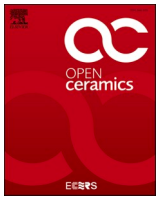
Open Ceramics 24 (2025) 100857

Contents lists available at ScienceDirect

Open Ceramics

journal homepage: www.sciencedirect.com/journal/open-ceramics



Machine learning for the optimization of porosity-hygroscopy correlations of porous geopolymers in indoor thermal comfort: A hybrid feature selection approach

Lauraine Tiogning-Djiogue [a,*], Herman Tcheneghon Motcheyo [b,*], Elie Kamseu [c,e,*], Sylvie Rossignol [d], Cristina Leonelli [e]

[a] National Advanced School of Engineering, University of Yaoundé 1, Yaoundé, P.O. Box 812, Cameroon
[b] Department of Computer Science, Faculty of Science, University of Yaoundé, 1, Yaoundé, P.O. Box 812, Cameroon
[c] Local Materials Promotion Authority, Nkolbisson, Yaoundé, P.O. Box 2393, Cameroon
[d] RCER, UMR 7315, University of Limoges, Limoges, 12 Rue Atlantis, France
[e] Department of Engineering "Enzo Ferrari", Modena, Via P. Vivarelli 10, 41125, Italy

ARTICLE INFO

Keywords:
Machine learning
Feature selection
Geopolymer matrice
Sustainability
Porosity-hygroscopy correlations
Thermal comfort
RReliefF. NSGA-II

ABSTRACT

Geopolymers are recognized as sustainable and environmentally friendly materials with a notable hygroscopic capacity that provides several advantages, particularly concerning thermal comfort. Optimizing the selection of variables with the most significant impact is essential for enhanced performance. However, conducting experimental tests to establish porosity hygroscopy correlations is costly regarding labor, time, and material resources. This study aims to employ a hybrid feature selection technique based on a multi-objective algorithm incorporating RReliefF and NSGA-II to streamline the geopolymer matrices by automatically selecting the most impactful and significant variables for their hygroscopic properties. Upon evaluating this feature selection method with laboratory-collected data, the Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE) results are significantly superior to those of other existing methods. These results underscore the importance of intrinsic parameters such as porosity, tortuosity, and pore diameter, along with external parameters like temperature and humidity, which directly affect hygroscopy. Consequently, this approach is expected to reduce experimental efforts and expedite the development of new geopolymer materials.

1. Introduction

Thermal comfort in buildings is a crucial factor for human well-being, productivity, and performance. It represents a major challenge for our current societies facing global warming and climate change. In the field of materials chemistry, porous geopolymer matrices are presented as materials with the capacity to regulate thermal comfort in buildings. Predicting the absorption/desorption capacity of these materials requires careful identification of the physicochemical parameters likely to affect their moisture control capacity. The study of parameters that influence the moisture control capacity of porous geopolymer matrices has so far been carried out experimentally in the literature [1, 2]. Although the experimental approach produces some interesting results, it presents a certain number of limitations, particularly in terms of time and cost of material resources, requiring significant budgets for the acquisition of raw materials and experimental management. In addition, the control of this approach is hampered by the possibility that unaccounted factors may influence the results obtained. Another limitation is the absence of formal mathematical equations governing the behavior of the porous geopolymer matrices regarding absorption/desorption. Exploring the data of porosity, relative humidity, temperature, and hygroscopic capacity using machine learning techniques provides an effective solution to evaluate porosity-hygroscopic correlations. The fundamental challenge of almost every data mining or modeling task in machine learning is to identify and characterize the relationships between one or more features and the target variable [3]. With the expansion and increasing complexity of data across various fields, data analysis has become progressively intricate. Generally, in these datasets,

* Corresponding authors.
E-mail addresses: lauraine.djiogue@univ-yaounde1.cm (L. Tiogning-Djiogue), herman.tcheneghon@facsciences-uy1.cm (H.T. Motcheyo), Kamseuelie2001@yahoo.fr (E. Kamseu).

https://doi.org/10.1016/j.oceram.2025.100857
Received 1 June 2025; Received in revised form 17 September 2025; Accepted 29 September 2025
Available online 2 October 2025

only a subset of the available features is relevant, meaning this subset can be used effectively to determine the endpoint value. The remaining irrelevant, rarely pertinent features in real-world problems are not informative but contribute to the overall dimensionality of the problem space. This increases the difficulty and calculations imposed on modeling methods and significantly affects model performance. Employing feature selection methods can effectively reduce noise, eliminate irrelevant information, and address redundancies. Feature selection can be broadly defined as the process of identifying relevant features while discarding those that are irrelevant [4]. There are two primary strategies in feature selection methods: the ranking method, which weights features according to their degree of relevance [4] and the subset selection method, which employs heuristic search strategies to derive a compact subset of features [5].

