The effectiveness of data augmentation in compressive strength prediction of calcined clay cements using linear regression learning

Yassine El Khessaimi, Youssef El Hafiane, Agnès Smith, Karim Tamine, Samir Adly, Moulay Barkatou

▶ **To cite this version:**

Yassine El Khessaimi, Youssef El Hafiane, Agnès Smith, Karim Tamine, Samir Adly, et al.. The effectiveness of data augmentation in compressive strength prediction of calcined clay cements using linear regression learning. 2023. hal-04166948

HAL Id: hal-04166948
https://hal.science/hal-04166948

Preprint submitted on 20 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

[CMSS23]

The effectiveness of data augmentation in compressive strength prediction of calcined clay cements using linear regression learning

Yassine El Khessaimi [a*], Youssef El Hafiane [a], Agnès Smith [a], Karim Tamine [b], Samir Adly [b], Moulay Barkatou [b]

[a] *Institute of Research in Ceramics, IRCER, UMR CNRS 7315, University of Limoges, 12 rue Atlantis, 87068 Limoges, France.*
[b] *MathIS, XLIM Laboratory, UMR CNRS 7252, University of Limoges, 123 Av. Albert Thomas, 87000 Limoges, France.*

Abstract

Cement production is a major contributor to global CO_2 emissions. To minimize its environmental impact while maintaining the required mechanical properties of cement, there is a pressing need for sustainable production processes. This paper explores the use of data augmentation techniques, specifically the copulas method, to improve the performance of linear regression models for linking the compressive strength of LC3 with its mix design. While data augmentation using copulas can be useful in augmenting tabular data, its effectiveness in improving linear regression performance may depend on the statistical characteristics of the original data. The method successfully generated additional data that preserved the original statistical properties, but it did not always lead to significant improvements in linear regression performance. The research highlights the potential of data-driven models for optimizing cement materials properties and emphasizes the importance of considering the statistical characteristics of the original data when applying data augmentation techniques.

Keywords: LC3 cement; compressive strength; artificial intelligence; data augmentation;

1. Introduction

Data-driven models based on artificial intelligence (AI) applied in the optimization of cement materials properties is an emerging research topic. Considering the number of current published papers about this subject, it confirms that the cement community is interested in these novel approaches. Data-driven approach for cement materials constitute a new paradigm to link the performance properties to their composition and process parameters. One example of cement for with there is an urgent need of data-driven approach is limestone calcined clay cement (LC3) (Scrivener et al., 2018). This cement is today considered as the next generation of building binders. When it is compared to the classical Portland cement, it shows a reduced carbon footprint of 25 to 35%, with equivalent or higher compressive strength (Berriel et al., 2016). The research studies on LC3 performance require some acceleration to reach rapidly

* corresponding author: yassine.el-khessaimi@unilim.fr

carbon neutrality of the cement production.

To link cement performance with its composition and process parameters, an empirical approach is applied (Canbek et al., 2022a; Riera et al., 2016; Saleem et al., 2021; Van Bunderen et al., 2021). This approach involves “idealized” models, and does not reflect the “real life” case. Whereas, data-driven approach does not enforce particular assumptions and can excel at treating complex and nonlinear links. Except, they require a large dataset for training and testing. Canbek et al. (Canbek et al., 2022b) linked the rheology of LC3 cements to the composition through support vector machine model showing high accuracy with R^2 =0.96, about 108 cement pastes were carried out to feed the model. Hafez et al. (Hafez et al., 2022) created a ML regression model to predict the performance of blended concretes including LC3. A database of 1650 data points was created to train and test the model. Even so, only few datapoints were relevant to mixes with LC3 cement. It is clear that there is a scarcity of research on the use of data-driven models to study LC3 cements. One reason to explain this lack of studies is the dataset availability. The challenge lies in organizing and standardizing large amounts of data. Additionally, the time-consuming and expensive process of characterizing a significant amount of samples poses a limitation for implementing ML algorithms (Zhang and Ling, 2018). There are several ways to do data augmentation of tabular values such as adding random noise to feature values, flipping the values of binary features, sampling random subsets of the data, standardizing the values by subtracting the mean and dividing by the standard deviation, transforming the values using a scaling function, creating new features by combining existing features or using domain knowledge, applying small transformations to the original data, and using Copulas technique, which involves modelling the dependencies between features and generating new samples that preserve these dependencies (Meyer et al., 2021; Peters et al., 2014; Silva et al., 2020). Tabular data augmentation is a technique that can be used to improve the performance of linear regression models (Cao et al., 2021). The basic idea is to generate new, synthetic data samples from the existing dataset by applying various transformations to the original data. This can help to increase the size of the dataset and add diversity to the data, which can help to improve the generalization performance of the linear regression model.

