES RESEARCH\origin\error.opju PE Folder: /error/Folder1/ Short Name: Graph3 |
| (c) |

Figure 10. Graphical representation of the absolute error in the predicted and experimental results: (a) *MOR*; (b) *E*; (c) *DM*

* 1. **External validation of the GEP model**

Several statistical external validation checks are also performed for the developed GEP models. Table 5 depicts the values of the criteria employed for external validation of the models. It has been proposed that the slopes of regression lines (k or k′) must be near one [26]. Furthermore, confirming indicator (Rm), a model's external predictability measure, was proposed by Roy [27]. When Rm > 0.5, this criterion is met. Table 5 demonstrates that all three models meet the requirements of the external validation, indicating that the proposed models are realistic and not merely a correlation of output and input parameters. Thus, the formulated GEP models have the potential to accurately and precisely predict values for *MOR*, *E*, and *DM*.

Table 5. Statistical metrics for external validation of the suggested models

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | *MOR* model | *E* model | *DM* model |
| k | 0.98 | 0.98 | 1.00 |
| k' | 1.00 | 1.00 | 1.00 |
| R2 | 1.00 | 0.99 | 1.00 |
| R20 | 0.98 | 0.98 | 1.00 |
| Rm | 0.96 | 0.93 | 0.97 |

* 1. **SHAP analysis**

ML models are considered to be complex since the explanation and interpretation of these models may be challenging. Lundberg and Lee's [30] Shapley Additive Explanations (SHAP) algorithm is one of the AI method that may be utilized to interpret these complicated models. The SHAP values are model-independent and provide a unifying way to illustrate the output of any ML model. Based on ML models SHAP develops surrogate models to provide local validity. Furthermore, SHAP offers a fast implementation for models which are tree-based and is widely used in ML model interpretation [31]. SHAP can be utilized to evaluate the proportional importance, significance, and contribution of each input parameter in estimating the output results [32], [33].

* + 1. **Parameter importance**

In Figure 8 the mean SHAP values are given which illustrate the relative significance of the input features on the outputs. It is observed that age is the most significant parameter for all three characteristics of concrete. The importance of AEA on *E* and *DM* is more than the importance of *MOR* as shown in Figure 8. Furthermore, the contribution of w:cm is lower than all other parameters for *MOR* and *E* while considerable it has on dynamic modulus. The slump has a significant contribution in determining the modulus of rapture, however, it has less contribution in *DM* and *E*. Similarly, the WRA is having relatively more contribution than another input variable in *MOR,* as illustrated in Figure (8-a).

* + 1. **SHAP summary plots**

The summary plots are shown in Figure 11 that are used to demonstrate the influences of input features, each dot shows the SHAP value of an attribute and a unique observation in the data set. It illustrates the related parameter pattern and the SHAP values distribution for a certain characteristic. The summary plot on the y-axis shows the used as inputs and their importance, while the x-axis shows each SHAP-value. The dots represent the data samples and the color of the dots indicates their magnitude, which varies from lite (blue) to dark (red), and the dots indicate data points. The x-axis shows the value of forecast in terms of SHAP-values for each attribute as the value of the input variable varies (red-blue). A higher attribute value for each feature shows that this input variable is improving the output result while on other hand, lower value shows less effect of the input variable, as illustrated in Figure 11. *MOR* has been positively influenced by increasing age and WRA, while the higher value of slump adversely affects the modulus of rupture of concrete. Furthermore, age, AEA, and WRA had a favorable impact on *E* and *DM*. The relation of w:cm with E and *DM* is inverse, as provided in Figure (9a-b).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  | |
| (c) | |
| Figure 11. Parameters importance of various input variables (a) *MOR*, (b) *E*, (c) *DM* | |

|  |
| --- |
|  |
| (a) |
|  |
| (b) |
|  |
| (c) |
| Figure 12. SHAP values of different parameters: (a) *MOR*, (b) *E*, (c) *DM* |

* 1. **Comparison of the GEP model with the Multiple Linear Regression (MLR) model**

Multiple linear regression (MLR) models are developed utilizing a similar database to predict the characteristics of concrete. The outcomes of MLR and GEP models are compared. The MAE, RSE, RMSE, and RRMSE values of the GEP model are significantly lower than MLR values for all three datasets. It is worth noticing from the figure, that the regression model failed in capturing the lower and higher value of *MOR*, *E* and *DM*.

