Construction and Building Materials 421 (2024) 135663

Contents lists available at ScienceDirect

Construction and Building Materials

journal homepage: www.elsevier.com/locate/conbuildmat

Explainable ensemble learning predictive model for thermal conductivity of cement-based foam

Celal Cakiroglu [a], Farnaz Batool [b], Kamrul Islam [c], Moncef L. Nehdi [d,*]

[a] Department of Civil Engineering, Turkish-German University, Istanbul, Turkey
[b] Department of Civil Engineering, NED University of Engineering & Technology, Pakistan
[c] Department of Civil, Geological and Mining Engineering, Polytechnique Montreal, Montreal, QC, Canada
[d] Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4L8, Canada

ARTICLE INFO

Keywords:
Cement-based foam
Thermal conductivity
Model prediction
Machine learning
Explainable
Ensemble Learning

ABSTRACT

Cement-based foam has emerged as a strong contender in sustainable construction owing to its superior thermal and sound insulation properties, fire resistance, and cost-effectiveness. To effectively use cement-based foam as a thermal insulation material, it is important to accurately predict its thermal conductivity. The current study aims at coining an accurate methodology for predicting the thermal conductivity of cement-based foam using state-of-the-art machine learning techniques. A comprehensive experimental dataset of 504 data points was developed and used for training ensemble learning models including XGBoost, CatBoost, LightGBM and Random Forest. The independent variables of this dataset affecting the thermal conductivity are the cast density, percentage of pozzolan, porosity, percentage of moisture, and duration of hydration in days. Using the Isolation Forest algorithm proved effective in detecting and eliminating outliers in the dataset. All the ensemble learning techniques explored in this study achieved superior predictive accuracy with a coefficient of determination greater than 0.98 on the test dataset. The influence of the input features on the thermal conductivity was visualized using the SHapley Additive exPlanations (SHAP) approach and individual conditional expectation (ICE) plots. The cast density had the greatest effect on thermal conductivity. The explainable machine learning models demonstrated superior accuracy, efficiency, and reliability in estimating the thermal insulation of cement-based foam, opening the door for wider acceptance of this material in sustainable energy efficient construction.

1. Introduction

Population growth, urbanization, and the associated colossal upsurge in energy consumption have increased the demand for sustainable and energy−efficient construction materials throughout the world. Hence, cement and energy production experienced a considerable rise. Cement production has contributed to the global increase in carbon dioxide emissions and is responsible for 7% of total CO_2 generated in the world [29]. Therefore, it has become an absolute necessity to reduce the carbon footprint of cement production.

Generally, about 50% of total generated electricity is consumed by households, which makes the availability of energy for the other sectors of the economy more challenging. This has led the construction industry to explore various insulating materials with reduced embodied carbon and less energy consumption. Recently, cement-based foam has emerged as a prospective building material owing to its ease of fabrication, good insulation, low density, and potential of reducing CO_2 emissions by utilizing agriculture/industrial byproducts as pozzolanic ingredients [6]. Cement−based foam is a porous and lightweight building material in which preformed foam is deliberately added to a base mixture made up of portland cement, water, and aggregates to create an air void network [5].

The cast density of cement−based foam ranges between 300 and 1600 kg/m^3; the lighter densities (<800 kg/m^3) are attractive for void-fill and insulation applications, while the heavier mixtures are used for structural purposes [5]. Other cement-based foam applications include producing lightweight blocks and pre-cast panels, fire insulation, thermal and acoustic insulation, road subbase, trench reinstatement, soil stabilization, and shock−absorbing barriers for airports and regular traffic [2]. Also, due to flowability features, cement−based foam is a superlative material for filling voids such as old sewers, storage tanks, basements, ducts, and voids under roadways caused by cliff of heavy

* Corresponding author.
E-mail address: nehdim@mcmaster.ca (M.L. Nehdi).

https://doi.org/10.1016/j.conbuildmat.2024.135663
Received 22 December 2023; Received in revised form 19 February 2024; Accepted 28 February 2024
Available online 4 March 2024
0950-0618/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

rains. In addition, foamed concrete applications are cost−effective at repairing and rehabilitating structures [2].

Jia et al. [14] conducted experiments to design an optimal composition of fluorine-free foam concrete with different amounts of nano-silica. Their results showed that nano-silica can reduce fluorine-free foam concrete's dry density and thermal conductivity, increase its impermeability and compressive strength, and improve the performance of fluorine-free foam concrete, making it a promising inorganic insulation material for building exterior walls. Li et al. [16] used hydroxy-terminated polydimethylsiloxane and tetraethyl orthosilicate to modify magnesium oxychloride cement and make it water-resistant and superhydrophobic. Hydrogen peroxide (H_2O_2) was used as a foaming agent to create pores in the magnesium oxychloride cement, reducing thermal conductivity and improving thermal insulation performance.

Jin et al. [15] investigated the effects of different volume fractions of basalt fiber and coir fiber, and different mixture ratios of cement, sand, and water on the physical, mechanical, and durability properties of hybrid natural fiber-reinforced roadbed foamed concrete. Using scanning electron microscopy and X-ray diffraction, the microstructure of hybrid natural fiber-reinforced roadbed foamed concrete was analyzed. The optimal fiber content and mixture ratios were deduced to be 0.3% and 1:2, respectively. Incorporating hybrid natural fibers was shown to be effective at improving the performance of foamed concrete for roadbed applications. Wu et al. [40] reported the results of flexible wall permeameter unconfined compressive strength, mercury intrusion porosimetry, and scanning electron microscopy tests on foamed cement-stabilized marine clay samples with different cement ratios, densities, and metakaolin additives. It was found that foamed cement-stabilized marine clay had significantly higher hydraulic conductivity than that of ordinary cement-stabilized clays, while maintaining adequate strength. The macro pore volume mainly influenced the hydraulic conductivity, while the strength was enhanced by the cementation of cement and metakaolin. It was concluded that foamed cement-stabilized marine clay can be a viable material for permeable cement-stabilized columns that can improve the settlement uniformity and long-term safety of soft clay foundations.

Mikulica et al. [28] presented foam concrete as a better alternative to slag and slag concrete for the rehabilitation of floor structures. The properties, production, and testing of various mixtures of foam concrete were examined. Zhao et al. [42] proposed a new material made of foamed cement, expandable polystyrene particles and polypropylene fiber as a sacrificial cladding for tunnel lining structures to provide blast mitigation effects. The impact resistance, compressive strength, and energy absorption performance of the foamed cement-base material with different ingredient proportions were investigated. Laboratory experiments and finite element analysis showed that the newly proposed foamed cement-base material was a viable option for the sacrificial cladding of tunnel lining structures for blast mitigation.

Among other building materials, foamed concrete has excellent thermal insulating properties. A reduction of 35% in thermal conductivity by reducing the density from 800 to 600 kg/m^3 and a further drop of 6% with the addition of pozzolans like fly ash, silica fume was reported [6]. Compressive strength of 0.6 MPa − 43 MPa for the density range of 300 − 1800 kg/m^3 at 28 days was achieved. Tan et al. [37] investigated the effects of aerogel and hydrogen peroxide on the density, compressive strength, thermal conductivity, sound absorption, porosity, and water absorption of foamed lightweight cementitious composites. It was reported that the optimum foamed lightweight cementitious composite mixture had 1% aerogel and 3% hydrogen peroxide, which resulted in low density, adequate strength, high porosity, and improved thermal and acoustic insulation properties. Raj et al. [34] studied foam concrete as a thermal insulation material for buildings. Cement, fly ash, and rice straw were used to produce foam concrete cladding with different foam volumes and fiber contents. Compressive strength and thermal conductivity tests showed 1% rice straw and 20% foam volume configuration as the optimum mixture.

Li et al. [17] investigated the thermal performance and economy of phase change material foamed cement walls for buildings in five different climate zones in China. It was found that phase change material foamed cement can moderate the indoor and outdoor temperature differences in all climate zones, and the optimal phase transition temperature was close to the local annual average temperature, except for the severe cold zone. Meng et al. [27] examined the preparation of phase change material foamed cement with different foaming rates. The thermal conductivity of the material was tested under different temperatures. It was shown that phase change material foamed cement roof can reduce the roof's internal surface temperature and attain better heat gain than a benchmark ordinary roof.

Cement-based foam is a promising construction material owing to its low thermal conductivity in comparison to conventional materials, which allows significant energy savings. However, predicting the thermal conductivity of cement-based foam has not been duly investigated. Accordingly, there is lack of reliable and widely adopted prediction tools for the thermal conductivity of cement-based foam. The current study aims to address this research gap by deploying state-of-the-art machine learning techniques such as XGBoost, LightGBM, CatBoost and Random Forest. Given their high accuracy and data-driven approach, which enables generalization, machine learning models are increasingly applied in predicting the behavior and properties of diverse construction materials (Ben [10,26,32,33,35,36,25]). The results of this study can be implemented towards better estimating the thermal conductivity of cement-based foams, thereby increasing the applicability of this sustainable construction material.

The subsequent sections present the statistical distribution of the original dataset and the isolation forest procedure used for enhancing the dataset quality via outlier detection and removal. The dataset was obtained from an experimental program the details of which are explained in the following section. After a brief overview of the gradient-boosting methodology used in model training, the predictions of four different ensemble learning models are visualized. The performances of these ensemble learning models are quantified using common accuracy metrics such as the coefficient of determination. The model hyperparameters are optimized using Cost-related Frugal Hyperparameter Optimization (CFO), and Economical Hyperparameter Optimization with Blended Search Strategy (BlendSearch) algorithms. The influence of different input features on the model predictions are then investigated using SHAP and individual conditional expectation plots.

2. Database development

The present study develops highly accurate machine learning models for the prediction of the thermal conductivity of cement-based foam. These models are based on a dataset of 504 samples compiled from experiments. In these experiments, cement-based foam with three cast densities of 800 kg/m^3, 600$kg/m^{3,}$ and 400 kg/m^3 was tested. SEM images of these concrete samples are shown in Fig. 1. Maintaining these densities, 21 mixtures were prepared by replacing fly ash, silica fume, and metakaolin with cement at a ratio of 10% and 20%, respectively. These cement-based foam samples were prepared by adding stable foam with density of 40 kg/m^3 generated using a synthetic foaming agent into a cement slurry. The mixing of stable foam primarily creates air void in the cementitious matrix, making it porous and lightweight. In this process, an optimal water-to-cement ratio is required to maintain the quality of the cement-based foam samples. Therefore, based on trials, a water/binder ratio of 0.69 was used throughout to maintain uniformity. The cement-based foam mixtures were cast into 150 mm×75 mm cylinders and stored in a water curing room for 28 days. After the completion of the curing period, the samples were left at room temperature of 21 ± 2 ° C and an average relative humidity of 65 ± 5%.

