PREDICTION OF THE COMPRESSIVE STRENGTH OF CONCRETE BY MACHINE LEARING REGRESSION MODELS

*Hana Schreiberová, **

Katedra betonových a zděných konstrukcí, Fakulta stavební,
České vysoké učení technické v Praze, Thákurova 7/2077, 166 29 Praha 6, Česká republika.
hana.schreiberova@fsv.cvut.cz

ABSTRAKT

Téma predikce mechanických vlastností cementových kompozitů se těší značnému zájmu, jelikož by díky ní mohlo dojít ke snížení potřeby nákladných a pracných laboratorních zkoušek. V tomto článku je použito několik modelů strojového učení (Lineární, Hřebenová, Lasso regrese a Metoda podpůrných vektorů), které jsou natrénovány a vyhodnoceny na veřejně dostupném datasetu, který obsahuje velké množství různých receptur a jejich tlakových pevností měřených v různých časech od betonáže. Studie ukázala, že Metoda podpůrných vektorů dosahuje nejvyšších pevností (průměrná absolutní odchylka byla 3.63 MPa). Natrénované modely pak byly aplikovány na další aktuálnější externí data. Bohužel se ukázalo, že žádný z modelů nedokáže nová data predikovat s dostatečnou přesností. Jedním z možných důvodů je zde uvedena i nedostatečná reprezentativnost staršího veřejného datasetu pro aktuálně používané směsi.

KLÍČOVÁ SLOVA

Predikce • Strojové učení • Tlaková pevnost • Beton • Modelování

ABSTRACT

Prediction of mechanical properties of cementitious composites is a topic of great concern as it could minimize the need for costly and laborious laboratory tests. In this paper, several machine learning models (Linear, Ridge, Lasso, and Support Vector Machine regression) are trained and evaluated on a publicly available dataset containing various concrete compositions and their compressive strength measured at different ages from casting. In this study, Support Vector Machine regression showed the highest accuracy when testing on the public dataset (mean absolute error 3.63 MPa). The trained models were also subsequently applied on additional more current data. Unfortunately, none of the models proved to be suitable which might be due to the low representativeness of the older public dataset for the currently used mixtures.

KEYWORDS

Prediction • Machine Learning • Compressive Strength • Concrete • Modeling

1. INTRODUCTION

Prediction of mechanical properties of concrete based on its composition is a topic of great concern in the field of building materials. Estimates are traditionally made based on empirical relationships; however, with the increasing variety of concrete compositions, the common approach is becoming insufficient and has restricted validity.

Several studies employed regression models to estimate chosen characteristics, such as compressive strength (Popovics and Ujhelyi 2008), shear strength (Slater, Moni, and Alam 2012), or tensile strength (Silva, de Brito, and Dhir 2015). However, further research concluded that more complex systems are needed, especially in the case of special mix designs such as high-performance concrete (Chou and Tsai 2012). For this reason, especially in the last decade, machine learning (ML) techniques have been employed for prediction tasks.

ML modeling starts with the identification of a target variable and explanatory variables for which we expect some dependency. To be able to evaluate the performance of the developed model, it is common practice to split available data into training sample on which model parameters are estimated, and a test sample which is left aside during development and used to assess how the model behaves when applied to unseen data.

A variety of ML models have been applied for the prediction of concrete properties, mostly its compressive strength. Linear and regularized regression (Kang, Yoo, and Gupta 2021), support vector regression and classification (Kang, Yoo, and Gupta 2021; Duan et al. 2020; Nguyen et al. 2021), boosting- and tree-based models (Kang, Yoo, and Gupta 2021; Kaloop et al. 2020; Duan et al. 2020; Nguyen et al. 2021; Vakharia and Gujar 2019). Further, studies using more advanced models such as artificial neural networks (ANNs) have been conducted (Sevim et al. 2021; Vakharia and Gujar 2019; Chou et al. 2014).

