minerals

MDPI

Article

On the Prediction of the Mechanical Properties of Limestone Calcined Clay Cement: A Random Forest Approach Tailored to Cement Chemistry

Taihao Han [1], Bryan K. Aylas-Paredes [1], Jie Huang [2], Ashutosh Goel [3], Narayanan Neithalath [4,*] and Aditya Kumar [1,*]

[1] Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA; thy3b@mst.edu (T.H.); b.aylasparedes@mst.edu (B.K.A.-P.)
[2] Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA; jieh@mst.edu
[3] Department of Materials Science and Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA; ag1179@soe.rutgers.edu
[4] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85287, USA
* Correspondence: narayanan.neithalath@asu.edu (N.N.); kumarad@mst.edu (A.K.); Tel.: +480-965-6023 (N.N.); +573-341-6994 (A.K.); Fax: +480-965-0057 (N.N.); +573-341-6934 (A.K.)

Abstract: Limestone calcined clay cement (LC^3) is a sustainable alternative to ordinary Portland cement, capable of reducing the binder's carbon footprint by 40% while satisfying all key performance metrics. The inherent compositional heterogeneity in select components of LC^3, combined with their convoluted chemical interactions, poses challenges to conventional analytical models when predicting mechanical properties. Although some studies have employed machine learning (ML) to predict the mechanical properties of LC^3, many have overlooked the pivotal role of feature selection. Proper feature selection not only refines and simplifies the structure of ML models but also enhances these models' prediction performance and interpretability. This research harnesses the power of the random forest (RF) model to predict the compressive strength of LC^3. Three feature reduction methods—Pearson correlation, SHapley Additive exPlanations, and variable importance—are employed to analyze the influence of LC^3 components and mixture design on compressive strength. Practical guidelines for utilizing these methods on cementitious materials are elucidated. Through the rigorous screening of insignificant variables from the database, the RF model conserves computational resources while also producing high-fidelity predictions. Additionally, a feature enhancement method is utilized, consolidating numerous input variables into a singular feature while feeding the RF model with richer information, resulting in a substantial improvement in prediction accuracy. Overall, this study provides a novel pathway to apply ML to LC^3, emphasizing the need to tailor ML models to cement chemistry rather than employing them generically.

Keywords: limestone calcined clay cement; compressive strength; feature reduction; feature enhancement; machine learning

Citation: Han, T.; Aylas-Paredes, B.K.; Huang, J.; Goel, A.; Neithalath, N.; Kumar, A. On the Prediction of the Mechanical Properties of Limestone Calcined Clay Cement: A Random Forest Approach Tailored to Cement Chemistry. *Minerals* **2023**, *13*, 1261. https://doi.org/10.3390/min13101261

Academic Editors: Vineet Shah and Anuj Parashar

Received: 29 August 2023
Revised: 23 September 2023
Accepted: 26 September 2023
Published: 27 September 2023

1. Introduction

Concrete stands as the most widely used human-made material in the world, an essential ingredient for construction. As urbanization continues across the globe, driven by ever-increasing population growth and ambitious infrastructure projects, the global demand for concrete is projected to rise by an additional 10% by the year 2050 [1]. However, the environmental toll of this new infrastructure is staggering: the production of Ordinary Portland Cement (OPC), the main component of concrete, contributes to ~8% of the world-wide carbon footprint [2,3]. In the face of this environmental conundrum, various strategies have been proposed to mitigate the detrimental impact of cement production, primarily

focusing on emission reductions, energy optimization, and the efficient use of materials. One of the most promising of these is using supplementary cementitious materials to partially replace clinkers. This approach holds particular significance for two primary reasons: (1) the energy efficiency of contemporary cement kilns has already been fine-tuned to near-maximum levels, meaning that there is limited scope for further reducing carbon emissions through clean energy solutions alone, and (2) the decomposition of limestone contributes to 60% of CO_2 emissions [4], which cannot be avoided during the OPC manufacturing process. Hence, any meaningful reduction in carbon emissions must directly address the OPC clinker itself. In recent years, researchers have formulated a new type of ternary-blended cement known as limestone calcined clay cement (LC^3). This innovative blend leverages calcined clay and limestone to significantly reduce the clinker content, allowing for formulations of binders with as little as 50% OPC. Preliminary studies indicate that LC^3 can meet or even exceed the performance metrics of OPC in various aspects, including strength, durability, and workability [5–8]. The emergence of LC^3, therefore, represents a promising milestone in the development of sustainable cementitious materials without compromising mechanical performance and durability.

LC^3 is formulated with a maximum of 30% calcined clay, 15% limestone, and 5% gypsum, thereby allowing the clinker content to be potentially reduced to as low as 50% [9]. Clay is typically calcined at 700–900 °C so as to convert crystalline aluminosilicate phases into amorphous ones. This is significantly lower than the manufacturing temperature (~1450 °C) compared with OPC. As a result, the LC^3 production process can achieve a remarkable 35–40% reduction in both energy consumption and CO_2 emissions [10]. Except from an energy-saving perspective, LC^3 also benefits from unique chemical synergies between its components. The primary chemical reaction involves the hydration of OPC, forming calcium silicate hydrate (C-S-H), portlandite, and other hydration products. Calcined clay, serving as a pozzolanic material, predominantly reacts with free portlandite to form additional C-S-H [11]. The reactivity of the clay is highly sensitive to the calcination temperature: insufficient heating fails to remove water and form amorphous phases, while temperatures exceeding 900 °C diminish reactivity because of recrystallization into spinel, mullite, or cristobalite [12]. Prior research [13] suggests that clay reactivity is influenced not just by the molecular structure but also by the alite and belite content in OPC. Limestone also plays a crucial role by providing additional surfaces for the nucleation of hydrates, thereby boosting OPC hydration kinetics, especially at early ages. Limestone can also react with alumina in clay (and OPC) to form carboaluminate hydrates [11]. All aforesaid hydrates are favorable because they can fill pores, serve as binding agents, and provide strength.

Compressive strength is a critical indicator of concrete quality, and a significant body of research has investigated the mechanical performance of LC^3. While a majority of studies conclude that the 28-day compressive strength of LC^3 is comparable to OPC, variations have been noted at other ages [6]. Dhandapani et al. [5] found that the 3-day strength of LC^3 and its associated concretes are slightly lower than OPC. By 7 days, however, the strength of LC^3 reportedly matches or even surpasses that of OPC [6]. Though indispensable, laboratory experiments aimed at understanding compressive strength are both labor-intensive and expensive. As a result, there is an urgent need for reliable numerical models to estimate compressive strength. Many numerical models have been developed to predict the compressive strength of cementitious systems [14–18]. These models effectively quantify the impact of various factors, such as the water-to-cement ratio, the degree of hydration, and curing ages, on the compressive strength of plain OPC pastes. However, these existing models fall short when applied to LC^3 for several reasons. Firstly, the data domains for these models differ from those of LC^3, thereby requiring the recalibration of coefficients. Secondly, while these models capture the chemical reactions in OPC, they fail to account for the complex mutual interactions between calcined clay, limestone, and OPC in LC^3. Lastly, assuming numerical models that encompass all these interactions are not only impractical but also dauntingly complex, such models would potentially require an extensive number

of coefficients, making them cumbersome—almost impractical—to use. Moreover, the underlying mechanisms affecting LC^3 strength have yet to be fully understood, adding another layer of complexity to the development of comprehensive models.

Machine learning is a promising solution to predicting the properties of multi-component materials. Although many studies have employed ML models to predict the properties of cementitious materials [19–21], only a few studies [22–24] have applied ML to LC^3. Thus, there is a technological gap associated with the limited development and use of ML applications in LC^3 systems, at least compared with other cementitious materials (e.g., OPC and alkali-activated cement). One major shortcoming in the current application of ML to cementitious materials is that researchers generally adopt ML models *as is* rather than customizing them to align with the unique features of cement chemistry, particularly during feature selection. Many studies [25–29] solely present Pearson correlation and SHapley Additive exPlanations (SHAP) to evaluate the influences of input variables on cement properties without deeper interpretation or without utilizing these metrics to refine input variables effectively. Feature refinement includes weeding out less significant variables to enhance prediction performance, forming a crucial juncture where data science intersects with cement chemistry. Since generic ML models are designed to be data-driven (rather than by theory) and applicable to a wide range of applications, certain features might contradict the foundational principles of cement chemistry. By investigating feature selection parameters, researchers can gain a more comprehensive understanding of the intricate relationships between mixture designs and properties, especially when introducing new materials to the cement system. Further, comparing the influences of input variables with established literature correlations can ensure that ML models adhere to core material principles. If contradictions are observed, researchers can gain insights into how to fix their models, rather than being left in the dark by the model's opaque nature.

This research harnesses the power of the random forest (RF) model to predict the compressive strength of LC^3 systems. To tailor the RF model to LC^3, three feature reduction methods—Pearson correlation, SHAP, and variable importance—are employed to analyze the influence of LC^3 components and mixture design on compressive strength. Direct comparison between methods and practical guidelines for utilizing these methods on cementitious materials are elucidated. Additionally, a feature enhancement method (i.e., topological constraint theory) is utilized, consolidating numerous input variables related to calcinated clay into a singular feature (*number of constraints*) while feeding the RF model with richer information, which includes not only the chemical composition but also reactivity. By evaluating the performance of feature reduction and feature enhancement methods, predictions with these two methods are compared with an outcome from the standalone model. While the methods proposed in this study are designed for LC^3, they provide potential applicability across a wide range of properties of various cementitious materials.

