The use of intelligent algorithms in the prediction of bonding strength in steel-concrete interfaces

Vanderci F. Arruda[1], Gray F. Moita[1], Eliene P. Carvalho[1], Priscila F. S. Silva[1]

[1] *Centro Federal de Educação Tecnológica de Minas Gerais*
Av. Amazonas, 5.253, Nova Suíça, CEP: 30.421-169, Belo Horizonte, MG, Brasil.
vanderci-engcivil@hotmail.com; gray@cefetmg.br; eliene@cefetmg.br; 201422800040@aluno.cefetmg.br

Abstract. The current study proposes the use of intelligent systems and a statistical technique to predict the strength of steel-concrete bond using a database of academic literature. The work used as the reference deals with the pull-out test that evaluated the steel-concrete bond behavior in thin bars. The experimental program employed concretes of class C25, C35 and C40, and CA-50 ribbed bars (with diameters of 6.3, 8.0, and 10 mm) and CA-60 notched bars (with diameters of 5.0, 6.0, 8.0, and 9.5 mm). The database was subjected to data mining strategies for statistical treatment. Conventionally, bond strength is obtained by pull-out and beam tests, as proposed by BS EN:10080, involving expensive and lengthy experimental tests. Alternative ways are the application of machine learning-based methods and the use of statistical techniques. Using these methods it is possible to assess their efficiency in predicting the maximum pull-out force. In the present research two particular methods are used to solve the problem. One technique is Multiple Linear Regression which is a generalization of Least Squares, where the minimization of the sums of the *n*-th powers of the residuals is considered. Multiple regression analysis is also very useful in experimental situations, where the experimenter can control the predictor variables. The other model is based on statistical learning theory and is called Support Vector Machines (SVM). It is a machine learning and computational intelligence technique, where it is possible to obtain a classification of data from the same domain in which the learning is performed. The method uses a principle called induction, where it is possible to draw generic conclusions from a training set. The main objective of this work is to propose an alternative way to predict the bond strength of thin bars using computational methods and a statistical technique. The methods used is compared via performance metrics to verify which one proves to be more reliable to predict steel-concrete interface bond, considering the safety coefficients used in engineering.

Keywords: Intelligent Systems, Multiple Linear Regression, Machine Learning, Pull-out test, Support Vector Machines.

1 Introduction

The constituents of concrete define its mechanical properties such as tensile and compressive strength, Poisson's coefficient, modulus of elasticity, among others. Other definers of such properties are the manufacturing process, application, and curing.

Amongst the concrete applications, the reinforced concrete, in which concrete and steel work in a solidary way, is largely used. This combination makes studies about the steel-concrete interface very relevant. Steel-concrete bond are essential for verifying the structural behavior of reinforced concrete parts. There are a variety of parameters that can influence the steel-concrete bond; in the case of bond stress, one can mention: bar diameter, the anchoring length of the bars, the compressive strength concrete, and the surface conformation coefficient. The steel-concrete bond is influenced by several factors, which explains its non-linear behavior.

Usually, bond strength is obtained by pull-out and beam tests, as proposed by BS EN:10080 [1], involving expensive and expensive experimental tests. As shown in Figure 1, the purpose of the pull-out test is to pull out a steel bar positioned in the center of a concrete cube.

Several studies on the bond strength of reinforced concrete have been proposed according to Carvalho et al. [2] and Carvalho et al. [3]. However, most of the studies are obtained from steel bars with diameters greater than 10mm according to Makni et al. [4] e Yartsev et al. [5], which shows the lack of studies with thin bars (diameters less than 10mm).