Feature selection can be generically defined as the process of identifying relevant features and eliminating irrelevant ones [6]. To the best of our knowledge, this is the first work that automatically identifies relevant features for efficient prediction of the absorption/desorption capacity, establishing real correlations between the pore network and hygroscopy. The problem of determining a minimal subset of features that best influence the hygroscopic capacity of porous geopolymers using a machine learning technique can be formalized as a multi-objective problem [7] consisting of two functions: minimizing the prediction error and minimizing the number of features. In this paper, the main contribution is to find the relevant features to predict the hygroscopic capacity of porous geopolymer matrices: (i) propose T-MOFSRRGA, a multi-objective feature selection approach for regression problems; (ii) show that a hybrid feature selection approach named MOFS-RFGA proposed by Yu et al. [8], based on ReliefF [3] and NSGA-2 [9], is applicable to the locally collected dataset; (iii) show that the intrinsic properties of the porous matrices are significant in predicting effective hygroscopy. The remainder of this paper is organized as follows: Section 2 presents the work related to porous geopolymer matrices and feature selection techniques, Section 3 presents the methodology, Section 4 presents the experimental design, and Section 5 presents the results and discussions.

2. Related work

2.1. Porous geopolymer matrice

Nowadays, mineral porous materials are preferred for their hygroscopic and insulating properties since they are low-energy, sustainable, and environmentally friendly [10]. The mix design and production of porous geopolymers have significantly advanced, making it possible to achieve hierarchical porosity with properly controlled volume and size distribution for efficient hygrothermal function and good insulation. Kamseu et al. [2] experimentally used a microstructural approach in porous geopolymers to understand the effects of pore size distribution and pore connectivity on heat and moisture flux transfer. By studying the non-linear relationship between tortuosity-porosity and tortuosity-thermal conductivity, they confirmed that pore volume is not the only parameter influencing heat flux transport. Zenabou et al. [1] proposed an innovative porous geopolymer whose volume, size, and distribution of the pores are controlled with the aim of optimizing their capacity to absorb and desorb moisture. In addition to these contributions, Gao et al. [11] investigated the durability of porous perlite geopolymers under hot and humid environments, revealing the crucial role of Na^{+}migration and its reaction with CO_2 or Cl^{-}in determining long-term stability. In another study, Gao et al. [12] optimized the thermal conductivity of porous geopolymers through theoretical modeling, demonstrating that increasing porosity while reducing the thermal conductivity of the solid skeleton is essential to reach values as low as 0.040 Wm^{-1} K^{-1}. Furthermore, Dai et al. [13] developed a kaolinite-based composite insulator reinforced with basalt fibers, achieving both enhanced compressive strength and reduced thermal conductivity after heat treatment. Similarly, Feng et al [14]. studied the incorporation of glass fibers into kaolinite-based porous insulators, confirming that glass fiber addition improves the strength-to-density ratio, although heat treatment leads to a trade-off between mechanical and thermal insulation performance.

Today, many methods have been proposed [15,16,2] for constructing porous materials for hygrothermal applications to ensure thermal comfort in buildings. For example, in various climate contexts, materials with high moisture absorption capacity can help regulate indoor humidity levels, contributing to better indoor air quality and comfort. However, finding the variables that impact the hygroscopic capacity of these materials is a very difficult experimental task. In the following section, we review the feature selection methods used in machine learning to determine the relevant features while our goal is to predict hygroscopic capacity.

