REVISTA MATÉRIA V.24 N.04
ISSN 1517-7076 article e-12527, 2019

Prediction of restrained shrinkage crack width of slag mortar composites using data mining techniques

Francisco Ferreira Martins [1], Aires Camões [2]

[1] ISISE, Department of Civil Engineering, University of Minho, Minho, Guimarães, Portugal.
e-mail: ffm@civil.uminho.pt
[2] CTAC, Department of Civil Engineering, University of Minho, Minho, Guimarães, Portugal.
e-mail: aires@civil.uminho.pt

ABSTRACT

The purpose of this study is to develop data mining models to predict restrained shrinkage crack widths of slag mortar cementitious composites. A database published by BILIR *et al.* [1] was used to develop these models. As a modelling tool R environment was used to apply these data mining (DM) techniques. Several algorithms were tested and analyzed using all the combinations of the input parameters. It was concluded that using one or three input parameters the artificial neural networks (ANN) models have the best performance. Nevertheless, the best forecasting capacity was obtained with the support vector machines (SVM) model using only two input parameters. Furthermore, this model has better predictive capacity than adaptative-network-based fuzzy inference system (ANFIS) model developed by BILIR *et al.* [1] that uses three input parameters.

Keywords: mortar; data mining; prediction; restrained shrinkage cracking.

1. INTRODUCTION

This study aims to estimate the drying shrinkage crack widths of mortars containing Granulated Blast Furnace Slag (GBFS) as fine aggregate. Mortars are cementitious materials currently used in construction industry. GBFS is commonly utilized as a fine aggregate substitute in mortar in order to reduce environmental problems related to aggregate mining and waste disposal. One of the most frequent problems in mortars is cracking that can be caused by several factors. One of these factor is the drying shrinkage that causes tensile stresses in the mortar due to restrictions to its free shortening. Drying shrinkage occurs after the mortar setting, being originated by the evaporation of free capillary water from the interior of the mortar that was not consumed in the hydration reactions of the cement.

Herewith, the estimation of drying shrinkage is based on data mining techniques. These techniques include decision trees, artificial neural networks (ANN), support vector machines (SVM) and k-nearest neighbours (k-NN). There are in scientific literature many applications of intelligent tools relative to mortars most of them based on artificial neural networks. Thus, artificial neural networks were used to predict compressive strength of mortar: for different cement grades (ESKANDARI *et al.* [2]); for mixtures containing different cement strength classes (ESKANDARI-NADDAF and KAZEMI [3]); containing metakaolin (SARIDEMIR [4]); using different saw waste for sand replacement (MAHZUZ *et al.* [5]); for different scoria percentages instead of sand (RAZAVI *et al.* [6]). Artificial neural networks were also used to predict rubberized mortar properties (TOPÇU and SARIDEMIR [7]), model the influence of salt on desorption isotherm and hygral state of cement mortar (KONIORCZYK and WOJCIECHOWSKI [8]), evaluate sand/cement ratio on mortar using ultrasonic transmission inspection (MOLERO *et al.* [9]) and establish a relationship between microstructural characteristics and compressive strength of cement mortar (ONAL and OZTURK [10]). Fuzzy logic was used to predict rubberized mortar properties [7] and compressive strength of mortars containing metakaolin [4]. Adaptive neuro-fuzzy methodology for prediction of sulfate expansion of PC mortar was used by İNAN *et al.* [11]. Optimized support vector machines were used to model the compressive strength of geopolymer paste, mortar and concrete by NAZARI and SANJAYAN [12] and a genetic algorithm–artificial neural network model to predict the compressive strength of cement mortar was created by AKKURT *et al.* [13].

Autor Responsável: Francisco Ferreira Martins Data de envio: 01/10/2018 Data de aceite:11/09/2019

10.1590/S1517-707620190004.0852

The previous studies do not include the application of intelligent tools neither to drying shrinkage nor to cracking of mortars. Therefore, there is a lack in this field as the scientific literature has only focused on the application of intelligent tools to concrete shrinkage [14, 15, 16, 17, 18] or cracking [19, 20, 21, 22, 23]. No mention in the literature is found related to the prediction of retrained shrinkage crack widths of mortars or concretes by data mining techniques.

