Research Article

Optimising the Selection of Input Variables to Increase the Predicting Accuracy of Shear Strength for Deep Beams

Mohammed Majeed Hameed,1 Faidhalrahman Khaleel,1 Mohamed Khalid Al Omar,1 Siti Fatin Mohd Razali,2 and Mohammed Abdulhakim AlSaadi3

1Department of Civil Engineering, Al-Maarif University College, Ramadi, Iraq
2Department of Civil Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia
3Natural and Medical Sciences Research Center (NMSRC), University of Nizwa Sultanate of Oman, Nizwa, Oman

Correspondence should be addressed to Mohammed Majeed Hameed; mohmmag1@gmail.com, Mohamed Khalid Al Omar; mohd.alomar@yahoo.com, and Mohammed Abdulhakim AlSaadi; m.hakim@unizwa.edu.om

Received 11 March 2022; Revised 19 May 2022; Accepted 28 May 2022; Published 4 July 2022

Academic Editor: M. Z. Naser

Copyright © 2022 Mohammed Majeed Hameed et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The deep beam in load transfer is very important as well as difficult to design due to its shear stress problems. Accurate estimation of shear stress would help engineers to get a safer design. One of the major obstacles in building an accurate prediction model is optimising the input variables. Therefore, developing an efficient algorithm to select the optimal input parameters that have the highest information content to represent the target and minimise redundant data is very important. The feature-section algorithm based on the combination of genetic algorithm and information theory (GAITH) was used to select the most important input combinations and introduce them into the prediction models. Four models were used in this study: locally weighted linear regression (LWLR) based on the radial basis kernel function, multiple linear regression (MLR), extreme learning machine (ELM), and random forest (RF). The study found that all applied models were significantly improved by the presence of the GAITH algorithm, except for the MLR model. The LWLR-GAITH model showed 29.15% to 47.88% higher performance accuracy in terms of root mean square error (RMSE) than the other hybrid models during the test phase. Moreover, the results of the standard models (without using the GAITH algorithm) proved the superiority of the LWLR model in reducing the RMSE by 34.51%, 55.17%, and 35.35% compared to RF, MLR, and ELM, respectively. Thus, the inclusion of the LWLR model with GAITH has demonstrated a reliable and applicable computer aid for modelling shear strength in deep beams.

1. Introduction

Since the 1950s, scientists have extensively studied the shear behaviour of beams made of deep reinforced concrete (RC) [1]. In general, it has long been known that deep beams can withstand far greater shear loads than slender beams due to their enormous ratio of shear span to depth ((a/d) ≤ 2.5). Therefore, deep beams are often used as transfer beams in structures, in bridge arches as cap beams, in foundations as pile caps, and in other highly loaded structural members. The main difference between slender and deep beams is that shear deformation in slender beams is minimal and can be neglected, while it must be taken into account in the design and analysis of deep beams, as failure in the latter is mainly due to shear stresses [2]. Moreover, from a modelling point of view, deep beams contradict the concept of the standard plane and require different models than slender beams [3].

In deep beams, the internal arch action generates the shear strength by transferring the load directly to the column through concrete struts. Several factors influence the structural system of deep beams, such as the ratio of shear span to depth, the compressive strength of the concrete, the...
yield strength of the horizontal and vertical reinforcement, and the ratio of the main reinforcement [4]. The RC behavior of deep beams and specifically the shear strength has been quantitatively studied over several decades using various models and empirical approaches such as the strut and sill model, which is considered the most commonly used approach and has been adopted by various codes such as ACI 318-14 [5], CSA A23.3-14 [6], and EC2 [7], or mechanism analysis based on the upper bound theory of plasticity and the finite element approach. However, the design methods of these approaches are linear and therefore unable to capture the complex relationship between the vectors affecting the shear strength and the shear strength itself, in which, in turn, the obtained predicted value of the shear strength differs significantly from the actual value [8].

Different researchers have proposed different design methods to calculate the ultimate shear strength of RC deep beams [9–11]. These design codes are able to capture the nonlinear relationship between the numerous parameters and the ultimate shear strength of deep beams. However, when the results of these methods are compared with the results of experimental tests, they are conservative at best and poor at worst [12]. Furthermore, developing a model that correctly approximates the mathematical shear strength is a major challenge due to this relationship. Due to the inherent limitations of classical models, the prediction of shear strength remains limited to experimental tests [13]. Therefore, the ability of designers to predict shear strength is limited because it is impossible to create an accurate model that can correctly estimate the shear strength capacity using a mathematical formula [14].

In the last two decades, data-driven models based on artificial intelligence (AI) have become increasingly crucial for structural analysis and designing in civil engineering [15, 16]. The most critical applications of AI are the analysis of data sets obtained from experimentally or numerically generated data sets to produce closed-form formulae or numerical tools that predict parameters related to structural response and mechanical behavior. Due to the low information content of the datasets and the high costs associated with their enrichment, it is essential to comprehensively analyze the available data and build the best possible prediction models, which can be done using AI modeling [17]. Moreover, AI models can capture relationships that are difficult to handle with conventional methods [18].

Table 1 provides some examples of the application of different AI and empirical models to predict shear strength. According to the work reviewed, researchers have focused too much on using various artificial and empirical models to estimate shear strength capacity. Since several predictors can affect the shear strength, the reviewed works used classical assumptions to determine the input combination. The most commonly used methods can be presented as follows:

(i) Trial and error approach (trying different combinations)

(ii) Linear correlation approach (selecting only the predictors with higher correlation with the shear strength capacity)

(iii) Using all available inputs

The use of the above methods has several drawbacks: (1) the applied model needs more time to complete the training and calibration phase, (2) the selection of input combinations may not represent all potential cases (primarily if the trial-and-error procedure is used), (3) the complexity of the modeling leads to difficulties in interpreting the model performance and the obtained results, and (4) the model is trained with lack or excessive and redundant information of the selected predictors, which may significantly affect the model performance and stability. Furthermore, the use of linear correlation may be misleading in some cases due to the complexity of the relationship between the shear strength of reinforcement beams and geometric, concrete, and steel parameters.