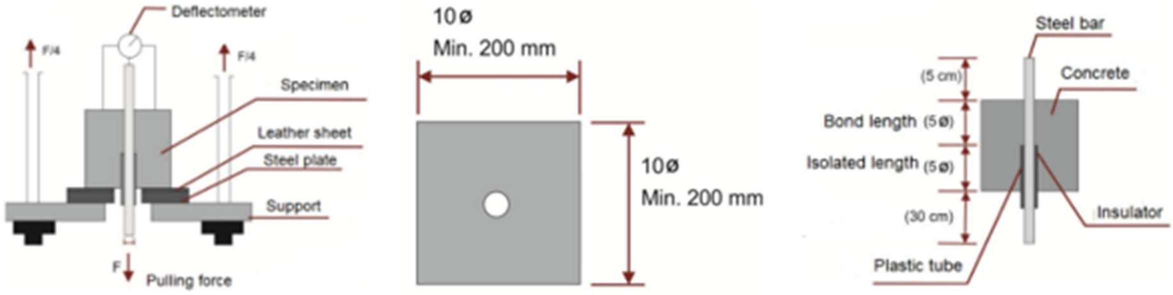

Figure 1. Pull-out test set-up [1]

The difficulty related with destructive testing makes the search for alternative methods pertinent. Hence, this paper proposes the use of computational intelligence method and statistical techniques in the prediction of the bond behavior of the steel-concrete interfaces, corroborating with other studies on the steel-concrete bond, especially with thin bars.

According to Rezende [6], the main techniques and methodologies used by intelligent systems are: knowledge acquisition, machine learning, artificial neural networks, fuzzy logic, evolutionary computing, agents and multi-agents, data, and text mining. In most, intelligent systems have demonstrated a strong ability to solve real-life problems, particularly for non-linear problems.

This work uses the intelligent algorithms called Support Vector Machine and statistical techniques called Multiple Linear Regression, which are evaluated by the following performance metrics: the mean square error (MSE), the root mean square error (RMSE), and the coefficient of determination (R^2). The methods used are compared via performance metrics to verify which one proves to be more reliable to predict steel-concrete interface bond, considering the safety coefficients used in engineering.

It is important to mention here that the current work is part of an ongoing research with the following publications in the subject: Carvalho et al. [2], Silva et al. [7] and Arruda et al. [8] used Artificial Neural Networks to deal with the problem; Silva et al. [9] used Support Vector Machine adopting none of these preprocessing strategies and using a database with 89 sample elements; and, Silva et al.[10] made use of Random Decision Trees and Random Forest to predict the bonding strength. The main contribution of this paper is the preprocessing strategies used on the dataset. The pre-processing technique applied was sample normality test adopting the Shapiro-Wilk test. A preliminary treatment and the Grubbs test were also performed to remove outliers.

2 Methodology

2.1 Data Description and Preprocessing

The current study proposes the use of intelligent systems to predict the steel-concrete bond strength using a database from the academic literature - the studies by Carvalho et al. [2]. The work used as reference deals with the pull-out test that evaluated the steel-concrete bond behavior in thin bars. The experimental program employed concretes of class C25, C35 and C40, and CA-50 ribbed bars (with diameters of 6.3, 8.0, and 10 mm) and CA-60 notched bars (with diameters of 5.0, 6.0, 8.0, and 9.5 mm).

The pull-out test in Carvalho's work [2] has as input parameters: surface conformal coefficient, concrete compressive strength, anchorage length, and bar diameter. And as output parameter the maximum pull-out force. The dataset has a total of 14 samples, both ribbed and notched, and the dataset contains 98 pull-out test results based on various input parameter compositions. The Shapiro-Wilk test was used to verify the normality of the database samples Table 1, for notched and ribbed bars.

Table 1- Shapiro-Wilk p-value for ribbed and notched bars

Sample	statistical value	Shapiro's p-value	Value-p
E1	0,91	0,26	0,05
E2	0,82	0,02	
E3	0,94	0,59	
E4	0,96	0,86	
E5	0,95	0,76	
E6	0,95	0,74	
E7	0,68	0,00	
E8	0,89	0,40	
E2SW	0,85	0,07	
E7SW	0,91	0,52	

Sample	statistical value	Shapiro's p-value	Value-p
N1	0,91	0,47	0,05
N2	0,79	0,08	
N3	0,84	0,18	
N4	0,98	0,94	
N5	0,90	0,20	
N6	0,83	0,03	
N6SW	0,83	0,05	

In Table 1, notched samples E1, E3, E4, E5, E6, and E8 have Shapiro's p-values greater than the reference value (p-value), concluding that the samples follow normal distribution. From samples E2 and E7, the spurious values were removed via preliminary treatment. Then, they were renamed to E2SW and E7SW, and the Shapiro's p-values obtained were greater than the reference value, also concluding that they follow normal distribution.