The main purpose of the present study is to construct predictive models for reliable estimation of drying shrinkage crack widths of mortars containing GBFS as fine aggregate using several data mining techniques. For this purpose, a database published by BILIR *et al.* [1] was used. By using this database several prediction models were constructed. A total of 358 registers including the replacement ratios of GBFS as fine aggregate in mortars (RR), the drying time of ring specimens (DT), the free shrinkage length changes of GBFS mortars (FS) and the crack widths of GBFS mortars (CW) exposed to ring test were used for constructing prediction models.

This paper has two main challenges. The first is to verify if, at least one of the DM techniques gives better results than ANFIS used by BILIR *et al.* [1]. The second is to check if one can get good results by reducing the number of input parameters.

2. DATA MINING

2.1 Definitions

The exponential development of the computational tools in recent decades has allowed the storage of large amounts of information. The need to extract useful knowledge from this information led to the emergence of the so-called Knowledge Discovery in Databases (KDD). KDD is an iterative process consisting of several steps, such as data selection, preprocessing of target data, data transformation, data mining and interpretation [24]. Data mining aims to extract useful patterns from data sets for decision-making purposes by applying automatic learning methods. DM can be applied to classification and regression tasks. The regression task consists of mapping several input variables to a numeric output. Usually in the DM process the dataset is splitted into two subsets denominated training set and testing set. The former set is used in a learning process of the algorithms whereas the latter is used to test the algorithms. During the learning process the various parameters of the algorithms are adjusted to optimize the results. The accuracy of the algorithms is assessed through metrics based on errors and the correlation coefficient. The validated algorithms are used as models to predict the value of the output variables.

2.2 DM algorithms

There are several DM algorithms such as Regression Trees (RT), Multiple Regressions (MR), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). A brief description of them will be provided next.

The Decision Trees [25] have an inverted tree structure composed of nodes and descendent branches. The result of a test performed at each node indicates the branch to continue the process. This process is repeated until the final decision can be made and a class is attributed to the register. Regression trees are a particular case of decision trees where classes are replaced by values (Figure 1).

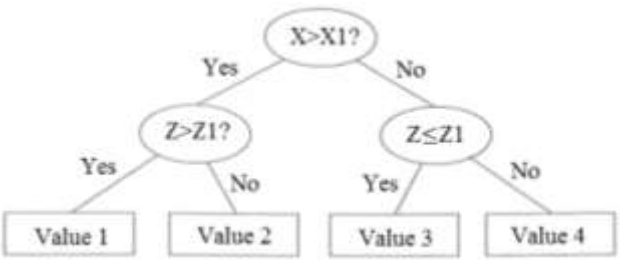

Figure 1: Example of a decision tree.

Multiple regressions are similar to simple regressions but with several independent variables instead of one independent variable.

ANN is a technique that seeks to mimic the way the human brain works. An artificial neural network has an input layer, some hidden layers and an output layer, each one consisting of several neurons. The neurons in two adjacent layers are linked with associated weights wi,j (i and j are neurons or nodes) that determine the importance of the input. Each neuron has an activation function that introduces a non-linear component [26, 27]. This study used a logistic activation function f given by 1/(1+e-x) and the following general equation:

$$\hat{y} = w_{o,0} + \sum_{j=I+1}^{o-1} f\left(\sum_{i=1}^{I} x_i w_{j,i} + w_{j,0}\right) w_{o,i} \tag{1}$$

where x_i are the input parameters or nodes, I is the number of input parameters and o is the output parameter.

The multilayer perceptron (feed forward network) is the most popular neural network architecture [27] (Fig.2). Thus, this type of neural network was adopted in this study with one hidden layer containing HN (hidden nodes) processing units. The ANN performance depend on the number of hidden nodes.

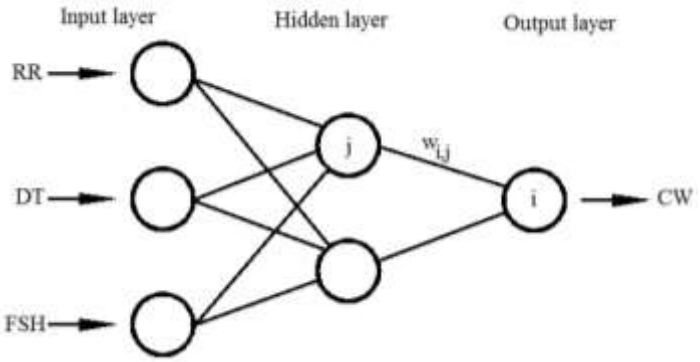

Figure 2: Example of a multilayer perceptron.