2.2. Feature selection techniques

Feature selection aims to identify and exclude irrelevant or redundant features from the initial dataset while retaining essential features to enhance the performance of the learning algorithm. It is a subpart of dimension reduction which involves transforming a $X \in \mathbf{R}^{n*p}$ representation into a $X^* \in \mathbf{R}^{n*m}$ representation with $m \ll p$ [17] where the X^* data will be easily analyzed, interpreted, and visualized [18]. There are a number of advantages to this technique, including reduced computational cost, improved data quality, simplification of the learning process, and elimination of noise. These elements improve the quality of the results [19]. Another subpart of dimensional reduction is the feature extraction which consists of transforming the initial data space

$X \in \boldsymbol{R}^{n*p}$ *into* $X^* \in \boldsymbol{R}^{n*m}$ where each observation $x_i^i \in R^m (m < p)$is calculated to maximize explained variance, maximize underlying structures, and reduce noise in the initial set. Feature selection techniques are divided into three main categories: filter, wrapper, and embedded methods.

Filter methods employ feature ranking techniques as the primary means of selecting variables through prioritization [20]. These ranking techniques are favored for their simplicity and have demonstrated effectiveness in real-world scenarios. A specific ranking criterion is applied to evaluate variables, with those falling below a certain threshold being eliminated from consideration [20]. Filter methods offer advantages including low computational efficiency, suitability for large datasets, and freedom from classifier bias, which reduces overfitting [21]. However, a limitation of the ranking methods is the possibility of acquiring a subset that is less than optimal because the interactions between the selected subsets and the performance of the induction algorithm are neglected, and redundant features could remain included in the chosen subset [21].

Wrapper methods evaluate feature subsets by treating the predictor as a black box and using its performance as an objective function to score candidate subsets [22]. The goal is to iteratively retain variable subsets, train models on these subsets, and select the configuration that maximizes predictive performance. Since an exhaustive search of the entire feature space is computationally infeasible, heuristic search strategies such as particle swarm optimization, genetic algorithms, and other metaheuristics are employed to efficiently navigate the subset space. These methods often outperform filter methods because they inherently account for feature dependencies and interactions (e.g., synergistic relationships between variables) and leverage the inductive bias of the underlying learning algorithm (e.g., a classifier's inherent assumptions). However, wrapper approaches face two major limitations:

- Computational Intensity: Repeated model training and validation, particularly with complex classifiers (e.g., SVMs, neural networks) or large datasets, becomes prohibitively expensive. This cost is further amplified when cross-validation is used to mitigate overfitting.

- Overfitting Risk: Feature subsets optimized for training data performance may generalize poorly to unseen data, especially with small datasets or highly flexible models.

The primary challenge lies in the significant computational burden required to identify optimal feature subsets [20]. Despite these drawbacks, wrapper methods remain valuable when model accuracy is prioritized and computational resources allow their use.

Embedded methods [23] aim to minimize the computational overhead associated with reclassifying different subsets, which is a characteristic of wrapper methods. The primary strategy involves integrating feature selection directly into the training process, so that the model selects its own features. Individually, these methods are insufficient to handle large datasets, as highlighted by the No-Free-Lunch theorem [24]. This theorem states that no single optimization technique can universally solve all variable selection problems.

2.3. Literature review of hybrid methods

To improve the efficiency of learning algorithms, recent research has concentrated on hybrid methods that integrate multiple feature selection approaches to enhance algorithmic performance. The majority of studies on hybrid methods emphasize the combination of filter and wrapper approaches. This section reviews recent advancements in hybrid feature selection techniques.

Yu Kun et al [25]. proposed a feature selection algorithm that integrates mutual information maximization and an adaptive genetic algorithm to effectively reduce dimensionality in gene expression data while eliminating redundant features for improved classification. The findings highlight the hybrid algorithm's superior performance and heightened efficiency in selecting features for microarray datasets compared to traditional method.

Deng Xiongshi et al. [26] introduced XGBoost-MOGA, a hybrid gene selection method that integrates wrapper and embedded techniques for cancer classification in microarray data. Initially, genes are ranked using importance scores derived from XGBoost. The XGBoost-MOGA framework then employs a multi-objective genetic algorithm to identify an optimal subset of genes from the highest-ranked candidates, effectively balancing classification accuracy and feature redundancy. In the same vein, Liu Xiao-Ying et al [27]. Proposed a hybrid genetic algorithm known as HGAWE, which integrates both wrapper and embedded approaches for gene selection.

