Using machine learning to predict concrete's strength: learning from small datasets

Engineering Research Express

PAPER

Using machine learning to predict concrete's strength: learning from small datasets

RECEIVED
3 November 2020

REVISED
27 January 2021

ACCEPTED FOR PUBLICATION
4 February 2021

PUBLISHED
17 February 2021

Boya Ouyang[1,2], Yu Song[1,3], Yuhai Li[1], Feishu Wu[1], Huizi Yu[1], Yongzhe Wang[1], Zhanyuan Yin[1], Xiaoshu Luo[1], Gaurav Sant[2,3,4] and Mathieu Bauchy[1,4]

[1] Physics of AmoRphous and Inorganic Solids Laboratory (PARISlab), Department of Civil and Environmental Engineering, University of California, Los Angeles, CA 90095, United States of America
[2] Department of Materials Science and Engineering, University of California, Los Angeles, CA, United States of America
[3] Laboratory for the Chemistry of Construction Materials (LC^2), Department of Civil and Environmental Engineering, University of California, Los Angeles, CA, United States of America
[4] Institute for Carbon Management, University of California, Los Angeles, CA, United States of America

E-mail: bauchy@ucla.edu

Keywords: concrete, strength, machine learning, stratification

Abstract

Despite previous efforts to map the proportioning of a concrete to its strength, a robust knowledge-based model enabling accurate strength predictions is still lacking. As an alternative to physical or chemical-based models, data-driven machine learning methods offer a promising pathway to address this problem. Although machine learning can infer the complex, non-linear, non-additive relationship between concrete mixture proportions and strength, large datasets are needed to robustly train such models. This is a concern as reliable concrete strength data is rather limited, especially for realistic industrial concretes. Here, based on the analysis of a fairly large dataset (>10,000 observations) of measured compressive strengths from industrial concretes, we compare the ability of three selected machine learning algorithms (polynomial regression, artificial neural network, random forest) to reliably predict concrete strength as a function of the size of the training dataset. In addition, by adopting stratified sampling, we investigate the influence of the representativeness of the training datapoints on the learning capability of the models considered herein. Based on these results, we discuss the nature of the competition between how accurate a given model can eventually be (when trained on a large dataset) and how much data is actually required to train this model.

1. Introduction

Concrete—which is by far the most manufactured material in the world—is made of stones (coarse aggregates) and sand (fine aggregates) that are glued together by the cement paste—which forms upon the reaction between cement and water [1]. The 28-day compressive strength is one of the most widely used metrics to characterize concrete's performance in its engineering applications (e.g., holding structural loads). Indeed, although this standardized index is primarily used to evaluate the ultimate strength of concretes [2], it can also be used to infer other critical mechanical properties, e.g., stiffness or tensile strength [3]. Accurate concrete strength predictions have a profound impact on construction projects since an insufficient concrete strength can be the culprit of a catastrophic failure of civil infrastructures. Conversely, concretes exhibiting an overdesigned strength leads not only to higher material expenses [4], but also to additional environmental burdens (e.g., cement production is associated with large CO2 emissions) [5].

Over the past decades, substantial efforts have been devoted to developing predictive models for correlating a given concrete mixture proportion to its associated strength [6]. Beyond this, an ideal predictive model would also provide important insights for designing new concrete formulas with better constructability and durability, and/or at [1, 2] a lower cost [7, 8]. Conventional approaches often seek to achieve these goals using physics or chemistry-based relationships [9–11]. Although the role played by major proportioning parameters—e.g.,

Table 1. Settings of the key hyperparameters adopted in the polynomial regression (PR), artificial neural networks (ANN), and random forest (RF) models investigated in this study. Unless specified, the other hyperparameters are kept as default in Scikit-learn [35].

Model type	Complexity parameters	Other parameters
PR	Polynomial degree: 3	N/A
ANN	Hidden layer: 1 Number of neurons: 7	Optimizer: LBFGS Activation function: Sigmoid Max iteration: 500
RF	Trees: 16 Maximum depth: 10 Minimum leaf size: 2	N/A

cement dosage, aggregate fraction, and air void content—has been extensively investigated, disentangling the combined effects of these features in real concretes can be a daunting task. In addition, the influence of other secondary factors is not always negligible (e.g., nature and dosage of chemical and mineral admixtures, aggregate size gradation, etc) [12]. Due to the limited understanding of these complex property-strength correlations, it is still extremely challenging to get a robust and universal concrete strength model using conventional approaches [13].