Equation (23) shows the empirical formulation to estimate the *MOR* of concrete using linear regression analysis. The outcomes of the linear regression model and GEP model of *MOR* or compared, as shown in Figure 13. The RRMSEof the GEP model for *MOR* is 58.81% lower than that of the linear regression model as shown in Figure 16, verifying the reliability and accuracy of the GEP model.

(23)

The formulation of linear regression for E is expressed in equation (24). The actual results of elastic modulus are compared with regression and the GEP model outcomes, as shown in Figure 14. The RRMSEof the GEP model for *E* is 69.91% lower than that of MLR which shows the excellent capability of the GEP model to precisely forecast the *E* of concrete.

] (24)

Similarly, based on the same dataset, linear regression analysis is conducted for the *DM* of concrete. Linear expression for *DM* is shown in equation (25). In figure x linear regression values of *DM* are compared with GEP predicted and the actual model of *DM*. The RRMSEvalue of the GEP model is 62.17% lower than that of the MLR model is less accurate.

(25)

Overall, these findings imply that GEP-based models perform better than MLR models with the same parameters as inputs for all three characteristics. The reason for this is that specific statistical regression procedures have limits, such as prediction by certain pre-defined equations may be the actual issue, and the assumption of residual normality [28]. However, the findings of GEP models, show that the models rapidly learn the linear and non-linear relation of input parameters with output properties, with a strong generalization potential and much less error values than the MLR.

|  |
| --- |
| Project Path: UNTITLED.opju PE Folder: /UNTITLED/Folder1/ Short Name: Graph2 |
| Figure 13. Comparison of *MOR* predicted by GEP with the linear regression model |

|  |
| --- |
| Project Path: UNTITLED.opju PE Folder: /UNTITLED/Folder1/ Short Name: Graph1 |
| Figure 14. Comparison of *E* predicted by GEP with the linear regression model |

|  |
| --- |
| Project Path: D:\COMSATES RESEARCH\origin\mlr dm.opju PE Folder: /mlr dm/Folder1/ Short Name: Graph1 |
| Figure 15. Comparison of *DM* predicted by GEP with linear regression model |

|  |  |  |
| --- | --- | --- |
| Project Path: D:\COMSATES RESEARCH\origin\R COMPARISION GEP&MLR.opju PE Folder: /R COMPARISION GEP&MLR/Folder1/ Short Name: Graph1 | Project Path: D:\COMSATES RESEARCH\origin\R COMPARISION GEP&MLR.opju PE Folder: /R COMPARISION GEP&MLR/Folder1/ Short Name: Graph2 | Project Path: D:\COMSATES RESEARCH\origin\R COMPARISION GEP&MLR.opju PE Folder: /R COMPARISION GEP&MLR/Folder1/ Short Name: Graph3 |
| (a) | (b) | (c) |
| Figure 16. Comparison of R, RSE, and RRMSE of GEP and MLR | | |

1. **Conclusion**

This study aimed to evaluate and demonstrate the application of GEP method to predict the modulus of rupture, elastic modulus, and dynamic modulus of concrete. The developed empirical model is based on a widespread data set of diverse variables established in the previous studies. The models are adaptable to a wide range of datasets and use the high influential characteristics of the strength characteristics of concrete. The extensive database has been categorized into three categories of testing, training, and validation stages to properly train the model on unseen data. To evaluate the efficiency of the GEP-predicted models, several error measures such as R, RMSE MAE, RRMSE and RSE were used. Additionally, the model has been evaluated using multiple external validations. SHAP analysis was conducted on all models to determine the effect of input features on the output characteristic of concrete. The GEP models are also compared to multiple linear regression (MLR) models, proving that the GEP model is more reliable and accurate than MLR. The major findings of the study are presented herein:

1. The developed GEP model showed higher accuracy as indicated by the values of R i.e., 0.96, 0.98, and 0.97 for *MOR*, *E*, and *DM*, respectively.
2. The error assessment of the proposed model proved that the model is valid and accurate and has significant generalizability to predict the strength characteristics of concrete.
3. The major concern of model overfitting is satisfactorily resolved as evidenced by lower values of the objective function.
4. The SHAP analysis revealed that age significantly influenced all three considered strength properties and higher age enhances these concrete properties.
5. Moreover, the comparison of the traditional regression model and the GEP model proved that the ML model has greater accuracy in predicting the output.
6. Finally, the proposed model will help the designer and builder to avoid costly and time-consuming laboratory tests.
7. The study also recommended using other machine learning methods such as multi expression programming, deep neural network, *k*-nearest neighbors, random forest and updated database for comparative analysis.