In this paper, selected ML models (Linear, Lasso, Ridge, and support vector machine regression) are trained and evaluated on a publicly available dataset containing 1030 concrete mix designs with/without blast furnace slag, fly ash, and superplasticizer and respective compressive strengths measured at various times from casting. Additionally, selected trained models are applied on unknown data obtained from more current papers, thus not present in the source dataset, in order to validate the model on mixtures with more detailed composition.

2. DATA

The experimental dataset was obtained from UCI Machine Learning Repository where it was donated by Prof. I.-C. Yeh (Yeh 1998) in 2007 for unlimited use. The dataset contains records of concrete compositions (with/without blast furnace slag, fly-ash, and superplasticizer) and their respective compressive strength values obtained from load tests. According to (Yeh 1998), the dataset was prepared using 17 different

* Školitel: prof. Ing. Alena Kohoutková, CSc.

sources and evaluated so the mixtures were fairly representative for all of the major parameters influencing the compressive strength of concrete. Further, some mixtures were omitted due to their atypical composition or curing conditions. As different studies have dealt with various specimen types for the compressive strength determination, the values were converted into 150 mm cylinders according to the relevant standards. As I.-C. Yeh points out in (Yeh 1998), in some cases, detailed information about individual components was missing e.g. the class of fly-ash or the exact chemical composition of superplasticizer. More detailed information concerning this dataset can be found in the aforementioned journal article.

2.1. Dataset Overview

The dataset contains 9 variables and 1030 entries. As the compressive strength value (f_c) is the targeted variable for the prediction task, the remaining 8 variables are used as the explanatory variables on which the target variable depends. Table 1 shows the ranges of the data. A similar overview can be found in (Yeh 1998), although different values are reported, possibly due to subsequent alternations of the dataset. The most frequent age of testing in the dataset is the standard 28th day from casting.

Table 1: *Ranges of the variables.*

Variable	Unit	Min	Max	Median
Cement	[kg/m^3]	102.0	540.0	272.9
Water	[kg/m^3]	121.8	247.0	185.0
Slag	[kg/m^3]	0.0	359.4	22.0
Fly-ash	[kg/m^3]	0.0	200.1	0.0
Superplast.	[kg/m^3]	0.0	32.2	6.4
Coarse agg.	[kg/m^3]	801.0	1145.0	968.0
Fine agg.	[kg/m^3]	594.0	992.6	779.5
Age	[days]	1	365	28
f_c	[MPa]	2.3	82.6	34.4

3. DATASET PREPARATION

3.1. Splitting of the Data into a Test and Train Dataset

As already briefly described in Introduction, in order to evaluate the model performance (i.e., estimate its generalization error using metrics described further), the data need to be split into a train and test set.

In this paper, 20 % of the available data was used as testing data, which is a common practice. Firstly, the dataset was split randomly, and the test set representativeness was verified as described further.

Based on the correlation between the target and explanatory variables (measured using Pearson's correlation coefficients which describes linear correlation), the cement dose was identified as the main driving factor influencing the targeted compressive strength. For this reason, the cement dose values were categorized and their relative representation in the test and train set was determined. As the ratios differed quite significantly, the data were further split by stratified sampling so the cement dose categories would be evenly represented in both the test and train set.

3.2. Data Transformation

In order to possibly improve the performance of the selected machine learning models, primary explanatory variables (i.e., concrete composition and age) were further transformed.

Primary analysis revealed that the target variable (compressive strength) expresses the most pronounced linear dependency on the amount of cement. For that reason, a variable set as the logarithm of the cement dose was added into the datasets in order to achieve its linearization.

Further additional variables were created as ratios of individual components to the amount of binder (this approach was also chosen in (Yeh 1998)). As binder, the sum of cement, fly-ash, and slag was considered. Table 2 illustrates the, in some cases, enhanced correlation coefficients (i.e., linear correlation) when the concrete composition is expressed in ratios.