As an alternative pathway, the recent development of machine learning (ML) techniques provides a novel data-driven approach to revisit the strength prediction problem. Importantly, ML-based predictions have been shown to significantly outperform those of conventional approaches, especially for non-linear problems [14]. As such, recent studies have established ML as a promising approach to predict concrete strength [15–18]. However, ML approaches often require a large dataset to 'learn' the relationship between inputs and outputs [19, 20]. This is a major concern for concrete applications, as strength data for industrial concretes are often difficult to access (i.e., data is not publicly available). In addition, reported concrete strength data are often incomplete (some important features that can impact concrete strength are often missing, e.g., mixing method, curing temperature, types of aggregates, etc). More generally, ML approaches require accurate and self-consistent data—which is often questionable for concrete strength data due to non-standardized measurements or inconsistencies in data recording [21]. For example, the strength of a given concrete can significantly vary when the testing protocol or specimen size is altered [22–24]. Although such difficulties can be filtered out with sufficiently large datasets, their significance tends to be exacerbated when the training datasets are small—which is often the case in applications involving engineering materials. For all these reasons, it is critical to assess how the reliability of ML approaches for concrete strength prediction applications depends on the number of training data points.

With the special goal for predicting concrete strength in mind, we carry out this study around three core questions: (i) how much data is sufficient for training a ML model, (ii) which ML algorithms are better suited to deal with small datasets, and (iii) what the role of data representativeness on the learning capability of ML models is. By building on our previous studies [18, 25], we explore these questions by taking the example of three archetypal ML algorithms, i.e., polynomial regression (PR), artificial neural network (ANN), and random forest (RF). We compare the ultimate accuracy of these algorithms and their learning efficiency as a function of data volume. These results are insightful for facilitating the adoption of ML for small datasets—as relevant to concrete engineering.

2. Background and methods

2.1. Dataset of concrete strength measurements

The dataset used in this study comprises the 28-day compressive strength of 10,264 commercial concretes and their associated mixture proportions [18]. All the mixtures were cast using ASTM C150 compliant Type I/II cement [26] and Class F fly ash compliant with ASTM C618 [27]—wherein fly ash is a by-product of coal power plants that can be used as supplementary cementitious material, that is, to replace cement in concrete [28]. The seven most influential features are considered in this study, namely, (1) water-to-cementitious ratio (in this case, the ratio between the mass of water and that of cement and fly ash), (2) cement fraction, (3) fly ash fraction, (4) fine aggregate fraction, (5) air-entraining admixture dosage (used for enhancing concrete durability), and (6) water-reducing admixture dosage (used for increasing concrete early-stage workability). For normalization

purposes, the features (2-to-4) are taken as the solid weight fractions. The fraction of coarse aggregates is excluded as it is redundant with features (2-to-4) since the four weight fractions always sum up to 100%.

2.2. Machine learning algorithms and hyperparameter optimization

We assess the performance of three archetypical ML algorithms (PR, ANN, and RF) as a function of the number of training data points. These three models are chosen as they belong to three distinct families of ML models, i.e., polynomial, network-based, and tree-based [29, 30]. All the hyperparameters of the ML models considered herein were optimized in a previous study so as to achieve an optimal balance between under- and overfitting (see [17]). The key hyperparameters of the three ML models are summarized in table 1 and also detailed in the following. First, we consider PR, which is essentially based on linear regression, wherein the model parameters designate an n-degree polynomial function [31]. Based on our previous work [17], the PR model adopted herein features a maximum polynomial degree of 3. Second, we explore the potential of ANN, which is a computational structure consisting of an input layer, an output layer, and one or several hidden layers bridging the two formers —wherein each layer comprises a collection of artificial neurons (i.e., computational units) [32]. Based on our previous work [17], the present ANN model exhibits 7 neurons in a single hidden layer. We adopt the sigmoid function as activation function and we use the backpropagation algorithm to optimize the model parameters [33]. Third, we consider RF, which is an enhanced bagging method. By using the majority-voting concept, this approach is typically more accurate than conventional single decision trees [34]. Here, based on our previous work [17], our RF model comprises 16 trees with a maximum depth of 10. Despite the different nature of these algorithms, their common goal is to predict a variable y (here, the 28-day strength) as a function of the input variables x (here, mixing proportions of concrete), while minimizing the difference between measured and predicted strength values (see [17] for details).

2.3. Model training, validation, and testing

Following common practices, 70% of the strength observations are randomly selected and used for model training (i.e., 'training set'). The remaining 30% of the data are kept hidden to the model, so as to assess the ability of the model to predict the strength of unknown concretes (i.e., 'test set'). A fraction of the initial training set then used as validation set to optimize the hyperparameters of the models. To this end, we adopt a five-fold cross-validation approach [36]. In detail, the initial training set is randomly split into five folds (each made of 20% of the training data). In each of the five rounds of analysis, the model is iteratively trained based on four folds and validated based on the remaining fold (i.e., 'cross-validation set').