The Hot Disk Thermal Constant Analyzer (Fig. 2), which operates on

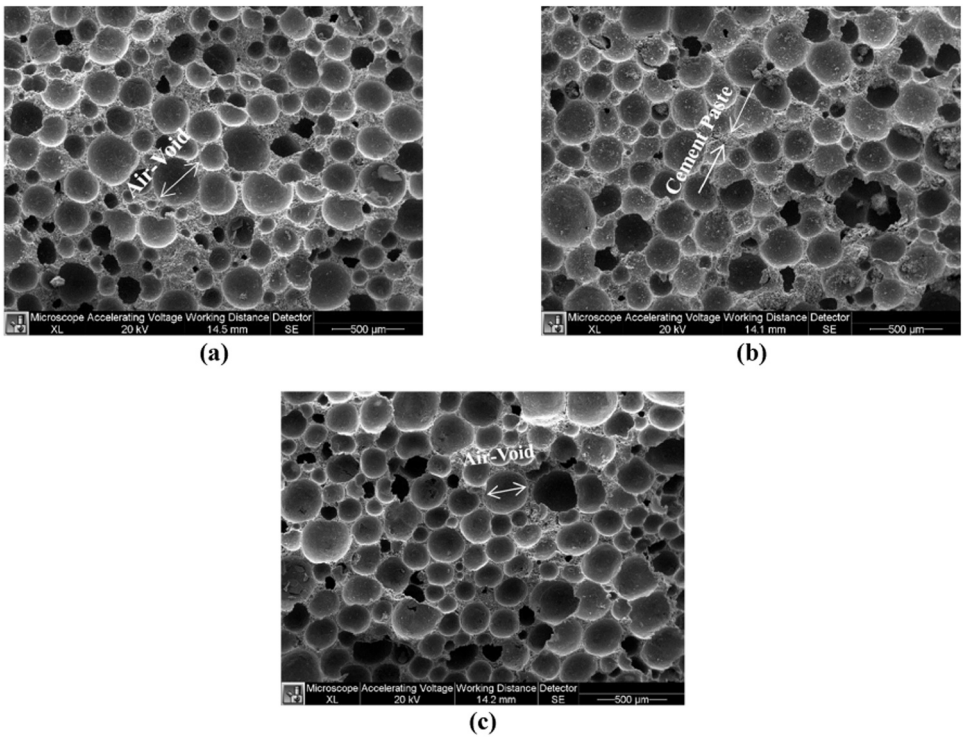

Fig. 1. Scanning electron microscopy (SEM) images of cement-based foam for the cast density of a) 800 kg/m^3, b) 600 kg/m^3, and c) 400 kg/m^3 [8].

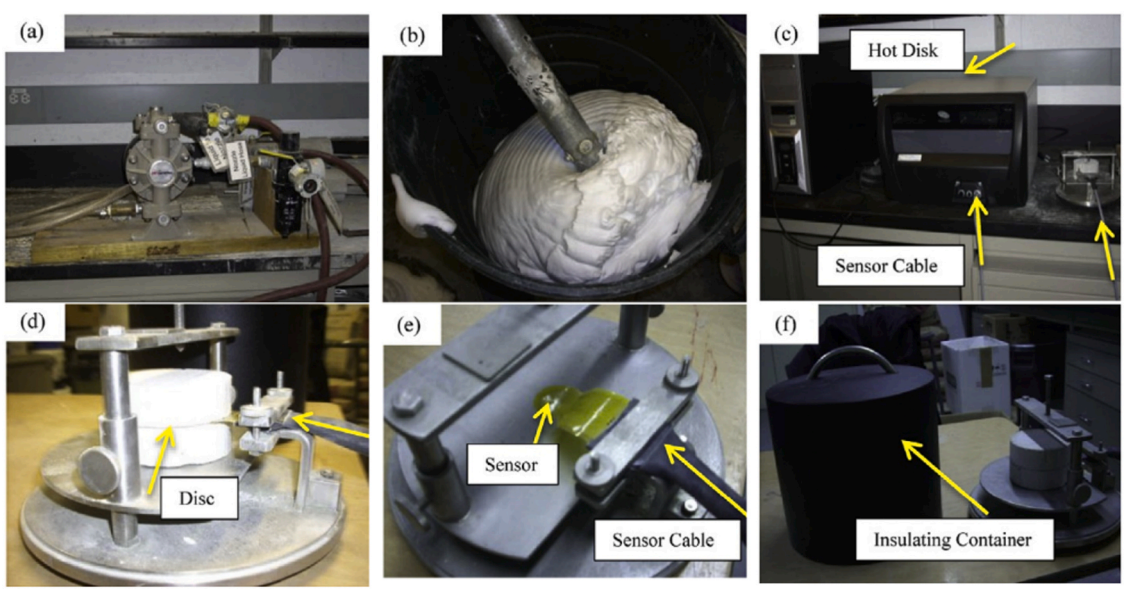

Fig. 2. a) Foam Generating Machine; b) Prepared Foam Ready to incorporate into Cementitious Slurry c) Thermal Plane Source Analyzer; d) Disc placed on the strand; e) Sensor; f) Insulating container [6].

the transient plane source technique, was used to measure the thermal conductivity of specimens from all these mixtures. The equipment needed in the application of this technique is depicted in Fig. 2. This analyzer consists of a hot disk sensor, which acts both as a heat source and resistance thermometer for increasing the temperature and recording the time-dependent temperature rise. Samples with 22 mm thickness and 75 mm diameter were used as they provide sufficient surface for placing the sensor. All specimens were examined at room temperature and curing ages of 60, 120, 210, and 300 days, respectively. The moisture content of specimens from all mixtures was measured through over-drying as per the guidance of ASTM C566 [4]. At the age of 300 days, X-ray tomography scanning was carried out and images were collected for all the mixture samples. Skyscan 1076 X-ray computed tomography machine, which consists of a high-performance micro-CT scanner with an X-ray source supply of 20–100 kV and X-ray detector of 4000 ×2672 pixels, was used. From the sample size of 35 mm x 35 mm and 14 mm height, a total of 1500 images were captured. These raw images were subsequently analyzed and quantified for porosity and air void distribution using Skyscan CT-analyzer software.

3. Statistics analysis of dataset

This section presents the statistical distribution of the features that constitute the dataset used in training the ML models. The correlation plot in Fig. 3 displays the distribution of each feature in its diagonal. These distributions show that the input features cast density and percentage of pozzolan are concentrated around three discrete values, whereas the number of days of hydration (curing age) is concentrated around four discrete values. On the other hand, the remaining input features, and the output feature (thermal conductivity) have a continuous distribution. The upper right triangular part of this plot contains the Pearson correlation coefficient between the features where the strength of correlation is shown with stars. The bivariate scatter plots of the features are shown in the lower left triangular part of Fig. 3, where the water-to-cement ratio was not included since this variable has a constant value of 0.69 in all samples. The magnitude range of each feature on a diagonal tile is shown in a horizontal and a vertical axis in Fig. 3.

It can be observed in Fig. 3 that the greatest correlation is between the cast density and thermal conductivity with a correlation coefficient of 0.92. A strong inverse correlation can be observed between the porosity and the cast density with a correlation coefficient of −0.91. Also, a relatively strong inverse correlation can be observed between the days of hydration and the moisture percentage. The Pearson correlation coefficient r_{xy} between two data sequences x and y is calculated as in Eq. (1) where n is the length of x and y.

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{1}$$

4. Outlier detection

Outlier detection is a significant part of a robust ML process, which aims to improve the quality of the dataset used to train predictive models. In this active research field, many novel techniques have been developed in recent years [3,9,13,30,39]. In this study the isolation forest and multivariate Gaussian distribution methods were applied for finding and removing the outliers in the dataset. The isolation forest algorithm differs from most statistical anomaly detection techniques in that it does not require the profiling of normal data points and the identification of samples that do not conform to the normal profiles as anomalies [18]. The approach adopted in the isolation forest algorithm is based on separating each data point using binary trees and using the tree length required to isolate each data point as an indicator of anomaly. The isolation of anomalous samples is possible since they are expected to be relatively small in number and to differ from the rest of the data set in their input feature values.

The process of data sample isolation is graphically illustrated in Fig. 4., which represents an isolation tree for a dataset with two input features denoted by x_1 and x_2. The isolation forest algorithm creates an ensemble of binary trees that isolate every data point. The binary trees in the ensemble differ from one another since each of them has a root node containing a randomly selected input feature and a random splitting point for that feature. As an example of this random selection, the tree in Fig. 4**b** starts with splitting the feature x_2. For each data point, the path length necessary to isolate the point is calculated and the average value of this path length in the entire ensemble of isolation trees is assigned to that data point as a measure of anomaly.

As shown in Fig. 4, the outliers tend to be closer to the root, while the inlier points tend to be closer to the deep end of the tree. The data points $\mathbf{X}_1$ and $\mathbf{X}_{10}$ that deviate from the dataset average are colored in red in Fig. 4**a**. It can be observed from Fig. 4**b** that these two points are also the closest points to the tree root with the least path lengths. These path lengths may be affected by the choice of the initial input feature and

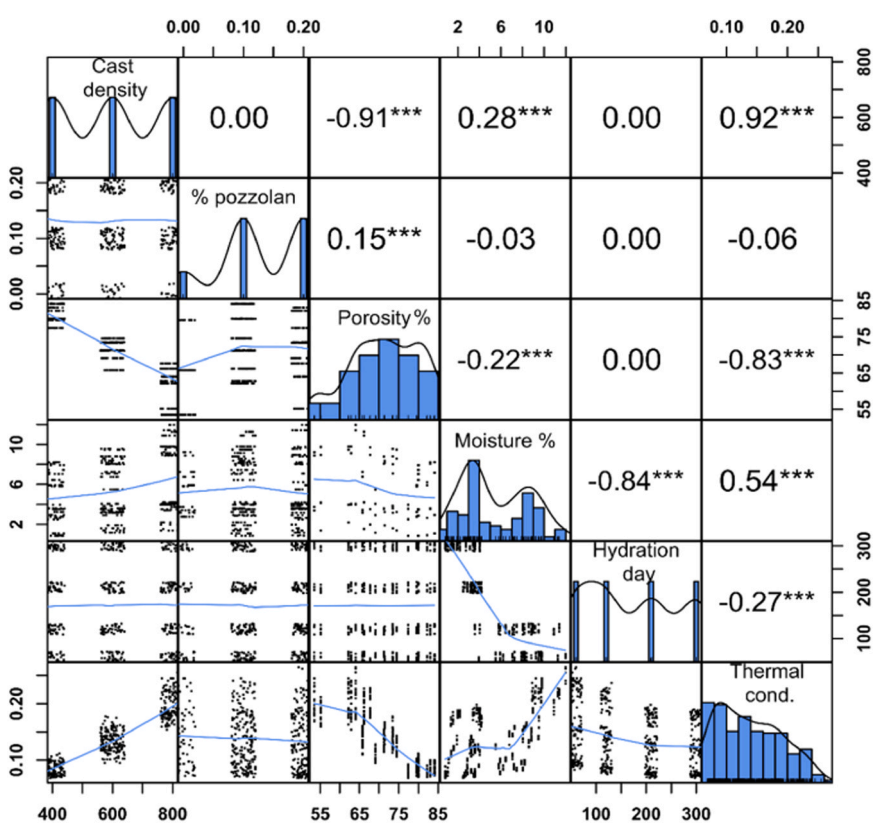

Fig. 3. Correlation plot of the model input features.