Table 2: *Comparison of the Pearson's correlation coefficient*

Variable	Corr. coef.	Variable	Corr. coef.
f_c	1.0	fc	1.0
Cement	0.5	–	–
Superplasticizer (Sp)	0.36	Sp/b	0.24
Slag (S)	0.11	S/b	0.01
Fly-ash (FlA)	-0.09	FlA/b	-0.16
Fine Agg. (FA)	-0.15	FA/b	-0.54
Coarse Agg. (CA)	-0.18	CA/b	-0.56
Water (w)	-0.28	w/b	-0.63

If necessary for the particular model application (regularized models and support vector machine regression), the datasets were transformed to have the same scale and unit variance of the resulting distribution. Standardization was performed according to Eq. 1:

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where z is the standardized value, x is the original value, μ is the mean value, and σ is the variance.

3.3. The Train Dataset Exploratory Analysis

To gain a greater understanding of the data, a brief exploratory analysis was performed. Only train data were used so the test data remained truly unknown and the unfavorable bias avoided.

As it is apparent from Figures 3,4, and 5, some tendencies between the target variable and composition ratios are observable; however, the variability outside the upper and lower quartiles is significant in all cases due to a large number of additional influencing variables.

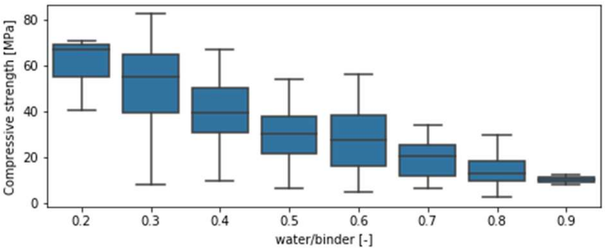

Figure 1: *Dependence of the compressive strength on water/binder ratio.*

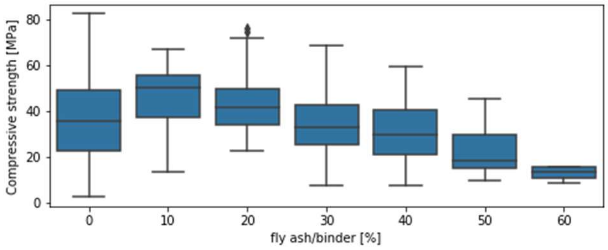

Figure 2: *Dependence of the compressive strength on fly-ash/binder ratio.*

4. MODEL PREPARATION

4.1. Applied Machine Learning Models

In this part of the paper, selected predictive models used for the task are briefly introduced.

4.1.1. Linear Regression

The Linear regression model describes the target value (scalar response) as a linear combination of the independent explanatory variables (features), as shown in Eq. 2:

$$y(\theta, x) = \theta_0 + \theta_1 x_1 + \ldots \theta_n x_n, \tag{2}$$

where y is the value of the target variable, n is the number of features, x_i is the ith feature value, and θ_j is the jth model parameter (where θ_0 is the bias term). This can be expressed in a vectorized form, as shown in Eq. 3:

$$\hat{y} = h_\theta(\mathbf{x}) = \theta^T.\mathbf{x}, \tag{3}$$

where h_θ is the regression function using the model parameters θ. In this paper, the model is fitted using the method of least squares where the mean squared error cost function is minimized by finding optimal parameter values as a solution to the problem shown in Eq. 4:

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\theta}.\boldsymbol{x} - y_i)^2. \tag{4}$$

For this problem, a closed solution exists, and it is given by Eq. 5. From the equation, we see that the inverse of the matrix needs to be computed. This can be computationally intractable for large numbers of predictors. This is not our case; however, it can be solved by methods as stochastic gradient descent.

$$\hat{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}.(\boldsymbol{X}^T\mathbf{y}), \tag{5}$$

where $\hat{\theta}$ is the parameter value which minimizes the cost function, X is the matrix of features, and $\boldsymbol{y}$ is the vector of values of the target variable.