The aim of this paper is to assess the effect of incorporating the Copulas method for enhancing tabular data on the precision of linear regression models used to link the compressive strength of LC3 with its mix design. We chose to use linear regression despite it being less powerful than other algorithms such as support vector machine or artificial neural network, in order to demonstrate the improvement that can be achieved with the application of tabular data augmentation technique. The research idea of the present work is illustrated in Figure 1.

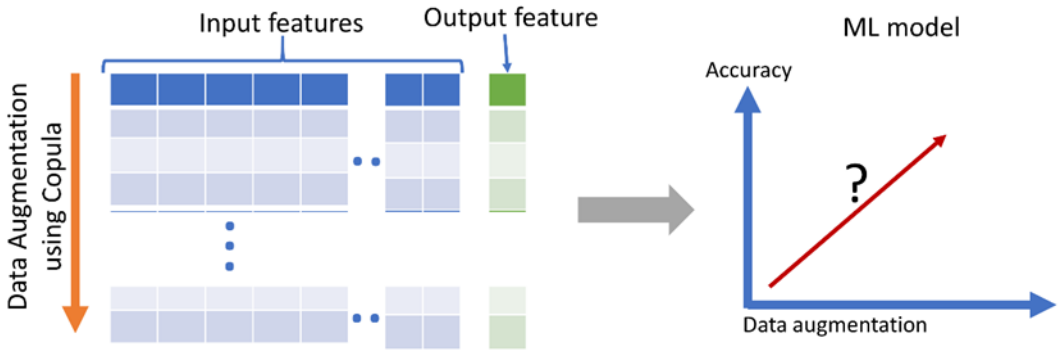

Figure 1: The research idea of the present work.

2. Materials and method

The followed work approach of the present study starts by the construction of a dataset from literature, then data is structured and missing values were handled in the preprocessing step. After this, data augmentation was carried out by applying the Copulas method. Thereafter, a linear regression ML model was applied to evaluate the efficiency of data augmentation and dimension reduction approaches.

2.1. Data collection and preprocessing

The size and the quality of the dataset are significant for the accuracy of the ML model (Zhang and Ling, 2018). An experimental database of 323 mix design (10692 data values), containing partial replacement of Portland cement with calcined clay and limestone, was compiled from previous studies that were reported in literature (Akindahunsi et al., 2020; Alujas et al., 2015; Antoni et al., 2012; Avet et al., 2016; Dhandapani and Santhanam, 2017; Dixit et al., 2021; Fernandez et al., 2011; Krishnan et al., 2018; Lin et al., 2021; Lorentz et al., 2020; Machner et al., 2017; Mishra et al., 2019; Msinjili et al., 2019). Data splitting is a usually used method for model validation, where the dataset is split into two separate parts: the first for training, and the second for testing (Larsen and Goutte, 1999). The data was randomly partitioned into training and testing sets: 80% of the data was used for training and the remaining 20% was used for testing. Table 1 shows a description and statistical parameters of the data features.

Table 1: Description and statistical parameters of the original data features.

Data items	Features	Symbol	Units	mean	std	min	max
Calcined clay	Proportion of calcined clay	CL%	wt.%	23.5	7.9	10	40
	BET surface area	CL_Ss	m^2/g	18.5	7.3	2.5	45.7
Calcination conditions of the clay	Temperature	T_calcin	°C	760.7	84.06	600	925
	Duration	time_calcin	hours	1.27	0.75	0.2	3
Portland Cement	Proportion of OPC	OPC%	wt.%	68.9	12.13	37.6	90
Limestone	Proportion Limestone	CC%	wt.%	7.5	7.7	0	31.1
Chemical composition of the binder	Reactivity ratio	RM	-	2.3	0.5	1.6	3.8
	Silica ratio	SM	-	2	0.6	1.2	4.3
	Alumina ratio	AM	-	3.9	2.5	1.6	17.8

	Hydraulic ratio	HM	-	1.2	0.3	0.7	2.1
Hardening conditions	Water to binder ratio	W/B	-	0.5	0.09	0.1	0.9
	Hardening temperature	T_cure	°C	22.6	6.0	5	50
	Hardening relative humidity	RH_cure	%	92.3	5.3	80	100
	Hardening age	age_D	days	30.7	41.3	1	270
Compressive strength	Compressive strength	R	MPa	39.2	16.6	5	75