The SVM technique was initially proposed for classification problems by Vladimir Vapnik and his coworkers [28]. It uses nonlinear mapping to transform inputs into a multidimensional feature space (Fig. 3). After this transformation, SVM finds the best linear separating hyperplane, related to a set of support vector points, within the feature space. The nonlinear mapping depends on a kernel function K(x,x'). This work uses the Gaussian kernel function, which is the most popular one:

$$k(x, x') = e^{\left(-\gamma \cdot \|x - x'\|^2\right)}, \quad \gamma > 0 \tag{2}$$

The application of SVM to regression problems was possible after the introduction of the ε-insensitive loss function [29]. SVM performance is affected by meta-parameters C, ε and the kernal parameter, γ. Parameter C determines the trade-off between the model complexity (flatness) and the degree to which deviations larger than ε are tolerated in optimization formulation [30]. Parameter ε controls the width of the ε-insensitive zone, used to fit the training data and its value can affect the number of support vectors used to construct the regression function [30]. In order to limit searching space, C and ε were set using heuristics proposed by [30]: C=3 and $\varepsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5 \times \sum_{i=1}^{N} (y_i - \hat{y}_i)$, $\hat{y}_i$ is the value predicted by a 3-nearest neighbor algorithm and N the number of examples. Under this setup, the search space was limited to the input values of γ.

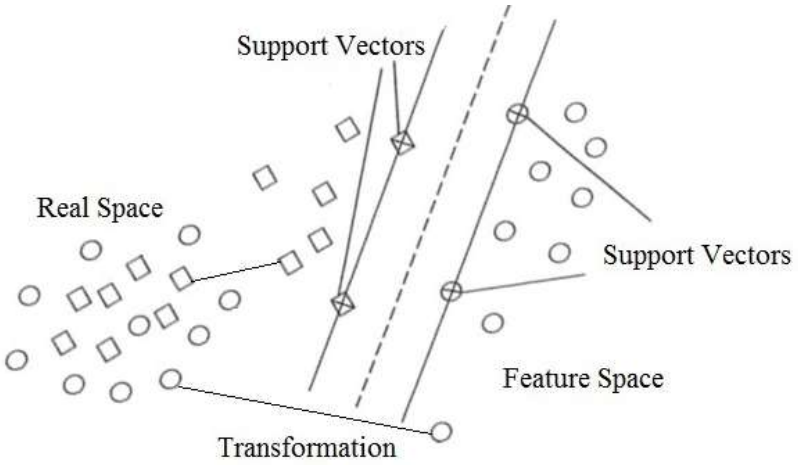

Figure 3: Example of a SVM transformation.

The k-Nearest Neighbor [31] is a simple supervised learning algorithm that can be used in classification and regression problems. In classification problems an instance query is classified according to its neighbors' classes (Figure 4). The dominant class among the nearest neighbors is attributed to the query instance. In regression problems the property value for the instance query is obtained as the average of the weighted values of the k nearest neighbors. This implies calculation of the distance between the target and its neighbors in the multidimensional space. Generally, weights are attributed according to distance. The closest neighbors are given more weight than more distant ones.

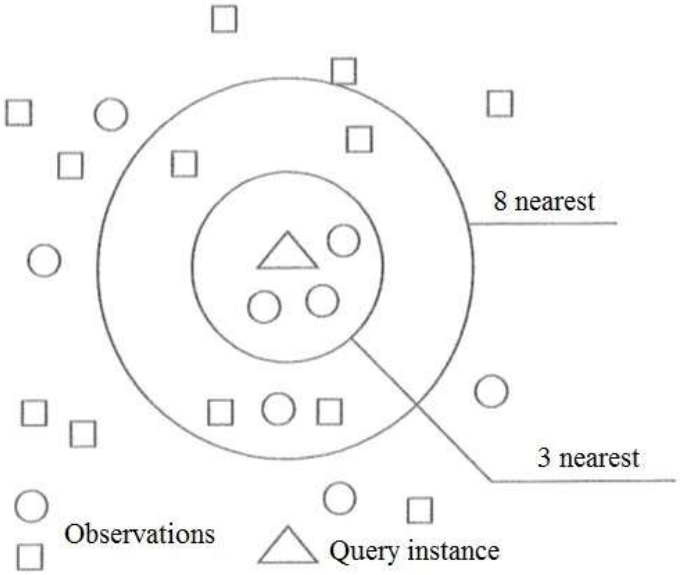

Figure 4: Example of k-NN.