4.1.2. Polynomial Regression

The primary analysis of the dataset indicated that the dependence of the target variable on the explanatory variables could be nonlinear. This relationship can be described by a special type of Linear regression – Polynomial regression. Although it is still a linear problem, as it is linear in the unknown parameters, the relationship between the target variable and explanatory variables is modeled as a pth degree polynomial in the explanatory variables, as shown in Eq. 6 for one explanatory variable:

$$y(\theta, x) = (\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots \theta_\mathrm{n} x_\mathrm{n})^p. \tag{6}$$

This more complicated model gives us the ability to alter the weight of each explanatory variable depending on the value(s) of one or more other independent variables thanks to the interaction effect.

Although the addition of a higher polynomial degree of features and their combination can be highly beneficial for nonlinear data, we must keep in mind that the transformation leads to a severe explosion of the feature numbers, possibly making the model too slow.

4.1.3. Regularized Regression (Ridge and Lasso)

In this study, regularized linear models (Ridge and Lasso) were used alongside with basic Linear regression model. In order to prevent overfitting of the model on training data, so it is able to sufficiently generalize on test data, models are regularized by the so-called regularization term which is added to the initial cost function. By regularization, we minimize the variance error without substantially increasing the bias error of the selected model. In both cases, the degree of regularization is given by a hyperparameter α. As the value of the hyperparameter increases, so does the degree of regularization. If the value is set to zero, the cost function equals the initial cost function in basic Linear regression.

In the case of Ridge regression, the regularization term is equal to $\alpha\sum_{i=1}^{n}\theta_i^2$. In Ridge, all of the parameters are equally constrained to take on only small values.

In the case of Lasso regression, the regularization term is equal to $\alpha\sum_{i=1}^{n}|\theta_i|$. Unlike Ridge regression, Lasso regression tends to completely eliminate the weights (parameters) of the least important features (i.e., set them to zero). By setting certain coefficients to zero, Lasso regression provides feature selection, thus improving the model interpretability.

Bayesian interpretation of the Lasso tendency to set parameters to zero was provided by Tibshirani (Tibshirani 1996). The study pointed out that in the case of Ridge regression, the coefficients have a normal distribution, whereas in Lasso regression they have double-exponential distribution (also known as Laplace distribution). As the double-exponential distribution puts more mass near zero and in the tails, the Lasso tends to produce estimates that are either large or zero.

For regularized models, standardized explanatory variables were used in all cases to ensure penalization of each member to the same extent and independence on units in which the variables were given.

4.1.4. Support Vector Machine Regression

Support vector machine (SVM) analysis is a Machine Learning model introduced by Vapnik *et al.* (Vapnik 1995) which is suitable primarily for classification tasks, but also for regression tasks as it is in our case.

Firstly, the SVM model will be briefly introduced on a classification problem for the sake of clarification. Simply put, the SVM analysis aims to determine a line or hyperplane, in the case of multidimensional space, that separates defined classes so new instances are classified (i.e., predicted) based on their position in relation to the line/hyperplane. The line/hyperplane is also accompanied by decision boundaries, defined by the nearest instances (the so-called support vectors), which determine the boundaries between positive and negative examples. The SVM analysis aims to fit the widest street (i.e., the area between the decision boundaries) between the classes with as few margin violations (i.e., instances on the street) as possible.

The aim of the SVM model in the case of a regression task is exactly the opposite. The model tries to fit the instances on the street while limiting the number of instances off the street. The SVM regression model has two hyperparameters – ε which determines the width of the street (i.e., determines the tolerable error), and C which determines the degree of regularization (the higher the C value, the less regularization).

In the case of nonlinear data, the application of the SVM model is possible as in the case of Linear regression. There are several approaches to Nonlinear SVR regression. Firstly, the addition of powers of features and their combination is possible in the same manner as described in the Polynomial regression. Further, the addition of the so-called Similarity features defined by Gaussian Radial Basis Function (RBF) is an option.

However, both of the mentioned approaches lead to a drastic increase in the number of features, slowing down the model greatly. The solution to the nonlinearity issue lies in the employment of special kernel functions (the so-called kernel trick) which replace the need for increasing the number of features while ensuring the same result. For example, the addition of feature polynomials can be substituted with Polynomial kernel while the Similarity features by RBF kernel.