According to the previous papers published in the literature, many researchers used ANN models to estimate the shear strength in the deep beam. Therefore, in this work, ELM was used as a comparable model. Notably, this model is considered the new variant of ANN and has the advantage of good performance, speed, and ability to generalize. However, while reviewing prior research, we discovered a group of highly motivated scientists to adopt sophisticated models based on regression tree-based models. Thus, we select RF as a robust assembling model. The reason for selecting MLR is to see the efficiency of the other models compared to simple models and, in other words, to investigate if there is a significant difference between the nonlinear-based models and simple models like MLR.

Furthermore, the structures of the three models (RF, MLR, and ELM) are somewhat lacking in flexibility. In other words, these parametric approaches derive a general or global model. Therefore, after finishing the training phase, there will be one complex function to present the targeting problem, and all the data samples are subjected to that function. However, the LWLR is not necessary to specify a function to fit a model to all data in a sample. Thus, LWLR is more flexible and can model complex processes for which no theoretical model exists.

The main objective of this study is to determine the ability of the locally weighted linear regression (LWLR) model based on the Gaussian kernel function to predict the ultimate shear strength of reinforced concrete beams with and without web reinforcement. Moreover, the LWLR model proposed in this study is evaluated by comparing its performance with comparable models such as extreme learning machine (ELM), random forest (RF), and classical multiple linear regression (MLR). Furthermore, the second objective of this study is to use an efficient feature selection tool to select the best input parameters. Selecting the inputs with the highest information with the target and the most diminutive relationship with each other is considered a vital step to achieve the desired prediction accuracy. In this context, information theory is combined with the genetic algorithm as a feature selection tool to remove the variable (s) that contain redundant information that negatively affects the model’s prediction ability. It is also the first time that the LWLR model has been used as a prediction tool in the concrete and structural fields.
2. Methodology

2.1. Shear Strength of RC Deep Beams and Data Collection.

Reinforced concrete (RC) deep beams are frequently used as load-bearing elements in bridge and building construction, so their mechanical behaviour should be carefully analysed and investigated. Since the span-to-depth ratio of RC deep beams is usually less than two, the load-bearing capacity of these structures is strongly influenced by the shear behaviour. It is well known that it is difficult to accurately model the shear behaviour of deep beams because the assumption of a plane cross section no longer holds. To overcome this difficulty, some researchers have applied well-known mechanics-driven models, including the soft truss model.

### Table 1: Examples of different SC models and empirical equations in modelling the shear strength.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Used method</th>
<th>Statistical parameter</th>
<th>Data division</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>[19]</td>
<td>ANN with training algorithms</td>
<td>$R$, $R^2$, MAE, MAPE</td>
<td>70% for training and 30% for testing, 106 data</td>
<td>ANN-GG has the best prediction accuracy.</td>
</tr>
<tr>
<td>[20]</td>
<td>NN</td>
<td>$R$</td>
<td>70% for training and 30% for testing, 233 data</td>
<td>The results of the proposed model show great similarity with the test results.</td>
</tr>
<tr>
<td>[3]</td>
<td>OSVM-AEW</td>
<td>$R$, $R^2$, MAE, MAPE</td>
<td>67% for training and 33% for testing using a triple cross-validation approach.</td>
<td>OSVM-AEW has the best prediction accuracy.</td>
</tr>
<tr>
<td>[21]</td>
<td>SVR-GA</td>
<td>$R^2$, RMSE, MAE, NSE, WI</td>
<td>70% for training and 30% for testing, 217 test records.</td>
<td>SVR-GA gives better predictions than the other models.</td>
</tr>
<tr>
<td>[22]</td>
<td>GEP</td>
<td>$R$, RMSE, MAE</td>
<td>70% for training, 10% for validation, and 20% for testing, 214 test records.</td>
<td>GEP has better prediction accuracy than ACI and CSA and shows very good agreement with the ANN model.</td>
</tr>
<tr>
<td>[4]</td>
<td>LS-SVR and SFA</td>
<td>$R$, RMSE, MAE, MAPE</td>
<td>70% for training and 30% for testing, 214 data set</td>
<td>LS-SVR with SFA has better prediction accuracy compared to the standard SVR.</td>
</tr>
<tr>
<td>[23]</td>
<td>ANN</td>
<td>Mean, STD, COV</td>
<td>—</td>
<td>ANN performed better than the ACI code, EURO code, zsutty method, and russo method.</td>
</tr>
<tr>
<td>[24]</td>
<td>Strut-and-tie model</td>
<td>AVG, COV</td>
<td>16 test specimens.</td>
<td>The results of the proposed model show great similarity with the test results.</td>
</tr>
<tr>
<td>[25]</td>
<td>CSTM</td>
<td>Mean, COV</td>
<td>355 test specimens.</td>
<td>The performance of CSTM is better than other STM models.</td>
</tr>
<tr>
<td>[26]</td>
<td>New analytical expression using strut-and-tie model (STM)</td>
<td>Mean, COV</td>
<td>111 test specimens.</td>
<td>The proposed model has better prediction accuracy compared to ACI 318-14 code.</td>
</tr>
<tr>
<td>[27]</td>
<td>GA</td>
<td>AVG, STD</td>
<td>50% for training and 50% for testing, 371 data records.</td>
<td>GA has better prediction accuracy compared to ACI 318-14 code.</td>
</tr>
<tr>
<td>[28]</td>
<td>Feedforward NN</td>
<td>Mean, STD</td>
<td>50% for training, 25% for validation, and 25% for testing, 433 data records.</td>
<td>The results of the proposed model are in agreement with the experimental and analytical data.</td>
</tr>
<tr>
<td>[29]</td>
<td>EMARS</td>
<td>$R^2$, RMSE, MAPE</td>
<td>10-fold cross-validation, 106 data set</td>
<td>The performance of EMARS is better than BPNN, RBFNN, and SVM.</td>
</tr>
<tr>
<td>[15]</td>
<td>AdaBoost</td>
<td>$R^2$, RMSE, MAE</td>
<td>10-fold cross-validation, 271 test records</td>
<td>The performance of the proposed models is better than the traditional machines single learning methods (DT, SVM, ANN).</td>
</tr>
<tr>
<td>[30]</td>
<td>SVR with RBF and polynomial kernel</td>
<td>$R$, RMSE</td>
<td>10-fold cross-validation</td>
<td>SVR with RBF and polynomial kernel gives better prediction accuracy compared to backpropagation neural network and empirical relations.</td>
</tr>
</tbody>
</table>
A total of 271 test data on RC deep beams were collected from the open-source literature to train the models used in this work. These include 52 specimens from Lu [36], 37 specimens from Ludwig and Nunes [37], 25 specimens from Hameed et al. [38], 53 specimens from Hameed et al. [39], 12 specimens from Naser and Alavi [40], 12 specimens from Ludwig et al. [41], 6 specimens by Nguyen et al. [42], 12 specimens by Yaseen et al. [43], 19 specimens by Zhang et al. [44], 39 specimens by Gong, and 4 specimens by Gandomi et al. [45]. It is worth noting that the database contains a wide range of RC deep beams to improve the generativity of the model. Moreover, the dataset contains four types of RC deep beams: beams with horizontal web reinforcement, beams without web reinforcement, beams with horizontal and vertical reinforcement, and beams with vertical web reinforcement.