In this work, the whole dataset was divided exactly in the same way used by BILIR *et al.* [1]. Thus, the training set consisted of approximately two-thirds of the total dataset (238 data) and the testing set was constructed with the remaining data (120 data).

In this study the ANN, SVM and k-NN parameters (H, γ and k) were optimized through a grid search of H {0, 2, 4, 6, ..., 20}, γ {2-15, 2-13, ..., 23} and k {2, 3, 4, ..., 12}.

The performance of the models was evaluated through a 10-fold cross-validation process by dividing the training dataset in 10 subsets of equal size [32]. A single subset is retained as the validation data for testing the model, and the 9 remaining subsets are used as training data. The cross-validation process is repeated 10 times, with each of the 10 subsamples used as the validation data. Under this scheme all the data are used for training and testing. By averaging the values obtained in each of the folds, a single value is obtained for

each of the considered performance measures. The model is retrained using all the training dataset whereas the testing dataset, composed of one third of the total data, is only used to validate the model.

The evaluation of the performance of the regression models can be done using different metrics. This study uses the root mean squared error (RMSE) and the coefficient of correlation (R), defined as:

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

$$R = \frac{\sum_{i=1}^{N}(y_i - \bar{y}) \times (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2} \times \sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{4}$$

where N denotes the number of examples, y_i the real value, $\hat{y}_i$ the value estimated by the model, $\bar{y}$ the mean of the real values and $\bar{\hat{y}}$ the mean of the estimated values.

3. DATA USED IN DATA MINING

The database used in this work was presented by BILIR *et al.* [1] and was composed of 358 data sets collected from free shrinkage tests determining the length changes and ring test determination of the restrained shrinkage cracks. These authors developed a model to predict crack width (CW) using this database and adaptative-network-based fuzzy inference system (ANFIS).

Tables 1 and 2 present some statistical data of the parameters used in the analyses. Comparing both tables, it is possible to observe that there are no significant differences between the coefficients of variation of the training and testing data. The coefficients of variation of RR and CW are very similar. The same happens between the coefficients of variation of DT and FSH. This means that the RR-CW and DT-FSH pairs present similar variability.

Table 1: Statistics of the input and the output parameters for the training dataset.

PARAMETERS		MINIMUM	MEAN	MAXIMUM	STD. DEV.	COEF. VAR. (%)
	RR (%)	0	42.94	100	30.74	71.58
Inputs	DT (day)	19	43.47	60	10.65	24.51
	FSH (mstr.)	0.000911	0.002515	0.003880	0.000682	27.12411
Output	CW (mm)	0.003	0.577	1.3	0.405	70.081

Table 2: Statistics of the input and the output parameters for the testing dataset.

PARAMETERS		MINIMUM	MEAN	MAXIMUM	STD. DEV.	COEF. VAR. (%)
	RR (%)	0	43.25	100	30.82	71.26
Inputs	DT (day)	20	43.38	59	10.64	24.53
	FSH (mstr.)	0.000910	0.002508	0.003877	0.000688	27.43314
Output	CW (mm)	0.005	0.572	1.300	0.408	71.234

4. RESULTS AND DISCUSSION

With the databases built, the predictive models were trained to forecast the drying shrinkage crack widths obtained from the ring test (CW).

The data mining models were tested using a single input variable and combinations of two or three variables (RR, DT and FSH). The mean values of the root mean square error (RMSE) and the coefficient of correlation (R) obtained during the training process are presented in Tables 3 and 4.

Table 3: Mean values of R obtained in the cross-validation scheme for different combination of input parameters.