2.4. Accuracy evaluation

We evaluate the accuracy of each model based on their mean-square error (MSE) and coefficient of determination (R^2), wherein the MSE is the averaged Euclidian distance between predicted and measured strength data. The root mean-square error (RMSE) is calculated as the square root of the MSE. The R^2 factor further quantifies the accuracy of the model predictions in terms of the degree of scattering around the fitted input-output relationship (wherein a perfect prediction is associated with $R^2 = 1$). We analyze the deviation between strength predictions and measurements by computing the error distribution—that is, the distribution of the differences between predicted and measured strength values for each concrete mixture in the test set. The error distribution yielded by each model then serves to calculate the 90 and 95% confidence intervals of a predicted strength falling into these ranges (see [17] for details).

2.5. Evaluation of the learning efficiency

To investigate how each model 'learns' how to predict concrete strength as it is exposed to increasing numbers of training examples, we compute their 'learning curve' [37]. This approach consists in plotting the training and validation accuracy of the model as it is exposed to an increasing number of training examples. Results are averaged over five train-validation splits by using the five-fold cross-validation method. In detail, we first split the whole dataset into five folds, wherein each fold iteratively serves as validation set while the remaining four folds are used as training set. Rather than exposing the model to the whole training set, a series of smaller subsets of the training set with increasing sizes are used to train the model iteratively (with 10% increments in each iteration). In turn, the validation set remains constant (with a fixed size) during the model training. At each training iteration, the MSE of the model on the (partial) training and (full) validation sets are recorded. Finally, the average and standard deviation of the MSE as a function of the size of the training set are determined based on the five folds. To ensure a consistent comparison, all the models considered herein are trained and evaluated based on the same training and validation sets.

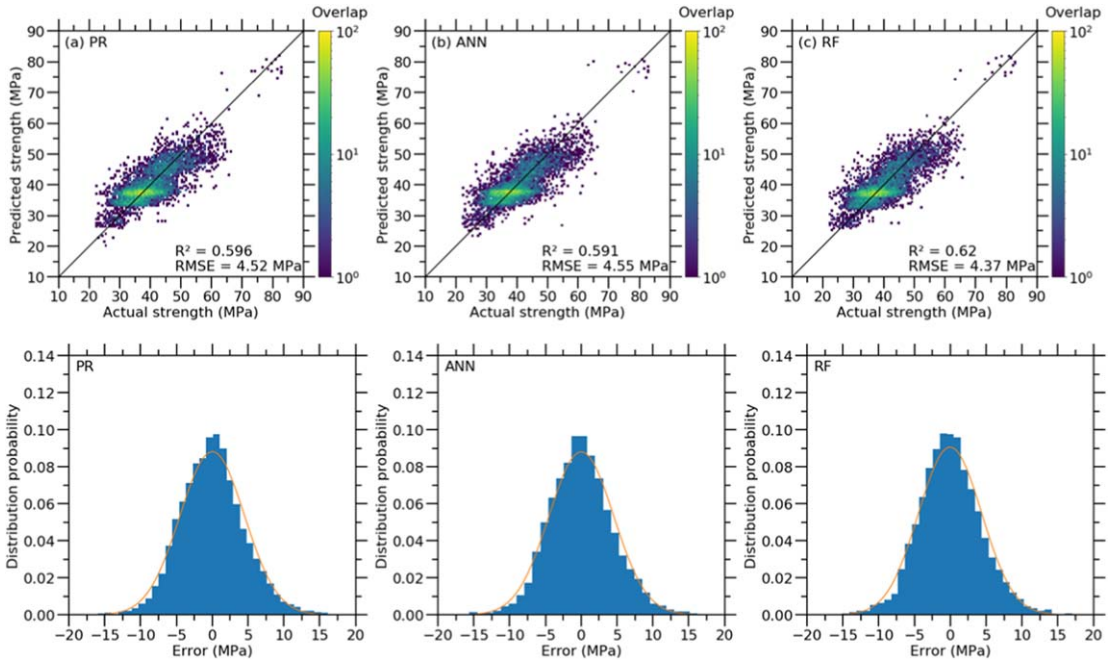

Figure 1. (top) Comparison between the predicted versus measured (ground-truth) 28-day strength and (bottom) error distribution for the test set offered by the (a) polynomial regression (PR), (b) artificial neural network (ANN), and (c) random forest (RF) models. The pixel colors in the top plots indicate the number of overlapped points. The error distributions are fitted by Gaussian distributions.