2.2. Multiple Linear Regression. Multiple linear regression (MLR) specifies the independent variable \(Y\) according to a linear equation in terms of more than two independent variables \(X_i\). The mathematical expression of MLR is as follows [46]:

\[
Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + E. \tag{1}
\]

Here, the MLR parameters (i.e., \(\beta_0\) and \(\beta_i\)) are calculated using the least-squares method (LS), and \(E\) is the unsystematic error. The fitness function should be defined to calculate the best fit line for the measured data set \(Y\). The following fitness function should be minimised.

\[
Fitness = \frac{1}{2N} \sum_{j=1}^{N} (Y_{obsj} - Y_{predj})^2, \tag{2}
\]

where \(N\) is the length of the data set used, and \(Y_{obsj} - Y_{predj}\) represent the actual or predicted values for the \(j^{th}\) sample. To simplify equation (2), a matrix can be derived from this equation as \((X\beta - Y)\). The derivative of the matrix with respect to \(\beta\) then gives \(X^T(X\beta - Y)\). By the LS method, the term \(X^T(X\beta - Y)\) should be zero and solved for \(\beta\). Finally, the parameter \(\beta\) can be given by the following equation:

\[
\beta = (X^TX)^{-1}X^TY, \tag{3}
\]

where \(X\) and \(Y\) are the training and target values.

2.3. Locally Weighted Linear Regression. The locally weighted linear regression (LWLR) model is an improved version of the MLR method. This method was invented by Chen et al. [47] to improve the efficiency of the classical MLR model. In the proposed technique, a weighting function is used to define the relationship between the input and output of the data sets. Moreover, the fitness function of LWLR can be defined as follows:

\[
Fitness = \frac{1}{2N} \sum_{j=1}^{N} W_j(Y_{obsj} - Y_{predj})^2, \tag{4}
\]

where the \(W\) in the above equation is the weight. Similar to MLR, the equation can be expressed as \(((X\beta - Y)^TW(X\beta - Y))\). When the fitness function of the LWLR technique is derived in terms of \(\beta\), the matrix \(X^TW(X\beta - Y)\) is obtained. To calculate \(\beta\), the matrix of \(X^TW(X\beta - Y)\) should be zero. Consequently, \(\beta\) is expressed by

\[
\beta = (X^TWX)^{-1}X^TWY. \tag{5}
\]

In the LWLR method, a kernel function is used as a weighted matrix. In this study, the radial basis function (RBF) is used to calculate the proposed matrix, which is expressed as follows [48]:

\[
W_{ik} = \exp\left(-\frac{1}{2}(X_i - X_j)^2\right), \tag{6}
\]
where \((-Y(X_i - X_j))\) is the difference between variables \(X\) in samples \(i\) and \(j\), while \(Y\) is a positive user-defined number.

2.4. Extreme Learning Machine. Extreme learning machine (ELM) is a novel machine learning approach introduced by Wakjira et al. [49] as a robust new learning algorithm for single hidden layer feedforward networks (SLFNs). ELM is thousands of times faster than conventional learning algorithms for feedforward networks and achieves higher generalisation performance [50]. Since ELM is encoded as SLFNs, many of the complications of gradient-based algorithms, such as learning rate, learning epochs, and local minima, are eliminated when using ELM. Moreover, even with randomly generated hidden nodes, ELM retains its ability to universally approximate SLFNs [51]. ELM networks consist of three layers: the input layer where the data are presented to the ELM network, the hidden layer where the basic computations are performed, and the output layer where the information from the hidden layer is transmitted to. In addition, the results of ELM are organised in the output layer. ELM randomly selects the input weights and biases for the hidden nodes and uses the least-squares solution to calculate the output weights analytically.