2.2. Linear regression model

Simpler models with fewer coefficients are preferable to complex ones. Li et al. (Li et al., 2022) emphasize the importance of avoiding the use of opaque and complex machine learning models, such as neural networks, when simpler and more interpretable models like linear regression can suffice. Model accuracy is determined by observed data, which may not accurately represent the ground truth if the data quality is insufficient. In concrete field, data quality is often impacted by cumulative random errors from experiments. Thus, it is recommended to begin with simple, interpretable models and gradually increase complexity while cautiously evaluating prediction performance. The simplest ML algorithm is the linear regression (LR). This later is a model in which the target value is expected to be a linear combination of the features noted x_1 to x_p (Pedregosa et al., 2011; Ray, 2019). In mathematical notation, $\hat{y}$ is the predicted value (equation (1)):

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \cdots + w_p x_p \qquad (1)$$

The vector $w = (w_1, \ldots, w_p)$ is the model coefficients and w_0 as intercept value of the model.

Ordinary Least Squares method aims to fit a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Mathematically it solves a problem of the form given by equation (2):

$$\min \|Xw - y\|_2^2 \qquad (2)$$

It is worth noting that the coefficient estimated for Ordinary Least Squares method relies on the independence of the features.

2.3. Copulas method for data augmentation

Before the application of Copulas data augmentation, the dataset was split according to hardening age into three parts: 3, 7 and 28 d. Because it is impractical to increase the sample size for time dependent data. As it was mentioned before, data augmentation is carried out using the gaussian copulas method (GCM) (Montanez, 2018). Intuitively, a copula is a mathematical function that allows us to describe the joint distribution of multiple random variables by analyzing the dependencies between their marginal distributions.

If the dataset can represent as a standard normal distribution, the corresponding probability density function $f(x)$ is given by:

$$f(x) = \frac{1}{2\pi} e^{-x^2/2} \quad (3)$$

The cumulative distribution function $F(x)$ which is defined as the integral of the probability density function is written as:

$$F(x) = \int f(x) \quad (4)$$

In probability theory, the probability integral transform relates to the result that data values that are modeled as being random variables from any given continuous distribution can be converted to random variables having a standard uniform distribution (Dodge et al., 2006). Suppose we have a random variable X that comes from a distribution with cumulative density function $F(X)$. Then, we can define a random variable Y which follows a uniform distribution over the interval [0,1]:

$$Y = F(X) \quad (5)$$

In mathematical terms, a copula is a distribution over the unit cube $[0,1]^d$ which is constructed from a multivariate normal distribution over $\mathbb{R}^d$ by using the probability integral transform.

3. Results and discussion

3.1. Data augmentation results using Copulas method

Table 2 presents a comparison of data statistics before and after augmentation for a 1000 augmentation rate of samples after 3 days of curing. It shows the original data statistics, including mean, standard deviation, minimum, and maximum values for each variable, and the corresponding statistics for the augmented data. The table shows that the means and standard deviations for most variables are consistent between the original and augmented data, with the exception of a few variables, such as RH_cure and HM, which have larger standard deviations in the augmented data. It's important to note that the effects of data augmentation can vary depending on the specific statistical characteristics of each variable (Perez et al., 2023). In the case of RH_cure and HM, the introduction of additional variability through data augmentation might reflect the complexity and sensitivity of these variables, leading to relatively larger standard deviations in the augmented data. The table suggests that the augmentation process did not significantly alter the statistical properties of the original data.

Table 2: Comparison of data statistics before and after augmentation for 1000 augmentation rate.

	Original data				1000 augmented data			
	Mean	Std	min	max	mean	Std	min	max
R	24.8	5.3	15	35	24.8	5.9	7.2	41.7
CL%	28.8	4.2	10	30	28.8	4.5	4.1	38.2
T_calcin	803.6	38.3	750	900	803.9	36.3	730.3	1014.4
CL_Ss	19.5	8.0	9.6	45.7	19.4	8.8	0	58.7
time_calcin	0.9	0.3	0.2	1.3	0.9	0.4	0	1.8
CC%	10.1	6.8	0	15	10.1	7.6	0	27.4
OPC%	61.1	8.4	55	85	61.1	9.3	40.0	97.1
W/B	0.5	0.0	0.45	0.5	0.5	0.0	0.4	0.5
T_cure	23.2	5.8	20	50	23.2	6.4	9.5	59.4
RH_cure	92.1	4.2	90	100	92.2	4.7	82.7	107.0
age_D	3.0	0.0	3	3	3.0	0.0	3.0	3.0
RM	2.0	0.2	1.62	2.62	2.0	0.2	1.6	2.6
SM	1.8	0.4	1.24	2.39	1.8	0.4	0.6	3.0
AM	3.5	1.8	1.78	7.86	3.4	1.8	1.8	8.0
HM	1.0	0.1	0.81	1.59	0.9	0.3	0	11.7