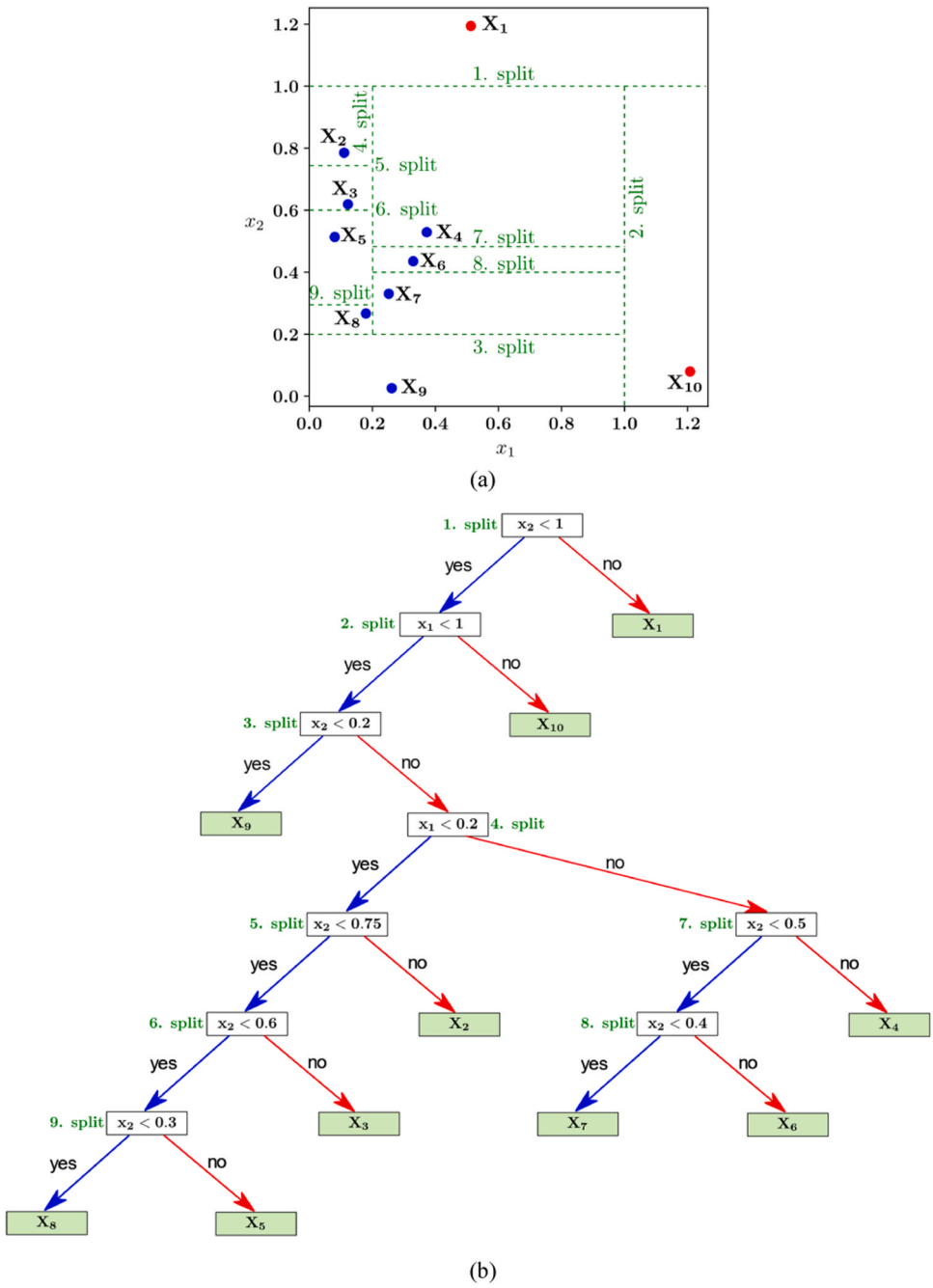

Fig. 4. a) Process of data sample isolation, and (b) example of isolation tree.

splitting points at each node in different trees of the ensemble. However, on average the outliers are expected to be closest to the root node. It should be noted that, as the number of trees in the ensemble increases, the average path length of each data point converges to a certain value [18,19].

5. Gradient boosting algorithms

Gradient boosting is a machine learning method that generates a prediction model in the form of a collection of weak prediction models, usually decision trees. Among the most widely used gradient boosting algorithms, Extreme Gradient Boosting (XGBoost), Random Forest, Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (CatBoost) are commonly deployed. The gradient boosting technique performs particularly well with large datasets in terms of prediction speed and accuracy. Gradient boosting starts with the training of an initial model using the training dataset. The prediction accuracy of this initial model is evaluated and the subsequent models in the ensemble of the decision trees improve the accuracy of the initial model. New corrective models are added to this ensemble until a certain level of accuracy, or the maximum number of decision trees has been reached [1]. The process of this incremental training of an ensemble model is illustrated in Fig. 5. In Fig. 5, the term $f_k(X)$ is the function that returns the output of the k-th decision tree, and the residuals of each decision tree are denoted with ε_k. Since each new tree that comes after the initial tree returns a correction value for the model consisting of the previous trees, the final model prediction can be expressed as the sum of all decision tree outputs.

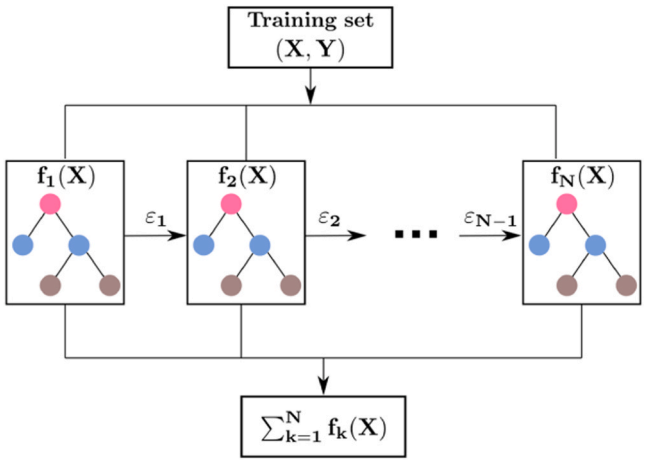

Fig. 5. Schematic illustration of gradient boosting model training.

6. Predictions of ML models

The models were trained on a dataset that initially consisted of 504 samples. This dataset has been further refined in terms of outliers and the final ML models have been trained after the removal of 48 outlier data points. The dataset was split into a training set (70%) and a test set (30%). The input features in this dataset are the cast density, water-to-cement ratio, percentage of pozzolan, porosity, moisture percentage, and hydration days, whereas the output feature is thermal conductivity. These input features were selected since they directly influence the thermal conductivity of cement-based foam, making it an effective insulation material. Cast density was selected as an input feature because it determines the amount of air voids in the foam, which are intentionally created to reduce weight and improve the insulating capacity. This is a direct consequence of air having low thermal conductivity. Porosity is another input feature that reflects the ratio of air voids to solid material in the foam. High porosity is an indicator of low density and low thermal conductivity. Percentage of pozzolan represents the amount of supplementary cementitious material that is added to the Portland cement, which can affect the thermal conductivity of the foam. Percentage of moisture is an input feature that indicates the water content, which can have an adverse effect on the thermal conductivity of the foam. Finally, the hydration days feature captures the time elapsed since the foam was mixed with water. It reflects the degree of hydration and evaporation of water in the foam, which influences the thermal conductivity. Fig. 6 displays the predictions of the XGBoost, CatBoost, Random Forest, and LightGBM models, showing plots of the predicted thermal conductivity values against the actual measured values. The scaling of the dataset was carried out prior to model training to bring all input features to the same order of magnitude, which improves model convergence. The scaling and outlier detection of the dataset as well as the subsequent predictive model training was carried out using the Scikit-learn library, which is available for the Python programming language. The predictions for the training set and test set are presented separately in different colors. The diagonal straight lines in Fig. 6 indicate a perfect match between the predicted and measured values, while the dotted lines indicate ±10% deviation from a perfect match.

Fig. 7 shows the third decision tree in the XGBoost model, which consists of 100 decision trees in total. The decision tree in Fig. 7 consists of a root node, 11 intermediate nodes, and 13 leaf nodes. Each intermediate node in Fig. 7 splits at a certain value of the input features f3

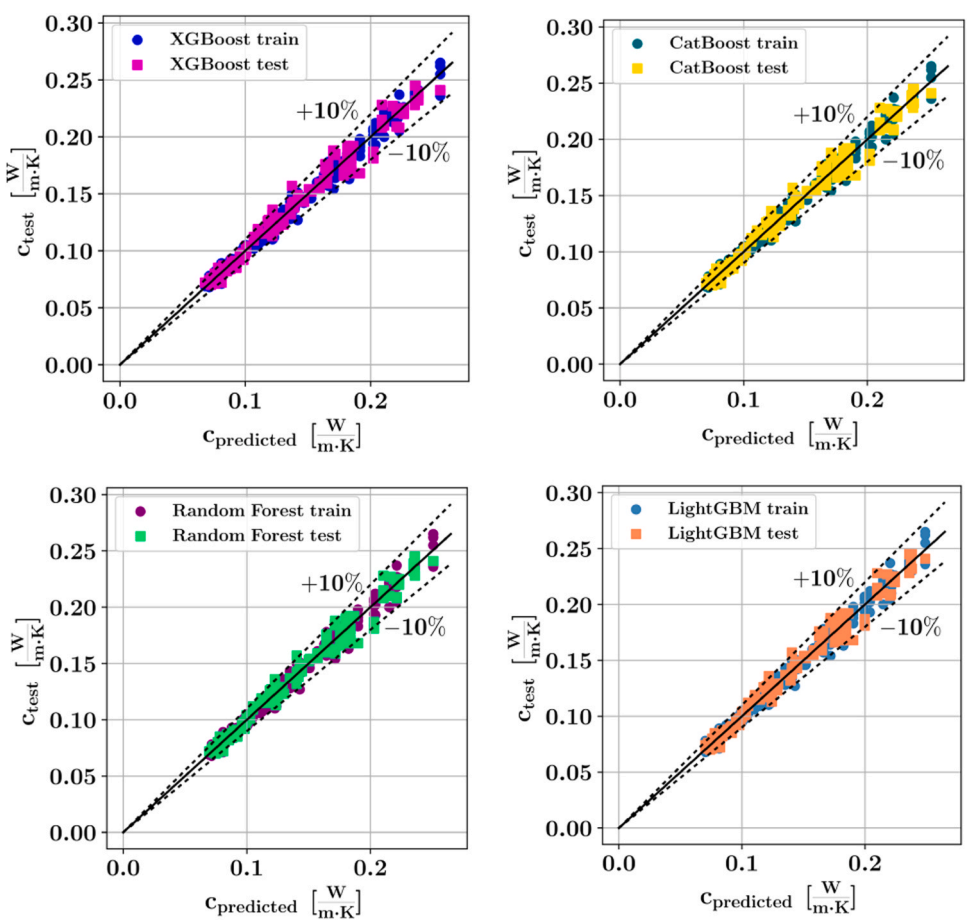

Fig. 6. Illustration of ML predictions of thermal conductivity of cement-based foam.