The ELM model can be mathematically expressed as

\[
V_i = \sum_{m=1}^{i} B_m g_m (\alpha_m \cdot x_m + \beta_m), m = 1, 2, \ldots, n, \tag{7}
\]

where \(V_i\) is the ELM target; \(i\) refers to the number of hidden nodes; \(B_m\) is the weight value connecting the \(m\)th hidden node with the output node; \(g_m(\alpha_m \cdot x_m + \beta_m)\) is the output function associated with \(i\) hidden nodes; and \((\alpha_m, \beta_m)\) are the hidden node parameters that are randomly initialised.

The above equation can be written compactly as follows:

\[
MR = T, \tag{8}
\]

where \(M\) is known as the output matrix of the hidden layer of the neural network.

\[
M(\alpha_1, \ldots, \alpha_i, \beta_1, \ldots, \beta_i, x_1, \ldots, x_i) = \begin{bmatrix}
g_m(\alpha_1 \cdot x_1 + \beta_1) & \ldots & g_i(\alpha_i \cdot x_1 + \beta_i) \\
\vdots & \ddots & \vdots \\
g_m(\alpha_1 \cdot x_m + \beta_1) & \ldots & g_i(\alpha_i \cdot x_m + \beta_i)
\end{bmatrix}_{N \times i}, \tag{9}
\]

R = \[
\begin{bmatrix}
R_1^T \\
\vdots \\
R_i^T
\end{bmatrix}, \tag{10}
\]

T = \[
\begin{bmatrix}
t_1^T \\
\vdots \\
t_i^T
\end{bmatrix}, \tag{11}
\]

where \((\cdot)^T\) stands for the transposition operator and the sigmoid transfer function used in this study. Figure 2 illustrates the structure of ELM with a single hidden layer.

2.5. Random Forest (RF). RF is a machine learning approach often used to solve problems related to classification and regression. Basically, it is an extension of the method CART (classification and regression trees). Decision tree models generally have many advantages, such as simplicity, ease of use, and interpretability, but also many disadvantages, such as poor performance and unsatisfactory robustness. Therefore, RF can overcome the shortcomings of conventional decision trees by combining the performance of many randomised decorrelated decision trees to perform prediction or classification tasks efficiently. In addition, RF uses a modified version of the bootstrap aggregation approach, called bagging, in which a considerable collection of decorrelated, noisy, approximately unbiased trees is constructed and averaged to minimise model variance and instability problems [52]. The RF strategy involves aggregating multiple trees to improve overall prediction accuracy while achieving low variance and bias. Figure 3 shows RF as a forest of \(n\) trees.

RF is not only able to model high-dimensional, non-linear relationships but is also resistant to overfitting, has relative robustness, estimates variable importance, and has few user-defined parameters [53]. The hyperparameters of the RF model strongly influence the performance of the models, so their values need to be determined precisely. The most critical hyperparameters are the number of regression trees, the proportion of the training dataset used to build the model, and the leaf nodes.

2.6. Feature Selection. Feature selection (FS) is a dimensionality technique used to eliminate the redundant and irreversible variables from the data set. This technique helps to use the minimum number of features that correctly describe a given problem in a given domain, resulting in more straightforward and accurate schemes. In machine learning, the tool FS is a fundamental concept that significantly impacts the performance of a prediction model. Machine learning models are highly influenced by the data features on
which they are trained. More specifically, the performance of
a model can be affected by features that are not relevant or
only partially relevant. The other advantages of using FS
before training a prediction model can be summarised as
follows:

(i) Overfitting is reduced by removing redundant data.
When there is less redundant data, it is more
challenging to make decisions based on noise data
(ii) Improves model accuracy; less misleading data
means better modelling accuracy
(iii) Reduces training time; a smaller number of features
reduces the algorithm’s complexity and speeds up training

2.6.1. Information Theory. Information theory was intro-
duced by Claude E. Shannon to study the quantitative as-
pects of information, including how coding affects
information transmission [54]. Information theory origi-
nated as a mathematical study of whether it is possible to
transmit information reliably and cheaply for a given source,
channel, and fidelity criteria. Shannon’s information theory
defines information as anything that reduces or eliminates
uncertainty. The model can achieve higher accuracy in
classification tasks if it receives more information because
the predicted classes of new instances are likely to match
their actual classes [55]. Mutual information (MI) is defined
in algorithmic information theory and has an “algorithmic”
relationship. MI is a dimensionless quantity, usually
expressed in bits, and can be viewed as a means of reducing uncertainty about one random variable given knowledge of another. The MI between two random variables indicates how much uncertainty has been reduced. The lower the mutual information, the greater the reduction, and zero means that the variables are strictly independent [56]. Figure 4 shows the structure of information theory.

2.6.2. Genetic Algorithm. The genetic algorithm (GA) is one of the most widely used metaheuristic algorithms, a stochastic optimisation technique inspired by natural evolution [57]. Crossing and mutating chromosomes are an essential part of the GA process. Each chromosome acts as an individual solution to the target problem, which is ultimately expressed in a binary string. The chromosome population is essential in the GA process, and its initial values are randomly selected. Then, a chromosome (the best one) that solves the given problem very well is selected for reproduction [58]. The optimisation process of GA consists of six steps: initialisation, fitness calculation (objective function), conditional termination, selection, crossover, and mutation. The detailed process of GA is shown in Figure 5. In addition, only chromosomes that function perfectly are retained for further reproduction during the fitness evaluation step. The process of selection and reproduction is repeated several times to obtain better chromosomes. After selecting the best chromosomes, these chromosomes can produce offspring during the crossover process by exchanging string parts and gene combinations, resulting in a new solution. A chromosome is nominated to change a randomly selected bit through a random exchange during the mutation process. The following step is to estimate the generated fitness and compare it with the termination criteria. The GA process is terminated when the termination criteria are met.