**References**

[1] D.-C. Feng *et al.*, “Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach,” *Constr. Build. Mater.*, vol. 230, p. 117000, Jan. 2020, doi: 10.1016/j.conbuildmat.2019.117000.

[2] F. K. Alqahtani and I. Zafar, “Characterization of processed lightweight aggregate and its effect on physical properties of concrete,” *Constr. Build. Mater.*, vol. 230, p. 116992, Jan. 2020, doi: 10.1016/j.conbuildmat.2019.116992.

[3] A. G. Khoshkenari, P. Shafigh, M. Moghimi, and H. Bin Mahmud, “The role of 0–2mm fine recycled concrete aggregate on the compressive and splitting tensile strengths of recycled concrete aggregate concrete,” *Mater. Des.*, vol. 64, pp. 345–354, Dec. 2014, doi: 10.1016/j.matdes.2014.07.048.

[4] B. P. Koya, S. Aneja, R. Gupta, and C. Valeo, “Comparative analysis of different machine learning algorithms to predict mechanical properties of concrete,” *Mech. Adv. Mater. Struct.*, vol. 29, no. 25, pp. 4032–4043, Oct. 2022, doi: 10.1080/15376494.2021.1917021.

[5] E. Vintzileou and E. Panagiotidou, “An empirical model for predicting the mechanical properties of FRP-confined concrete,” *Constr. Build. Mater.*, vol. 22, no. 5, pp. 841–854, May 2008, doi: 10.1016/j.conbuildmat.2006.12.009.

[6] M. Jalal, “Soft computing techniques for compressive strength prediction of concrete cylinders strengthened by CFRP composites,” *Sci. Eng. Compos. Mater.*, vol. 22, no. 1, pp. 97–112, Jan. 2015, doi: 10.1515/secm-2013-0240.

[7] X. Xu, N. Fallahi, and H. Yang, “Efficient CUF-based FEM analysis of thin-wall structures with Lagrange polynomial expansion,” *Mech. Adv. Mater. Struct.*, vol. 29, no. 9, pp. 1316–1337, Apr. 2022, doi: 10.1080/15376494.2020.1818331.

[8] V. Nilsen, L. T. Pham, M. Hibbard, A. Klager, S. M. Cramer, and D. Morgan, “Prediction of concrete coefficient of thermal expansion and other properties using machine learning,” *Constr. Build. Mater.*, vol. 220, pp. 587–595, Sep. 2019, doi: 10.1016/j.conbuildmat.2019.05.006.

[9] and J. W. K. Smarsly, K. Dragos, “No Title,” in *Machine learning techniques for structural health monitoring*, 2016.

[10] and A. R. V.-K. A. Mosavi, T. Rabczuk, “Reviewing the novel machine learning tools for materials design,” in *In: Proceedings of the Recent Advances in Technology Research and Education, Springer International Publishing, Cham*, 2018.

[11] X. Xu and H. Yang, “Vision Measurement of Tunnel Structures with Robust Modelling and Deep Learning Algorithms,” *Sensors*, vol. 20, no. 17, p. 4945, Sep. 2020, doi: 10.3390/s20174945.

[12] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, “Machine learning in materials informatics: recent applications and prospects,” *npj Comput. Mater.*, vol. 3, no. 1, p. 54, Dec. 2017, doi: 10.1038/s41524-017-0056-5.

[13] H. Alabduljabbar, M. Khan, H. H. Awan, S. M. Eldin, R. Alyousef, and A. M. Mohamed, “Predicting Ultra-High-Performance Concrete Compressive Strength Using Gene Expression Programming Method,” *Case Stud. Constr. Mater.*, p. e02074, Apr. 2023, doi: 10.1016/j.cscm.2023.e02074.

[14] P. Chopra, R. K. Sharma, M. Kumar, and T. Chopra, “Comparison of Machine Learning Techniques for the Prediction of Compressive Strength of Concrete,” *Adv. Civ. Eng.*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/5481705.

[15] M.-C. Kang, D.-Y. Yoo, and R. Gupta, “Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete,” *Constr. Build. Mater.*, vol. 266, p. 121117, Jan. 2021, doi: 10.1016/j.conbuildmat.2020.121117.