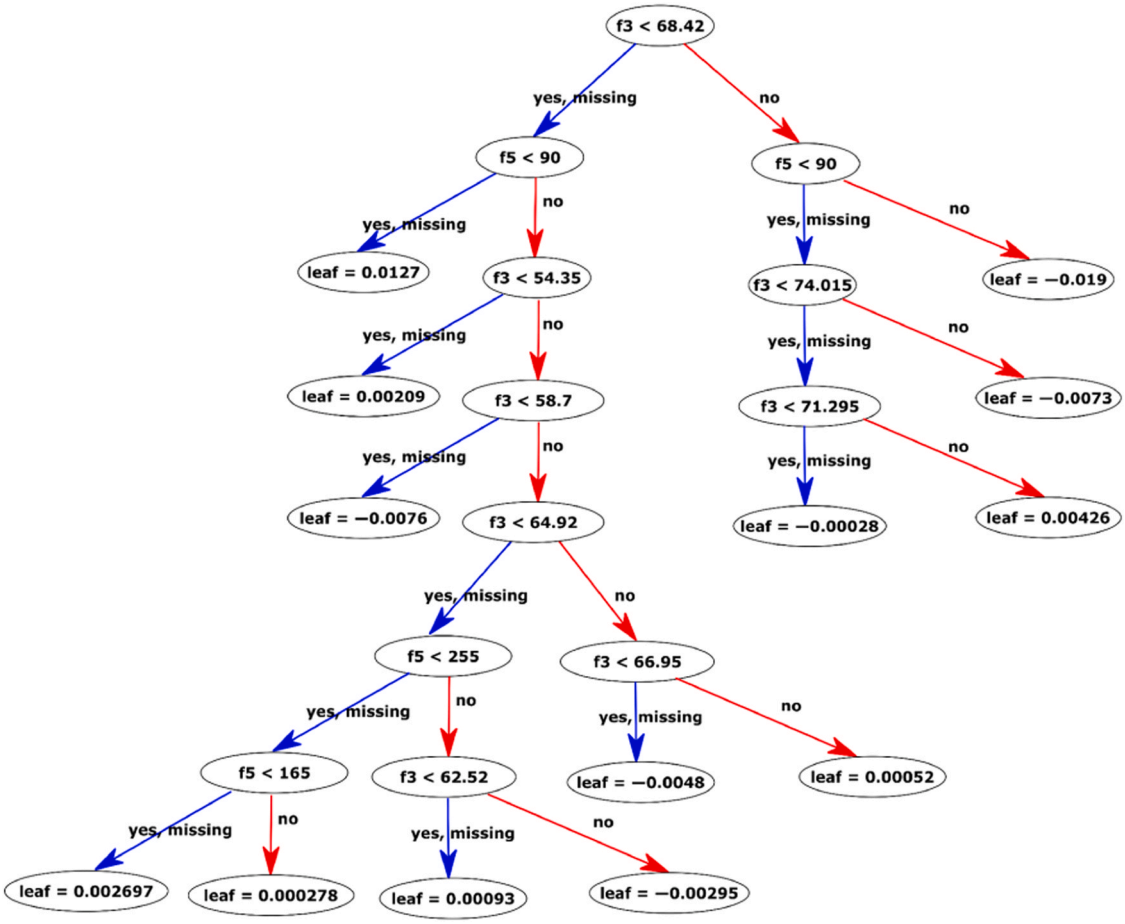

Fig. 7. Third decision tree in the XGBoost model.

and f5, which denote porosity and hydration days, respectively. The intermediate nodes branch to the left-hand side with a blue arrow if the condition inside the node is satisfied or the feature value is missing in the current sample, and they branch to the right-hand side with a red arrow otherwise. Depending on the values of f3 and f5, a leaf value is reached for any sample. These leaf values contain positive or negative corrections to the predicted thermal conductivity values in the first decision tree of the model. The sum of all decision tree leaf values constitutes the final prediction of the XGBoost model.

The effect of outlier identification and removal on the predictive performance of ML models was investigated in Fig. 8 for all the models. In Fig. 8 the variation of the model performance for each model in terms of the R^2 score on the test set has been plotted against the contamination variable of the isolation forest methodology. The isolation forest algorithm was applied using the scikit-learn library of the Python programming language. The contamination variable quantifies the threshold value for the classification of a data point as an outlier based on the average distance from the root node. An increased contamination value indicates a lower threshold value and a larger number of data samples being categorized as outliers. The ML models have been trained for 5 different contamination values in addition to the case without outlier detection and their performances have been measured on the test set. The test set consists of randomly selected samples and constitutes 30% of the entire dataset.

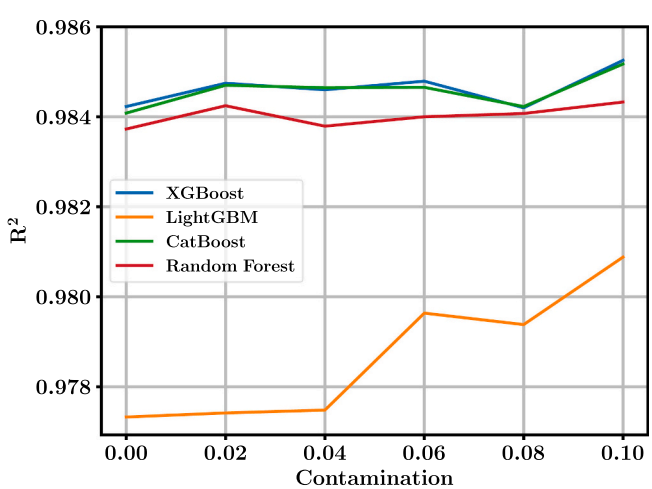

Fig. 8. Variation of the R^2 score with respect to contamination.

Fig. 8 shows that increasing the contamination value in general tends to improve the model accuracy. This tendency is more clearly visible in the LightGBM algorithm, whereas in the remaining algorithms, only a slight increase in the model accuracy can be observed as the contamination value increased. The curves in Fig. 8 were generated by taking the average R^2 scores on the test set corresponding to 10 different random states of the algorithms. Fig. 9 visualizes the outliers of the dataset corresponding to the contamination value of 0.1.

Fig. 10 displays a comparison of the predicted and actual target values, and the percentage errors in the predictions. The close overlap between the predictions and target values can be observed in Fig. 10**a** where a vertical dashed red line separates the training and test sets. Similarly, in Fig. 10**b** the percentage errors are shown separately for the training and test sets, where it can be observed that the error percentages remain within ±10% range for a large majority of the samples. Fig. 10**c and** Fig. 10**d** display the distribution of the error percentages of

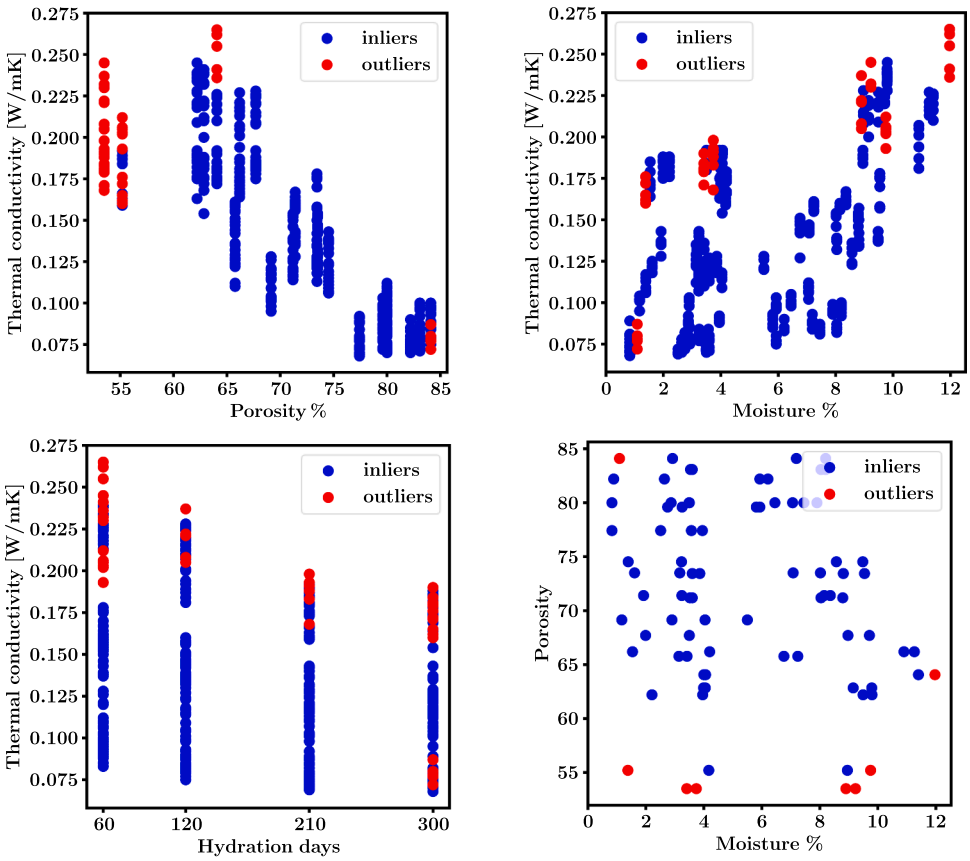

Fig. 9. Illustration outliers in the dataset.