2.6.3. Genetic Algorithm-Based Information Theory. Information theory is generally used for input variable selection. Variable selection attempts to maximise the mutual
Input: \( I \left( X_{h}^{\text{var}}, X_{m}^{\text{var}} \right) \) and \( I \left( X_{h}^{\text{var}}, y_{\text{target}} \right) \)
\( n = (1, \ldots, m) \): statistical data determined by the previous algorithm (Figure 4);
\( N_{\text{no}} \): the desired number of predictors;
\( A \): selection pressure;
\( G_{\text{max}} \): maximum number of generations;
\( N_{P} \): size of the population; and
Output: \([j]\) set the indexes of the selected predictors.

(i) Generate a set of \( N_{P} \) chromosome \((CR)\) for the initial population. Each chromosome is a vector \( CR = [j_{1}, j_{2}, \ldots, j_{N_{\text{no}}}] \) containing the indices of \( N_{\text{no}} \) neuron \( j \) generated randomly without repeating elements.

(ii) For \( n = 1: G_{\text{max}} \) do

(iii) Evaluate the population.

(iv) For \( idx = 1: N_{P} \) do

(v) Calculate \( V \) and \( P \) for each \( CR \) using the following formula by the calculated values of mutual information for all elements of chromosomes \( CR \).

\[
V = \left(1/N_{\text{no}}\right)\sum_{i=1}^{N_{\text{no}}} I(X_{h,i}; y)
\]

\[
P = \left(1/N_{\text{no}}\right)\sum_{i=1}^{N_{\text{no}}} \sum_{k=1}^{N_{\text{no}}} I(X_{h,i}; y_{k})
\]

(vi) \( \phi_{dx} = (V - P) \): storing the fitness of each \( idx \);

(vii) end for (loop);

(viii) Rank the individuals according to their fitness \( \phi_{dx} \).

(ix) Store the genes of the best individual into \([j]\).

(x) Perform the crossover

(xi) \( k \leftarrow 0 \).

(xii) For \( idx = 1: N_{\text{no}} \) do

(xiii) \( k \leftarrow k + 1 \).

(xiv) Choose the indices of the parents randomly using the asymmetric distribution [60].

(xv) \( \omega_{ij} = (1, 2) \rightarrow \text{random number } \epsilon \in [0, 1] \text{ with uniform distribution} \)

(xvi) parent, \( i = (1, 2) \rightarrow \text{round} \left( N_{P} \left( e^{\omega_{ij}} - 1 \right)/(e^{1} - 1) \right) \) [60];

(xvii) Storing the indices missing in both parents in \([j_{\text{abs}}]\).

(xviii) Assembling the chromosome \( CR_{n}^{\text{pop}} \).

(xix) For \( n = 1: N_{\text{no}} \) do

(xx) Randomly select a parent (i.e., between parent 1 and 2) to get the \( n^{\text{th}} \) gene for the \( k^{\text{th}} \) of the individual in the new generation.

(xx) \( CR_{n}^{\text{pop}} \leftrightarrow CR_{(\text{parent 1 or 2})} \)

(xxii) Considering the constraint [40],

(xxiii) If there is duplicity of indices in \( CR_{n}^{\text{pop}} \), then

(xxiv) Pick up a new index for \( CR_{n}^{\text{pop}} \) from \([j_{\text{abs}}]\).

(xxv) end if

(xxvi) end for

(xxvii) end for

(xxviii) end for

Algorithm 1: Selection of the best predictors by GA and mutual information.

### 2.7. Model Development and Performance Evaluation

In this study, two scenarios are created to develop more reliable models to predict the shear strength values of the deep beams. Both scenarios are created using a dataset containing geometric parameters, steel, and concrete properties. Table 2 shows the statistical description of each parameter. The data set is divided into two groups: the training phase comprises two-thirds of the data, while the rest is used to test the accuracy of the models. All the input and output variables are normalised (between 0 and 1) to remove the influence of dimensions and improve the capabilities of the predicting models [61]. This method prevents numerical difficulties
Taylor diagrams, scatter plots, line plots, and bar charts are error (RMSE). In addition, various visualisations such as Sutcliffe (NSE), Willmott index (WI), mean absolute error (MAE), and Nash-Sutcliffe (NSE) are used to evaluate the results of both scenarios. However, the first scenario involves the use of feature selection to choose the most optimal input combination by removing the redundant information. The second scenario of this study involves the use of standard models such as LWLR, MLR, RF, and ELM. However, the second scenario involves the use of standard models such as LWLR, MLR, RF, and ELM. However, the second scenario of this study involves the use of feature selection to choose the most optimal input combination by removing the redundant information. However, the results of both scenarios are evaluated using various statistical measures such as coefficient of determination (R²), Nash-Sutcliffe (NSE), Willmott index (WI), mean absolute error (MAE), uncertainty interval (U₁₀), and root mean square error (RMSE). In addition, various visualisations such as Taylor diagrams, scatter plots, line plots, and bar charts are produced to provide more information about the best prediction model and to allow better comparison between the models used. The mathematical expressions of the statistical matrices are explained as follows [62, 63]:

\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_{\text{obsi}} - Y_{\text{predi}})^2}, \]

\[ \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |Y_{\text{obsi}} - Y_{\text{predi}}|, \]

\[ R^2 = 1 - \frac{\sum_{i=1}^{N} (Y_{\text{obsi}} - Y_{\text{predi}})^2}{\sum_{i=1}^{N} (Y_{\text{obsi}} - \mu)^2}, \]

\[ \text{WI} = 1 - \frac{\sum_{i=1}^{N} (Y_{\text{obsi}} - Y_{\text{predi}})^2}{\sum_{i=1}^{N} ([Y_{\text{predi}} - \mu] + [Y_{\text{obsi}} - \mu])^2}, \]

\[ \text{NSE} = 1 - \frac{1/n \sum_{i=1}^{N} |Y_{\text{obsi}} - Y_{\text{predi}}|^2}{\sum_{i=1}^{N} (Y_{\text{predi}} - \mu)^2}, \]

\[ U_{95} = \frac{1.96}{n} \sqrt{\frac{n}{\sum_{i=1}^{n} (Y_{\text{obsi}} - \alpha)^2 + \sum_{i=1}^{n} (Y_{\text{obsi}} - Y_{\text{predi}})^2}, \]

where \( Y_{\text{obsi}} - Y_{\text{predi}} \) is the measured and predicted shear strength value of the \( i^{\text{th}} \) sample, \( M \) is the number of samples, and \( \mu, \alpha \) is the mean predicted and observed shear strength values.