	M1			M2			M3
	RR	DT	FSH	RR&DT	RR&FSH	DT&FSH	RR&DT&FSH
RT	0.921	0.024	0.772	0.941	0.937	0.937	0.943
MR	0.911	0.035	0.326	0.931	0.923	0.923	0.937
ANN	**0.927**	0.076	0.754	0.987	0.974	0.974	**0.993**
SVM	0.918	0.103	0.552	**0.994**	0.980	0.980	0.989
k-NN	0.846	-0.188	0.769	0.987	0.977	0.977	0.984

Table 4: Mean values of RMSE obtained in the cross-validation scheme for different combination of input parameters.

	M1			M2			M3
	RR	DT	FSH	RR&DT	RR&FSH	DT&FSH	RR&DT&FSH
RT	0.158	0.410	0.258	0.137	0.141	0.301	0.134
MR	0.167	0.405	0.382	0.147	0.155	0.384	0.141
ANN	**0.152**	0.415	0.267	0.066	0.092	0.256	**0.049**
SVM	0.161	0.409	0.345	**0.044**	0.081	0.319	0.062
k-NN	0.216	0.437	0.261	0.065	0.086	0.333	0.071

From the analysis of Tables 3 and 4 it can be seen that the best result using only one input parameter was obtained with the ANN and RR input. The best model with two input variables is the SVM with RR and DT inputs. ANN is the best model when all the input parameters are used.

Among all the combinations and models, the best performance was obtained with the SVM model using RR and DT as input variables. This model was fitted with all the training set and the result is graphically presented in Fig. 5, which also presents the values obtained with the testing set. Figures 6 and 7 presents the results for the best models with one and two input parameters. It must be highlighted that the testing data were not used in generating the model.

Figures 5 and 6 confirm the good predictive capacity of the best models that use two or three input parameters. Figure 7 shows a set of scattered points above the 45 degrees line and some sets of points arranged in steps. This confirms the poor predictive capacity of models with only one input variable.

The model developed by BILIR *et al.* [1] presented a good performance. The values of R and RMSE extracted from Figures 5 to 7 and obtained with the BILIR *et al.* [1] model are summarized in Table 5. The values presented in this table confirm the best predictive capacity of the SVM model with RR and DT input parameters. Furthermore, the best models developed in this study with two or three input parameters have a better performance than the model developed by BILIR *et al.* [1].

Table 5: Measures of performance of the best models of this study and the model of BILIR *et al.* [1].

Input		RR	RR&DT	RR&DT&FSH	
Model		ANN	SVM	ANN	ANFIS*
R	Training	0.932	0.999	0.998	0.998
	Testing	0.928	0.997	0.995	0.992
RMSE	Training	0.146	0.021	0.024	0.024
	Testing	0.151	0.029	0.042	0.050

* BILIR *et al.* [1]

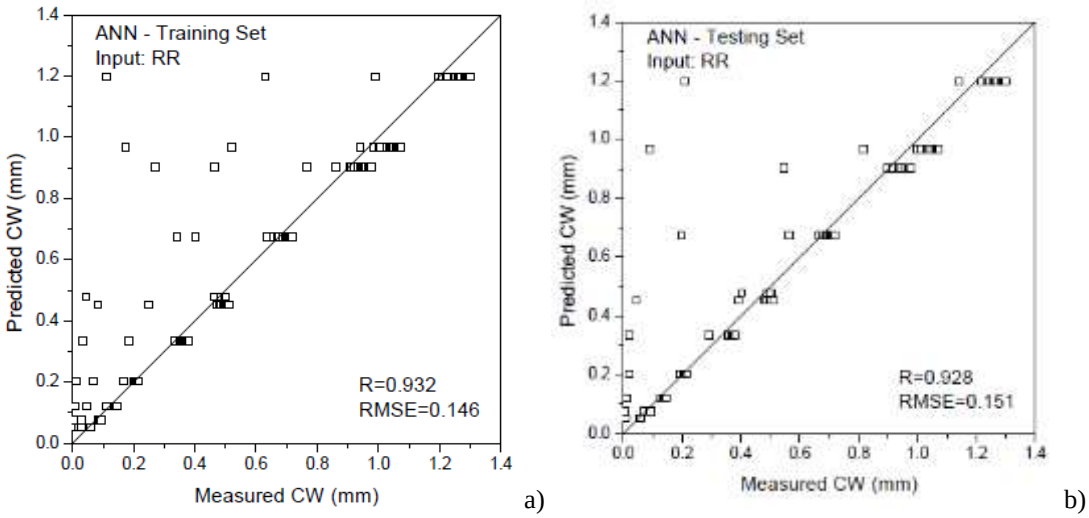

Figure 5: Performance of the SVM model using the combination RR and DT with: a) training data set; b) testing data set.