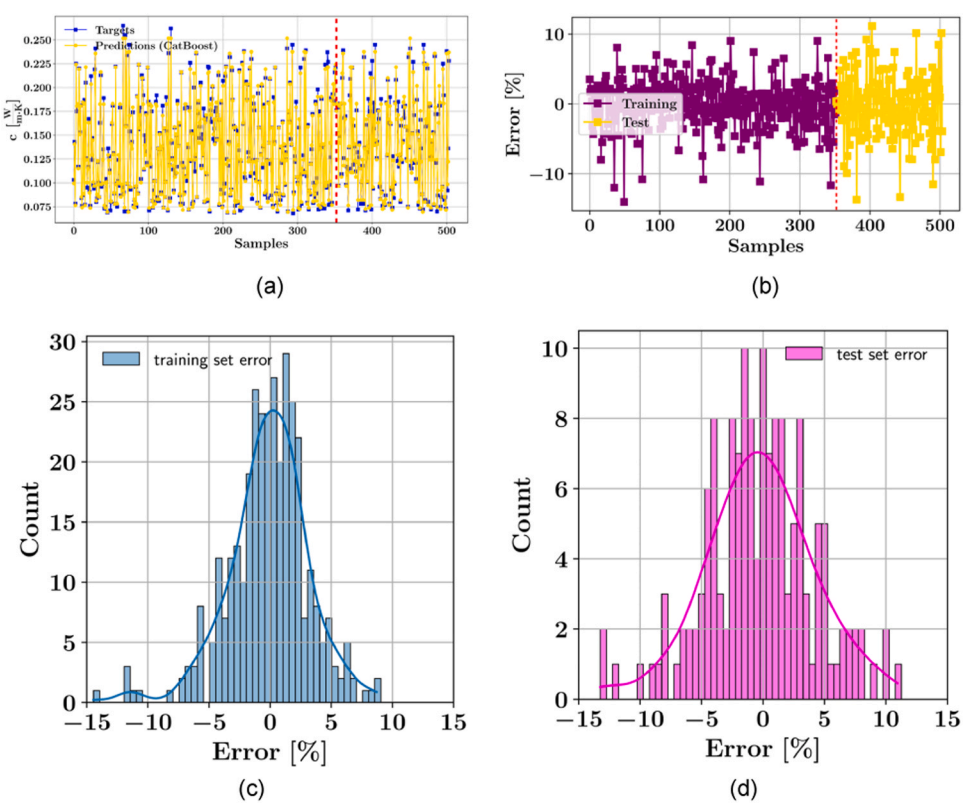

Fig. 10. Error distributions on the training and test datasets.

the XGBoost model on the training and test sets. The vertical axes in Fig. 10**c** and Fig. 10**d** show the number of samples for which the percentage error is in a certain interval and the error percentage intervals are shown on the horizontal axes. The length of each vertical bar in the histogram plots represents the number of samples in a 0.5% error interval. A comparison of Fig. 10**c** and Fig. 10**d** reveals that a greater number of samples are clustered in the ±5% error interval for the training set.

The performances of the ML models are quantified using error metrics and duration of training the ML models which also includes the 10-fold cross-validation process, and the results are listed in Table 1. According to Table 1, the XGBoost model performed best on the test set in terms of the R^2 score, whereas the LightGBM model performed best in terms of *RMSE* and *MAE*. It should be noted that all ML algorithms demonstrated high accuracy with R^2 scores greater than 0.98 on the test set and 0.99 on the training set. In terms of computational speed, the best performance was obtained from the XGBoost model while CatBoost performed worse than the remaining models. The accuracies of the ML models are also visualized in Fig. 11 in a radar chart.

The performance of the ML models has also been investigated in terms of the corresponding standard deviation and Pearson correlation coefficient with respect to the measured values. The results were visualized in a Taylor diagram as depicted in Fig. 12. The Pearson correlation is a measure of the linear correlation between data sequences. Fig. 12 shows that all ML predictions had Pearson correlation values greater than 0.99. A standard deviation of 0.05 W/(mK) can be observed for the original dataset and the predictions.

The hyperparameters of the ML models were optimized using Cost-related Frugal Hyperparameter Optimization (CFO) and Economical Hyperparameter Optimization with Blended Search Strategy (BlendSearch) algorithms, which are part of the FLAML package. The CFO algorithm is based on randomized direct search and applied in local search. On the other hand, BlendSearch is a global search technique and prevents the optimization process from being trapped in local minima. A list of the hyperparameters optimized with CFO and BlendSearch is given in Table 2. Fig. 13 shows the optimization steps of the LightGBM hyperparameters and the convergence of the R^2 score. The details of the CFO and BlendSearch algorithms can be found in [38,41].

In order to further validate the ML models, additional data points were collected from the literature [7]. In Table 3, "nan" stands for non-numerical or missing feature values. It can be observed that the moisture percentage information is missing in this additional dataset. As shown in Fig. 7, the LightGBM, CatBoost and XGBoost models were capable of dealing with missing feature values, whereas the Random Forest regressor is not designed for making predictions on data points with missing values. It should be noted that, missing feature values may adversely affect the ML model performance. The performances of the ML models on the dataset of Table 3 were listed in Table 4 and the predicted thermal conductivity values were plotted against the true target values in Fig. 14. It can be observed that the XGBoost model was able to predict the thermal conductivity with an R^2 score greater than 0.92. Although the XGBoost, CatBoost and LightGBM models were able the predict the thermal conductivity with relatively high accuracy, the overall model performances were adversely affected due to the presence of missing feature values.

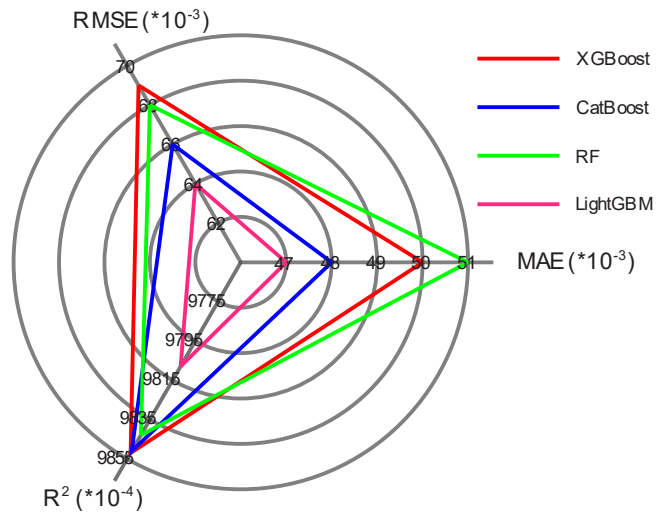

Fig. 11. Radar chart of the prediction accuracies.

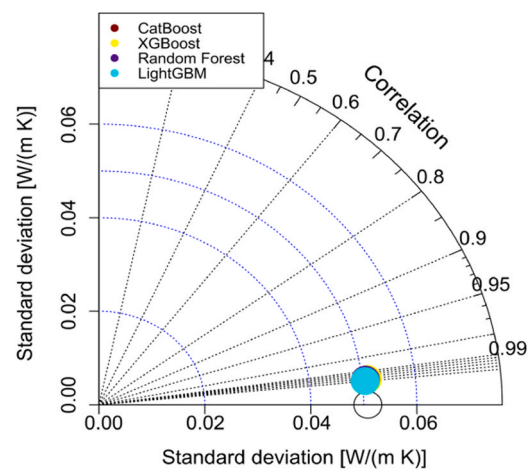

Fig. 12. Taylor diagram of the ML predictions.

7. SHAP analysis of ML model results

SHAP (SHapley Additive exPlanations) provides a methodology for visualizing the importance and impact of each input feature. The SHAP methodology assigns importance values to features in a machine learning model thereby quantifying the contribution of each feature to the model output and offering insights into the model's decision-making process. The calculation of the SHAP values is represented in Eq. (2), where the SHAP value of the feature with index i is denoted with ϕ_i. F is the set of all input features and S is a subset of the feature set which does not contain the feature i. The ML model output is represented with the function f, and the impact of a feature is determined by the difference of the model outputs when that feature is included in the feature vector x and withheld from that vector [20,21,23,24].

Table 1
Accuracy performances of the ML models.

	R^2		RMSE		MAE		Duration [s]
	Train	Test	Train	Test	Train	Test	
XGBoost	0.9914	0.9852	0.0047	0.0069	0.0033	0.0050	5.62
CatBoost	0.9921	0.9851	0.0045	0.0066	0.0032	0.0048	31.5
Random Forest	0.9918	0.9843	0.0046	0.0068	0.0033	0.0051	18.6
LightGBM	0.9917	0.9822	0.0046	0.0064	0.0033	0.0047	16.5

Table 2
Hyperparameters of the machine learning models.

Model	Optimum hyperparameter values	
Random Forest	max_features = 0.5632	max_leaf_nodes = 175
	n_estimators = 12	n_jobs = −1
XGBoost	colsample_bylevel = 1.0	colsample_bytree = 1.0
	learning_rate = 0.3555	enable_categorical = False
	max_leaves = 17	grow_policy = 'lossguide'
	importance_type = None	n_estimators = 59
	min_child_weight = 15.57	n_jobs = −1
LightGBM	n_estimators = 55	learning_rate = 0.2042
	colsample_bytree = 0.5568	min_child_samples = 4
	num_leaves = 11	reg_alpha = 0.00097
	reg_lambda = 0.08185	log_max_bin = 10
CatBoost	early_stopping_rounds = 11	learning_rate = 0.1346
	n_estimators = 80	

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (2)$$

A ranking of the input features with respect to their impacts on the model output is visualized in Fig. 15 where the average of the absolute SHAP values of a feature is used to quantify the importance of that feature. The average absolute SHAP values in Fig. 15 are obtained from the CatBoost model. According to Fig. 15 the cast density was the most impactful input feature, whereas the pozzolan percentage was the least impactful input feature.

The impacts of the different input features on the model predictions can also be visualized using SHAP summary plots. Fig. 16 shows the SHAP summary plot obtained from the CatBoost model, where each data point is represented with a dot and this representation is repeated for every input feature. The color of these dots corresponds to the numerical value of an input feature in a particular data point, whereas the horizontal distance of a point from the zero SHAP value line corresponds to the impact of the feature on the model output for that data point. The

