ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

ISSN 2454-9940
www.ijasem.org
INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
Vol 19, Issue 1, 2025

Enhancing Compressive Strength Prediction of Calcined Clay Cements with Data Augmentation in Linear Regression Models

[1] Sarojini Rani, [2] P Radhika krupalini, [3] N.Jeevan Jyothi

[1] Asst Professor Dept of CSE, T.K.R Engineering College,Hyderabad

[2] Asst Professor Dept of CSE, Ideal Institute of Technology

[3] Asst Professor Dept of CSE, B.V.C Engineering College, odalarevu

Abstract

Cement production is a major contributor to global CO_2 (Carbon dioxide) emissions. To minimize its environmental impact while maintaining the required mechanical properties of cement, there is a pressing need for sustainable pro- duction processes. This paper focuses on developing sustainable cement produc- tion processes by optimizing the mechanical properties of limestone calcined clay cement (LC3) using data-driven models based on artificial intelligence. The study 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 improve- ments 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

Limestone calcined clay cement, Compressive strength, Artificial intelli- gence, Data augmentation

Introduction

Data-driven models based on artificial intelligence applied in the optimiza- tion 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 prop- erties to their composition and process parameters. One example of cement for with there is an urgent need of data-driven approach is LC3 [1]. 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 [2]. The research studies on LC3 performance require some acceleration to reach rapidly carbon neutrality of the cement production.

To link cement performance with its composition and process parameters, an empirical approach is applied [3-6]. This approach involves “idealized” mod- els and does not reflect the “real life” case. Whereas data-driven approach does not enforce particular assumptions and can excel at treating complex and non- linear links. Except, they require a large dataset for training and testing. Canbek et al. [7] linked the rheology of LC3 cements to the composition through support vector machine model showing high accuracy with R^2 = 0.96, about 108 ce- ment pastes were carried out to feed the model. Hafez et al.[8] 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-driv- en 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. Addi- tionally, the time-consuming and expensive process of charac- terizing a significant number of samples poses a limitation for implementing ML algorithms [9]. 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 devia- tion, 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 modeling the dependencies between features and generating new sam- ples that preserve these dependencies [10-12]. Tabular

ISSN 2454-9940
www.ijasem.org
INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
Vol 19, Issue 1, 2025

data augmentation is a technique that can be used to improve the performance of linear regression models [13]. The basic idea is to generate new, synthetic data samples from the existing dataset by applying various transformations to the original data.

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.

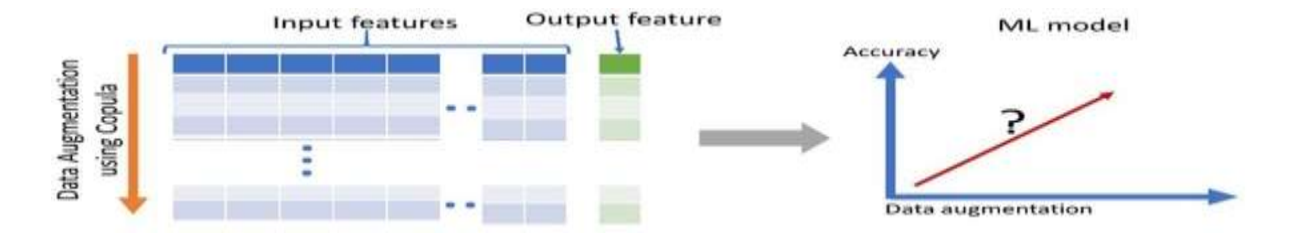

Figure 1: The research idea of the present work.

Materials and Method

The followed work approach of the present study starts by the construction of a dataset from literature, then data is struc- tured, 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. All used materials are nanoscale materials.

Data collection and preprocessing

The size and the quality of the dataset are significant for the accuracy of the ML model [9]. An experimental database of 323 mix design (10692 data values), containing partial re- placement of Portland cement with calcined clay and lime- stone, was compiled from previous studies that were reported in literature [14-26]. Data splitting is a usually used method for model validation, where the dataset is split into two sepa- rate parts: the first for training, and the second for testing [27]. 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 sta- tistical parameters of the data features.

Linear regression model

Simpler models with fewer coefficients are preferable to complex ones. Li et al. [28] emphasize the importance of avoiding the use of opaque and complex machine learn- ing 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 ac- curately represent the ground truth if the data quality is in- sufficient. In concrete field, data quality is often impacted by cumulative random errors from experiments. Thus, it is rec- ommended to begin with simple, interpretable models and gradually increase complexity while cautiously evaluating pre- diction performance. The simplest ML algorithm is the linear regression. 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 [29, 30]. In mathematical notation, $\hat{y}$ is the predicted 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 [32]. 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]:

$Y = F(X)$

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 sam- ples 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.

Conclusion and Recommendations

This paper demonstrates that the use of data augmenta- tion 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 regres- sion performance may vary based on the statistical character- istics of the original data. This contribution adds to the grow- ing body of knowledge in data-driven modeling for studying LC3 cements and suggests further exploration of alternative augmentation methods and their application to different ce- ment materials.

Funding

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

1. Scrivener K, Martirena F, Bishnoi S, Maity S. 2018. Calcined clay limestone cements (LC3). *Cem Concr Res* 114: 49-56. https://doi. org/10.1016/j.cemconres.2017.08.017
2. Berriel SS, Favier A, Domínguez ER, Machado IS, Heierli U, et al. 2016. Assessing the environmental and economic potential of lime- stone calcined clay cement in Cuba. *J Clean Prod* 124: 361-369. https:// doi.org/10.1016/j.jclepro.2016.02.125
3. Riera JD, Miguel LFF, Iturrioz I. 2016. Evaluation of the discrete el- ement method (DEM) and of the experimental evidence on concrete behaviour under static 3D compression. *Fatigue Fract Eng Mater Struct* 39(11): 1366-1378. https://doi.org/10.1111/ffe.12453
4. Saleem MA, Saleem MM, Ahmad Z, Hayat S. 2021. Predicting com- pressive strength of concrete using impact modulus of toughness. *Case Stud Constr Mater* 14: e00518. https://doi.org/10.1016/j.cscm.2021. e00518
5. Van Bunderen C, Benboudjema F, Snellings R, Vandewalle L, Cizer Ö. 2021. Experimental analysis and modelling of mechanical properties