2. Modeling Methods

2.1. Database Collection

A compressive strength database for LC^3 was compiled from the existing literature [11,30–45], comprising 430 distinct data records, each with 18 inputs and a single output. Through the random selection of data records, this database is split into two subsets: a training dataset and a testing dataset. The training dataset containing 75% of the data records trains the RF model, while the testing dataset containing the remaining 25% of the data records is employed to assess the model's performance. The evaluation process utilizes five key statistical metrics: coefficient of determination (R^2), Pearson correlation coefficient (R), mean absolute error (*MAE*), root mean squared error (*RMSE*), and mean absolute percentage error (*MAPE*). The input variables of the database are as follows: clay content ($\%_{\text{mass of LC3}}$); SiO_2 in clay ($\%_{\text{mass of clay}}$); Al_2O_3 in clay ($\%_{\text{mass of clay}}$); CaO in clay ($\%_{\text{mass of clay}}$); calcination temperature (°C); calcination time (hour); limestone content ($\%_{\text{mass of LC3}}$); CaO in limestone ($\%_{\text{mass of limestone}}$); OPC content ($\%_{\text{mass of LC3}}$); SO_3 in OPC ($\%_{\text{mass of OPC}}$); CaO in OPC ($\%_{\text{mass of OPC}}$); SiO_2 in OPC ($\%_{\text{mass of OPC}}$); Al_2O_3 in OPC

($\%_{\text{mass of OPC}}$); Fe_2O_3 in OPC ($\%_{\text{mass of OPC}}$); water-to-binder ratio (unitless); age (day); curing temperature (°C); and relative humidity during curing (%). The output is compressive strength (MPa). Though previous studies have provided other oxide compositions in clay, these were deliberately omitted from this database, as their contribution to strength is minimal. When integrating experimental results into a database, it is vital to apply expert knowledge to filter out irrelevant features. This narrows the degree of freedom of the database, simplifies the prediction process of machine learning models, and helps to avert overfitting on irrelevant information. Next, limestone can provide carbonate ions to react with calcined clay. The required number of carbonate ions allowing for an effective reaction depends on OPC and clay compositions [13]. The database includes low-quality limestones (CaO < 50%), and the impact of quality on the compressive strength remains ambiguous. Consequently, the CaO content in limestone was included in the database to shed light on this aspect. The statistical parameters pertaining to input and output variables are shown in Table 1, which exhibits the data domain and data distribution.

Table 1. Statistical parameters interpreting the data domain for 18 inputs and 1 output (bold) for the LC3 compressive strength database.

Attribute	Unit	Min.	Max.	Mean	Std. Dev.
Clay Content	$\%_{\text{mass of LC3}}$	10	60	25.29	9.66
SiO_2 in Clay	$\%_{\text{mass of clay}}$	34.10	79.63	55.98	10.52
Al_2O_3 in Clay	$\%_{\text{mass of clay}}$	10.55	46.99	31.57	10.33
CaO in Clay	$\%_{\text{mass of clay}}$	0	5.89	0.53	0.85
Calcination Temperature	°C	550	925	762.01	77.39
Calcination Time	hour	0.20	3	1.27	0.71
Limestone Content	$\%_{\text{mass of LC3}}$	0	31.13	8.53	7.67
CaO in Limestone	$\%_{\text{mass of limestone}}$	29.05	100	70.70	24.90
OPC Content	$\%_{\text{mass of LC3}}$	25	90	65.93	13.35
SO_3 in OPC	$\%_{\text{mass of OPC}}$	0.67	9.49	3.26	1.21
CaO in OPC	$\%_{\text{mass of OPC}}$	16.37	34.07	20.86	3.22
SiO_2 in OPC	$\%_{\text{mass of OPC}}$	1.52	7.35	5.11	1.07
Al_2O_3 in OPC	$\%_{\text{mass of OPC}}$	52.17	68.48	62.39	3.49
Fe_2O_3 in OPC	$\%_{\text{mass of OPC}}$	0.20	7.69	3.01	1.06
Water-to-Binder Ratio	unitless	0.10	0.90	0.47	0.08
Age	day	1	270	28.75	37.93
Curing Temperature	°C	5	50	22.53	5.38
Relative Humidity	%	80	100	92.95	4.94
Compressive Strength	**MPa**	4.60	75	36.66	16.11

Figure 1 illustrates the feature selection methods adopted in this study. Those methods are used to refine the input variables used in the LC^3 database and enhance the performance of the RF model. It is important to recognize the value of domain knowledge in simplifying the database for LC^3 and similar cementitious databases. These databases usually contain complex mixture design and processing parameters, but some input features are irrelevant to some properties. Using domain knowledge can identify and eliminate irrelevant features to reduce the degree of freedom of the database. Moreover, drawing upon the perspective of data science, it is noteworthy that, while certain input features may appear to mathematically correlate with specific properties, this might be a result of a data domain limitation. Contrarily, from a cement chemistry viewpoint, these features might not genuinely correlate with target properties. The inclusion of such illusory correlations can lead to the overfitting of ML models, thereby compromising their generalizability. Section 2.3 will delve into three feature reduction methods (i.e., Pearson correlation, SHAP value, and variable importance). The core objective of these methods is to rank input variables based on their influences on target properties. Post-ranking, the less significant variables are removed from subsequent analyses. This judicious reduction ensures that the ML models reduce the processing time without scarifying prediction accuracy, and in many instances,

the accuracy is bolstered. Section 2.4 introduces the feature enhancement technique. A key advantage of this approach is its ability to merge multiple input variables into a singular, more informative feature. Therefore, ML models can learn more useful correlations from a reduced number of inputs while consuming fewer computational resources. This not only enhances the models' learning capability but also substantially curtails computational complexity (and, thus, the time required to train the models). Furthermore, this method saves computational power, paving the way for improved prediction performance.

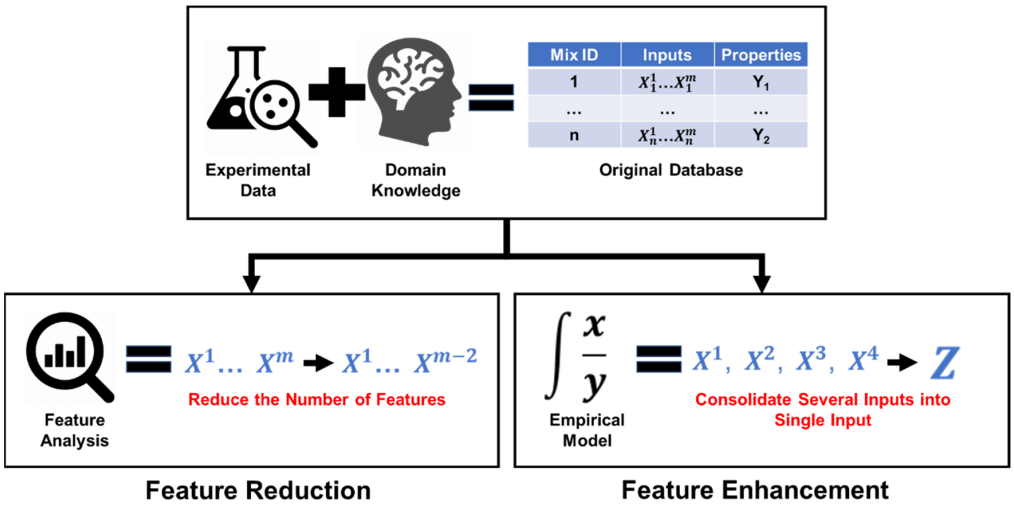

Figure 1. Schematic representation of feature selection method proposed in this study. Researchers utilize their knowledge to pre-filter irrelevant input variables while consolidating the database. The feature reduction method can be used to further reduce the degree of freedom of the database. The feature enhancement method can use a new singular input variable to represent information performed by several input variables.

2.2. Random Forest (RF)

The RF model builds upon the conventional classification-and-regression tree (CART) model to deliver more accurate and robust predictions. Unlike CART, RF incorporates the bagging algorithm [46,47] and a two-step randomization [47,48] process to create a forest consisting of independent decision trees. During training, RF constructs hundreds of these trees, each grown from a randomly selected subset of the parent training dataset, with repeated selection permitted. The unselected data records are defined as an out-of-bag (OOB) sample. Notably, the sub-dataset for training each tree must equal the size of the parent training dataset, preserving diversity. While the CART model evaluates all input variables at each node, RF introduces further randomness by selecting only a certain number of input variables to determine the optimal split. Trees in RF grow until the homogeneity of the last tree node cannot be further improved by splitting, ensuring diversity within the forest. Unlike CART, pruning and smoothing algorithms are not applied in RF, allowing each tree to grow as deeply as possible. At the testing stage, RF leverages the bagging algorithm to collect and average the outputs from individual trees, producing the final prediction. This combination of bagging and two-step randomization effectively reduces both variance and bias errors, enhancing the model's reliability [49,50]. To avoid overfitting and underfitting, the hyperparameters of RF are optimized via the 10-fold cross-validation (CV) [51,52] and grid-search methods [51,53].

Tree-based models provide a unique feature: they can rank the importance of variables without an additional algorithm. This ranking capability allows researchers to filter out insignificant or irrelevant input variables, enhancing models' computational efficiency and performance. When the RF model processes numerical data, variable importance [54–57] can be detailed as follows. For each individual tree, t, within the forest, there is a corresponding OOB sample, OOB_t. The OOB_t sample comprises data points that are not included in the

bootstrap sample used to grow the tree, t. When tree, t, produces predictions about OOB_t, the error (mean absolute error) is denoted as $errOOB_t$. To calculate the importance of a variable, the values of the target input variables of OOB_t are randomly permuted to obtain a new sample, denoted $OOB_t{}^j$. Then, the model evaluates the prediction performance ($errOOB_t{}^j$) of the $OOB_t{}^j$ sample. The variable importance is defined as

$$Variable\ Importance = \frac{1}{ntree}\sum_t \left(errOOB_t^j - errOOB_t\right) \quad (1)$$

where *ntree* represents the number of trees, t, in the forest. In the end, the RF model ranks input variables based on their importance. Researchers can use it as a guideline to remove insignificant variables and reduce the complexity of the tree structure.