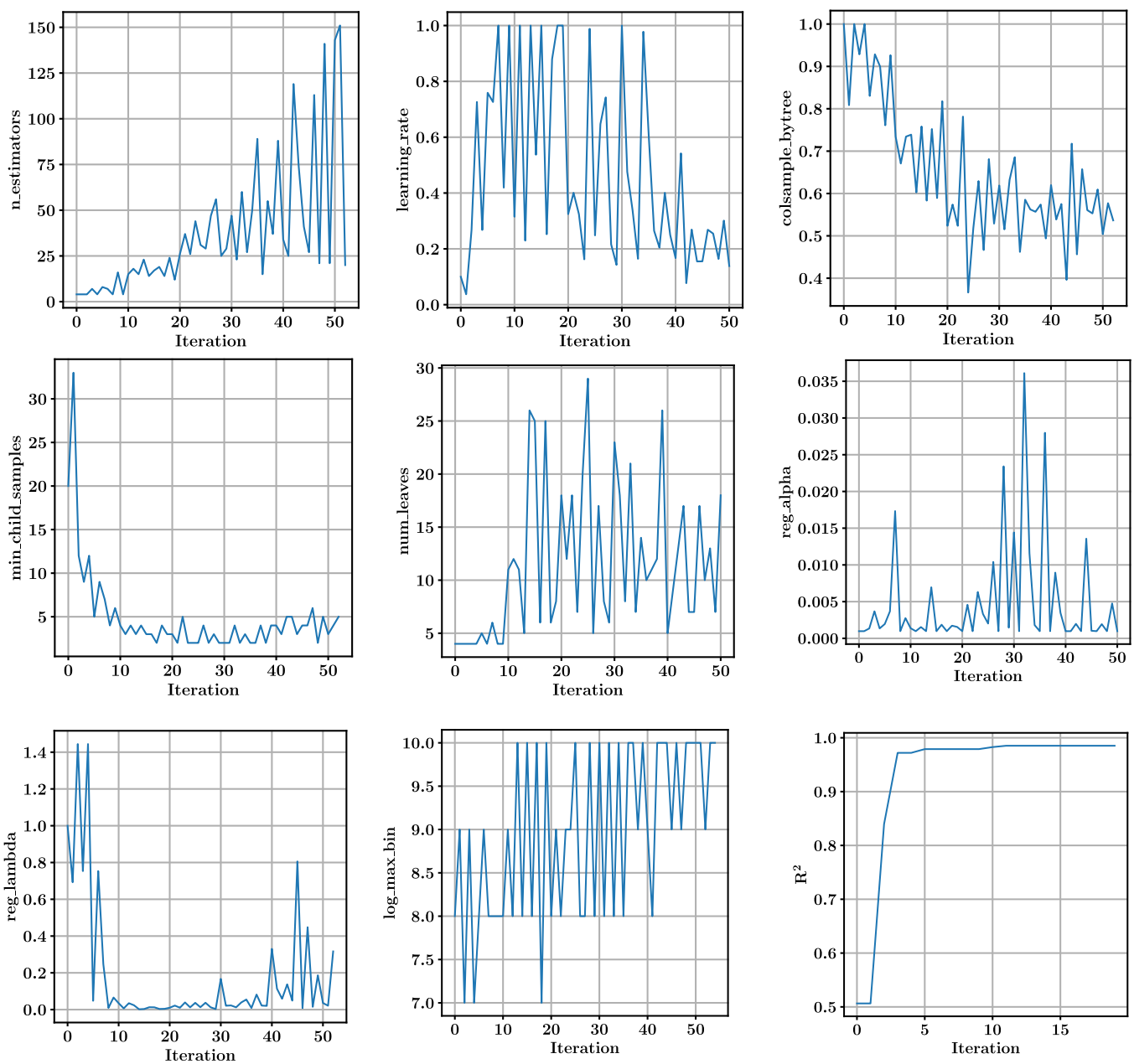

Fig. 13. Optimization steps of the hyperparameters (LightGBM).

Table 3
Additional data points for model validation [7].

Cast density (kg/m3)	Pozzolan (%)	Porosity (%)	Moisture (%)	Hydration days	Thermal conductivity (W/mK)
800	0	54	nan	35	0.234
800	0	54	nan	35	0.231
800	0	54	nan	35	0.263
600	0	66	nan	35	0.175
600	0	66	nan	35	0.191
600	0	66	nan	35	0.172
400	0	80	nan	35	0.104
400	0	80	nan	35	0.122
400	0	80	nan	35	0.13
800	0.2	68	nan	35	0.232
800	0.2	68	nan	35	0.243
800	0.2	68	nan	35	0.25
600	0.2	74	nan	35	0.17
600	0.2	75	nan	35	0.167
600	0.2	74	nan	35	0.17
400	0.2	82	nan	35	0.108
400	0.2	82	nan	35	0.108
400	0.2	80	nan	35	0.102
800	0.2	59	nan	35	0.24
800	0.2	59	nan	35	0.242
800	0.2	59	nan	35	0.249
600	0.2	67	nan	35	0.162
600	0.2	68	nan	35	0.154
600	0.2	68	nan	35	0.153
400	0.2	75	nan	35	0.104
400	0.2	75	nan	35	0.11
400	0.2	74	nan	35	0.105
800	0.2	68	nan	35	0.239
800	0.2	66	nan	35	0.238
800	0.2	68	nan	35	0.237
600	0.2	74	nan	35	0.179
600	0.2	73	nan	35	0.183
600	0.2	73	nan	35	0.19
400	0.2	82	nan	35	0.116
400	0.2	82	nan	35	0.117
400	0.2	82	nan	35	0.117

Table 4
ML model performances on the validation dataset.

Algorithm	RMSE	R^2	MAE
XGBoost	0.0146	0.9257	0.0120
LightGBM	0.0195	0.8683	0.0171
CatBoost	0.0212	0.8439	0.0166

shades of red and blue correspond to high and low values of an input feature as shown in the color bar. A horizontal position at the right-hand side of the zero line indicates an increasing effect of the input feature on the model prediction, whereas a position to the left-hand side indicates a decreasing effect on the model prediction. According to Fig. 16, the cast density feature has the greatest magnitude of SHAP values and the greatest impact on the model predictions.

In addition to the SHAP summary plots, the feature dependance plots give further insights into the relationships between the input features and their impacts on the predictive model output. Fig. 17 presents the feature dependence plots for each one of the input features. In these plots, in addition to the influence of an input feature on the model output, the color-coded values of another variable that is most dependent on that feature are displayed. The feature dependence plots represent each data point with a dot. The colors of these dots are based on the values of the most dependent variables. For instance, Fig. 17**a** shows that porosity is the variable most dependent on the cast density. Fig. 17**a** also shows that, for cast density values of 600$\mathrm{kg/m^3}$ and less, the cast density has a negative SHAP value and a decreasing effect on the model output.

Clearly, the cast density values are clustered at three different levels. Similarly, the percentage of pozzolan in the mixtures is clustered at the 0%, 10% and 20% levels. It can be observed in Fig. 17**b** that the percentage increase of the pozzolan had mostly a decreasing effect on the model output. Furthermore, the moisture percentage was the most dependent variable on the pozzolan percentage. However, there was no clear relationship between the increase of moisture and the impact of pozzolan percentage. At 20% pozzolan level this feature had mostly a decreasing effect on the model output, whereas removing the pozzolan from the mixture (0%) increased the model output.

Fig. 17**c** and Fig. 17**d** show the dependence plots of porosity and moisture percentage, respectively. It can be observed that increasing the moisture tended to increase the SHAP values. At moisture percentages less than 6% the SHAP values of the moisture percentage were less than zero, which indicates a decreasing effect on the model output. On the other hand, at moisture levels greater than 6%, the SHAP values increased with the moisture content. According to Fig. 17**c**, the effect of porosity on the model outputs was opposite to that of moisture. At porosity levels less than 70%, the SHAP values of porosity were greater than zero, whereas at greater than 70% porosity levels, the SHAP values were less than zero and tended to decrease with increasing porosity.

The ICE plots in Fig. 18 illustrate how thermal conductivity varied depending on the cast density, percentage of pozzolan, porosity, and moisture percentage. In these ICE plots, each data point was represented with a different curve [12]. For instance, in the ICE plot of cast density, for each data point the value of cast density varied between its upper and lower bounds, while every other input feature was kept constant. The horizontal axes of the ICE plots in Fig. 18 also display histogram distribution of the input features in light gray color. From these histograms it is visible that the cast density values were clustered around 400, 600, and 800 $\mathrm{kg/m^3}$, while the pozzolan percentages were clustered around 0%, 10%, and 20%. Due to the clustering of values around 3 distinct levels, the ICE plots of the cast density and the pozzolan percentage had stepwise variations.

The thermal conductivity as a function of cast density, denoted with $f(x)|$cast density, increased with respect to the cast density in all data samples. On the other hand, thermal conductivity as a function of the pozzolan percentage, denoted as $f(x)|$% of pozzolan, decreased with respect to the pozzolan percentage in all data samples. These variations of the thermal conductivity are also supported by the research output of Mydin [31], Chinnu et al. [11] and Maglad et al. [22]. In the ICE plots, the average prediction of all data points was plotted with a thick blue curve. A more irregular variation of the thermal conductivity can be observed with respect to porosity and moisture percentage in Figs. 18**c** and 18**d**. On average, a decrease with respect to porosity can be observed, whereas the moisture percentage had an increasing effect on the thermal conductivity.

8. Conclusions

Cement-based foam has several advantages compared to conventional concrete, including better thermal and sound insulation, fire resistance, eco-friendliness, and lighter weight. The present study harvests experimental research carried out for determining the thermal conductivity of cement-based foam as a function of its cast density, porosity, pozzolan percentage, moisture content, and hydration days. The results of these experiments were used for developing predictive machine learning models of thermal conductivity. This paper presents the application of the isolation forest methodology in data outlier identification and shows that this experimental data further processing allowed to attain superior model predictive accuracy. State-of-the-art ensemble machine learning algorithms were deployed to predict the thermal conductivity of foamed concrete. The main conclusions of this study are drawn below.

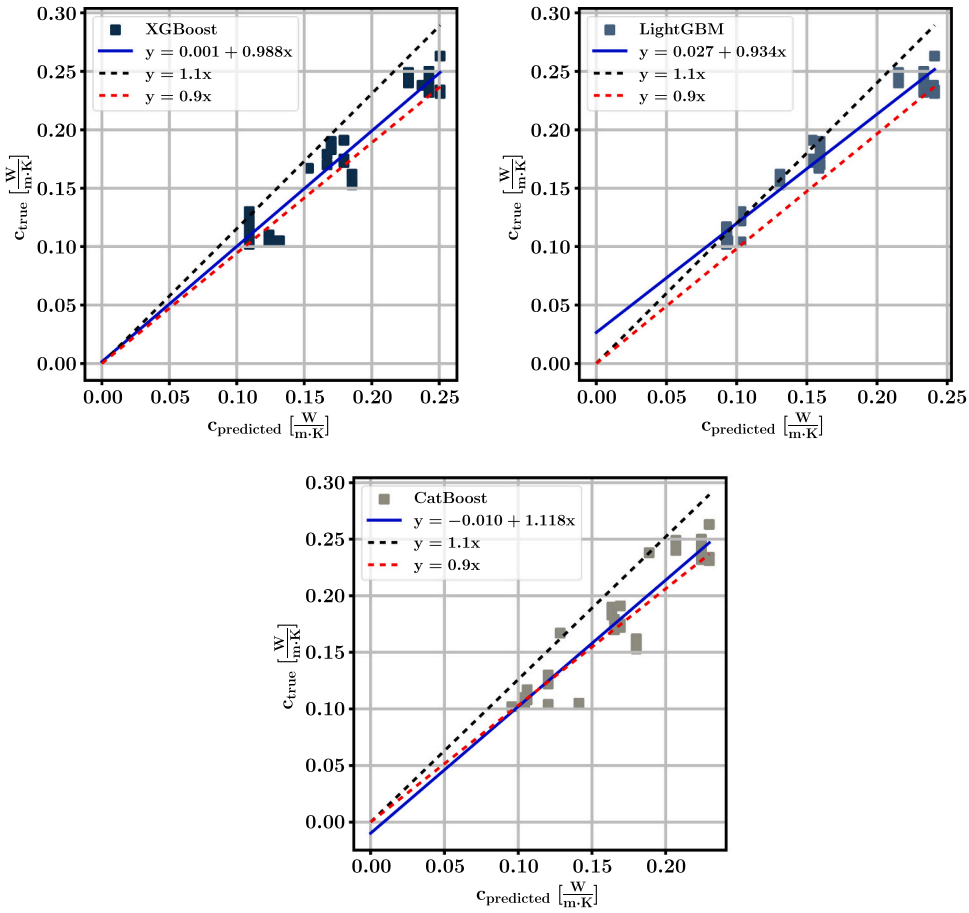

Fig. 14. ML predictions on the validation dataset.