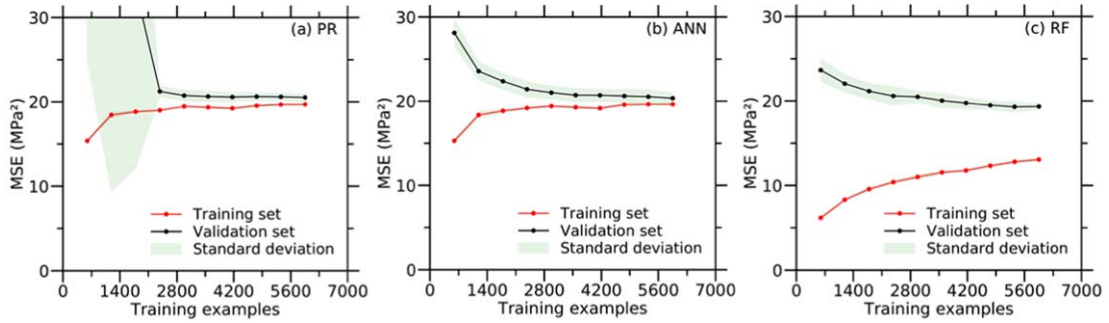

Figure 2. Learning curves showing the MSE of the training and cross-validation sets as a function of the size of the training set for the (a) polynomial regression (PR), (b) artificial neural network (ANN), and (c) random forest (RF) models. The green shadowed area indicates the standard deviation of the validation set MSE.

2.6. Stratification of the dataset

We then assess the role played by the representativeness of the datapoints in small training sets. Indeed, a common issue of using small datasets for ML applications is that the extracted training and validation sets do not offer a proper representation of the data distribution in the entire dataset. This issue is often encountered when the datapoints in the dataset vary within a large range or follow a certain statistical pattern—such that randomly extracted subsets (i.e., the training and validation sets) may not alone be representative of the entire feature space that is covered in the dataset due to the limited density of datapoints. Note that this problem is typically more exacerbated for validation sets (or test sets) since they are typically smaller than the associated training sets. Such lack of representativeness may yield biased validation sets, which do not fairly assess the ability of models to generalize.

In that regard, stratified sampling is a data preprocessing technique that aims to improve the representativeness of training and validation sets [38, 39]. In detail, data stratification divides the whole dataset into homogeneously distributed subgroups based on the statistical distribution of the datapoints. As such, when compared to random training-validation splits, data stratification is a more efficient data sampling technique as it ensures that both the training and validation sets homogeneously sample the entire dataset. This technique can be combined with *k*-fold cross-validation (i.e., 'stratified cross-validation'), wherein each of the *k* folds is constructed so as to present the same data distribution as that of the entire dataset.

Table 2. Coefficient of determination (R^2) and confidence interval values offered by each model for the test set (when trained based on the entire training set), and minimum number of training data that is needed for each model to achieve an average validation set MSE that is less than one standard deviation away from its final validation set MSE.

Accuracy analysis				
		Confidence interval (MPa)		Learning analysis
Model type	R^2	90%	95%	Minimum number of training data
PR	0.596	±7.43	±8.86	2680
ANN	0.591	±7.45	±8.88	3010
RF	0.620	±7.22	±8.60	4070

Here, we explore the effect of stratified cross-validation on the learning capability of the ML models considered herein. To this end, we first split the raw concrete dataset into ten even classes (with an equal number of datapoints), wherein the datapoints are ranked based on the 28-day strength value (i.e., the output). We then iteratively randomly extract 20% of the samples from each class to construct each of the five folds—so that the folds comprise the same number of datapoints than in the case of random training-validation splits. To investigate the influence of stratification on the learning efficiency of the different ML models considered herein, we subsequently conduct the learning curve analysis detailed in section 2.5, wherein the training and validation MSE values are computed after progressively training the model with increasing numbers of (stratified) samples.

3. Results

3.1. Accuracy of the machine learning models

We first compare the final accuracy offered by each ML model when trained based on the entire training set. Figure 1 shows, for each model, the predicted versus measured strengths for the entire test set and the associated error distributions. The accuracy analysis is summarized in table 2. In detail, we find that RF features the highest degree of accuracy, which manifests itself by a minimum RMSE, maximum R^2, and minimum confidence intervals.

3.2. Gradual learning upon increasing training set size

Having shown that RF offers the highest final accuracy when trained based on the entire training set, we now focus on the learning curve exhibited by each model—to assess their ability to quickly learn the input-output relationship as they become exposed to a gradually increasing number of training examples (see figure 2). As expected, all the models exhibit a fairly similar trend, that is, (i) the MSE of the training set increases with increasing training set size since it becomes increasingly difficult for the model to perfectly interpolate the training set with a fixed number of model parameters and (ii) the MSE of the cross-validation set decreases with increasing training set size as the model gradually manages to learn the input-output relationship and, hence, eventually shows an increased ability to generalize, that is, to predict the strength of unknown concretes.