### 3. Result and Discussion

In this part of the study, the results of four prediction models for predicting the ultimate shear strength of deep beams are presented. In this work, two scenarios are created to achieve...
the objective of this work. In the first scenario, all the models used are presented based on all the available input variables. Furthermore, in this scenario, four standard prediction models have been developed to predict the shear strength capacity of reinforced concrete deep beams with and without web reinforcement. The models are extreme learning machine (ELM), random forest (RF), multiple linear regression (MLR), and locally weighted linear regression (LWLR). The second scenario is created using assumed feature selection (FS) to select the most impressive input variables. Then, these variables are introduced into the prediction models to estimate the ultimate shear capacity of the deep bars. In this case, five input combinations are created based on the assumed FS (see Table 3). The main purpose of using GAITH as a feature selection tool is to remove the repetitive information from the dataset and thus improve the prediction accuracy. Finally, to evaluate the performance of the adopted models, we used various statistical parameters, including error measures and accuracy indices, as well as a graphical representation. According to the error measures, the proposed model (LWLR) of this study yielded lower prediction errors than comparable models (MAE = 13.249, RMSE = 22.563, NSE = 0.974, WI = 0.993). From this table, it can be inferred that the LWLR model provided more accurate estimates on all statistical measures, followed by the RF, ELM, and MLR models.

3.1. First Scenario: Standard Models. This part of the study looks at the performance results obtained with the standard models using all input parameters. The performances of the models used, such as extreme learning machine (ELM), random forest (RF), multiple linear regression (MLR), and locally weighted linear regression (LWLR) during the training and testing phases, are presented in Table 4. In general, all the models used performed well in the training set, yielding high values for NSE and WI. In this phase, WI ranged between 0.967 and 0.993, while NSE varied between 0.879 and 0.974. From these parameters, it can be seen that the LWLR model performed better than the other models, while the MLR model had lower prediction accuracy. To select the best model, the higher performance of this model in the training phase is not sufficient because, in this step, the model receives input variables and their corresponding target. However, the testing phase is crucial and more reliable to assess the performance of a model. According to the results shown in Table 4, the LWLR model provided higher and desired accuracy in predicting the shear strength capacity compared to other models. This model achieved a higher agreement with the actual values with WI of 0.984, NSE of 0.941, and less prediction error (MAE = 33.933, RMSE = 57.776). However, the MLR model performed very poorly and therefore provided undesirable estimation accuracy (MAE = 61.165, RMSE = 89.651, RMSE = 89.651.

### Table 4: The statistical description of the proposed models: first scenario.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (kN)</th>
<th>RMSE (kN)</th>
<th>NSE</th>
<th>WI</th>
<th>MAE (kN)</th>
<th>RMSE (kN)</th>
<th>NSE</th>
<th>WI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWLR</td>
<td>13.249</td>
<td>22.563</td>
<td>0.974</td>
<td>0.993</td>
<td>33.933</td>
<td>57.776</td>
<td>0.941</td>
<td>0.984</td>
</tr>
<tr>
<td>RF</td>
<td>18.138</td>
<td>28.613</td>
<td>0.958</td>
<td>0.988</td>
<td>49.068</td>
<td>77.712</td>
<td>0.892</td>
<td>0.964</td>
</tr>
<tr>
<td>MLR</td>
<td>36.924</td>
<td>48.692</td>
<td>0.879</td>
<td>0.967</td>
<td>61.165</td>
<td>89.651</td>
<td>0.857</td>
<td>0.953</td>
</tr>
<tr>
<td>ELM</td>
<td>23.194</td>
<td>31.045</td>
<td>0.951</td>
<td>0.987</td>
<td>51.999</td>
<td>78.200</td>
<td>0.891</td>
<td>0.968</td>
</tr>
</tbody>
</table>

### Table 5: A comparison between the efficiency of the proposed models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training phase</th>
<th>Testing phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>LWLR</td>
<td>62.78</td>
<td>87.78</td>
</tr>
<tr>
<td>RF</td>
<td>54.44</td>
<td>80.00</td>
</tr>
<tr>
<td>MLR</td>
<td>22.22</td>
<td>43.33</td>
</tr>
<tr>
<td>ELM</td>
<td>35.56</td>
<td>61.67</td>
</tr>
</tbody>
</table>