In Table 1, ribbed samples N1, N2, N3, N4, and N5 have Shapiro's p-values greater than the reference value (p-value), concluding that the samples follow normal distribution. From sample N6, the spurious values were removed via preliminary treatment. The sample was renamed to N6SW, and a Shapiro's p-value equal to the reference value was obtained, i.e., it also follows normal distribution.

Datasets may possess values that have different scales, which causes difficulty for visualization and may even worsen the predictive performance of some of the computational methods. Unstandardized data can reduce the speed of implementation or even prevent the convergence of estimators, Wang, Wang and Alexander [11].

Some models are built by the assumption that their parameters take on values close to zero, i.e., at comparable scales. Estimators using metrics and gradient assume that the data are standardized [11]. However, preprocessing techniques prove effective and can improve the performance of computational models.

Due to the above, the database was subjected to data mining strategies for statistical treatment. This treatment was used to remove outliers using a preliminary treatment that takes into account the mean and standard deviation of the samples, as shown in Arruda et al. [8]. With this treatment, from the total of 98 elements of the raw sample, 15 sample elements were removed, and the sample was reduced to 83 elements, which were used in the following computational methods.Furthermore, the dataset was scaled to comparable scales using StandardScaler as a technique that standardizes features by removing the mean and scaling to unit variance.

2.2 Performance Metrics

As mentioned, MSE, given by Equation (1), RMSE, as per Equation (2), and R^2, given by Equation (3), are used as performance metrics:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{1}$$

$$RMSE = \sqrt{MSE} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3}$$

In the above equations, y_i is the observed value, $\bar{y}$ is the mean of the observations, $\hat{y}$ is the predicted value of y_i and N is the total number sampling in the data set.

2.3 Support Vector Machine

Support Vector Machines (SVMs) are a machine learning technique based on statistical learning theory. Its purpose is to obtain a class of data from the same domain in which the learning took place. According to Lorena and Lorena and Carvalho [12], machine learning techniques employ a principle called induction, where it is possible to draw generic conclusions from a particular set of examples. SVMs can achieve superior results to those obtained by other learning algorithms.

SVMs are learning algorithms that work on classification and regression problems. In addition to performing linear regression and classification, SVMs also work well on nonlinear data [13]. In the case of classification, in SVMs the problem reduces to obtaining a hyperplane in order to maximize the separation between two classes. Support vector regression (SVR), on the other hand, is a nonparametric regression technique that relies exclusively on kernel functions [14].

The most commonly used kernels are:

- Linear $\boldsymbol{K(X_i, Y_j) = X_i^T Y_j}$
- Polynomial $\boldsymbol{K(X_i, Y_j) = [(X_i^T Y_j) + 1]^d}$
- Radial basis function $\boldsymbol{K(X_i, Y_j) = e^{-\gamma|X_i - Y_j|^2}}$
- Sigmoid $\boldsymbol{K(X_i, Y_j) = \tanh(\beta_0 X_i^T Y_j + \beta_1)}$

The sequence of the study was as follows: the dataset was chosen, outliers were removed via preliminary treatment, and then a preprocessing strategy was chosen (StandartScaler) using the Train Split Test as the training and testing method. The models were implemented with different kernels, and finally the models were evaluated using the performance metrics obtained from each kernel, arriving at the chosen model.

2.4 Multiple Linear Regression

Multiple Linear Regression is a generalization of least squares, where the minimization of the sums of the *n*-th powers of the residuals are considered. According to Andrews [15], the Least Squares Method is an ideal procedure when the errors in a regression model have a Gaussian distribution or when linear estimates are required.

According to Almeida [16], the versatility of the least squares method makes it a tool with potential in mathematical modeling and in the use of computational resources, speeding up and promoting significant improvements in the practicality of the method. The problem of the Least Squares Method consists in relating the variables involved in a certain phenomenon, from a set of data, and thus finding a mathematical model that satisfactorily expresses this phenomenon. Thus, the method proposes a way to approximate functions.