Nevertheless, we find that, although the final accuracy offered by the models shows only minor differences (see table 2), their learning curves exhibit significantly distinct features. In detail, in agreement with the data presented in table 2, we find that RF eventually features the lowest MSE for the validation set, as well as for the training set. As compared with PR and ANN—wherein both the training and validation set accuracy converge evidently as a function of the training set size—the RF model does not present a clear plateau and finally shows a gap between its training and validation accuracy values (when trained on the entire training set). This indicates that the RF model features a greater potential to eventually achieve even higher accuracy if additional data points can be added to the training set. Interestingly, we also note that the MSE of the validation set exhibits a faster decrease in the case of PR and ANN, as compared to the more gradual decrease obtained in the case of RF. In addition, we further note that the models exhibit significantly different initial degrees of stability when trained based on extremely small numbers of samples. In particular, the PR model considered herein present a level of fluctuation that is notably larger than that shown by the other more complex models when the number of training samples is small (i.e., $<$2600 samples). This is likely related to the fact that, unlike the other models (ANN and RF), the PR model is not regularized. As such, PR is more sensitive to the noise that is presented in the training dataset, especially when its size is not statistically representative. This point is further discussed in the next section.

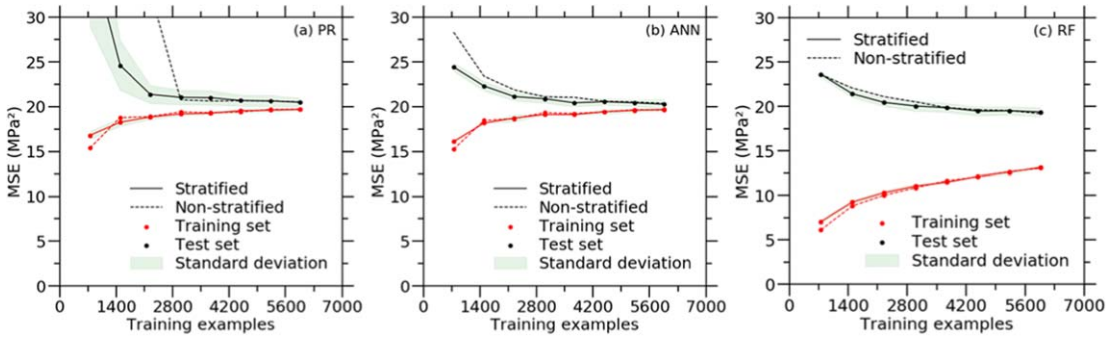

Figure 3. Learning curves showing the MSE of the training and cross-validation sets as a function of the size of the training set for the (a) polynomial regression (PR), (b) artificial neural network (ANN), and (c) random forest (RF) models in the case of stratified cross-validation. For reference, the data are compared with the non-stratified learning curves from figure 2, which are represented as dashed lines. The green shadowed area indicates the standard deviation of the validation set MSE.

Overall, since various factors can affect the minimum number of data points that is needed to reliably train a given model, we adopt herein the following criterion: we define the minimum training set size as the lowest number of training data points that is needed for the model to achieve an average validation set MSE that is less than one standard deviation away from its final validation set MSE (i.e., when trained based on the entire training set), wherein the standard deviation is calculated based on the MSE obtained for each validation fold during cross-validation. In case of the three models investigated in this study, we find that PR (and, to a lesser extent, ANN) features an increased ability to quickly learn how to predict concrete strength from small datasets as compared to RF (see table 2).

3.3. Gradual learning assisted by stratified sampling

Next, we assess the role of the representativeness of the training samples in controlling the learning capability of each model. To this end, we implement stratified cross-validation (see section 2.6 for details) [40]. By following the general analysis discussed in section 3.2, we repeat the model training by gradually exposing more stratified datapoints to the three models investigated herein (see figure 3). We first observe that, for the training set, the learning curves obtained with stratified cross-validation are fairly similar to the non-stratified ones shown in figure 2 (which, for reference, appear as dashed lines in figure 3). Similarly, we note that the final (when trained based on the entire training set) validation set MSE achieved by the three models is unaffected by stratification, which indicates that the level of representativeness of the training samples does not notably affect the final accuracy offered by the ML models.