2.3. Feature Reduction Methods

This section presents three feature reduction techniques, emphasizing their capacity to prioritize input variables based on their impact on the output. By harnessing this knowledge, insignificant input variables can be systematically removed, thereby reducing the dimensionality of the LC^3 database. While the descriptions of Pearson correlations and SHAP values are demonstrated herein, the variable importance is detailed in the previous section.

The Pearson correlation coefficient [58], often represented as R, is a statistical measure used to quantify the linear relationship between two variables. Its value can range from -1 to 1. A value of 1 signifies a perfect positive linear relationship, indicating that as one variable increases, the other does as well in a directly proportional manner. Conversely, a value of -1 implies a perfect negative linear relationship, meaning that as one variable increases, the other decreases in a directly inverse proportion. A coefficient of 0 suggests no linear correlation between the variables. The calculation for this coefficient is derived from the following formula:

$$R = \frac{\sum_{I=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{I=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{I=1}^{n}(y_i - \overline{y})^2}} \quad (2)$$

where x and y represent individual data points, and $\overline{x}$ and $\overline{y}$ are the means of the respective datasets. It is crucial to understand that the Pearson correlation coefficient strictly measures linear relationships. Hence, nonlinear relationships might not be effectively captured by R. Additionally, a correlation value of 0 does not necessarily indicate the variables are independent; it simply denotes the absence of a linear relationship. Furthermore, it is pivotal to remember that this coefficient does not equate to causation. A high correlation between two variables does not inherently suggest that changes in one cause changes in the other; other factors or underlying variables could influence the observed relationship. In summary, while the Pearson correlation coefficient offers valuable insight into the linear dependence between two variables, its interpretation demands careful consideration and often warrants further analysis.

SHapley Additive exPlanations (SHAP) is a method developed by Lundberg and Lee [59] to reveal the importance and effects of input features. It is built on the concept of the Shapley value by unifying the additive feature attribution method, game theory, and local explanations. Using principles from cooperative game theory, SHAP calculates the Shapley value for each feature, representing the average marginal contribution of that feature across all possible combinations of features. To be specific, in a database with the input variables $x = (x_1, x_2, \ldots, x_n)$, where n is the number of input variables, SHAP creates

simplified inputs, x', which map into x through $x = h_x(x')$. Based on the x', the original model, *f(x)*, can be approximated with a linear function:

$$f(x) = g(x') = \varphi_0 + \sum_{i+1}^{M} \varphi_i x_i' \tag{3}$$

M represents the number of input features; φ_0 is the constant when all inputs are missing; φ_i is the feature attribution value expressed by

$$\varphi_i = \sum_{z' \in x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \backslash i)] \tag{4}$$

$$f_x(z') = f\left(h_x^{-1}(x')\right) = E[f(x)|\, x_{z'}] \tag{5}$$

$|z'|$ represents the number of non-zero entries in x', and φ_i is the SHAP value. Given these structures, the SHAP value inherits the properties of additivity, local accuracy, missingness, and consistency [60,61].

The two unique advantages of the SHAP value are its dual levels of interpretability—both global and local. Unlike many traditional feature importance metrics in machine learning, SHAP not only discerns the significance of each input feature but also ascertains its positive or negative influence. While global interpretability provides an overarching view of the model, highlighting general feature influences on predictions, local interpretability delves deeper, examining individual instances. Moreover, the SHAP value enhances the interpretability of ML models by consistently explaining interaction effects between features for individual predictions.

2.4. Feature Enhancement Method

This section introduces a feature enhancement method that consolidates multiple input variables into a single parameter that embodies extensive information. For clay, calcination removes water from clay and transforms crystalline structures into amorphous ones, enhancing their reactivity [6]. Yang et al. [62] found that the reactivity of amorphous calcium aluminosilicate materials can be envaulted by a singular parameter—*number of constraints*—which can be calculated with topological constraint theory [63,64]. Our preceding research [65–67] further validated how this parameter can reliably estimate the reactivity of various families of aluminosilicate-rich cementitious materials. Additionally, our studies highlighted the potential of the *number of constraints* to replace various input variables used in ML and enhance prediction accuracy. The benefit of using this parameter is twofold: it simplifies the dataset for machine learning models and encapsulates vital information on the molecular structure and aqueous reactivity of aluminosilicate-based cementitious materials. Consequently, the *number of constraints* is utilized in this study to replace the chemical composition and processing parameters of clays. Most clays in our database underwent calcination at temperatures exceeding 600 °C for over an hour, ensuring their largely amorphous nature.

The fundamental constituents of the clay framework are SiO_2, CaO, and Al_2O_3, disregarding any minor components. The normalized chemical composition is represented as $(Cao)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$, where x and y denote the molar fractions. Two chemical constraints found in amorphous calcium aluminosilicate materials are angular bond-bending (BB) and radial bonding–stretching (BS) constraints [62,68,69]. Si/Al tetrahedrons contribute 4 BS and 5 BB constraints. While O atoms linked to Si/Al tetrahedrons add 1 BB constraint, those connected to Ca atoms provide 1 BS constraint [62,69–72]. Calcium aluminosilicate materials can be categorized into three groups based on their chemical compositions:

Depolymerized regime ($y - x \leq -2/3$): Dominated by Ca atoms, leading to the isolation of Si and Al tetrahedrons due to non-bridging oxygens (NBOs). NBOs promote aqueous reactivity.

Partially depolymerized regime ($-2/3 \leq y - x \leq 0$): Dominated by Si atoms. Contains both bridging oxygen (BOs) atoms and NBOs, leading to increased crystallinity and reduced reactivity.

Fully polymerized regime ($0 \leq y - x$): Dominated by Al atoms, resulting in a rigid structure with minimal reactivity due to scarce NBOs.

In this research, all examined clays fall within the fully polymerized regime, which demonstrates the least reactivity. The formula to determine the number of constraints is presented in Equation (6).

$$n_c = \frac{11 + 13y - 13x}{3 - 2x + 2y} \tag{6}$$

3. Results and Discussion

The LC^3 database was partitioned into training and testing datasets, with the former encompassing 75% of the primary database and the latter constituting the remaining 25%. Figure 2 shows the predictions of compressive strength as yielded by the RF model, compared with the measurements from the testing dataset. The statistical parameters representing the accuracy of the predictions—i.e., prediction performance—are itemized in Table 2. A cursory glance at both the figure and the table reveals impressive reliability in the compressive strength predictions, underscored by an *R*-value of 0.94 and an *RMSE* of 5.64 MPa. The experimental measurement error for compressive strength for cementitious materials stands at approximately 5 MPa [73]. Remarkably, the deviation in our prediction closely mirrors this experimental error. This implies that the RF model can yield highly accurate predictions of the compressive strength of LC^3. Such excellent performance holds significant promise for cement scientists, empowering them to rapidly identify promising mixture designs and evaluate their compressive strength rather than experimenting with an expansive array of mixture designs. It is hardly surprising that the RF model exhibits such excellent performance. A retrospective look at our past research [65–67,74,75] demonstrates that the RF model consistently produces reliable predictions of compressive strength for various cementitious materials. These publications also elucidate the reasons that the RF model—when contrasted with analytical models or other ML models—can achieve such excellent performance with cementitious materials.

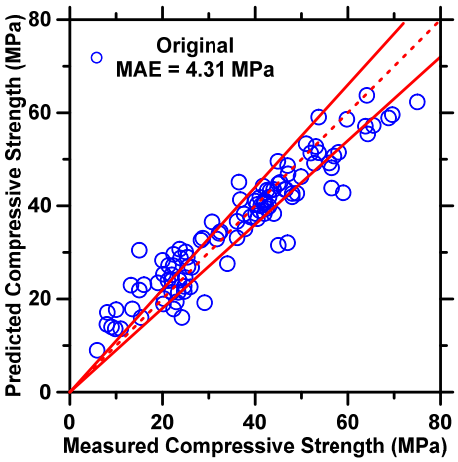

Figure 2. The prediction performance of the compressive strength of LC3 on the testing dataset as produced by the RF model with original input variables. The mean absolute error (MAE) for overall predictions is shown in the legend. The solid lines show 10% error bounds, and the dashed line is the ideal prediction.

Table 2. Prediction accuracy (represented by five statistical parameters) of compressive strength of LC3 as produced by the RF model with original inputs, feature reduction, and feature enhancement methods.

ML Model	R	R^2	MAE	MAPE	RMSE
	Unitless	Unitless	MPa	%	MPa
Original	0.9453	0.8936	5.641	16.56	5.641
Feature Reduction	0.9421	0.8875	4.243	15.52	5.548
Feature Enhancement	0.9588	0.9194	3.431	11.30	4.608

After evaluating the performance of the RF model in predicting the compressive strength of LC^3, the study now shifts its focus to understanding feature selection techniques. Figure 3 illustrates the Pearson correlation coefficient between input and output variables for the LC^3 compressive strength database. Such techniques are commonly employed during data pre-processing to identify and eliminate irrelevant variables, thereby reducing the dimensionality of the dataset. In terms of coefficient *R*, a value close to 1 indicates a strong positive correlation; while one variable increases, the other does too. Conversely, a value close to −1 implies a strong negative correlation; while one variable increases, the other decreases. A value near 0 indicates that no linear correlation is found between the two variables. However, this does not necessarily mean the variables are independent; nonlinear correlations might still exist.