When using data augmentation for linear regression, it is important to keep in mind that the goal is to generate new samples that are representative of the underlying data distribution. Thus, the augmentation techniques should be chosen carefully to ensure that they do not introduce any bias or unrealistic samples that can negatively affect the model's performance (Meyer et al., 2021). The copula method for data augmentation demonstrates advantages over other techniques. While variational approaches have their merits, such as their popularity and flexibility, they can involve complex training processes and make strong assumptions (Tagasovska et al., 2019). In contrast, the training of copulas is relatively easy and robust, requiring less guesswork in terms of hyperparameters and network architecture (Tagasovska et al., 2019). The copula method provide a direct representation of statistical distributions, offering interpretability and ease of adjustment. Copulas-based models have proven effective in generating synthetic data, including for privacy protection purposes (Patki et al., 2016).

3.2. Application of linear regression and comparisons

Figure 2 shows the performance of a regression model trained on datasets with varying degrees of data augmentation using copulas. The model's performance is evaluated using the R^2 metric, both on the training and testing data. The results suggest that moderate levels of data augmentation, up to 500 augmented samples, can improve the model's performance on the testing data, with R^2 scores ranging from 0.4 to 0.47. However, further increasing the number of augmented samples does not consistently improve the model's performance and may even lead to overfitting, as indicated by decreasing R^2 scores on the testing data for some of the larger augmentation rates.

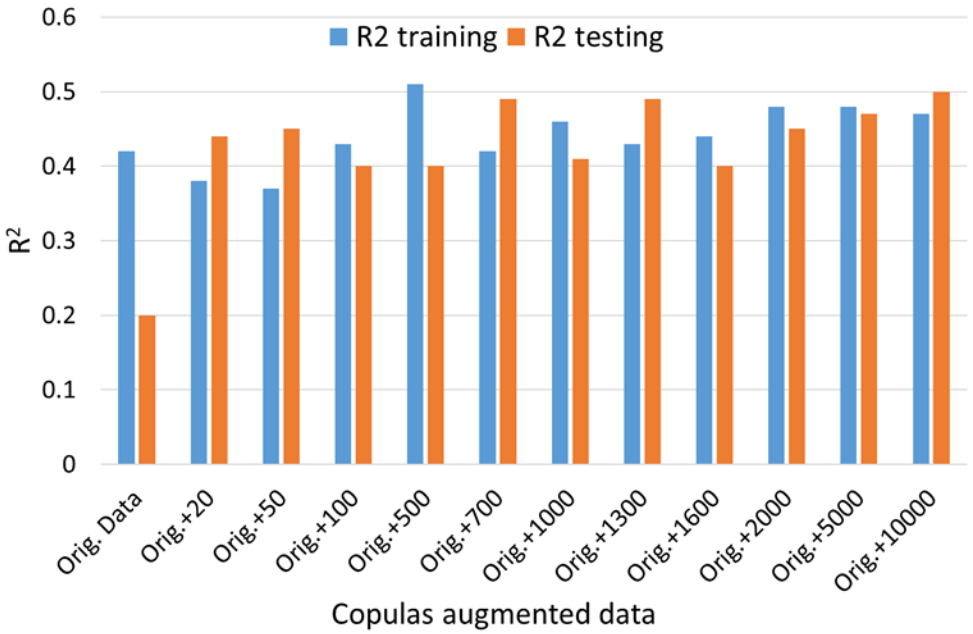

Figure 2: R^2 metric according to the Copula augmented data.

Interestingly, the optimal number of augmented samples that leads to the best model performance varies depending on the dataset, with some datasets showing improved performance with lower or higher augmentation rates. The results suggest that data augmentation using copulas can be a useful technique to augment tabular data, but it may not always lead to significant improvements in linear regression performance. The results suggest that the statistical characteristics of the original data play an important role in the effectiveness of data augmentation using copulas in improving the results of linear regression. A comparative study between Copula method and Generative Adversarial Networks confirmed this finding (Xu and Veeramachaneni, 2018), indicating that copula-based models have limitations in terms of the available distribution functions, which restricts the range of representations and consequently affects the fidelity of the generated synthetic data. Therefore, the success of data augmentation using copulas depends on understanding the statistical properties of the original data and selecting appropriate augmentation techniques accordingly.