However, we find that stratification has a significant effect on learning when the training set is small. Indeed, at low number of training examples, the validation set MSE values achieved by stratified cross-validation are systematically smaller than the non-stratified ones. In addition, we observe that stratification systematically results in a decrease in the standard deviation of the validation set MSE (see figures 2 and 3). This indicates that data representativity plays a key role in governing the ability of the ML models to learn the input-output relationship when they are exposed to small training sets. In the case of the present concrete strength dataset, data representativity is important as the distribution of the output strength values is highly non-uniform. In particular, very few datapoints are available in the low- and high-strength domains (i.e., below 30 MPa and above 60 MPa, respectively, wherein the degree of overlapping of data points is far below 10 in figure 1), while the majority of the datapoints present intermediate strength values (i.e., between 30 and 60 MPa). Due to the sparsity of the present dataset, conventional cross-validation exhibits a high propensity to result in the formation of folds that are not representative of the entire dataset—for instance, such folds might be ignoring some poorly-populated regions of the feature space. In turn, models that are trained/validation based on poorly representative folds tend to require additional training samples to achieve the same accuracy as those that are trained/validation based on representative folds.

We find that the effect of stratification is especially pronounced in the case of PR and ANN, whereas stratification has little effect, if any, on the learning of the RF model. This indicates that, due to its non-analytical form (i.e., based on decision trees), RF models are less affected by the distribution of the points in the dataset [41]. In contrast, the PR model features the largest improvement in its accuracy when trained by stratified cross-validation. This is a consequence of the fact that the datapoints that are located at the boundaries of the populated region of the feature space (which are likely to be associated with very small or very large strength values) have a strong influence on the value of the PR model parameters, especially those associated with high degree terms—since the high degree components of PR models tend to quickly diverge toward $\pm\infty$ at the limits

of the populated region of the feature space. As such, whether or not such boundary datapoints are included in the training set has a strong influence on the ability of the model to properly generalize toward the edges of the feature space. Overall, these results demonstrate that stratified cross-validation has the potential to significantly accelerate the learning of ML model, that is, to reduce the number of datapoints that is needed for the models to achieve their final accuracy.

4. Discussion

Overall, we find that the model offering the highest final degree of accuracy (i.e., RF) requires the largest training set to be trained, whereas, in turn, the models presenting the lowest final accuracy (i.e., PR and ANN) require the smallest training set. These results highlight the existence of competition between (i) the ultimate ability of a model to accurately learn the input-output relationship when trained based on an abundance of training examples and (ii) the ability of a model to quickly learn this relationship when trained based on a small dataset. This competition can be rationalized in terms of the intrinsic 'flexibility' of the model.

On the one hand, PR and ANN are constrained, poorly-flexible models—since PR relies on a fixed analytical form, while the present ANN model exhibits a limited ability to capture complex input-output relationships as it comprises a single hidden layer. This lack of flexibility limits the final accuracy that is achievable by these models. Although the degree of complexity of these models (i.e., maximum polynomial degree for PR and number of hidden neurons for ANN) is already tuned to achieve the best balance between under- and overfitting [17], the fact that the MSE of the training and validation sets both plateau toward the same value suggests that these models are too simple and lack some degrees of freedom. For a given amount of data, this limitation could potentially be mitigated by carefully increasing the complexity of these models (while avoiding overfitting)—for instance, by increasing the number of hidden layers in ANN [42]. In turn, the constrained nature of the PR and single-layer ANN models allows them to quickly achieve their maximum accuracy—since only a limited number of parameters (i.e., polynomial coefficients for PR and neuron-neuron connection weights for ANN) need to be parameterized [43]. This makes it possible for these algorithms to handle small, sparse datasets. However, it is clear from figure 2 that these models have already achieved their maximum accuracy and, hence, would not benefit from being trained with any additional data. In addition, it can be noted from figure 3 that the level of data representativity in the training set has a large impact on those models' performance. Clearly, stratified sampling is able to significantly reduce the critical (minimum) number of datapoints that is needed for the models to achieve their final accuracy.

On the other hand, RF is, in contrast, more flexible as it is not constrained by any analytical formulation. Indeed, in contrast to PR (which intrinsically yields a smooth, continuous, and differentiable relationship between inputs and output due to its analytical form), the tree-based structure makes it possible for the RF model to capture rough, less-continuous, and less-differentiable functions [44]. This flexibility enables RF to eventually reach a higher final degree accuracy once trained based on the entire training set. In turn, such complexity comes at a cost, namely, a large number of training data points is needed to properly parameterize the RF model. This is well illustrated by the facts that, unlike the cases of PR and ANN, (i) the validation set MSE of the RF model does not reach a plateau and continues to decrease upon increasing training set size and (ii) the final validation set MSE is significantly higher than the final training set MSE. Both of these learning curve features suggest that the RF model has not yet finished its training and, hence, would further be improved if exposed to an increased number of data.