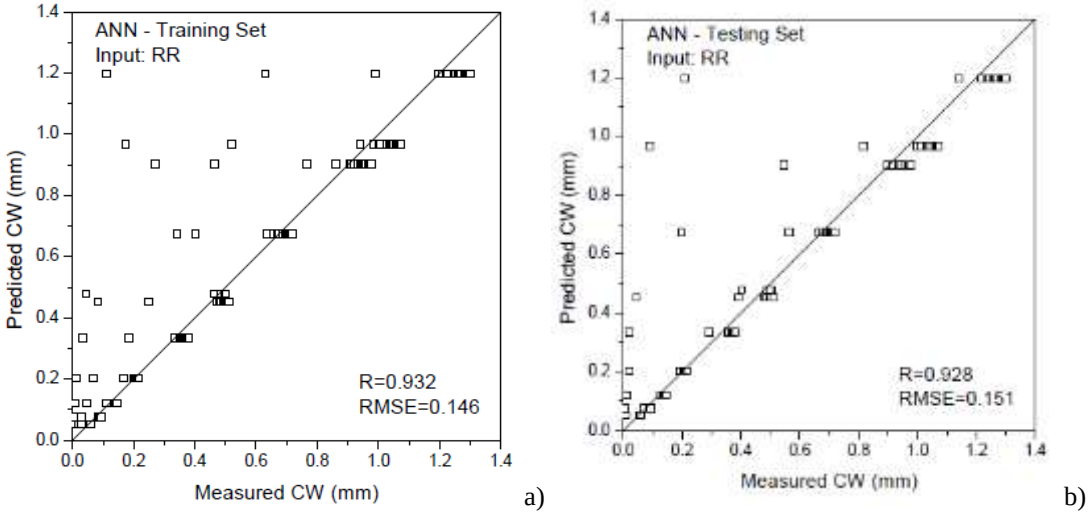

Figure 6: Performance of the ANN model using all the input parameters with: a) training data set; b) testing data set.

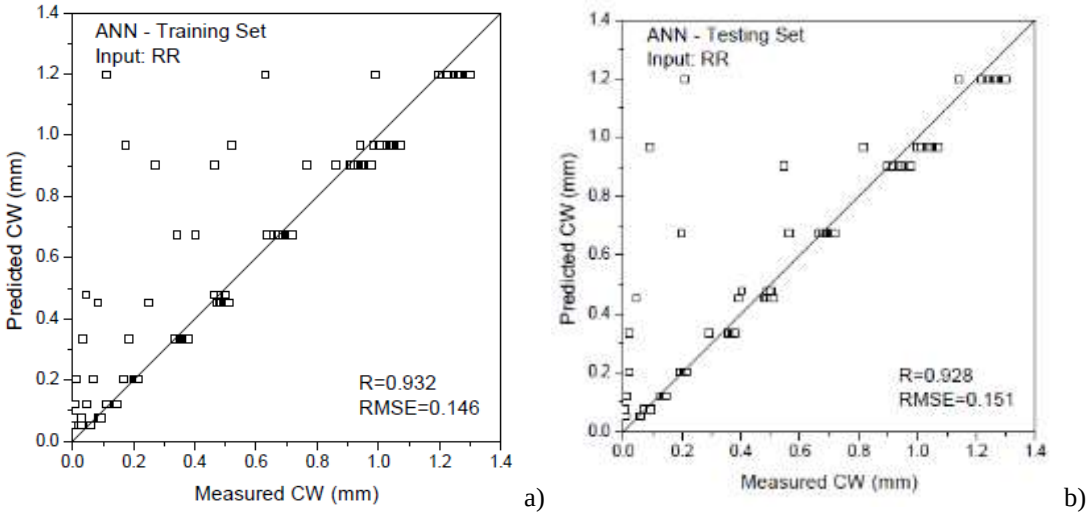

Figure 7: Performance of the ANN model using RR with: a) training data set; b) testing data set.