4. Conclusion and recommendations

This paper demonstrates that the use of data augmentation techniques, particularly the Copula method, enhances the performance of linear regression models in linking the compressive strength of LC3 with its mix design.

The research findings highlight the potential of data augmentation using copulas to augment tabular data while preserving its statistical properties. However, the impact on improving linear regression performance may vary based on the statistical characteristics of the original data. This contribution adds to the growing body of knowledge in data-driven modelling for studying LC3 cements and suggests further exploration of alternative augmentation methods and their application to different cement materials.

List of Abbreviations

LC3: Limestone Calcined Clay Cement

AI: Artificial Intelligence

GCM: Gaussian Copulas Method

Conflict of Interest Statement

The authors declare that they have no conflict of interest regarding the subject matter of this publication.

Acknowledgements

The authors thank Prof. Claire Peyratout for her contribution in the completion of this paper.

Funding Source Statement

This work is supported by institutional grants from the National Research Agency under the Investments for the future program with the reference ANR 10 LABX 0074 01 Sigma LIM.

References

Akindahunsi, A.A., Avet, F., Scrivener, K., 2020. The Influence of some calcined clays from Nigeria as clinker substitute in cementitious systems. Case Stud. Constr. Mater. 13, e00443.

Alujas, A., Fernández, R., Quintana, R., Scrivener, K.L., Martirena, F., 2015. Pozzolanic reactivity of low grade kaolinitic clays: Influence of calcination temperature and impact of calcination products on OPC hydration. Appl. Clay Sci. 108, 94–101.

Antoni, M., Rossen, J., Martirena, F., Scrivener, K., 2012. Cement substitution by a combination of metakaolin and limestone. Cem. Concr. Res. 42 12 , 1579–1589.

Avet, F., Snellings, R., Diaz, A.A., Haha, M.B., Scrivener, K., 2016. Development of a new rapid, relevant and reliable (R3) test method to evaluate the pozzolanic reactivity of calcined kaolinitic clays. Cem. Concr. Res. 85, 1–11.

Berriel, S.S., Favier, A., Domínguez, E.R., Machado, I.S., Heierli, U., Scrivener, K., Hernández, F.M., Habert, G., 2016. Assessing the environmental and economic potential of Limestone Calcined Clay Cement in Cuba. J. Clean. Prod. 124, 361–369.

Canbek, O., Washburn, N.R., Kurtis, K.E., 2022a. Relating LC3 microstructure, surface resistivity and compressive strength development. Cem. Concr. Res. 160, 106920.

Canbek, O., Xu, Q., Mei, Y., Washburn, N.R., Kurtis, K.E., 2022b. Predicting the rheology of limestone calcined clay cements (LC3): Linking composition and hydration kinetics to yield stress through Machine Learning. Cem. Concr. Res. 160, 106925.

Cao, P., Bao, W., Wang, K., Yang, T., 2021. A timing prediction framework for wide voltage design with data augmentation strategy, in: Proceedings of the 26th Asia and South Pacific Design Automation Conference. pp. 291–296.

Dhandapani, Y., Santhanam, M., 2017. Assessment of pore structure evolution in the limestone calcined clay

cementitious system and its implications for performance. Cem. Concr. Compos. 84, 36–47.

Dixit, A., Du, H., Dai Pang, S., 2021. Performance of mortar incorporating calcined marine clays with varying kaolinite content. J. Clean. Prod. 282, 124513.

Dodge, Y., Cox, D., Commenges, D., 2006. The Oxford dictionary of statistical terms. Oxford University Press on Demand.

Fernandez, R., Martirena, F., Scrivener, K.L., 2011. The origin of the pozzolanic activity of calcined clay minerals: A comparison between kaolinite, illite and montmorillonite. Cem. Concr. Res. 41 1 , 113–122.

Hafez, H., Teirelbar, A., Kurda, R., Tošić, N., de la Fuente, A., 2022. Pre-bcc: A novel integrated machine learning framework for predicting mechanical and durability properties of blended cement concrete. Constr. Build. Mater. 352, 129019.

Krishnan, S., Kanaujia, S.K., Mithia, S., Bishnoi, S., 2018. Hydration kinetics and mechanisms of carbonates from stone wastes in ternary blends with calcined clay. Constr. Build. Mater. 164, 265–274.