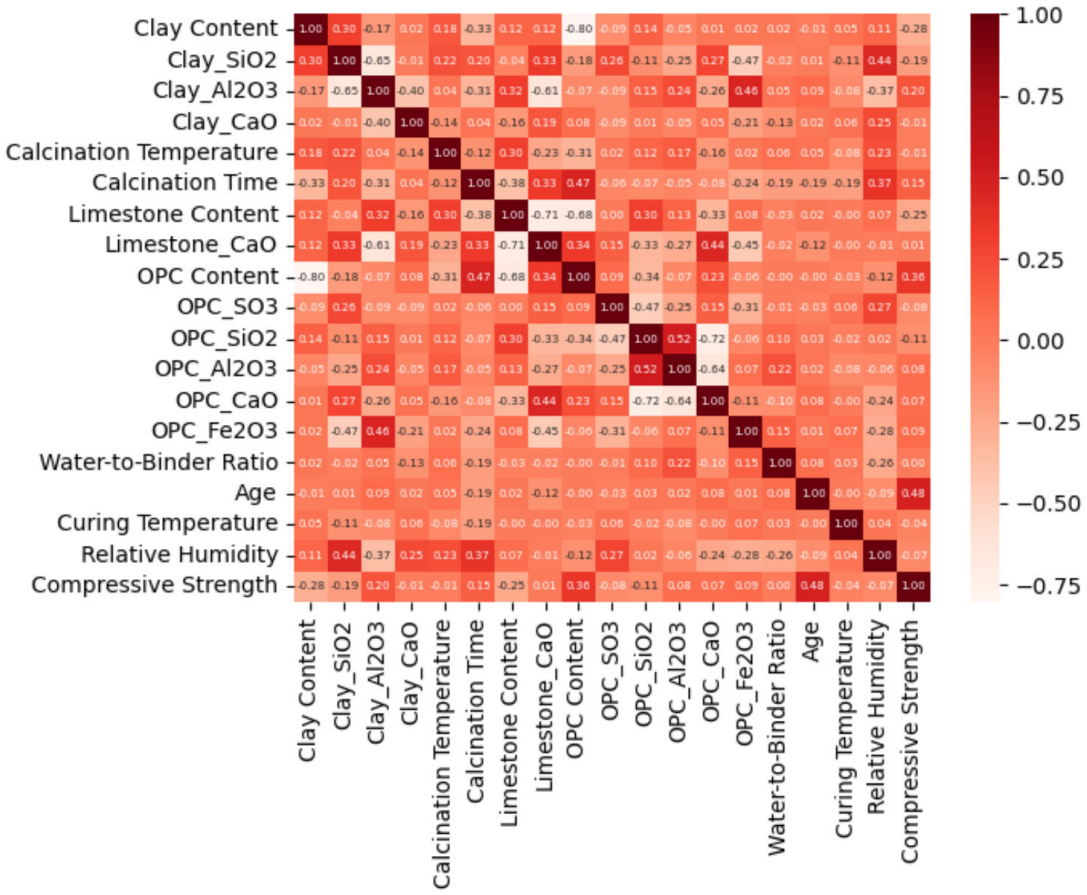

Figure 3. Pearson correlation coefficients between LC3 components, processing parameters, and compressive strength. The dark color represents positive correlations, and the lighter color represents negative correlations.

An analysis of Figure 3 reveals that the absolute value of *R* between most input variables remains below 0.5, suggesting that these variables are relatively independent. In some cement studies, variables such as water content, cement content, and the water-to-binder ratio are included. The absolute value of *R* among these three variables may be high. Consequently, researchers might contemplate excluding one of these to prevent potential overfitting in ML models. This caution arises because high correlations may assign additional weights to certain parameters. However, the removal of any variable should be approached judiciously. Some variables may exhibit strong mathematical correlations—for instance, an *R*-value of 0.47 between OPC content and calcination time—but they are independent in real experiments. Such discrepancies can be attributed to the data distribution in the sampled database. Incorporating a larger and more diverse database might drive such correlation coefficients closer to 0. Considering the *R*-values between OPC content, clay content, and limestone content in LC^3, these variables are negatively correlated. This is anticipated, as their measurements are in the $\%_{\text{mass of LC3}}$; an increase in one implies a decrease in the others. Some researchers may remove one of these three input variables owing to their strong correlations. However, it is imperative to retain all three variables in the database since they significantly influence the compressive strength. ML models, by their nature, do not understand these three parameters collectively, accounting for 100%. Without applying constraints, the model could establish incorrect correlations and fail to optimize the mixture design of the new LC^3.

After interpreting Pearson correlations between input variables, we shift our attention to the relationships between inputs and output. The underlying assumption is that input variables should exhibit a discernible relationship with the output. If certain input variables demonstrate little-to-no correlation, they might be pruned from the database. However, this principle is not universal. To further elaborate this concept, the chemical compositions of clay, limestone, and OPC do not manifest direct linear correlations with compressive strength. These chemical parameters fundamentally define these three raw materials. The relationship between chemical composition and compressive strength becomes clearer when considered in tandem with raw material content. When researchers review experimental results from prior research, they will find that certain correlations between LC^3 parameters and compressive strength are already established. By comparing these known experimental correlations with Pearson correlations, discrepancies may be identified. If Pearson correlations appear to contradict experimental findings, it could lead to doubts regarding the database's reliability and its data diversity. Figure 3 reveals the robust positive correlation between age and compressive strength. This observation aligns with prior findings showing that compressive strength tends to increase monotonically with age. However, an unexpected insight is the negligible correlation observed between the water-to-binder ratio, curing conditions, and compressive strength. Conventionally, lower water content is associated with higher compressive strength. However, exceedingly low water levels can hamper the hydration reaction, thereby undermining the compressive strength. Moreover, optimal curing conditions, like elevated temperatures and high-humidity environments, are known to accelerate hydration and enhance compressive strength. This divergence between the database and experimental findings is attributed to the fact that the majority of LC^3 samples share similar water content and curing conditions, which dilutes their influences on compressive strength. Although Pearson correlation presents some limitations in feature selection, it provides invaluable insights into data selection when introducing new materials and complex materials (e.g., fly ash) to LC^3 systems. Given that the interactions between these novel materials and LC^3 are not extensively studied, Pearson correlation offers an initial framework to elucidate potential relationships. Compared with other feature section techniques, Pearson correlation is easy to apply to any database without the need for in-depth machine learning or programming expertise. This approach, therefore, can be a powerful tool in efficiently filtering out insignificant variables.

Figure 4 demonstrates SHAP values corresponding to each input variable for individual predictions. This visualization aids in understanding the relative influence of each

variable on the RF model's predictions. The variables are arranged hierarchically, with the most influential ones positioned at the top. The color-coding—red and blue—is indicative of the magnitude of the input values. Specifically, a red dot represents high input values, while blue signifies lower values. The positioning of these colors relative to the zero baseline provides insights into their impact on the output. For instance, when most red dots are situated on the positive side, it denotes that the higher values of that input variable tend to increase the output. Conversely, if more red dots are on the negative side, it signifies that higher values lead to a decrease in the output properties. Blue dots are interpreted similarly but with the opposite value behavior in mind. Compared with Pearson correlations, the SHAP value method has several advantages. While Pearson correlation primarily provides global relationships between variables, SHAP values provide additional information for interpreting the influences of input variables. They not only highlight the significance of each variable for specific predictions but also elucidate the quantitative influence an input variable has on the output. This presents a detailed picture beyond just a generalized correlation coefficient. SHAP values can also be instrumental in developing analytical models, which allow end-users to predict properties without the need for advanced programming expertise. The magnitude of both positive and negative correlations between inputs and outputs provides valuable insights into determining weight assignments within these models. By setting these weights appropriately, coefficients can be fitted in refined ranges, leading to accurate prediction performance. Nonetheless, it is essential to recognize that SHAP values only evaluate the correlations between inputs and outputs. As a result, SHAP values do not effectively determine whether or not a given input variable has the potential to cause overfitting.

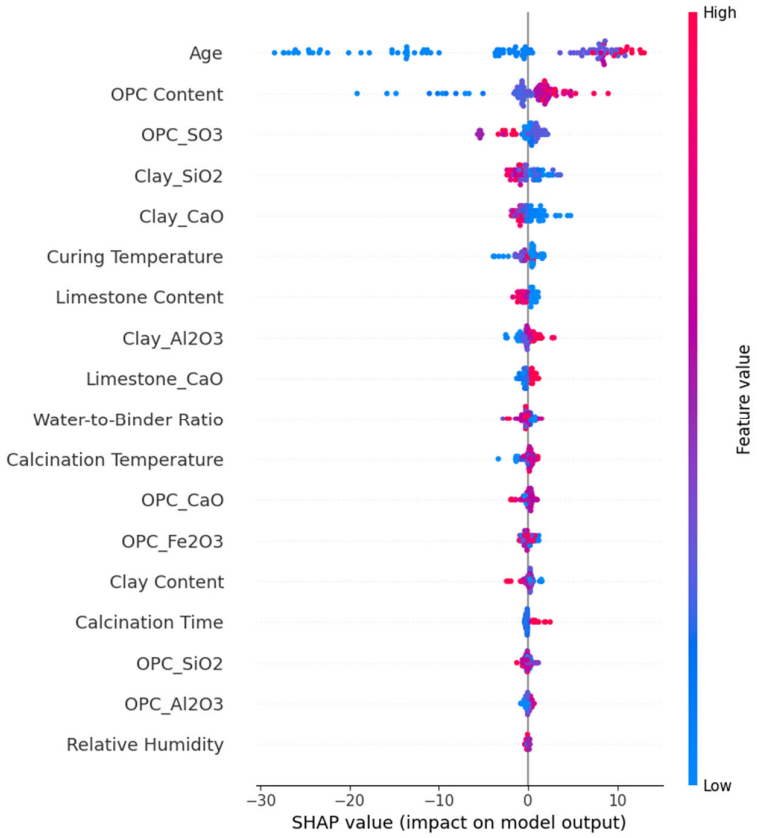

Figure 4. SHAP values of LC3 components and processing parameters for each prediction of compressive strength. The most influential variable is ranked at the top. The red color represents positive correlations, and the blue color represents negative correlations.