[16] J. . Lee, D. . Almond, and B. Harris, “The use of neural networks for the prediction of fatigue lives of composite materials,” *Compos. Part A Appl. Sci. Manuf.*, vol. 30, no. 10, pp. 1159–1169, Oct. 1999, doi: 10.1016/S1359-835X(99)00027-5.

[17] I.-C. Yeh and L.-C. Lien, “Knowledge discovery of concrete material using Genetic Operation Trees,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5807–5812, Apr. 2009, doi: 10.1016/j.eswa.2008.07.004.

[18] J. R. Koza, *Genetic Programming, On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[19] A. Gholampour, A. H. Gandomi, and T. Ozbakkaloglu, “New formulations for mechanical properties of recycled aggregate concrete using gene expression programming,” *Constr. Build. Mater.*, vol. 130, pp. 122–145, Jan. 2017, doi: 10.1016/j.conbuildmat.2016.10.114.

[20] C. Ferreira, “Gene expression programming: mathematical modeling by an artificial intelligence,” vol. 21, 2006.

[21] M. Sarıdemir, “Genetic programming approach for prediction of compressive strength of concretes containing rice husk ash,” *Constr. Build. Mater.*, vol. 24, no. 10, pp. 1911–1919, Oct. 2010, doi: 10.1016/j.conbuildmat.2010.04.011.

[22] A. Nazari and F. Pacheco Torgal, “Modeling the compressive strength of geopolymeric binders by gene expression programming-GEP,” *Expert Syst. Appl.*, vol. 40, no. 14, pp. 5427–5438, Oct. 2013, doi: 10.1016/j.eswa.2013.04.014.

[23] A. H. Gandomi, S. K. Babanajad, A. H. Alavi, and Y. Farnam, “Novel Approach to Strength Modeling of Concrete under Triaxial Compression,” *J. Mater. Civ. Eng.*, vol. 24, no. 9, pp. 1132–1143, Sep. 2012, doi: 10.1061/(ASCE)MT.1943-5533.0000494.

[24] S. M. C. J.B. Effinger, R. Li, J. Silva, “Laboratory Study of Concrete Properties to Support Implementation of the New AASHTO Mechanistic Empirical Pavement Design Guide,” 2012.

[25] A. H. Gandomi and D. A. Roke, “Assessment of artificial neural network and genetic programming as predictive tools,” *Adv. Eng. Softw.*, vol. 88, pp. 63–72, Oct. 2015, doi: 10.1016/j.advengsoft.2015.05.007.

[26] A. Golbraikh and A. Tropsha, “Beware of q2!,” *J. Mol. Graph. Model.*, vol. 20, no. 4, pp. 269–276, Jan. 2002, doi: 10.1016/S1093-3263(01)00123-1.

[27] P. P. Roy and K. Roy, “On Some Aspects of Variable Selection for Partial Least Squares Regression Models,” *QSAR Comb. Sci.*, vol. 27, no. 3, pp. 302–313, Mar. 2008, doi: 10.1002/qsar.200710043.

[28] A. H. Gandomi, A. Faramarzifar, P. G. Rezaee, A. Asghari, and S. Talatahari, “NEW DESIGN EQUATIONS FOR ELASTIC MODULUS OF CONCRETE USING MULTI EXPRESSION PROGRAMMING,” *J. Civ. Eng. Manag.*, vol. 21, no. 6, pp. 761–774, Jun. 2015, doi: 10.3846/13923730.2014.893910.

[29] ILDIK~ E. FRANK and ROBERTO TODESCHINI, “The data analysis handbook , vol. 14. ELSEVIER SCIENCE B.V.,” 1994.

[30] S. Lundberg, S., & Lee, “A Unified Approach to Interpreting Model Predictions. ArXiv. /abs/1705.07874,” 2017, doi: 1705.07874.

[31] Molnar C, “Interpretable machine learning - A Guide for Making Black Box Models Explainable.,” 2019.

[32] A. S. Bakouregui, H. M. Mohamed, A. Yahia, and B. Benmokrane, “Explainable extreme gradient boosting tree-based prediction of load-carrying capacity of FRP-RC columns,” *Eng. Struct.*, vol. 245, p. 112836, Oct. 2021, doi: 10.1016/j.engstruct.2021.112836.

[33] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, “Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach,” *Eng. Struct.*, vol. 219, p. 110927, Sep. 2020, doi: 10.1016/j.engstruct.2020.110927.