4.2. Evaluation of the Model Performance

In this paper, four metrics were used for evaluation of the model accuracy – Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, and R-squared.

Mean Squared Error (MSE) measures the variance of the residuals, as shown in Eq. 7:

$$MSE = \frac{1}{m}\sum_{j=1}^{m}(y_i - \hat{y}_i)^2, \tag{7}$$

where *m* is the number of instances, y_i is the ith measured target value, and $\hat{y}_i$ is the ith predicted value.

Root Mean Squared Error (RMSE) measures the standard deviation of residuals, as shown in Eq. 8:

$$RMSE = \sqrt{MSE}. \tag{8}$$

Mean Absolute Error (MAE) measures the average of the absolute difference between the actual and predicted values in the dataset, as shown in Eq. 9:

$$MAE = \frac{1}{m}\sum_{j=1}^{m}|y_i - \hat{y}_i|, \tag{9}$$

where *m* is the number of instances, y_i is the ith measured target value, and $\hat{y}_i$ is the ith predicted value. Although RMSE and MAE have the same units as the predicted variable, their values can differ. MAE is a linear score, thus weighting all of the individual differences equally. RMSE, on the other hand, gives high weights to large errors due to their exponentiation. It implies that the larger their difference, the higher the variability of the errors.

R-squared (R^2) represents the proportion of the variance in the dependent variable, as shown in Eq. 10:

$$R^2 = 1 - \frac{\sum_{j=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{j=1}^{m}(y_i - \bar{y}_i)^2}, \tag{10}$$

where *m* is the number of instances, y_i is the ith measured target value, $\hat{y}_i$ is the ith predicted value, and $\bar{y}$ is the mean value. The R^2 takes values less than or equal to 1 where 1 means a perfect correlation. It is important to acknowledge that by adding more independent explanatory variables, the R2 score tends to rise. Thus, it may lead to the introduction of redundant variables in the model.

In all of the cases, models were firstly evaluated using K-fold cross-validation. Thus, the training set was split into k subsets (in our paper k = 5). The models were then trained on k-1 subsets and evaluated on the remaining one, the so-called validation set. This process was repeated by switching the validation set. In our study, we obtained five R^2, MSE, and MAE values on the training set. This procedure is generally intended to determine the model prediction accuracy while reducing the impact of the specific test set selection. The evaluation metrics were also determined using the prediction on the test set.

4.3. Hyperparameters Tuning

As described in chapter 4.1, regularized regression models and SVM regression are defined by certain hyperparameters which influence the fitting of the model, thus its achievable prediction accuracy.

So that it was not necessary to search for optimal hyperparameter values manually, the so-called Grid Search cross-validation was employed. By its application, the hyperparameter values in the specified range and all of their possible combinations are evaluated automatically and, based on the evaluation scores (R^2), their optimal values are determined.

5. RESULTS

5.1. Training and Performance of the Models

In this part of the paper, the performance of the applied ML models is presented. In Table 3, an overview of all of the evaluation metrics of the models using various datasets is given. Further, based on the obtained information, the most suitable model and hyperparameter values are herein determined.

Table 3: *An overview of the trained models and their metrics from cross-validation (CV) and test data prediction.*

Model	Nonlin.	Hyperparam.	CV Mean(R^2) Best CV R^2	R^2	MSE	RMSE	MAE
			[-]	[-]	[MPa]	[MPa]	[MPa]
Linear	–	–	0.62	0.40	114.79	10.71	8,49
Poly.	2nd order	–	0.75	0.79	57.66	7.59	5.59
Ridge	2nd order	α = 0.24e-2	0.79	0.81	50.52	7.11	5.35
Lasso	–	α = 0.24e-1	0.62	0.40	114.30	10.69	8.46
SVM	rbf kernel	C = 1000; ε = 3	0.87	0.89	27.96	5.29	3.63

As expected, based on the clearly non-linear relationships between features and the target variable, basic Linear regression did perform rather poorly. The low R^2 values and high mean errors (MAE 8.49 MPa) on the test data suggest that the Linear regression model is underfitting as it is not complex enough to describe the dependencies between the data sufficiently. A comparison of the predicted and measured compressive strength values can be seen in Figure 3.