5. Conclusions

Overall, these results establish machine learning as a promising approach to predict the strength of commercial concrete based on the sole knowledge of their mixture proportions. Although all the models considered herein present only minor differences in their final accuracy (i.e., when trained based on the entire dataset), they exhibit notable differences in their abilities to quickly learn the input-output mapping when exposed to small, sparse datasets—especially when trained with stratified cross-validation. We find that simple, more constrained models (e.g., polynomial regression) offer limited final accuracy, but can quickly achieve their maximum accuracy while trained based on small training sets. In contrast, less constrained, more flexible models (e.g., random forest) require larger training sets, but can eventually feature a higher final prediction accuracy. This highlights the importance of properly comparing the performance offered by various machine learning models when considering small datasets—which are common in engineering applications wherein each additional datapoint comes with a significant time and cost burden.

Acknowledgments

The authors acknowledge some financial support for this research provided by the US Department of Transportation through the Federal Highway Administration (Grant #: 693JJ31950021) and the US National Science Foundation (DMREF: 1922167).

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

ORCID iDs

Yu Song https://orcid.org/0000-0001-6218-3234
Gaurav Sant https://orcid.org/0000-0002-1124-5498
Mathieu Bauchy https://orcid.org/0000-0003-4600-0631

References

[1] Taylor W H 1967 *Concrete Technology and Practice* 4/E (United Kingdom: McGraw Hill Education)
[2] Rodríguez de Sensale G 2006 Strength development of concrete with rice-husk ash *Cem. Concr. Compos.* **28** 158–60
[3] Ashour S A and Wafa F F 1993 Flexural behavior of high-strength fiber reinforced concrete beams *Struct. J.* **90** 279–87
[4] Purnell P and Black L 2012 Embodied carbon dioxide in concrete: variation with common mix design parameters *Cem. Concr. Res.* **42** 874–7
[5] Vance K, Falzone G, Pignatelli I, Bauchy M, Balonis M and Sant G 2015 Direct carbonation of Ca(OH)2 using liquid and supercritical CO_2: implications for carbon-neutral cementation *Ind. Eng. Chem. Res.* **54** 8908–18
[6] Moutassem F and Chidiac S E 2016 Assessment of concrete compressive strength prediction models *KSCE J. Civ. Eng.* **20** 343–58
[7] Biernacki J J *et al* 2018 Cements in the 21st century: challenges, perspectives, and opportunities *J. Am. Ceram. Soc.* **100** 2746–73
[8] Provis J L 2015 Grand challenges in structural materials *Front. Mater.* **2**
[9] Powers T C 1960 *Physical Properties of Cement Paste Proc. First International Symposium on the Chemistry of Cement* 2 (Washington, DC, 1960, US Department of Commerce, National Bureau of Standards, Monograph 4) pp 577–613
[10] Popovics S 1998 History of a mathematical model for strength development of portland cement concrete *Mater. J.* **95** 593–600
[11] Zain M F M and Abd S M 2009 Multiple regression model for compressive strength prediction of high performance concrete *J. Appl. Sci.* **9** 155–60
[12] Wild S, Sabir B B and Khatib J M 1995 Factors influencing strength development of concrete containing silica fume *Cem. Concr. Res.* **25** 1567–80
[13] Burris L E, Alapati P, Moser R D, Ley M T, Berke N and Kurtis K E 2015 Alternative cementitious materials: challenges and opportunities *Int. Workshop on Durability and Sustainability of Concrete Structures* (Bologna, Italy)
[14] Yeh I-C 1998 Modeling of strength of high-performance concrete using artificial neural networks *Cem. Concr. Res.* **28** 1797–808
[15] Rafiei M H, Khushefati W H, Demirboga R and Adeli H 2016 Neural network, machine learning, and evolutionary approaches for concrete material characterization *ACI Mater. J.* **113** 781–9
[16] DeRousseau M A, Kasprzyk J R and Srubar W V III 2018 Computational design optimization of concrete mixtures: a review *Cem. Concr. Res.* **109** 42–53
[17] Ouyang B, Li Y, Wu F, Yu H, Wang Y, Sant G and Bauchy M 2020 Predicting Concrete's Strength by Machine Learning: Balance between Accuracy and Complexity of Algorithms *ACI Mater. J.* **117** 125–33
[18] Young B A, Hall A, Pilon L, Gupta P and Sant G 2019 Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: new insights from statistical analysis and machine learning methods *Cem. Concr. Res.* **115** 379–88
[19] Liu H, Zhang T, Krishnan N M A, Smedskjaer M M, Ryan J V, Gin S and Bauchy M 2019 Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning *Npj Mater. Degrad.* **3** 1–12
[20] Bishnoi S, Singh S, Ravinder R, Bauchy M, Gosvami N N, Kodamana H and Krishnan N M A 2019 Predicting Young's modulus of oxide glasses with sparse datasets using machine learning *J. Non-Cryst. Solids* **524** 119643
[21] Pani A K, Amin K G and Mohanta H K 2012 Data driven soft sensor of a cement mill using generalized regression neural network *2012 Int. Conf. on Data Science Engineering (ICDSE) 2012 Int. Conf. on Data Science Engineering (ICDSE)* pp 98–102
[22] Rejeb S K 1996 Improving compressive strength of concrete by a two-step mixing method *Cem. Concr. Res.* **26** 585–92
[23] Hemalatha T, Sundar K R, Murthy A R and Iyer N R 2015 Influence of mixing protocol on fresh and hardened properties of self-compacting concrete *Constr. Build. Mater.* **98** 119–27
[24] Elhakam A A, Mohamed A E and Awad E 2012 Influence of self-healing, mixing method and adding silica fume on mechanical properties of recycled aggregates concrete *Constr. Build. Mater.* **35** 421–7
[25] Oey T, Jones S, Bullard J W and Sant G 2020 Machine learning can predict setting behavior and strength evolution of hydrating cement systems *J. Am. Ceram. Soc.* **103** 480–90
[26] ASTM C150 / C150M-20 2020 *Standard Specification for Portland Cement* (ASTM International: United States of America) (www.astm.org) (https://doi.org/10.1520/C0150_C0150M-20)
[27] ASTM C618-19 2019 *Standard Specification for Coal Fly Ash and Raw or Calcined Natural Pozzolan for Use in Concrete* (West Conshohocken, PA: ASTM International) (www.astm.org) (https://doi.org/10.1520/C0618-19)
[28] Conn R E, Sellakumar K and Bland A E 1999 *Utilization of CFB Fly Ash for Construction Applications* (Livingston, NJ: Foster Wheeler Development Corp.)