Figure 4 reveals that age is the most significant variable, exerting a positive influence on the compressive strength of LC^3. This observation aligns with foundational principles

in cement chemistry, wherein longer hydration periods translate into greater compressive strengths [6,76,77]. Such correlations can be used to debug ML models. For instance, a SHAP analysis indicates a diminishing compressive strength with increasing age, while Pearson correlation suggests the contrary (which means that the database is error-free). The monotonous, directly proportional relationship between the age and compressive strength of LC^3 is well known. Such a discrepancy suggests that the ML model may have learned an incorrect correlation. In such cases, the solution might involve adjusting the model's hyperparameters and re-training or even embedding certain constraints to guide the model in establishing accurate correlations. The content of OPC demonstrates a pronounced positive correlation with compressive strength. This is anticipated, given that OPC serves as the primary constituent responsible for providing strength. Meanwhile, the SO_3 content in OPC is placed in the third rank, displaying an inverse relationship with compressive strength. Earlier research showed that even a small amount of gypsum can substantially delay the hydration reaction, leading to a notable dip in compressive strength during the initial 3-day period [65,78–80]. Compounds such as ettringite and monosulfoaluminate, which form from SO_3, contribute minimally to compressive strength. Given the vast range of SO_3 content variations, the RF model can sufficiently learn the influences of SO_3 on compressive strength. Further down the rankings, SiO_2 in clay exhibits a strong negative correlation. Higher SiO_2 levels imply a more rigid clay molecular structure, resulting in a reduced dissolution rate and reactivity. Interestingly, other components of OPC compositions and relative humidity seem to exert minimal influence. This could be attributed to narrow ranges and the limited variability of these input variables. Such unforeseen outcomes also highlight the potential limitations of SHAP values. Although a SHAP value can be utilized to evaluate the influence of input variables across diverse ML models, it might be inefficient when the model assigns less weight to an input. This is because SHAP values primarily assess the shifts in predicted values prompted by incremental changes in specific input variables. When a variable holds minimal weight, it corresponds to only slight variations in prediction, potentially obscuring its true impact.

Figure 5 presents the variable importance derived from the RF model for each input variable. These variables are systematically arranged: the variables exerting the most-to-least influence are positioned from left to right. It is noteworthy that the ranking of variables may differ between the SHAP value and variable importance; this discrepancy arises from the distinct mechanisms underlying each method. The SHAP value calculates predictions that fluctuate when a specific variable is altered. Essentially, it aggregates local data to quantify the global influence of an input variable. The performance is heavily reliant on the dataset in use, which means that a wide range of highly varied input variables could have strong influences. Conversely, variable importance is determined by shuffling a particular input variable and then measuring its impact on the overall prediction performance, making this method more contingent on the model's features and structures than the database. Given its direct correlation with prediction performance, variable importance is especially adept at pinpointing and tailoring inconsequential variables. Meanwhile, the variable importance provides critical knowledge to develop analytical models. Our previous studies [53,65,67,74,81–83] successfully harnessed this tool to craft user-friendly, closed-form analytical models for different materials.

Figure 5 illustrates that the composition of OPC exerts great influence on compressive strength, a finding that seems contradictory to the results derived from the SHAP analysis. This discrepancy is understandable. While the variability and data range for OPC compositions might be narrow, they undeniably play a pivotal role in shaping the prediction accuracy of the RF model. CaO in clay, the calcination temperature, and CaO in limestone show minimum impacts on compressive strength. One might consider omitting these from the database to decrease its complexity. Nevertheless, any decision to remove them must be grounded in cement chemistry insights. Past research has illuminated that, although only a minor fraction of limestone reacts with the alumina phases in cement and clay, forming the carboaluminate phase, most of it persists as an inert filler [5,6,34]. Given

its minimal chemical influence on hydration product formation, the variable related to limestone quality can be discarded. Furthermore, as clay is typically calcined between 700–800 °C [9], kaolinite begins its decomposition, transitioning into amorphous structures at temperatures as low as 500 °C [84]. Therefore, the calcination temperature might also be deemed redundant, especially since all clays in the database underwent calcination at temperatures exceeding 500 °C. However, caution must be exercised when considering the removal of CaO from clay. CaO is one of the key factors that determines reactivity. While our study predominantly features clays with low CaO content, in practical scenarios, some clays might exhibit higher CaO content. To ensure the generalization, this variable ought to be retained.

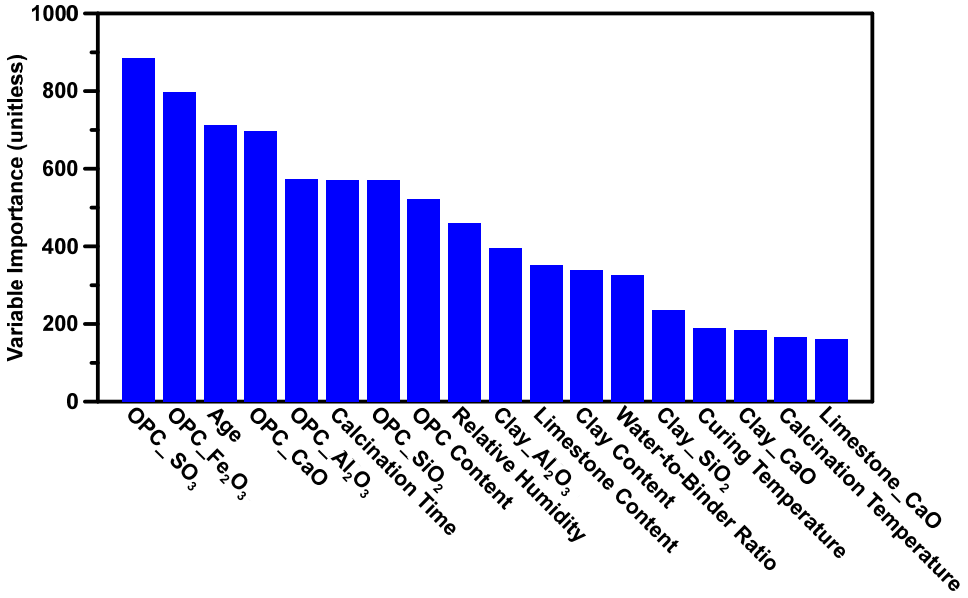

Figure 5. Quantitative evaluation of impacts of LC3 components and processing parameters on compressive strength. The most influential variable is ranked on the left.

By excluding two variables, the RF model discovers underlying correlations for LC^3 with only sixteen input variables. Figure 6 illustrates the RF model's predictions of the compressive strength of LC^3, now optimized through feature reduction. A detailed account of prediction errors from testing datasets is presented in Table 2.

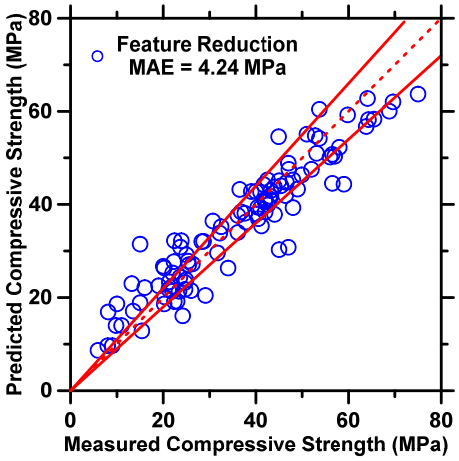

Figure 6. The prediction performance of the compressive strength of LC3 based on the testing dataset produced by the RF model while the feature reduction method is applied. The mean absolute error (MAE) for the overall predictions is shown in the legend. The solid lines show 10% error bounds, and the dashed line is the ideal prediction.

An examination of both Figure 7 and Table 2 demonstrates reliable predictions of compressive strength, especially when fine-tuned using the feature reduction method. From a quantitative standpoint, the predictions have an *R* of 0.94, coupled with an *RMSE* of 4.54 MPa. Training the model with these 16 variables trims the training time by nearly 10% in comparison with the 18 input variables, and yet, the predictive accuracy is superior. This reinforces the efficiency of the variable importance method in not only reducing the complexity of the database but also maintaining robust prediction reliability. While the SHAP value method was explored to prune input variables, it led to a noticeable slash in prediction accuracy. Given this outcome, its results have been omitted from this study.

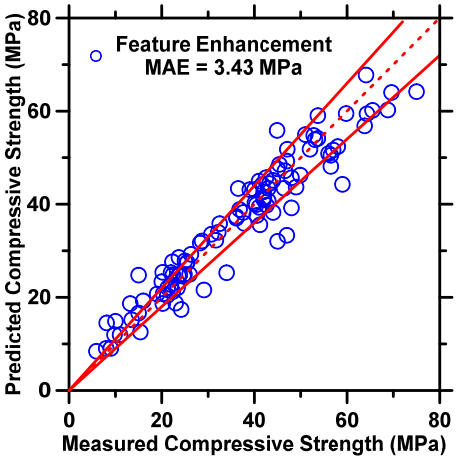

Figure 7. The prediction performance of the compressive strength of LC3 based on the testing dataset as produced by the RF model while both the feature reduction and feature enhancement methods are applied. The mean absolute error (MAE) for overall predictions is shown in the legend. The solid lines show 10% error bounds, and the dashed line is the ideal prediction.

After applying the feature reduction method, the feature enhancement method is utilized to replace the chemical composition and processing parameters of the clay with the *number of constraints*. As a result of implementing both methodologies, the RF model only needs to learn input–output correlations from 13 input variables. This simplification notably reduces both computational memory usage and the time required for training and testing. Figure 7 illustrates the RF model's predictions of the compressive strength of LC^3 when informed by feature reduction and enhancement techniques. A detailed account of prediction errors from the testing datasets is presented in Table 2.