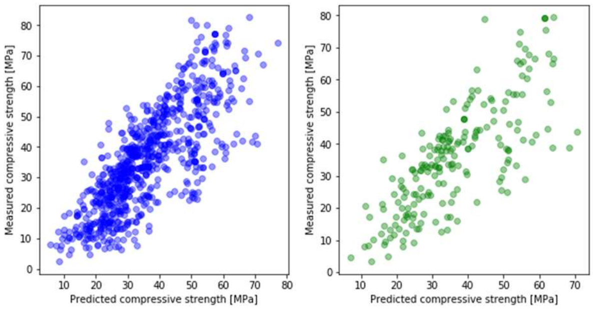

Figure 3: *Measured/predicted compressive strength values by Linear regression on the train set (left) and test set (right).*

The addition of squares of the explanatory variables and their combination in the case of Polynomial linear regression improved the model performance radically (MAE 5.59 MPa on the test set), as illustrated in Figure 4.

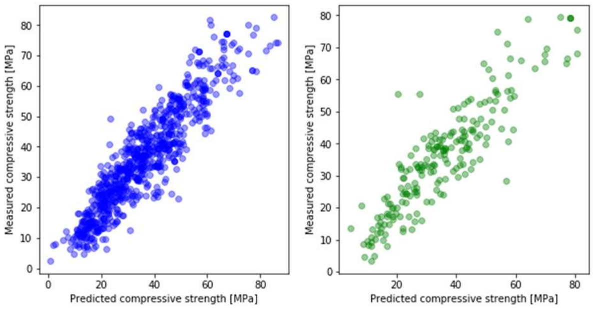

Figure 4: *Measured and predicted compressive strength values by Polynomial Linear regression on the train set (left) and test set (right).*

The Ridge regression model showed higher accuracy when the explanatory features were polynomially transformed in an identical way as in the case of Polynomial regression. As it is apparent from Table 3, the optimal value of alpha (a hyperparameter determining the regularization degree) was determined by the searching algorithm to be rather low. Thus, only a small difference can be expected compared to the unregularized Polynomial regression. The results proved this assumption, as can be seen in Figure 5 and Table 3.

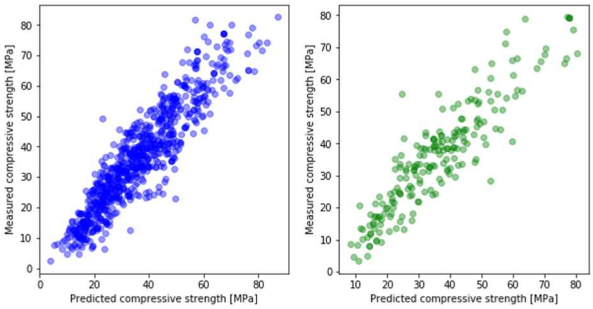

Figure 5: *Measured/predicted compressive strength values by Ridge regression on the train set (left) and test set (right).*

In the case of Lasso regression, the polynomial transformation of features led to a drastic drop in the model accuracy. When considering the 1st order features, the model performance varied from the basic Linear regression only negligibly as the optimal values alpha were determined to be close to zero. Thus, the parameters estimated by the Lasso model were not far from the ones estimated by the basic Linear regression model. Based on these findings the Lasso regression model proved to be unsatisfactory as the regularization did not seem to improve the performance of the Linear regression model (Figure 6).