[29] Anoop Krishnan N M, Mangalathu S, Smedskjaer M M, Tandia A, Burton H and Bauchy M 2018 Predicting the dissolution kinetics of silicate glasses using machine learning *J. Non-Cryst. Solids* **487** 37–45
[30] Liu H, Fu Z, Yang K, Xu X and Bauchy M 2019 Machine learning for glass science and engineering: a review *J. Non-Cryst. Solids* X **4** 100036
[31] Sinha P 2013 Multivariate polynomial regression in data mining: methodology *International Journal of Scientific & Engineering Research* **4** 962
[32] Wasserman P D 1993 *Advanced Methods in Neural Computing* (New York, NY, United States of America: Wiley)
[33] Li J, Cheng J, Shi J and Huang F 2012 Brief introduction of back propagation (BP) neural network algorithm and its improvement *Advances in Computer Science and Information Engineering Advances in Intelligent and Soft Computing* ed D Jin and S Lin (Berlin, Heidelberg: Springer) pp 553–8
[34] Liaw A and Wiener M 2002 *Classification and Regression by random Forest* **2** 18–22
[35] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R and Dubourg V 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
[36] Stone M 1974 Cross-validatory choice and assessment of statistical predictions *J. R. Stat. Soc. Ser. B Methodol.* **36** 111–33
[37] Anzanello M J and Fogliatto F S 2011 Learning curve models and applications: literature review and research directions *Int. J. Ind. Ergon.* **41** 573–83
[38] Breiman L, Friedman J, Olshen R and Stone C 1984 *Classification and Regression Trees* 37 (Belmont, CA: Wadsworth Int. Group) pp 237–51
[39] Jablonka K M, Ongari D, Moosavi S M and Smit B 2020 Big-data science in porous materials: materials genomics and machine learning *Chem. Rev.* **120** 8066–8129
[40] Zeng X and Martinez T R 2000 Distribution-balanced stratified cross-validation for accuracy estimation *J. Exp. Theor. Artif. Intell.* **12** 1–12
[41] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[42] Ravinder R, Sridhara K H, Bishnoi S, Grover H S, Bauchy M, Kodamana H and Krishnan N A 2020 Deep learning aided rational design of oxide glasses *Mater. Horiz.* 1819–27
[43] Liu H, Fu Z, Li Y, Sabri N F A and Bauchy M 2019 Balance between accuracy and simplicity in empirical forcefields for glass modeling: insights from machine learning *J. Non-Cryst. Solids* **515** 133–42
[44] Yang K, Xu X, Yang B, Cook B, Ramos H, Krishnan N M A, Smedskjaer M M, Hoover C and Bauchy M 2019 Predicting the Young's modulus of silicate glasses using high-throughput molecular dynamics simulations and machine learning *Sci. Rep.* **9** 8739