Observing both Figure 7 and Table 2, it is evident that the RF model, when augmented with the aforementioned methods, yields accurate predictions of compressive strength. Quantitatively, the *R* and *RMSE* values for the predictions stand at 0.95 and 4.61 MPa, respectively. This figure demonstrates the superiority of predictions implemented with a combination of feature reduction and enhancement over those generated solely by the RF model or just with feature reduction. This can be attributed to the enhanced scope of information that the RF model receives. Unlike the standalone model, which is solely informed by the chemical composition and processing parameters of clay, the *number of constraints* provides the RF model with insights into the chemostructural properties of clay. This includes details like the quantities of various chemical bonds. Such data act as an effective proxy for representing the reactivity of clay—a facet not directly discernible from just the chemical composition. Clay with high reactivity readily interacts with free portlandite, water, and sulfate, leading to the formation of C-A-S-H, ettringite, and monosulfoaluminate [85,86]. These compounds play a pivotal role in reducing the binder's porosity, thereby enhancing compressive strength. In essence, such information obtained from the *number of constraints* empowers the RF model to robustly discover correct underlying input–output correlations for LC^3.

In conclusion, the feature reduction and feature enhancement methods have demonstrated their robust potential in trimming down the degree of freedom within the LC^3 database and enhancing the prediction performance. The abovementioned guidelines not only apply to LC^3 but can also be extrapolated to encompass other cementitious materials. Such tailored approaches are pivotal, as they demonstrate the importance of fine-tuning ML models to better fit the principles of cement chemistry rather than employing these models generically. Furthermore, the feature reduction methodologies serve a dual purpose. Firstly, they enhance the interpretability of ML models. This heightened transparency aids researchers in diagnosing potential issues within ML models and, if necessary, incorporating new features to refine predictions. Secondly, these methods pave the way for more informed decisions in the realm of cementitious material experiments. By discerning which components considerably influence a particular property, manufacturers and researchers can adjust formulations more precisely, ensuring optimal performance and efficiency in the resulting product.

4. Conclusions and Perspectives

Reducing its carbon footprint has placed the cement industry at the forefront of research initiatives. LC^3 emerges as a promising alternative to OPC, with a significantly reduced carbon footprint. The inherent compositional heterogeneity in select components of LC^3, combined with their convoluted chemical interactions, poses challenges to conventional analytical models when predicting mechanical properties. ML provides a promising solution for predicting the properties of multicomponent materials (e.g., LC^3). However, the generic applications of ML on cementitious materials may violate some laws of cement chemistry. This underscores a need for deeper explorations into tailoring ML models that can seamlessly integrate with cement chemistry's intricacies. This highlights the ongoing need for further research to fully understand ML models and integrate knowledge of cement chemistry into them.

In this study, an RF model was employed to predict the compressive strength of LC^3 in a high-fidelity manner. The database comprises over 400 data records, marking it sizable in comparison with most cement databases. Nevertheless, from a broader data science perspective, this scale would still be classified as relatively small. Most data science databases contain thousands to billions of data records, allowing for a richer understanding of input–output correlations. Gathering such vast amounts of data is not practical in cement research given the extensive costs and prolonged durations associated with data collection, especially for properties like long-term strength and durability. The solution lies in fostering a culture of collaborative data sharing within the cement research community. Such collaboration is commonplace in data science, where numerous repositories exist for researchers to share and access databases. Regrettably, the cement community currently lacks a dedicated platform for data communication. The development of an open-source repository for cement research is urgently required. Such a platform would not only encourage researchers to share data and ML algorithms but also ensure standardized data quality through the implementation of specific sharing protocols. With the inception of such a repository, the evolution of ML techniques in cement research would experience a significant boost. Concurrently, it would empower scientists to innovatively design new cement formulas more efficiently and at reduced costs.

Furthermore, three data reduction (i.e., Pearson correlations, SHAP value, and variable importance) and one data enhancement (i.e., topological constraint) methods were explored in this study. To aid in their application, this research provides an in-depth breakdown and step-by-step guidelines on how to leverage these data reduction methods to analyze and understand the intricate relationships between inputs and output. Each technique has a unique set of strengths and potential pitfalls. For this reason, a robust data analysis strategy would be better anchored on a combination of these methods rather than overly depending on just one. For instance, while one method might be good at identifying weaker correlations, another might be adept at understanding nonlinear relationships. After

identifying insignificant variables, it is crucial to overlay this understanding with domain knowledge regarding cement chemistry. This ensures a rational decision-making process on whether to retain or discard a given input variable. Venturing into data enhancement, the method amalgamates multiple input variables into a more enriched and informative single entity. Such an approach not only reduces the complexity of the database but also presents ML with more potent correlations to analyze and learn from.

Both the data reduction and enhancement strategies signify a pivotal shift from a broad, one-size-fits-all approach to ML to more tailored, cement-chemistry-based ML. Looking to the future, there is an evident trajectory toward further refining this symbiosis between ML and cement chemistry, starting with science-informed ML, where input variables are rooted in established scientific principles, and then, a transition toward ML models constrained and guided by material laws can occur, where these models would be adept at learning specific trends across diverse scenarios. The zenith of this evolution would be the development of ML models highly integrated with thermodynamic or kinetic frameworks. Such models would encapsulate material laws at every juncture of prediction, magnifying the reliability of their outputs.

To conclude, it is undeniable that ML has revolutionized research related to cement science, ushering in the conceptualization and development of innovative cementitious materials. While this paper merely scratches the surface of the potential intersections between ML and cement chemistry, but it ignites a robust dialog focused on customizing ML to cement science. The rapid evolution of AI has brought forth the emergence of generative AI as a cutting-edge field of exploration. Currently, its applications span a myriad of domains, from content creation in writing and image generation to advanced video synthesis. However, the potential of integrating generative AI with cement chemistry remains largely untapped. Imagine a scenario where generative AI is harnessed to learn from cement databases. This AI model could then extrapolate and design novel cementitious formulas that not only diverge from known databases but also amalgamate insights across them. Such an approach could inspire researchers to explore unthought realms. Generative AI could be profound, potentially fast-tracking the development of sustainable cement toward a future of carbon neutrality.

Author Contributions: Conceptualization, A.K.; methodology, T.H., B.K.A.-P. and J.H.; software, T.H. and A.K.; validation, T.H. and A.G.; formal analysis, T.H.; investigation, B.K.A.-P. and A.G.; resources, A.K. and N.N.; writing—original draft, T.H.; writing—review and editing, J.H., A.G., A.K. and N.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by the National Science Foundation (NSF-DMR: 2034856 and NSF-DMR: 2034871); the Kummer Institute (Missouri S&T) Ignition Grant; and the Federal Highway Administration (Award no: 693JJ31950021).

Data Availability Statement: The data used in this study are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Material Economics. *Industrial Transformation 2050-Pathways to Net-Zero Emissions from EU Heavy Industry; Stockholm, Sweden*. 2019. Available online: https://materialeconomics.com/publications/publication/industrial-transformation-2050 (accessed on 1 September 2023).
2. Shah, I.H.; Miller, S.A.; Jiang, D.; Myers, R.J. Cement Substitution with Secondary Materials Can Reduce Annual Global CO_2 Emissions by up to 1.3 Gigatons. *Nat. Commun.* **2022**, *13*, 5758. [CrossRef]
3. Miller, S.A.; Horvath, A.; Monteiro, P.J.M. Readily Implementable Techniques Can Cut Annual CO_2 Emissions from the Production of Concrete by over 20%. *Environ. Res. Lett.* **2016**, *11*, 074029. [CrossRef]
4. Pamenter, S.; Myers, R.J. Decarbonizing the Cementitious Materials Cycle: A Whole-systems Review of Measures to Decarbonize the Cement Supply Chain in the UK and European Contexts. *J. Ind. Ecol.* **2021**, *25*, 359–376. [CrossRef]
5. Dhandapani, Y.; Sakthivel, T.; Santhanam, M.; Gettu, R.; Pillai, R.G. Mechanical Properties and Durability Performance of Concretes with Limestone Calcined Clay Cement (LC3). *Cem. Concr. Res.* **2018**, *107*, 136–151. [CrossRef]
6. Sharma, M.; Bishnoi, S.; Martirena, F.; Scrivener, K. Limestone Calcined Clay Cement and Concrete: A State-of-the-Art Review. *Cem. Concr. Res.* **2021**, *149*, 106564. [CrossRef]