### Table 6: A comparison of the performance outcomes for the uncertainty ($U_{95}$) analysis.

<table>
<thead>
<tr>
<th>Combination</th>
<th>LWLR</th>
<th>RF</th>
<th>MLR</th>
<th>ELM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>21.83</td>
<td>22.48</td>
<td>23.07</td>
<td>22.22</td>
</tr>
<tr>
<td>M2</td>
<td>19.73</td>
<td>22.33</td>
<td>23.04</td>
<td>22.27</td>
</tr>
<tr>
<td>M3</td>
<td>21.77</td>
<td>22.99</td>
<td>23.06</td>
<td>22.21</td>
</tr>
<tr>
<td>M4</td>
<td>20.62</td>
<td>20.79</td>
<td>21.89</td>
<td>20.69</td>
</tr>
<tr>
<td>M5</td>
<td>23.38</td>
<td>22.70</td>
<td>24.48</td>
<td>23.00</td>
</tr>
<tr>
<td>Without feature selection tool (all inputs)</td>
<td>21.87</td>
<td>22.23</td>
<td>22.76</td>
<td>22.27</td>
</tr>
</tbody>
</table>
The other important observation can be drawn from the same table (Table 4): RF and ELM performed well compared to the MLR model and provided lower prediction accuracy than the LWLR model. However, the RF model provided a slightly better prediction than ELM with an RMSE of 77.712, MAE of 49.068, WI of 0.964, and an NSE of 0.892. Furthermore, ELM is considered the third best model and provides much better estimates of shear strength capacity than the MLR model (MAE = 51.999, RMSE = 78.22, NSE = 0.891, WI = 0.968). After quantitative evaluation, the superiority of the LWLR model is determined by its ability to reduce the RMSE criteria during the test phase. Specifically, the results showed a 34.51%, 55.17%, and 35.35% improvement in estimation by the LWLR model compared to the RF, MLR, and ELM models, respectively.

Table 5 evaluates the performance of the adopted models in terms of their efficiency in reducing the absolute relative error. According to the reported results, the LWLR model

Figure 6: A comparison between the objectives of the models: (A)–(D) represent the LWLR, RF, MLR, and ELM models, respectively. Figures (a) and (b) show scatter plots and line plots for the training and test data, respectively, while figure (c) is a Taylor plot based on the test data. (a) Training data. (b) Testing data. (c) Taylor diagram.
20% was 75%, 59.78%, and 68.48% for RF, MLR, and ELM, respectively. The data that had an absolute relative error of less than 20% performed excellently in the test series. 85.87% of the data performs excellently in the test series. 85.87% of the data that had an absolute relative error of less than 20% was 75%, 59.78%, and 68.48% for RF, MLR, and ELM, respectively.

A more comprehensive statistical analysis can be seen in Table 6. The primary purpose of uncertainty analysis is to restrict the predicted range in which the actual value of an experiment’s outcome lies. The uncertainty interval describes the estimated range as an interval in this context. The results provided in that table showed that the LWLR-M2 model had the lowest value for uncertainty with a 95% confidence level (U95 = 19.73).

Visualisation assessment is critical to see how each model handles a single sample. Scatter plots and line plots provide essential information about the behaviour of a model and show the deviation between the actual and predicted values of shear strength (see Figures 6(a) and 6(b)).
Figure 6 shows that the performance of all models generally gives a satisfactory prediction. However, the testing phase was crucial, and some models gave poor predictions. It is important to note that the model proposed in this study (LWLR) was superior to the others in estimating the shear strength capacity with the highest accuracy ($R^2 = 0.945$). The RF models showed good prediction accuracy with $R^2 = 0.943$, followed by the ELM model with $R^2 = 0.895$ and the MLR model ($R^2 = 0.892$). Another important observation is that the LWLR model showed an excellent ability to predict extreme values compared to the other models. Moreover, all the proposed models except the LWLR model showed that...
several predicted samples are far from the ideal line. The Tayler diagram is one of the most important figures for visually assessing the performance of a particular prediction model. This figure summarises three important statistical criteria: root mean square error (RMSE), standard deviation, and correlation coefficient. Figure 6(c) shows that the LWLR model is close to the actual data set compared to the other models. It has the highest correlation coefficient, the lowest...
3.2. Second Scenario: Selection of Features Based on Prediction Models.

This section focuses on using GAITH, a feature selection tool based on a mixture of information theory and genetic algorithms. The main advantage of using GAITH is to select the most efficient input variables that have the most significant impact on shear strength capacity and to minimise redundant information between variables. As mentioned earlier, five input combinations were selected using the GAITH algorithm. Table 7 summarises the performance of the models used based on GAITH. From Table 7, it can be

RMSE, and a lower standard deviation than the observed data set.
seen that the GAITH-based LWLR model (GAITH−LWLRM2) showed excellent performance in predicting shear strength capacity over comparable models in the training and testing phases. For example, GAITH−LWLRM2 gave the lowest measured error (RMSE = 49.519, MAE = 33.617) and higher prediction accuracy (NSE = 0.954 and WI = 0.988). Based on the evaluations shown in this table, the RF model with the m4 combination gave a good estimate (RMSE = 69.895, MAE = 43.132, NSE = 0.913, and WI = 0.972) but relatively lower than that of the LWLR model. However, the MLR model is found to provide the worst estimates as it could not account for the nonlinear relationship between the shear strength capacity and the geometric and concrete and steel properties. On the other
hand, ELM proved to be much better than MLR for the first combination (m1) and provided good prediction accuracy (RMSE = 58.095, MAE = 37.052, NSE = 0.940, WI = 0.98), which was, however, slightly lower than that of the RF model. For further evaluation, the cumulative percentage of absolute relative error was calculated and summarised in Table 8. The main result is that the proposed model (LWLR) produced a large percentage of data with less than the absolute relative benchmark error in each input combination (i.e., 5%, 10%, 15%, and 20%). For the M2 combination, the proposed model (GAITH – LWLRM5) showed efficient performance, and 86.78% of the data set had an absolute relative error of less than 20%. Moreover, for the combination m4, the table shows that more than 78% of the shear

![Graphs and tables showing performance metrics for different models and combinations.](image-url)
strength capacity estimated with the LWLR had an absolute relative error of less than 15%, while the corresponding values for the RF, MLR, and ELM models were 67.26%, 51.09%, and 66.30%, respectively. In this respect, MLR showed undesirable accuracy in all cases compared to the other prediction models.