Larsen, J., Goutte, C., 1999. On optimal data split for generalization estimation and model selection, in: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98TH8468). IEEE, pp. 225–234.

Li, Z., Yoon, J., Zhang, R., Rajabipour, F., Srubar III, W.V., Dabo, I., Radlińska, A., 2022. Machine learning in concrete science: applications, challenges, and best practices. Npj Comput. Mater. 8 1 , 1–17.

Lin, R.-S., Lee, H.-S., Han, Y., Wang, X.-Y., 2021. Experimental studies on hydration–strength–durability of limestone-cement-calcined Hwangtoh clay ternary composite. Constr. Build. Mater. 269, 121290.

Lorentz, B., Zhu, H., Stetsko, Y., Riding, K.A., Zayed, A., 2020. Feasibility Study for Calcined Clay Use in the Southeast USA, in: Calcined Clays for Sustainable Concrete. Springer, pp. 27–36.

Machner, A., Zajac, M., Haha, M.B., Kjellsen, K.O., Geiker, M.R., De Weerdt, K., 2017. Portland metakaolin cement containing dolomite or limestone–Similarities and differences in phase assemblage and compressive strength. Constr. Build. Mater. 157, 214–225.

Meyer, D., Nagler, T., Hogan, R.J., 2021. Copula-based synthetic data augmentation for machine-learning emulators. Geosci. Model Dev. 14 8 , 5205–5215.

Mishra, G., Emmanuel, A.C., Bishnoi, S., 2019. Influence of temperature on hydration and microstructure properties of limestone-calcined clay blended cement. Mater. Struct. 52 5 , 1–13.

Montanez, A., 2018. SDV: an open source library for synthetic data generation (PhD Thesis). Massachusetts Institute of Technology.

Msinjili, N.S., Gluth, G.J., Sturm, P., Vogler, N., Kühne, H.-C., 2019. Comparison of calcined illitic clays (brick clays) and low-grade kaolinitic clays as supplementary cementitious materials. Mater. Struct. 52 5 , 1–14.

Patki, N., Wedge, R., Veeramachaneni, K., 2016. The synthetic data vault, in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 399–410.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perez, J., Arroba, P., Moya, J.M., 2023. Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks. Appl. Intell. 53 2 , 1469–1486.

Peters, G.W., Dong, A.X., Kohn, R., 2014. A copula based Bayesian approach for paid–incurred claims models for non-life insurance reserving. Insur. Math. Econ. 59, 258–278.

Ray, S., 2019. A quick review of machine learning algorithms, in: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, pp. 35–39.

Riera, J.D., Miguel, L.F.F., Iturrioz, I., 2016. Evaluation of the discrete element method (DEM) and of the experimental evidence on concrete behaviour under static 3D compression. Fatigue Fract. Eng. Mater. Struct. 39 11 , 1366–1378.

Saleem, M.A., Saleem, M.M., Ahmad, Z., Hayat, S., 2021. Predicting compressive strength of concrete using impact modulus of toughness. Case Stud. Constr. Mater. 14, e00518.

Scrivener, K., Martirena, F., Bishnoi, S., Maity, S., 2018. Calcined clay limestone cements (LC3). Cem. Concr. Res. 114, 49–56.

Silva, D., Leonhardt, S., Antink, C.H., 2020. Copula-Based Data Augmentation on a Deep Learning Architecture for Cardiac Sensor Fusion. IEEE J. Biomed. Health Inform. 25 7 , 2521–2532.

Tagasovska, N., Ackerer, D., Vatter, T., 2019. Copulas as high-dimensional generative models: Vine copula autoencoders. Adv. Neural Inf. Process. Syst. 32.

Van Bunderen, C., Benboudjema, F., Snellings, R., Vandewalle, L., Cizer, Ö., 2021. Experimental analysis and modelling of mechanical properties and shrinkage of concrete recycling flash calcined dredging sediments. Cem. Concr. Compos. 115, 103787.

Xu, L., Veeramachaneni, K., 2018. Synthesizing tabular data using generative adversarial networks. ArXiv Prepr. ArXiv181111264.

Zhang, Y., Ling, C., 2018. A strategy to apply machine learning to small datasets in materials science. Npj Comput. Mater. 4 1 , 1–8.

Figure captions:

Figure 1: The research idea of the present work.

Figure 2: R^2 metric according to the Copula augmented data.

Table captions:

Table 1: Description and statistical parameters of the original data features.

Table 2: Comparison of data statistics before and after augmentation for 1000 augmentation rate.