7. Hassan, A.M.S.; Shoukry, H.; Perumal, P.; Abd El-razik, M.M.; Aly, R.M.H.; Alzahrani, A.M.Y. Evaluation of the Thermo-Physical, Mechanical, and Fire Resistance Performances of Limestone Calcined Clay Cement (LC3)-Based Lightweight Rendering Mortars. *J. Build. Eng.* **2023**, *71*, 106495. [CrossRef]
8. Alghamdi, H.; Shoukry, H.; Abadel, A.A.; Khawaji, M. Performance Assessment of Limestone Calcined Clay Cement (LC3)-Based Lightweight Green Mortars Incorporating Recycled Waste Aggregate. *J. Mater. Res. Technol.* **2023**, *23*, 2065–2074. [CrossRef]
9. Scrivener, K.; Martirena, F.; Bishnoi, S.; Maity, S. Calcined Clay Limestone Cements (LC3). *Cem. Concr. Res.* **2018**, *114*, 49–56. [CrossRef]
10. Gettu, R.; Patel, A.; Rathi, V.; Prakasan, S.; Basavaraj, A.S.; Palaniappan, S.; Maity, S. Influence of Supplementary Cementitious Materials on the Sustainability Parameters of Cements and Concretes in the Indian Context. *Mater. Struct.* **2019**, *52*, 10. [CrossRef]
11. Antoni, M.; Rossen, J.; Martirena, F.; Scrivener, K. Cement Substitution by a Combination of Metakaolin and Limestone. *Cem. Concr. Res.* **2012**, *42*, 1579–1589. [CrossRef]
12. Amin, N.-; Alam, S.; Gul, S.; Muhammad, K. Activation of Clay in Cement Mortar Applying Mechanical, Chemical and Thermal Techniques. *Adv. Cem. Res.* **2012**, *24*, 319–324. [CrossRef]
13. Krishnan, S.; Bishnoi, S. A Numerical Approach for Designing Composite Cements with Calcined Clay and Limestone. *Cem. Concr. Res.* **2020**, *138*, 106232. [CrossRef]
14. Ahmed, H.U.; Abdalla, A.A.; Mohammed, A.S.; Mohammed, A.A. Mathematical Modeling Techniques to Predict the Compressive Strength of High-Strength Concrete Incorporated Metakaolin with Multiple Mix Proportions. *Clean. Mater.* **2022**, *5*, 100132. [CrossRef]
15. Elaty, M.A.A.A. Compressive Strength Prediction of Portland Cement Concrete with Age Using a New Model. *HBRC J.* **2014**, *10*, 145–155. [CrossRef]
16. Cui, H.Z.; Lo, T.Y.; Memon, S.A.; Xing, F.; Shi, X. Analytical Model for Compressive Strength, Elastic Modulus and Peak Strain of Structural Lightweight Aggregate Concrete. *Constr. Build. Mater.* **2012**, *36*, 1036–1043. [CrossRef]
17. Ouyang, X.; Wu, Z.; Shan, B.; Chen, Q.; Shi, C. A Critical Review on Compressive Behavior and Empirical Constitutive Models of Concrete. *Constr. Build. Mater.* **2022**, *323*, 126572. [CrossRef]
18. Ohemeng, E.A.; Ekolu, S.O.; Quainoo, H.; Kruger, D. Model for Predicting Compressive Strength and Elastic Modulus of Recycled Concrete Made with Treated Coarse Aggregate: Empirical Approach. *Constr. Build. Mater.* **2022**, *320*, 126240. [CrossRef]
19. Ben Chaabene, W.; Flah, M.; Nehdi, M.L. Machine Learning Prediction of Mechanical Properties of Concrete: Critical Review. *Constr. Build. Mater.* **2020**, *260*, 119889. [CrossRef]
20. Li, Z.; Yoon, J.; Zhang, R.; Rajabipour, F.; Srubar III, W.V.; Dabo, I.; Radlińska, A. Machine Learning in Concrete Science: Applications, Challenges, and Best Practices. *NPJ Comput. Mater.* **2022**, *8*, 127. [CrossRef]
21. Mohtasham Moein, M.; Saradar, A.; Rahmati, K.; Ghasemzadeh Mousavinejad, S.H.; Bristow, J.; Aramali, V.; Karakouzian, M. Predictive Models for Concrete Properties Using Machine Learning and Deep Learning Approaches: A Review. *J. Build. Eng.* **2023**, *63*, 105444. [CrossRef]
22. El Khessaimi, Y.; El Hafiane, Y.; Smith, A.; Peyratout, C.; Tamine, K.; Adly, S.; Barkatou, M. Machine Learning-Based Prediction of Compressive Strength for Limestone Calcined Clay Cements. *J. Build. Eng.* **2023**, *76*, 107062. [CrossRef]
23. Sui, H.; Wang, W.; Lin, J.; Tang, Z.Q.; Yang, D.-S.; Duan, W. Spatial Correlation and Pore Morphology Analysis of Limestone Calcined Clay Cement (LC3) via Machine Learning and Image-Based Characterisation. *Constr. Build. Mater.* **2023**, *401*, 132721. [CrossRef]
24. Canbek, O.; Xu, Q.; Mei, Y.; Washburn, N.R.; Kurtis, K.E. Predicting the Rheology of Limestone Calcined Clay Cements (LC3): Linking Composition and Hydration Kinetics to Yield Stress through Machine Learning. *Cem. Concr. Res.* **2022**, *160*, 106925. [CrossRef]
25. Li, Y.; Li, H.; Jin, C.; Shen, J. The Study of Effect of Carbon Nanotubes on the Compressive Strength of Cement-Based Materials Based on Machine Learning. *Constr. Build. Mater.* **2022**, *358*, 129435. [CrossRef]
26. Lyngdoh, G.A.; Zaki, M.; Krishnan, N.M.A.; Das, S. Prediction of Concrete Strengths Enabled by Missing Data Imputation and Interpretable Machine Learning. *Cem. Concr. Compos.* **2022**, *128*, 104414. [CrossRef]
27. Cakiroglu, C.; Aydın, Y.; Bekdaş, G.; Geem, Z.W. Interpretable Predictive Modelling of Basalt Fiber Reinforced Concrete Splitting Tensile Strength Using Ensemble Machine Learning Methods and SHAP Approach. *Materials* **2023**, *16*, 4578. [CrossRef]
28. Silva, V.P.; de Carvalho, R.A.; da Rêgo, J.H.S.; Evangelista, F. Machine Learning-Based Prediction of the Compressive Strength of Brazilian Concretes: A Dual-Dataset Study. *Materials* **2023**, *16*, 4977. [CrossRef] [PubMed]
29. Jiang, Y.; Li, H.; Zhou, Y. Compressive Strength Prediction of Fly Ash Concrete Using Machine Learning Techniques. *Buildings* **2022**, *12*, 690. [CrossRef]
30. Dhandapani, Y.; Santhanam, M. Assessment of Pore Structure Evolution in the Limestone Calcined Clay Cementitious System and Its Implications for Performance. *Cem. Concr. Compos.* **2017**, *84*, 36–47. [CrossRef]
31. Yu, T.; Zhang, B.; Guo, H.; Wang, Q.; Liu, D.; Chen, J.; Yuan, P. Calcined Nanosized Tubular Halloysite for the Preparation of Limestone Calcined Clay Cement (LC3). *Appl. Clay Sci.* **2023**, *232*, 106795. [CrossRef]
32. Msinjili, N.S.; Gluth, G.J.G.; Sturm, P.; Vogler, N.; Kühne, H.-C. Comparison of Calcined Illitic Clays (Brick Clays) and Low-Grade Kaolinitic Clays as Supplementary Cementitious Materials. *Mater. Struct.* **2019**, *52*, 94. [CrossRef]
33. Lin, R.-S.; Lee, H.-S.; Han, Y.; Wang, X.-Y. Experimental Studies on Hydration–Strength–Durability of Limestone-Cement-Calcined Hwangtoh Clay Ternary Composite. *Constr. Build. Mater.* **2021**, *269*, 121290. [CrossRef]