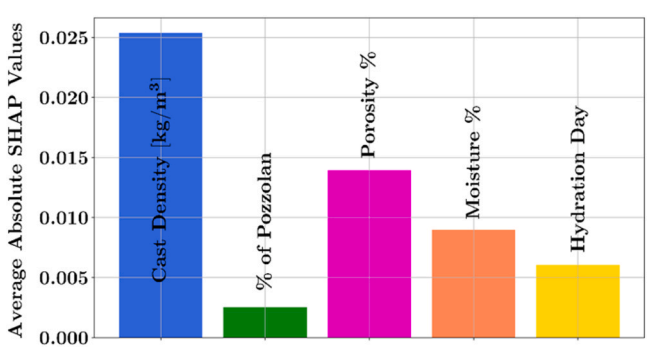

Fig. 15. Feature importance using average absolute SHAP values.

- The isolation forest algorithm was shown to effectively eliminate outliers in the dataset, thereby considerably optimizing the performance of the predictive machine learning models.
- Cost-related Frugal Hyperparameter Optimization (CFO), and Economical Hyperparameter Optimization with Blended Search Strategy (BlendSearch) algorithms have been demonstrated as effective methodologies for hyperparameter optimization of ensemble models.
- The XGBoost model performed best on the test dataset in terms of the R^2 score, whereas the LightGBM model scored best in terms of the mean absolute error and root mean squared error metrics. However, it should be noted that all the employed algorithms predicted the thermal conductivity with a coefficient of determination greater than 0.98 on the test dataset.

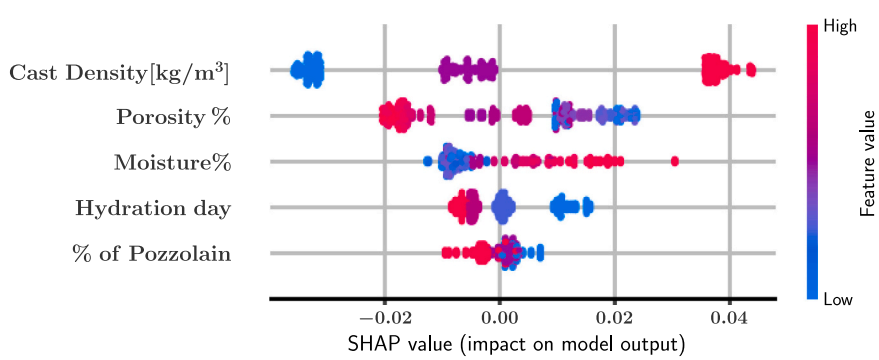

Fig. 16. SHAP summary plot of the CatBoost model.

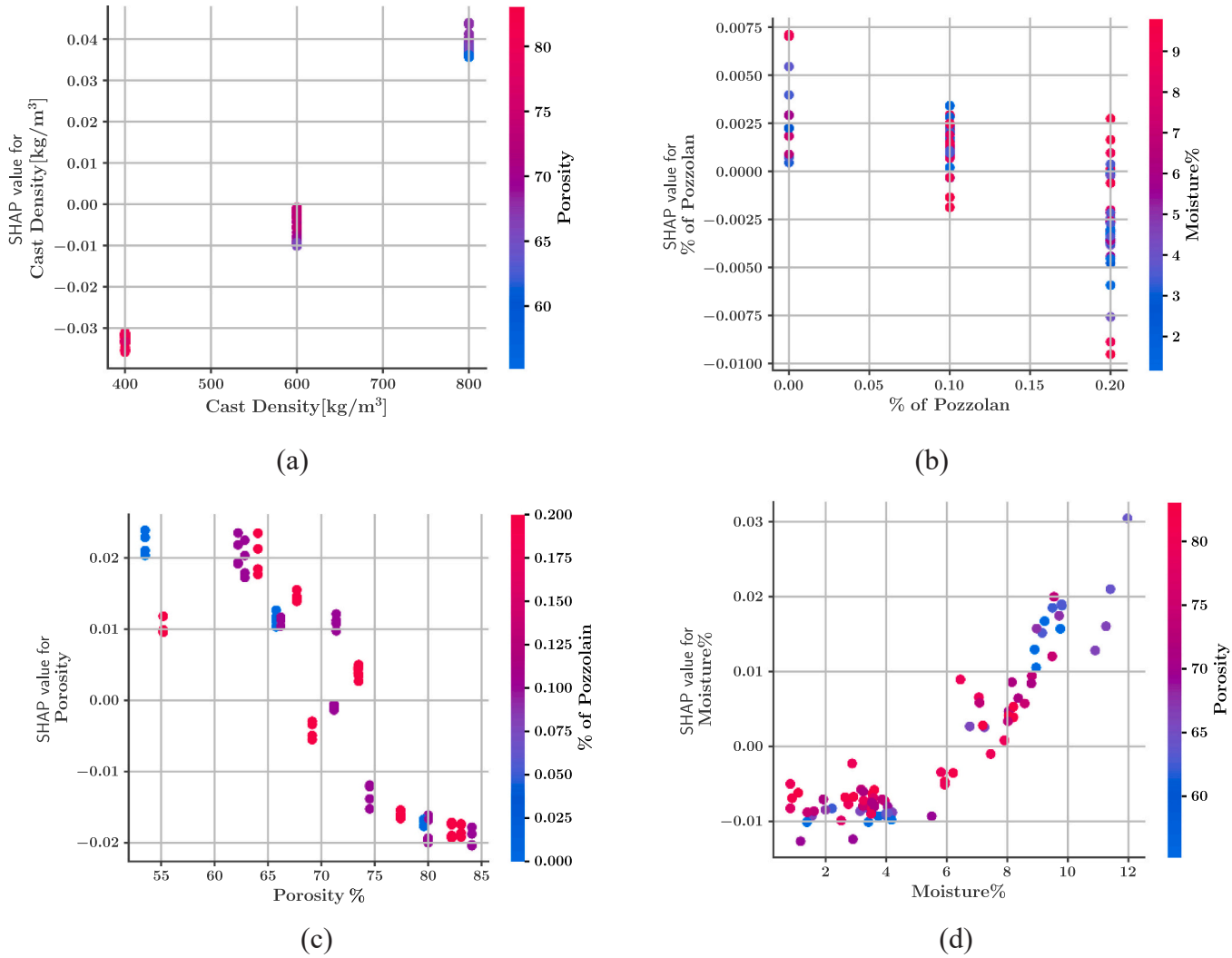

Fig. 17. SHAP feature dependence plots (CatBoost) for (a) cast density, (b) percentage of pozzolan, (c) porosity, and (d) moisture percentage.

- Similar performances could be obtained from all of the ensemble models in terms of RMSE and MAE, with an average RMSE of 0.0067 W/(mK) and an average MAE of 0.0049 W/(mK). The model performances have been visually summarized in the form of a radar chart.
- The linear correlation between the ensemble model predictions and the target values have been visualized in terms of the Pearson correlation coefficient in a Taylor diagram. All predictive models had Pearson correlation values greater than 0.99 and were closely positioned on the Taylor diagram.
- The outcome of the SHAP analysis showed that the cast density had the greatest influence on the ML model predictions, followed by porosity. On the other hand, the percentage of pozzolan in the mixture had the least effect on the model output.
- The relationship between different input features and the thermal conductivity was further visualized using ICE plots, where thermal conductivity increased with increasing cast density and moisture. Conversely, conductivity decreased as porosity and the percentage of pozzolan increased.

Since the present study only deals with densities lower than 800 kg/m^3, there is need for future research in this area for the development of models for cement-based foam with densities greater than 1000 kg/m^3 and possibly incorporating nanoparticle addition. While the models delivered highly accurate results, they are limited by the input feature ranges and distributions that were tested in the experimental program. More advanced predictive models can be developed in the future when more pertinent data becomes available from further experimental research. The results obtained from the verified numerical models can be used to further enhance the dataset used for machine learning model training. Since the quality of a data-driven model is determined by the size and quality of the dataset on which it has been trained, it is important to enhance the size and variable ranges of the dataset used in predictive model training to attain broader applicability of the model.

CRediT authorship contribution statement

Farnaz Batool: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Celal Cakiroglu:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Kamrul Islam:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Moncef L. Nehdi:** Writing – review & editing, Validation, Project administration, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

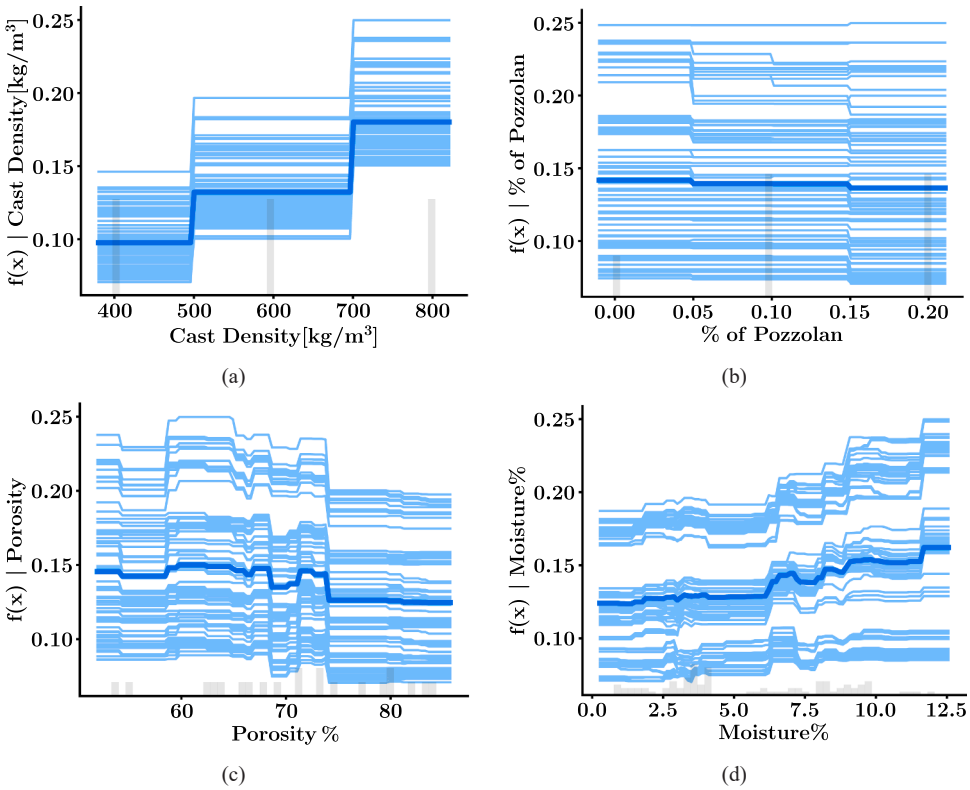

Fig. 18. Individual conditional expectation (ICE) plots for the variables in the prediction of the thermal conductivity: (a) cast density, (b) percentage of pozzolan (c) porosity, and (d) moisture percentage.

the work reported in this paper.

Data availability

We created a link to make the data available and accessible in the public domain. The link is given in the manuscript.