Multiple Linear Regression is a statistical approach to describe simultaneous associations between several variables. An essential item in this modeling includes the estimation, inference and selection of the variables that build the model, as well as, the evaluation of the model fitness [17]. Much of the statistical computing is done on linear regression models. The choice of the predictor variables that affect the response variable is crucial, because each variable affects the regression model in a distinct and important way.

Multiple regression analysis is also highly useful in experimental situations, where the experimenter can control the predictor variables. An experimenter usually wants to investigate several predictor variables simultaneously, because almost always more than one predictor variable influences the response. Multiple linear regression models can be used for any data observation or for experimental data of a completely randomized design. According to Tranmer and Elliot [18], the Multiple Linear Regression method is a deductive model, that is, it is based on the real-world understanding of the problem to be modeled and is grounded in theory.

Laboratory tests, measurements and observations are commonly subject to errors and approximations, i.e. even if a careful laboratory test is carried out, errors are frequent [19].

According to Swearingen [20], multiple regression analysis is an approach that examines the relationship of a series of independent variables to a single dependent variable, and is a model widely used in statistical analysis. According to the author, the relationship between the variables can be linear or curvilinear, and the analysis of the method provides an assessment of how well a set of predictors contribute to the estimation of the target data. The method evaluates the individual contribution of the predictors and how they affect the output parameter. Multiple linear regression is an extension of simple linear regression, however, with multiple predictor variables and an output with *n* observation units.

In Equation (4) the multiple predictor variables and one outcome $(x_{i1}, x_{i2}, \dots, x_{i,n}, y_i)$ for $i = 1, 2, \dots, n$ observation units are presented, thus formalizing the simultaneous statistical relationship between the single continuous outcome Y and the predictor variables X_k $(k = 1, 2, \dots, n)$.. The term β_0 represents the intersection to the axis, and each β_k represents a slope with respect to X_k, and are called partial regression coefficients. It is important to note that the assumptions of the method are the same as for simple regression, among them: the y_i are independent values among themselves and the y_i values follow normal distribution.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{i,n} + \varepsilon_i \tag{4}$$

The sequence of execution of the multiple linear regression method is as follows:

i. For each predictor, check through the graphs whether there is a linear relationship between the target data and each predictor;
ii. Estimate the linear regression model;
iii. Assess whether the model provides an adequate fit for the data;
iv. Use the model to make inferences about the regression coefficients;
v. Reassess through diagnostics whether the model provides an adequate fit to the data.

According to Scott and Holt [21], most statistical methods were developed for data collected in controlled experiments, where it is reasonable to assume that the observations are independent.

3 Presentation and analysis of the results

3.1 Support Vector Machine

When using machine learning techniques such as SVR, it is paramount to choose the best parameters to use in the implementation. In the present work, the GridSearchCV method was employed to help in this choice. Once the pre-processing strategy and the kernel are defined, GridSearchCV is used to choose the best parameters for the model. The output of the GridSearchCV defines the "optimal" parameters for each Kernel and the implementation with these parameters results in the best inputs for the method. The input parameters of GridSearchCV are placed in Table 2 and it is possible to check the ones with the best performance in the same Table.

Table 2. GridSearchCV parameters

	Parameters				Chosen
Kernel	RBF	Poly	Linear	Sigmoide	***Poly***
C	1_1000	1_1000	1_1000	1_1000	***100***
Degree	-	2_5	-	-	***4***
Epsilon	0.1_1	0.1_1	0.1_1	0.1_1	***0.1***
Gamma	scale/auto	scale/auto	scale/auto	scale/auto	***auto***

The results obtained in this work are described in Table 3.

Table 3. Performance metrics of SVRs

Support Vector Regression				
Kernel	RBF	LIN	POLY	SIG
RMSE (N)	1602.65	2946.,66	1251.94	3710.30
R^2	0.90	0.67	0.94	0.47

The use of the SVR machine learning technique proved to be effective in predicting the bond strength. From Table 3, it is possible to verify among the results obtained, the best performance metrics were obtained with the Polynomial Kernel, obtaining an RMSE of 1251.94 N and a coefficient of determination of 0.94.