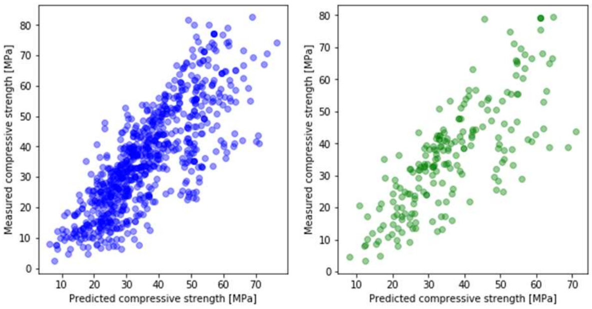

Figure 6: *Measured/predicted compressive strength values by Lasso regression on the train set (left) and test set (right).*

In our paper, the SVM regression had the best prediction accuracy from the applied models (MAE 3.63 MPa), see Figure 7. Grid Search cross-validation determined the optimal values of the hyperparameters C, ε, and the suitable kernel function. RBF kernel function was determined to be the most appropriate. Thus, the non-linearity of the data was handled by the addition of the Similarity features defined by Gaussian Radial Basis Function (as described in chapter 4.1).

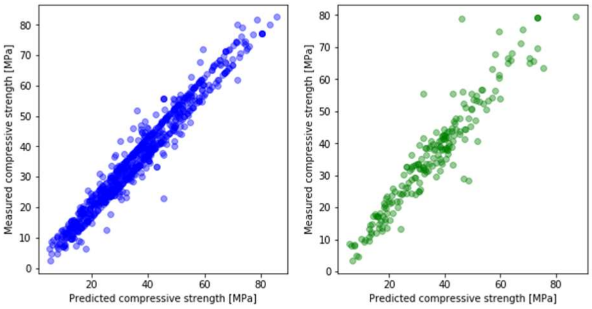

Figure 7: *Measured/predicted compressive strength values by SVM regression on the train set (left) and test set (right).*

5.2. Prediction on Unknown External Data

In this part of the paper, the trained models were confronted with external data obtained from current journal papers. The comparison of the predicted compressive strengths by the models trained in this study and the real measured values should determine the representativeness of the dataset of currently used concrete mix designs.

As the external data, mix designs from (Fantu et al. 2021) and selected mix designs (REF, POP10, POP20, and POP30) from (Bily et al. 2020) were used. The values of compressive strengths were converted from cubic to cylindrical to match the original training dataset. Further, the Linear regression, Polynomial regression, and SVM regression models were applied to predict the compressive strength values.

Unfortunately, despite the overall satisfactory accuracy of the models on the original dataset, the models performed poorly on the external data, as can be seen in Table 4. The less complex Linear regression model consistently underestimated the results; however, due to its lower variance, the predictions did not reach unrealistic values. Further, although the original dataset did not involve such high values of the compressive strength (the maximum was 82.60 MPa), the Linear regression tried to extrapolate from the known values.

On the contrary, the more flexible models (i.e., Polynomial regression and SVM regression) with more degrees of freedom showed clear signs of overfitting of the model on the train set as some values were predicted rather accurately while others were dramatically far from the measured value (MAE 36.4 and 36.1 MPa for Polynomial and SVM regression respectively).

Table 4: Comparison of the measured and predicted compressive strength values using external data.

Measured [MPa]	**Linear.** [MPa]	**Polynom.** [MPa]	**SVM** [MPa]
108.02	56.66	171.76	13.40
101.27	60.46	142.76	18.56
114.76	68.38	15.26	23.18
119.04	69.94	-18.24	22.62
61.28	43.96	46.26	50.20
59.09	47.11	41.51	72.99
60.11	47.41	42.51	70.70
60.90	47.62	44.51	66.54
58.71	47.73	48.26	61.10
57.67	47.73	51.51	54.88
57.45	47.60	58.01	48.27
54.15	47.32	64.51	41.53

6. DISCUSSION AND CONCLUSIONS

This paper aims to introduce several Machine Learning models for the prediction of the compressive strength based on the concrete composition. The models were trained and evaluated on a publicly available dataset and further applied on new external data.

Our findings were in line with the previous research which considered basic linear regression models to be insufficient for the complex non-linear relationships between material composition and strength (Ben Chaabene, Flah, and Nehdi 2020). In our study, the application of the most flexible model – Support Vector Machine regression, led to the most accurate prediction of the target value on the test data (MEA as low as 3.63 MPa). In practice this could be considered to be a highly satisfactory result, as the tolerable deviation when classifying mixtures is typically in the order of percentage units.