References

[1] M.A. Aladsani, H. Burton, S.A. Abdullah, J.W. Wallace, Explainable machine learning model for predicting drift capacity of reinforced concrete walls, Acids Struct. J. 119 (3) (2022), https://doi.org/10.14359/51734484.

[2] D. Aldridge, Introduction to foamed concrete: what, why, how?. Use of Foamed Concrete in Construction: Proceedings of the International Conference Held at the University of Dundee, Scotland, UK on 5 July 2005 Thomas Telford Publishing, 2005, pp. 1–14.

[3] R. Alsini, A. Almakrab, A. Ibrahim, X. Ma, Improving the outlier detection method in concrete mix design by combining the isolation forest and local outlier factor, Constr. Build. Mater. 270 (2021) 121396, https://doi.org/10.1016/j.conbuildmat.2020.121396.

[4] ASTM C566, Standard Test Method for Total Evaporable Moisture Content of Aggregate by Drying, ASTM International, West Conshohocken, PA, 2013.

[5] F. Batool, V. Bindiganavile, Air-void size distribution of cement based foam and its effect on thermal conductivity, Constr. Build. Mater. 149 (2017) 17–28, https://doi.org/10.1016/j.conbuildmat.2017.05.114.

[6] F. Batool, V. Bindiganavile, Quantification of factors influencing the thermal conductivity of cement-based foam, Cem. Concr. Compos. 91 (2018) 76–86, https://doi.org/10.1016/j.cemconcomp.2018.04.015.

[7] Batool, F., & Bindiganavile, V. (2013). Thermal conductivity of cement based foams and applicability of predictive models. in: Proceedings of the CONSEC Conference. 13, Nanjing, China, September 23–25, 2013.

[8] F. Batool, N.N. Prasad, V. Bindiganavile, Statistical modeling of thermal conductivity for cement-based foam, J. Build. Eng. 19 (2018) 449–458, https://doi.org/10.1016/j.jobe.2018.05.022.

[9] Burzawa, A., Bodet, L., Dhemaied, A., Dangeard, M., Pasquet, S., Vitale, Q., Boisson-Gaboriau, J., & Cui, Y.J. Detecting Mechanical Property Anomalies Along Railway Earthworks by Bayesian Appraisal of MASW Data. Construction and Building Materials, 404, 133224, https://doi.org/10.1016/j.conbuildmat.2023.133224.

[10] W.B. Chaabene, M. Flah, M.L. Nehdi, Machine learning prediction of mechanical properties of concrete: critical review, Constr. Build. Mater. 260 (2020) 119889, https://doi.org/10.1016/j.conbuildmat.2020.119889.

[11] S.N. Chinnu, S.N. Minnu, A. Bahurudeen, R. Senthilkumar, Reuse of industrial and agricultural by-products as pozzolan and aggregates in lightweight concrete, Constr. Build. Mater. (2021), https://doi.org/10.1016/j.conbuildmat.2021.124172.

[12] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking Inside the Black Box: visualizing statistical learning with plots of individual conditional expectation, J. Comput. Graph. Stat. 24 (1) (2015) 44–65.

[13] Y. Gu, G. Li, J. Gu, J.J. Jung, Graph embedding-based anomaly localization for HVAC system, J. Build. Eng. 77 (2023) 107511, https://doi.org/10.1016/j.jobe.2023.107511.

[14] H. Jia, B. Cui, G. Niu, J. Chen, Y. Yang, Q. Wang, C. Tang, Experimental and mechanism study on the impermeability and thermal insulation of foam concrete regulated by nano-silica and fluorine-free foam, J. Build. Eng. 64 (2023) 105675, https://doi.org/10.1016/j.jobe.2022.105675.

[15] Y. Jin, X. Wang, W. Huang, X. Li, Q. Ma, Mechanical and durability properties of hybrid natural fibre reinforced roadbed foamed concrete, Constr. Build. Mater. 409 (2023) 134008, https://doi.org/10.1016/j.conbuildmat.2023.134008.

[16] K. Li, Y. Wang, X. Zhang, J. Wu, X. Wang, A. Zhang, Intrinsically hydrophobic magnesium oxychloride cement foam for thermal insulation material, Constr. Build. Mater. 288 (2021) 123129, https://doi.org/10.1016/j.conbuildmat.2021.123129.

[17] Q. Li, Z. Ju, Z. Wang, L. Ma, W. Jiang, D. Li, J. Jia, Thermal performance and economy of PCM foamed cement walls for buildings in different climate zones, Energy Build. 277 (2022) 112470, https://doi.org/10.1016/j.enbuild.2022.112470.

[18] F.T. Liu, K.M. Ting, Z.-H. ZhouIsolation Forest. ICDM'08: Proceedings of the 2008 Eighth IEEE International Conference on DataMining, 413–422, 10.1109/ICDM.2008.17.2008.

[19] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, ACM Trans. Knowl. Discov. Data 6.1 (2012) 3, https://doi.org/10.1145/2133360.2133363.

[20] S.M. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st Conference on Neural Information Processing Systems, NIPS, Long Beach, CA, USA, 2017, 4–9 December 2017.

[21] G.A. Lyngdoh, M. Zaki, N.A. Krishnan, S. Das, Prediction of concrete strengths enabled by missing data imputation and interpretable machine learning, Cem. Concr. Compos. 128 (2022) 104414, https://doi.org/10.1016/j.cemconcomp.2022.104414.

[22] A.M. Maglad, M.A.O. Mydin, S.S. Majeed, B.A. Tayeh, S.A. Mostafa, Development of eco-friendly foamed concrete with waste glass sheet powder for mechanical, thermal, and durability properties enhancement, J. Build. Eng. 80 (2023) 107974, https://doi.org/10.1016/j.jobe.2023.107974.

[23] S. Mangalathu, S.H. Hwang, J.S. Jeon, Failure mode and effects analysis of RC members based on machine-learning-based shapley additive exPlanations (SHAP) approach, Eng. Struct. 219 (2020) 110927, https://doi.org/10.1016/j.engstruct.2020.110927.
[24] S. Mangalathu, H. Shin, E. Choi, J.S. Jeon, Explainable machine learning models for punching shear strength estimation of flat slabs without transverse reinforcement, J. Build. Eng. 39 (2021) 102300, https://doi.org/10.1016/j.jobe.2021.102300.
[25] A. Marani, E. Geranfar, L. Zhang, M.L. Nehdi, Deep learning-assisted calculation of apparent activation energy for cement-based systems incorporating microencapsulated phase change materials, Constr. Build. Mater. 404 (2023) 133324, https://doi.org/10.1016/j.conbuildmat.2023.133324.
[26] A. Marani, M.L. Nehdi, Machine learning prediction of compressive strength for phase change materials integrated cementitious composites, Constr. Build. Mater. 265 (2020) 120286, https://doi.org/10.1016/j.conbuildmat.2020.120286.
[27] E. Meng, J. Yang, B. Zhou, C. Wang, J. Li, Preparation and thermal performance of phase change material (PCM) foamed cement used for the roof, J. Build. Eng. 53 (2022) 104579, https://doi.org/10.1016/j.jobe.2022.104579.
[28] K. Mikulica, M. Labaj, R. Hela, Rehabilitation of floor structures using foam concrete, Procedia Eng. 195 (2017) 108–113, https://doi.org/10.1016/j.proeng.2017.04.531.
[29] N. Mohamad, K. Muthusamy, R. Embong, A. Kusbiantoro, M.H. Hashim, Environmental impact of cement production and solutions: a review, Mater. Today. Proc. 48 (2022) 741–746, https://doi.org/10.1016/j.matpr.2021.02.212.
[30] J. Moon, Y. Noh, S. Jung, J. Lee, E. Hwang, Anomaly detection using a model-agnostic meta-learning-based variational auto-encoder for facility management, J. Build. Eng. 68 (2023) 106099, https://doi.org/10.1016/j.jobe.2023.106099.
[31] M.A.O. Mydin, Effective thermal conductivity of foamcrete of different densities, Concr. Res. Lett. 2 (1) (2011) 181–189.
[32] I. Nunez, A. Marani, M. Flah, M.L. Nehdi, Estimating compressive strength of modern concrete mixtures using computational intelligence: a systematic review, Constr. Build. Mater. 310 (2021) 125279, https://doi.org/10.1016/j.conbuildmat.2021.125279.
[33] J. Rahman, K.S. Ahmed, N.I. Khan, K. Islam, S. Mangalathu, Data-driven shear strength prediction of steel fiber reinforced concrete beams using machine learning approach, Eng. Struct. 233 (2021) 111743, https://doi.org/10.1016/j.engstruct.2020.111743.
[34] A. Raj, D. Sathyan, K. Balaji, K.M. Mini, Heat transfer simulation across a building insulated with foam concrete wall cladding, Mater. Today. Proc. 42 (2021) 1442–1446, https://doi.org/10.1016/j.matpr.2021.01.242.
[35] A. Soleymani, H. Jahangir, M.L. Nehdi, Damage detection and monitoring in heritage masonry structures: systematic review, Constr. Build. Mater. 397 (2023) 132402, https://doi.org/10.1016/j.conbuildmat.2023.132402.
[36] H. Song, A. Ahmad, F. Farooq, K.A. Ostrowski, M. Maślak, S. Czarnecki, F. Aslam, Predicting the compressive strength of concrete with fly ash admixture using machine learning algorithms, Constr. Build. Mater. 308 (2021) 125021, https://doi.org/10.1016/j.conbuildmat.2021.125021.
[37] T.H. Tan, S.N. Shah, C.C. Ng, A. Putra, M.N. Othman, K.H. Mo, Insulating foamed lightweight cementitious composite with co-addition of micro-sized aerogel and hydrogen peroxide, Constr. Build. Mater. 360 (2022) 129485, https://doi.org/10.1016/j.conbuildmat.2022.129485.
[38] Wang, C., Wu, Q., Huang, S., & Saied, A. (2021). Economic hyperparameter optimization with blended search strategy. In Ninth International Conference on Learning Representations (ICLR 2021) | May 2021, (Virtual Only).
[39] R. Wang, H. Qiu, X. Cheng, X. Liu, Anomaly detection with a container-based stream processing framework for industrial internet of things, J. Ind. Inf. Integr. 35 (2023) 100507, https://doi.org/10.1016/j.jii.2023.100507.
[40] J. Wu, Y. Deng, X. Zheng, Y. Cui, Z. Zhao, Y. Chen, F. Zha, Hydraulic conductivity and strength of foamed cement-stabilized marine clay, Constr. Build. Mater. 222 (2019) 688–698, https://doi.org/10.1016/j.conbuildmat.2019.06.164.
[41] Wu, Q., Wang, C., Huang, C. Frugal Optimization for Cost-related Hyperparameters. The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21) (Virtual Only).
[42] H. Zhao, H. Yu, Y. Yuan, H. Zhu, Blast mitigation effect of the foamed cement-base sacrificial cladding for tunnel structures, Constr. Build. Mater. 94 (2015) 710–718, https://doi.org/10.1016/j.conbuildmat.2015.07.076.