5. CONCLUSIONS

Data mining techniques have the capacity to learn from examples. In this study several data mining techniques were used to predict drying shrinkage crack widths of mortars containing GBFS as fine aggregate. The cross-validation scheme indicates that with one or three input parameters the ANN models provide the best prediction results. Among all the developed models the SVM using only two input parameters (RR and DT) leads to the best results. Furthermore, this model and ANN with three input variables have better performances than the model ANFIS developed by BILIR *et al.* [1].

6. ACKNOWLEDGMENTS

This work was partly financed by FEDER funds through the Competitivity Factors Operational Programme - COMPETE and by national funds through FCT – Foundation for Science and Technology within the scope of the project POCI-01-0145-FEDER-007633".

7. BIBLIOGRAPHY

[1] BILIR, T., GENCEL, O., TOPÇU İ. B., "Prediction of restrained shrinkage crack widths of slag mortar composites by Takagi and Sugeno ANFIS models", *Neural Comput & Applic*, v.27, pp. 2523-2536, 2016.

[2] ESKANDARI, H., NIK, M. G., EIDI, M. M., "Prediction of mortar compressive strengths for different cement grades in the vicinity of sodium chloride using ANN", *Procedia Engineering*, v.150, pp. 2185-2192, 2016.

[3] ESKANDARI-NADDAF, H., KAZEMI, R., "ANN prediction of cement mortar compressive strength, influence of cement strength class", *Construction and Building Materials*, v.138, pp.1–11, 2017.

[4] SARIDEMIR, M., "Predicting the compressive strength of mortars containing metakaolin by artificial neural networks and fuzzy logic", *Advances in Engineering Software*, v.40, pp. 920-927, 2009.

[5] MAHZUZ, H. M. A., AHMED, M., HOSSAIN, M. M., *et al.*, "Evaluation of strength of mortar using different saw waste with sand and model development", *KSCE Journal of Civil Engineering*, v.20, n.7, 2822-2831, 2016.

[6] RAZAVI, S. V., JUMAAT, M. Z., EI-SHAFIE, A. H., *et al.*, "Artificial neural networks for mechanical strength prediction of lightweight mortar", *Scientific Research and Essays*, v.6, n.16, pp. 3406-3417, 2011.

[7] Topçu, İ. B., Sarıdemir, M., "Prediction of rubberized mortar properties using artificial neural network and fuzzy logic", *Journal of Materials Processing Technology*, v.199, pp.108-118, 2008.

[8] KONIORCZYK, M., WOJCIECHOWSKI, M., "Influence of salt on desorption isotherm and hygral state of cement mortar – Modelling using neural networks", *Construction and Building Materials*, v.23, pp. 2988-2996, 2009.

[9] MOLERO, M., SEGURA, I., IZQUIERDO, M. A. G., *et al.*, "Sand/cement ratio evaluation on mortar using neural networks and ultrasonic transmission inspection", *Ultrasonics*, v. 49, pp. 231-237, 2009.

[10] ONAL, O., OZTURK, A. U., "Artificial neural network application on microstructure–compressive strength relationship of cement mortar", *Advances in Engineering Software*, v.41, pp. 165-169, 2010.

[11] İNAN, G., GÖKTEPE, A. B., RAMYAR, K., SEZER, A., "Prediction of sulphate expansion of PC mortar using adaptive neuro-fuzzy methodology", *Building and Environment*, v. 42, pp. 1264-1269, 2007.

[12] NAZARI, A., SANJAYAN, J. G., "Modelling of compressive strength of geopolymer paste, mortar and concrete by optimized support vector machine", *Ceramics International*, v. 41, pp. 12164-12177, 2015.

[13] AKKURT, S., OZDEMIR, S., TAYFUR, G., AKYOL, B., "The use of GA–ANNs in the modelling of compressive strength of cement mortar", *Cement and Concrete Research*, v. 33, pp. 973-979, 2003.

[14] BAL, L., BUYLE-BODIN, F., "Artificial neural network for predicting drying shrinkage of concrete", *Construction and Building Materials*, v. 38, pp. 248-254, 2013.