Unfortunately, the models did not perform well on the additionally acquired more current data as all of the evaluation metrics were highly unsatisfactory. However, this finding does not automatically mark the model's parameters to be poorly estimated. Rather, it indicates that the original data are not representative enough for the current mix designs. Furthermore, the original publicly available dataset contained only a limited amount of information about the compositions and testing procedures.

This finding can be highlighted as an important point in this study, as it outlines the need for continuous, correct, and comprehensive data collection. The application of Machine Learning algorithms is a highly promising approach to numerous issues in civil engineering; however, it is not feasible with a sufficient data supply.

ACKNOWLEDGEMENTS

This paper was prepared thanks to the support project "Durability of concrete structure and assessment of its life cycle" (SGS19/149/OHK1/3T/11) and "Analysis of the resistance of selected critical infrastructure structures exposed to accidental loading" (SGS21/040/OHK1/1T/11)

References

Bily, P., *et al.* (2020), Micromechanical Characteristics of High-Performance Concrete Subjected to Modifications of Composition and Homogenization, *Mag. of Civ. Eng.* 94 (2), 145–57

Chaabene, Wassim, *et al.* (2020) Machine Learning Prediction of Mechanical Properties of Concrete: Critical Review, *Const. and Build. Mat.* 260, 119889

Chou, Jui-Sheng, *et al.* (2012) Concrete Compressive Strength Analysis Using a Combined Classification and Regression Technique, *Automat. in Const.* 24, 52–60

Chou, Jui-Sheng, *et al.* (2014) Machine Learning in Concrete Strength Simulations: Multi-Nation Data Analytics." *Const. and Build. Mat.* 73, 771–80

Duan, Jin, *et al.* (2020) A Novel Artificial Intelligence Technique to Predict Compressive Strength of Recycled Aggregate Concrete Using ICA-XGBoost Model, *Eng. with Comp.*

Fantu, *et al.* (2021) Materials Today : Proceedings Experimental Investigation of Compressive Strength for Fly Ash on High Strength Concrete C-55 Grade, *Mat. Today: Proceedings*

Kaloop, Mosbeh R, *et al.* (2020) Compressive Strength Prediction of High-Performance Concrete Using Gradient Tree Boosting Machine, *Const. and Build. Mat.* 264, 120198

Kang, Min-Chang, *et al.* (2021) Machine Learning-Based Prediction for Compressive and Flexural Strengths of Steel Fiber-Reinforced Concrete. *Const. and Build. Mat.* 266, 121117

Nguyen, Hoang, *et al.* (2021) Efficient Machine Learning Models for Prediction of Concrete Strengths, *Const. and Build. Mat.*266, 120950

Popovics, Sandor, *et al.* (2008) Contribution to the Concrete Strength versus Water-Cement Ratio Relationship. *J. of Mat. in Civil Eng.* 20 (7), 459–63

Sevim, Umur Korkut, *et al.* (2021) Compressive Strength Prediction Models for Cementitious Composites with Fly Ash Using Machine Learning Techniques, *Const. and Build. Mat.* 271, 121584

Silva, R V, *et al.* (2015) Tensile Strength Behaviour of Recycled Aggregate Concrete, *Const. and Build. Mat.* 83, 108–18

Slater, Emma *et al.* (2012) Predicting the Shear Strength of Steel Fiber Reinforced Concrete Beams, *Const. and Build. Mat.* 26 (1): 423–36

Tibshirani, Robert. (1996) Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1), 267–88.

Vakharia, Vinay *et al.* (2019) Prediction of Compressive Strength and Portland Cement Composition Using Cross-Validation and Feature Ranking Techniques, *Const. and Build. Mat.* 225: 292–301

Vapnik, Vladimir N. (1995) *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag.

Yeh, I.-C. (1998) Modeling of Strength of High-Performance Concrete Using Artificial Neural Networks, *Cem. and Conc. Research* 28 (12): 1797–1808