Several graphs were generated during the training and testing datasets to further evaluate the impact of the GAITH approach on the predicted models (Figures 7–11). These figures provide essential information about the performance of each proposed model based on different input parameters. In general, the LWLR, in combination with GAITH, provided much more accurate estimates than other comparable models. It can also be visually seen that the M2 combination performed best in terms of the highest prediction accuracy. Among all the proposed models built with five input combinations, the GAITH−LWLRM2 model stood out as the best model. The accuracy of this model was the highest with $R^2 = 0.962$ during the testing phase in terms of shear strength prediction. The Taylor diagram in Figure 8(c) shows a satisfactory agreement between the actual and estimated data to illustrate the comparison between the adopted models better. It also shows that the estimates for GAITH−LWLRM2 are closest to the point corresponding to the actual data. Thus, these results prove that the GAITH−LWLRM2 model was the best in terms of generalisation abilities and performed satisfactorily in both the training and testing phases.

The improvement of each model during the training and testing phase through GAITH is summarised in Figures 12–14. It is important to note that the MLR showed very poor estimates and therefore should not be used as a comparative model at this stage. In general, LWLR improved more significantly than other models due to the presence of the GAITH algorithm. More specifically, the performance of the hybrid model (GAITH−LWLRM2) outperformed the standard model (LWLR) during the training phase in terms of reducing the predicted errors by 45.08% and 23.69% for MAE and RMSE, respectively. Looking at the performance of the RF model before and after using the feature selection tool, a slight improvement is observed after using the GAITH algorithm, with the reduction in RMSE and MAE being 4.87% and 3.83%, respectively. However, the ELM model was not run efficiently with fewer input parameters. This model required an extensive data set and complete input vectors to learn well, while other models performed efficiently with fewer input parameters. Nevertheless, all models used in the test phase improved their capacity prediction with the presence of the GAITH algorithm. The LWLR and GAITH algorithm combination provided the most accurate predictions compared to the other models. The superiority of GAITH−LWLRM2 was clearly shown in the reduction of RMS parameters by 29.15% and 47.88% compared to the GAITH−RFM4 and GAITH − ELMM1 models, respectively. Finally, Taylor plots (see Figures 15 and 16) are produced throughout the training and testing phase, showing that there is perfect agreement between the shear strength data predicted by the GAITH−LWLRM2 model and the actual values.
Figure 13: Comparison between the proposed models before and after using feature selection algorithm: testing phase.

Figure 14: Improvement of the correlation of determination in the proposed models using the feature selection algorithm.

Figure 15: The Taylor diagram shows the improvement in prediction due to the presence of a feature selection algorithm: training phase.
4. Conclusion

As the reinforced concrete deep beams are essential components for load distribution, the design is still challenging due to the problem of shear stress. Several nonlinear factors influence the behaviour of the shear strength capacity, making the accurate estimation of this parameter challenging. Accurate estimation of shear stress for deep beams would help designers to design a better and safer structure to prevent structural failure, thus saving lives and property. One of the most critical factors for an efficient and accurate prediction model is the selection of input combinations. In this study, GAITH is introduced based on the integration of genetic algorithm and mutual information to determine the most influential input parameters. This method is developed to overcome some of the shortcomings of classical data-driven input selection. Instead of the trial-and-error technique or the linear correlation, this study presented a robust method for selecting input combinations for prediction models. The structure of the proposed models includes locally linear regression (LWLR) based on the radial basis kernel function, multiple linear regression (MLR), random forest (RF), and extreme learning machine (ELM). The integration between the GAITH algorithm and the aforementioned data-driven models yields promising results. Moreover, the performance of the models (except MLR) is significantly improved by the GAITH algorithm in terms of shear strength prediction. More specifically, the LWLR-GAITH model achieved the highest prediction accuracy in reducing root mean square error by 29.15% to 47.88% compared to the applied models (MLR – GAITH, ELM – GAITH, and RF – GAITH). The main reason for the improvement in prediction is the presence of the GAITH algorithm, which selects the most influential input combination that contains a minimum of redundant data and a maximum of useful information. Redundant data complicate the model’s training process and have a negative impact on the generalisation of the model. Another important finding of this study is that the LWLR – GAITH model needs only two input variables to achieve the best prediction accuracy, while the other comparable models need four predictors. It is essential to mention that among the different parameters (i.e., geometry, concrete, and steel properties), the effective height ($h_0$) and $f_y$ are the most critical parameters that greatly influence the shear strength capacity. In conclusion, this study recommends the application of the methodology used (LWLR – GAITH) to solve various problems related to the structure.

Abbreviations

- $b$: Beam width
- $h$: Beam height
- $A$: Shear span
- $h_0$: Effective height
- $l_o$: Beam span
- $a/h_0$: Shear span to-effective height ratio
- $l_o/h_0$: Beam span to-effective height ratio
- $S_v$: Vertical reinforcement spacing
- $f_yv$: Vertical reinforcement strength
- $\rho_v$: Vertical reinforcement ratio
- $S_h$: Horizontal reinforcement spacing
- $f_yh$: Horizontal reinforcement strength
- $\rho_h$: Horizontal reinforcement ratio
- $f_yl$: Longitudinal reinforcement strength
- $\rho_l$: Longitudinal reinforcement ratio
- $f_y$: Concrete strength
- $V_u$: Beam shear strength
- GA: Genetic algorithm
- ANN: Artificial neural network
- LM: Levenberg–marquardt
- QN: Quasi-Newton method
- GG: Conjugate gradient
- GD: Gradient descent
- NN: Neural network
- OSVM- AEW: Optimized support vector machines with adaptive ensemble weighting
- LS-SVM: Least-squares support vector machine
- SOS: Symbiotic organisms search
- $R$: Coefficient of correlation
- $R^2$: Coefficient of determination
- MAE: Mean absolute error
- MAPE: Mean absolute percentage error
- RMSE: Root mean square error
- SVR: Support vector regression
References


[5] A. C. Institute, 318-14: Building Code Requirements for Structural Concrete and Commentary, American Concrete Institute, Indianapolis, IN, United States, 2014.