A similar paper was proposed by Silva et al [22], where Support Vector Regression was used in the database present in Carvalho's study [2]. However, in this work the removal of outliers took into account the experience of the authors and a criterion based on the coefficient of variation. As output for the GridSearchCV, in the study the parameters C = 500, Degree = 3, Epsilon = 0.1 and Kernel RBF were used. The results obtained in the authors' study were RMSE equal to 1.305 N and R^2 equal to 0.95.

3.2 Multiple Linear Regression

The first step in performing the Multiple Linear Regression method is to choose the dataset, so the same database was used. The model is estimated using the least squares criterion: the partial regression coefficient that minimizes the sum of the squared distances between the observed value and the value obtained by the model is chosen. At the end of the MLR execution, the partial regression coefficients of the process were calculated.

Equation (5) is an analytical expression and was obtained by the implemented Multiple Linear Regression.

$$y_i = 4448.84x_{i1} + 28.59x_{i2} - 134.98x_{i3} + 142.97x_{i4} \tag{5}$$

The values x_{ij} are the normalized values for each parameter. shown in Table 3.

Once the analytical equation that describes the phenomenon of obtaining the bond strength was calculated, the performance parameters obtained by the method were calculated, according to Table 4.

Table 4. Performance metrics of MLR

Multiple Linear Regression	
RMSE (N)	3520.76
R^2	0.83

3.3 Comparison between the results

Finally, the comparison between the results presented in this work and in Silva et al. [9] obtained by the SVR and the Multiple Linear Regression in this work are presented in Table 5.

Table 5. Performance metrics

Research	Performance Metrics	MLR	SVR
This paper	RMSE (N)	3520.76	1251.94
	R^2	0.83	0.94
Silva et al [9]	RMSE (N)	-	1305.00
	R^2	-	0.95

4 Conclusion

The general objective of this work was to present an analysis of computational methods in the prediction of bond strength, using Multiple Linear Regression and Support Vector Machines, and employing the dataset obtained by the study of Carvalho et al [2].

From this work it was possible to understand the mathematical and computational tools related to machine learning and statistical techniques in order to predict bond strength without the need for experimental testing. Thus, the proposed methods were judged in comparison based on the chosen performance parameters. This study aimed to corroborate in the area of civil engineering, with the purpose of proposing an alternative way to obtain a mechanical characteristic of steel-concrete without the need for destructive testing. The mathematical and computational modeling techniques employed differ from those commonly used to obtain the mechanical characteristics of concrete.

The assessment of the normality of the data is an important prerequisite for the application of statistical regression techniques. Therefore, it is suggested that these tests always precede the application of these techniques.

The analysis of the database for the removal of outliers is also an important prerequisite for the performance of the computational models. If this analysis is performed using well-founded techniques, it becomes more robust and well justified.

It should be noted that these considerations are even more important when there is little technical knowledge about the data being treated. In the case of the Silva et al. [9], the preprocessing strategy involved only the analysis of technicians with high knowledge on the subject. This analysis also led to good results, but with some limitations.

Preprocessing strategies were also used in order to improve the performance of the proposed algorithms. Another means of obtaining better results is the use of GridSearchCV which was employed to perform an exhaustive search on the dataset in order to determine the best input parameters for the SVM.

The results obtained show that SVM performed the best with an RMSE value of 1251.94 N and $R^2 = 0.94$. The SVMs had an overall error rate that can be considered low. The obtained MLR result was as expected due to the non-linear steel-concrete bond behavior.

Thus, it is possible to conclude that the use of an intelligent algorithm to obtain the bond strength was satisfactory. Therefore, the use of these methods should be encouraged, because they can estimate a certain mechanical characteristic of the steel-concrete interface, in this case, the bond strength, with reliability, eliminating the use of time-consuming and expensive tests.

References

[1] EN, B. S. 10080: 2005. *Steel for the reinforcement of concrete–Weldable reinforcing steel–general," British Standards*, p. 54-60, 2005.