[15] GEDAM, B. A., BHANDARI, N. M., UPADHYAY, A., "An apt material model for drying shrinkage and specific creep of HPC using artificial neural network", *Structural Engineering and Mechanics*, v. 52, n.1, 97-113, 2014.

[16] KARTHIKEYAN, J., UPADHYAY, A., BHANDARI, N. M., "Artificial neural network for predicting creep and shrinkage of high performance concrete", *Journal of Advanced Concrete Technology*, v. 6, n.1, pp. 135-142, 2008.

[17] LIU, J., YAN, K. Z., ZHAO, X., *et al*., "Prediction of autogenous shrinkage of concretes by support vector machine", *International Journal of Pavement Research and Technology*, v. 9, n.3, pp. 169-177, 2016.

[18] MERMERDAŞ, K. AND ARBILI, M. M., "Explicit formulation of drying and autogenous shrinkage of concretes with binary and ternary blends of silica fume and fly ash", *Construction and Building Materials*, v.94, pp. 371-379, 2015.

[19] DARAIN, K. M., SHAMSHIRBAND, S., JUMAAT, M. Z., *et al*., "Adaptive neuro fuzzy prediction of deflection and cracking behavior of NSM strengthened RC beams", *Construction and Building Materials*, v.98, pp. 276-285, 2015.

[20] LI, G., ZHAO, X., DU, K., *et al*., "Recognition and evaluation of bridge cracks with modified active contour model and greedy search-based support vector machine", *Automation in Construction*, v.78, pp.51-61, 2017.

[21] ROSALES, M. B., FILIPICH, C. P., BUEZAS, F. S., "Crack detection in beam-like structures", *Engineering Structures*, v.31, pp. 2257-2264, 2009.

[22] SAEED, R. A., GALYBIN, A. N. AND POPOV, V., "Crack identification in curvilinear beams by using ANN and ANFIS based on natural frequencies and frequency response functions", *Neural Comput & Applic*, v.21, pp. 1629-1645, 2012.

[23] SULEIMAN, A.R. AND NEHDI, M. L., "Modelling self-healing of concrete using hybrid genetic algorithm–Artificial Neural Network", *Materials*, v. 10, n.135, pp. 1-15, 2017.

[24] FAYYAD, U., PIATESKY-SHAPIRO, G., SMYTH, P., "From data mining to knowledge discovery: an overview", In: Fayyad et al. (eds) *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press, Cambridge MA, pp. 471-493, 1996.

[25] QUINLAN, J., "Induction of decision trees", *Machine Learning*, Kluwer Academic Publishers 1, pp. 81-106, 1986.

[26] CORTEZ, P., "Data mining with neural networks and support vector machines using the R/rminer tool", Proceedings of *10th Industrial Conference on Data Mining, Advances in Data Mining, Applications and theoretical aspects*, (P. Pernered.), Lecture Notes in Computer Science, Springer, Berlin, Germany, pp. 572-583, 2010.

[27] HAYKIN, S., *Neural networks - a comprehensive foundation*, 2nd ed. New Jersey, Prentice-Hall, 1999.

[28] CORTES, C., VAPNIK, V., "Support vector networks", *Machine Learning*. Kluwer Academic Publishers, v. 20, n.3, pp. 273-297, 1995.

[29] SMOLA, A., SCHOLKOPF, B., "A tutorial on support vector regression", *Stat Computers*, v.14, n. 3, pp. 199-222, 2004.

[30] CHERKASSKY, V., MA, Y., "Practical selection of SVM parameters and noise estimation for SVM regression", *Neural Networks*, v.17, n.1, pp. 113-26, 2004.

[31] HECHENBICHLER, K., SCHLIEP, K., "Weighted k-nearest-neighbor techniques and ordinal classification", Discussion Paper 399, *SFB 386*, Ludwig-Maximilians University Munich, 2004.

[32] EFRON, B., TIBSHIRANI, R., *An introduction to the bootstrap*, Chapman & Hall, 1993.

ORCID

Francisco Ferreira Martins https://orcid.org/0000-0001-6088-7860
Aires Camões https://orcid.org/0000-0002-9677-3627