34. Krishnan, S.; Kanaujia, S.K.; Mithia, S.; Bishnoi, S. Hydration Kinetics and Mechanisms of Carbonates from Stone Wastes in Ternary Blends with Calcined Clay. *Constr. Build. Mater.* **2018**, *164*, 265–274. [CrossRef]
35. Kafodya, I.; Basuroy, D.; Marangu, J.M.; Kululanga, G.; Maddalena, R.; Novelli, V.I. Mechanical Performance and Physico-Chemical Properties of Limestone Calcined Clay Cement (LC3) in Malawi. *Buildings* **2023**, *13*, 740. [CrossRef]
36. Yu, T.; Zhang, B.; Yuan, P.; Guo, H.; Liu, D.; Chen, J.; Liu, H.; Setti Belaroui, L. Optimization of Mechanical Performance of Limestone Calcined Clay Cement: Effects of Calcination Temperature of Nanosized Tubular Halloysite, Gypsum Content, and Water/Binder Ratio. *Constr. Build. Mater.* **2023**, *389*, 131709. [CrossRef]
37. Hay, R.; Celik, K. Performance Enhancement and Characterization of Limestone Calcined Clay Cement (LC3) Produced with Low-Reactivity Kaolinitic Clay. *Constr. Build. Mater.* **2023**, *392*, 131831. [CrossRef]
38. Shoukry, H.; Perumal, P.; Abadel, A.; Alghamdi, H.; Alamri, M.; Abdel-Gawwad, H.A. Performance of Limestone-Calcined Clay Cement Mortar Incorporating High Volume Ferrochrome Waste Slag Aggregate. *Constr. Build. Mater.* **2022**, *350*, 128928. [CrossRef]
39. Dixit, A.; Du, H.; Pang, S.D. Performance of Mortar Incorporating Calcined Marine Clays with Varying Kaolinite Content. *J. Clean. Prod.* **2021**, *282*, 124513. [CrossRef]
40. Aramburo, C.H.; Pedrajas, C.; Talero, R. Portland Cements with High Content of Calcined Clay: Mechanical Strength Behaviour and Sulfate Durability. *Materials* **2020**, *13*, 4206. [CrossRef]
41. Machner, A.; Zajac, M.; Ben Haha, M.; Kjellsen, K.O.; Geiker, M.R.; De Weerdt, K. Portland Metakaolin Cement Containing Dolomite or Limestone–Similarities and Differences in Phase Assemblage and Compressive Strength. *Constr. Build. Mater.* **2017**, *157*, 214–225. [CrossRef]
42. Alujas, A.; Fernández, R.; Quintana, R.; Scrivener, K.L.; Martirena, F. Pozzolanic Reactivity of Low Grade Kaolinitic Clays: Influence of Calcination Temperature and Impact of Calcination Products on OPC Hydration. *Appl. Clay Sci.* **2015**, *108*, 94–101. [CrossRef]
43. Lin, R.-S.; Oh, S.; Du, W.; Wang, X.-Y. Strengthening the Performance of Limestone-Calcined Clay Cement (LC3) Using Nano Silica. *Constr. Build. Mater.* **2022**, *340*, 127723. [CrossRef]
44. Akindahunsi, A.A.; Avet, F.; Scrivener, K. The Influence of Some Calcined Clays from Nigeria as Clinker Substitute in Cementitious Systems. *Case Stud. Constr. Mater.* **2020**, *13*, e00443. [CrossRef]
45. Fernandez, R.; Martirena, F.; Scrivener, K.L. The Origin of the Pozzolanic Activity of Calcined Clay Minerals: A Comparison between Kaolinite, Illite and Montmorillonite. *Cem. Concr. Res.* **2011**, *41*, 113–122. [CrossRef]
46. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *Winst.-Salem For.* **2001**, *23*.
49. Biau, G.; Devroye, L.; Lugosi, G. Consistency of Random Forests and Other Averaging Classifiers. *J. Mach. Learn. Res.* **2008**, *9*, 2015–2033.
50. Chen, X.; Ishwaran, H. Random Forests for Genomic Data Analysis. *Genomics* **2012**, *99*, 323–329. [CrossRef]
51. Cook, R.; Lapeyre, J.; Ma, H.; Kumar, A. Prediction of Compressive Strength of Concrete: A Critical Comparison of Performance of a Hybrid Machine Learning Model with Standalone Models. *ASCE J. Mater. Civ. Eng.* **2019**, *31*, 04019255. [CrossRef]
52. Schaffer, C. Selecting a Classification Method by Cross-Validation. *Mach. Learn.* **1993**, *13*, 135–143. [CrossRef]
53. Han, T.; Ponduru, S.A.; Cook, R.; Huang, J.; Sant, G.; Kumar, A. A Deep Learning Approach to Design and Discover Sustainable Cementitious Binders: Strategies to Learn From Small Databases and Develop Closed-Form Analytical Models. *Front. Mater.* **2022**, *8*, 796476. [CrossRef]
54. Archer, K.J.; Kimes, R.V. Empirical Characterization of Random Forest Variable Importance Measures. *Comput. Stat. Data Anal.* **2008**, *52*, 2249–2260. [CrossRef]
55. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
56. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable Selection Using Random Forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
57. Díaz-Uriarte, R.; Alvarez de Andrés, S. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinform.* **2006**, *7*, 3. [CrossRef]
58. Sedgwick, P. Pearson's Correlation Coefficient. *BMJ* **2012**, *345*, e4483. [CrossRef]
59. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
60. Mangalathu, S.; Hwang, S.-H.; Jeon, J.-S. Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive ExPlanations (SHAP) Approach. *Eng. Struct.* **2020**, *219*, 110927. [CrossRef]
61. Lundberg, S.M.; Erion, G.G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2019**, arXiv:1802.03888.
62. Yang, K.; Hu, Y.; Li, Z.; Krishnan, N.M.A.; Smedskjaer, M.M.; Hoover, C.G.; Mauro, J.C.; Sant, G.; Bauchy, M. Analytical Model of the Network Topology and Rigidity of Calcium Aluminosilicate Glasses. *J. Am. Ceram. Soc.* **2021**, *104*, 3947–3962. [CrossRef]
63. Mauro, J.C. Topological Constraint Theory of Glass. *Am. Ceram. Soc. Bull.* **2011**, *90*, 31–37.
64. Phillips, J.C. Topology of Covalent Non-Crystalline Solids I: Short-Range Order in Chalcogenide Alloys. *J. Non-Cryst. Solids* **1979**, *34*, 153–181. [CrossRef]

65. Han, T.; Bhat, R.; Ponduru, S.A.; Sarkar, A.; Huang, J.; Sant, G.; Ma, H.; Neithalath, N.; Kumar, A. Deep Learning to Predict the Hydration and Performance of Fly Ash-Containing Cementitious Binders. *Cem. Concr. Res.* **2023**, *165*, 107093. [CrossRef]
66. Bhat, R.; Han, T.; Akshay Ponduru, S.; Reka, A.; Huang, J.; Sant, G.; Kumar, A. Predicting Compressive Strength of Alkali-Activated Systems Based on the Network Topology and Phase Assemblages Using Tree-Structure Computing Algorithms. *Constr. Build. Mater.* **2022**, *336*, 127557. [CrossRef]
67. Han, T.; Gomaa, E.; Gheni, A.; Huang, J.; ElGawady, M.; Kumar, A. Machine Learning Enabled Closed-Form Models to Predict Strength of Alkali-Activated Systems. *J. Am. Ceram. Soc.* **2022**, *105*, 4414–4425. [CrossRef]
68. Bauchy, M. Deciphering the Atomic Genome of Glasses by Topological Constraint Theory and Molecular Dynamics: A Review. *Comput. Mater. Sci.* **2019**, *159*, 95–102. [CrossRef]
69. Oey, T.; Kumar, A.; Pignatelli, I.; Yu, Y.; Neithalath, N.; Bullard, J.W.; Bauchy, M.; Sant, G. Topological Controls on the Dissolution Kinetics of Glassy Aluminosilicates. *J. Am. Ceram. Soc.* **2017**, *100*, 5521–5527. [CrossRef]
70. Bauchy, M.; Abdolhosseini Qomi, M.J.; Bichara, C.; Ulm, F.-J.; Pellenq, R.J.-M. Nanoscale Structure of Cement: Viewpoint of Rigidity Theory. *J. Phys. Chem. C* **2014**, *118*, 12485–12493. [CrossRef]
71. Bauchy, M.; Micoulaut, M. Atomic Scale Foundation of Temperature-Dependent Bonding Constraints in Network Glasses and Liquids. *J. Non-Cryst. Solids* **2011**, *357*, 2530–2537. [CrossRef]
72. Oey, T.; Frederiksen, K.F.; Mascaraque, N.; Youngman, R.; Balonis, M.; Smedskjaer, M.M.; Bauchy, M.; Sant, G. The Role of the Network-Modifier's Field-Strength in the Chemical Durability of Aluminoborate Glasses. *J. Non-Cryst. Solids* **2019**, *505*, 279–285. [CrossRef]
73. Nawy, E. (Ed.) *Concrete Construction Engineering Handbook*; CRC Press: Boca Raton, FL, USA, 2008; ISBN 978-0-8493-7492-0.
74. Ponduru, S.A.; Han, T.; Huang, J.; Kumar, A. Predicting Compressive Strength and Hydration Products of Calcium Aluminate Cement Using Data-Driven Approach. *Materials* **2023**, *16*, 654. [CrossRef]
75. Gomaa, E.; Han, T.; ElGawady, M.; Huang, J.; Kumar, A. Machine Learning to Predict Properties of Fresh and Hardened Alkali-Activated Concrete. *Cem. Concr. Compos.* **2021**, *115*, 103863. [CrossRef]
76. Canbek, O.; Washburn, N.R.; Kurtis, K.E. Relating LC3 Microstructure, Surface Resistivity and Compressive Strength Development. *Cem. Concr. Res.* **2022**, *160*, 106920. [CrossRef]
77. Marangu, J.M. Physico-Chemical Properties of Kenyan Made Calcined Clay -Limestone Cement (LC3). *Case Stud. Constr. Mater.* **2020**, *12*, e00333. [CrossRef]
78. Tang, F.J.; Gartner, E.M. Influence of Sulphate Source on Portland Cement Hydration. *Adv. Cem. Res.* **1988**, *1*, 67–74. [CrossRef]
79. Frigione, G. *Gypsum in Cement. In Advances in Cement Technology*; Ghosh, S.N., Ed.; Pergamon Press Ltd.: Oxford, UK, 1983; pp. 485–535. ISBN 978-0-08-028670-9.
80. Irassar, E.F.; Violini, D.; Rahhal, V.F.; Milanesi, C.; Trezza, M.A.; Bonavetti, V.L. Influence of Limestone Content, Gypsum Content and Fineness on Early Age Properties of Portland Limestone Cement Produced by Inter-Grinding. *Cem. Concr. Compos.* **2011**, *33*, 192–200. [CrossRef]
81. Han, T.; Ponduru, S.A.; Reka, A.; Huang, J.; Sant, G.; Kumar, A. Predicting Dissolution Kinetics of Tricalcium Silicate Using Deep Learning and Analytical Models. *Algorithms* **2023**, *16*, 7. [CrossRef]
82. Han, T.; Huang, J.; Sant, G.; Neithalath, N.; Kumar, A. Predicting Mechanical Properties of Ultrahigh Temperature Ceramics Using Machine Learning. *J. Am. Ceram. Soc.* **2022**, *105*, 6851–6863. [CrossRef]
83. Xu, X.; Han, T.; Huang, J.; Kruger, A.A.; Kumar, A.; Goel, A. Machine Learning Enabled Models to Predict Sulfur Solubility in Nuclear Waste Glasses. *ACS Appl. Mater. Interfaces* **2021**, *13*, 53375–53387. [CrossRef]
84. Scrivener, K.; Avet, F.; Maraghechi, H.; Zunino, F.; Ston, J.; Hanpongpun, W.; Favier, A. Impacting Factors and Properties of Limestone Calcined Clay Cements (LC3). *Green Mater.* **2019**, *7*, 3–14. [CrossRef]
85. Silva, A.S.; Gameiro, A.; Grilo, J.; Veiga, R.; Velosa, A. Long-Term Behavior of Lime–Metakaolin Pastes at Ambient Temperature and Humid Curing Condition. *Appl. Clay Sci.* **2014**, *88–89*, 49–55. [CrossRef]
86. Tironi, A.; Trezza, M.A.; Scian, A.N.; Irassar, E.F. Assessment of Pozzolanic Activity of Different Calcined Clays. *Cem. Concr. Compos.* **2013**, *37*, 319–327. [CrossRef]