[2] E. P. Carvalho et al. Experimental investigation of steel-concrete bond for thin reinforcing bars. *Latin American Journal of Solids and Structures*, v. 14, p. 1932-1951, 2017.

[3] E. P. Carvalho et al. Comparison of test methodologies to evaluate steel-concrete bond strength of thin reinforcing bar. *Construction and Building Materials*, v. 183, p. 243-252, 2018.

[4] M. Makni et al. Artificial neural network for the prediction of the steel–concrete bond behaviour. *European journal of environmental and civil engineering*, v. 18, n. 8, p. 862-881, 2014.

[5] V. P Yartsev; A. N. Nikolyukin and A. O. Korneeva, Neural network modeling of concrete bond strength to reinforcement. In: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, 2019. p. 033011.

[6] S. O. Rezende, *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003.

[2] E. P. Carvalho et al., "Experimental investigation of steel-concrete bond for thin reinforcing bars". *Latin American Journal of Solids and Structures*, v. 14, p. 1932-1951, 2017.

[7] P.F.S. Silva, G.F. Moita, E.P. Carvalho, "Predicting the bond strength of thin steel rebars in concrete by means of artificial neural networks", *5th World Congress on Civil, Structural, and Environmental Engineering (CSEE'20), pp.* 165-1-165–8. 2020.

[8] V. F. Arruda et al., "Statistical techniques for experimental data pre-processing for use in intelligent algorithms". *XXXIX Jornadas Sulamericanas de Engenharia Estrutural,* 2022 (accepted for publication).

[9] P.F.S. Silva, G.F. Moita, E.P. Carvalho, V.F. Arruda, "A Computational Approach to Predict the Bond Strength of Thin Steel Rebars in Concrete by Means of Support Vector Machine", in *Proceedings of the joint XLII Ibero-Latin-American Congress on Computational Methods in Engineering and III Pan-American Congress on Computational Mechanics, ABMEC-IACM* , 2021.

[10] P.F.S. Silva et al., "A Computational approach to predict the pull-out force of thin steel bars in concrete using random decision trees and random forest". *XXXIX Jornadas Sulamericanas de Engenharia Estrutural,* 2022

[11] L. Wang, G. Wang and C. Alexander, "Big data and visualization: methods, challenges and technology progress". *Digital1Technologies*, v. 1, n. 1, p. 33-38, 2015.

[12] A. C. Lorena and A. Carvalho, "Uma introdução às Support Vector Machines". *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43-67, 2007.

[13] C. Lin and S. Wang, "Fuzzy support vector machines". *IEEE transactions on neural networks*, v. 13, n. 2, p. 464-471, 2002.

[14] H. Drucker et al., "Support vector regression machines". *Advances in neural information processing systems*, v. 9, 1996.

[15] D. F. Andrews, "A robust method for multiple linear regression". *Technometrics*, v. 16, n. 4, p. 523-531, 1974.

[16] R. N. Almeida, O método dos mínimos quadrados: estudo e aplicações para o ensino médio. *Campos dos Goytacazes: UENF*, 2015.

[17] R. G. Lomax and D. L. Hahs-Vaughn, *Statistical Concepts-A Second Course*. Routledge, 2013.

[18] M. Tranmer and M. Elliot, "Multiple linear regression". *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, v. 5, n. 5, p. 1-5, 2008.

[19] O. Helene, *Métodos dos Mínimos Quadrados*. Editora Livraria da Física, 2006.

[20] C. J. Swearingen, *Handbook for Clinical Research*. 2. ed. John Wiley & Sons, 2018. 542 p.

[21] A. Scott and D. Holt, "The effect of two-stage sampling on ordinary least squares methods". *Journal of the American Statistical Association*, v. 77, n. 380, p. 848-854, 1982.

[22] P. F. S. Silva et al., "A Computational Approach to Predict the Bond Strength of Thin Steel Rebars in Concrete by Means of Support Vector Machine". *Proceedings of the joint XLII Ibero-Latin-American Congress on Computational Methods in Engineering and III Pan-American Congress on Computational Mechanics*, ABMEC-IACM, 2021