Hammami Marwa et al [28]. Introduced a multiobjective hybrid feature selection method that combines two filter techniques with a wrapper approach. The method operates in two stages: first, the filter techniques evaluate the performance of the entire feature population, and second, a local search refines only the top-ranked features. Experimental results demonstrate superior performance compared to existing methods, particularly in balancing feature subset quality and computational efficiency.

Jiahao Li et al [29]. Employed a hybrid filter-wrapper approach to address high-dimensional data challenges. They proposed a filter–wrapper feature selection method based on an improved multi-objective artificial bee colony algorithm (IMOABC). The filter method evaluates features individually using statistical dependency measures, distance metrics, or information-theoretic principles, independent of the learning model. While computationally efficient, this approach often overlooks feature interactions. In contrast, the wrapper method assesses the relevance of candidate feature subsets by directly measuring their impact on the predictive performance of a specific learning algorithm

Beijia Zhao et al. [30] proposed a novel transformer fault diagnosis model that integrates a hybrid filter-wrapper feature selection method with an AdaBoost-enhanced Weighted Broad Learning System (AdaBoost-WBLS). The proposed method demonstrates superior accuracy and more balanced performance in power transformer fault classification compared to the conventional diagnostic approaches.

Xianguo Wu et al [31]. Proposed a hybrid intelligence framework, integrating random forest (to identify the ten most influential factors) and NSGA-II to capture non-linear relationships among shield construction parameters (SCPs). The framework establishes a Mult objective optimization model, termed RF-NSGA II, which determines the optimal Pareto front to tune SCPs. The results demonstrate that the RF-NSGA-II framework not only achieves precise compliance with surface settlement and driving speed targets but also serves as a robust decision-support tool for real-time optimization and control of SCP, outperforming conventional methods.

Ghareb et al [32] have proposed several hybrid feature selection methods that combine the strengths of filter approaches with an enhanced genetic algorithm within an envelope framework to address the challenges of high dimensionality in the feature space and enhance classification performance.

Zhiwei Li et al [33] combined filter and wrapper methods for indoor temperature prediction, achieving high accuracy with relatively low computational overhead.

Yu Xue et al [8]. introduced a hybrid feature selection algorithm named MOFS-RFGA which is based on a multi-objective approach combined with the Relief method. This algorithm integrates the strengths of both filter and wrapper techniques to improve its effectiveness in tackling feature selection challenges. MOFS-RFGA has been compared with seven advanced multi-objective feature selection algorithms across 20 datasets. The results indicate that MOFS-RFGA effectively leverages the strengths of both filter and wrapper methods, outperforming the comparison algorithms in numerous datasets while ensuring robust classification performance and significantly reducing the number of features.

The main advantage of hybrid method is its robustness in terms of stability and reliability, which makes it particularly well suited for handling high-dimensional data. Following this study, we recognized that most existing methods have been developed for classification problems and applied, in particular, in various fields, especially health. However, no prior research has explored the use of machine learning techniques to investigate how porosity hygroscopy properties can enhance thermal comfort, particularly within a regression-based framework. Although these methods work effectively, determining a practical algorithm for feature selection is still an open issue.

2.4. RReliefF

RReliefF [34], for regressionnal ReliefF is an extension of ReliefF algorithm adapted for regression problems. It works in a similar way to ReliefF. Initially, the algorithm select random instance O_i and its k nearest instances O_j. Since the nearest neighbor of an instance O_i belonging to the same class as O_i is no longer applicable, several assumptions have been made regarding the calculation of neighbors:

- P_{dY} represents the weights for different prediction values y,
- $P_{dA}[A]$ represents the weights for different attribute values,
- $P_{dY\ \&dA}[A]$ represents the weights for different prediction values and different attributes

These weigths will be set to zero firstly and iteratively, the algorithm selects random instance O_i and its k nearest instances O_j.

$$P^{i}_{dY} = P^{i-1}_{dY} + \gamma_y(O_i, O_j) * d_{ij} \tag{1}$$

$$P^{i}_{dA}[A] = P^{i-1}_{dA}[A] + \gamma_y(O_i, O_j) * d_{ij} \tag{2}$$

$$P^{i}_{dY\&dA}[A] = P^{i-1}_{dY\&dA}[A] + \gamma_y(O_i, O_j) * \gamma_A(O_i, O_j) * d_{ij}$$

The final estimation of each attribute is giving by this formula: (3)

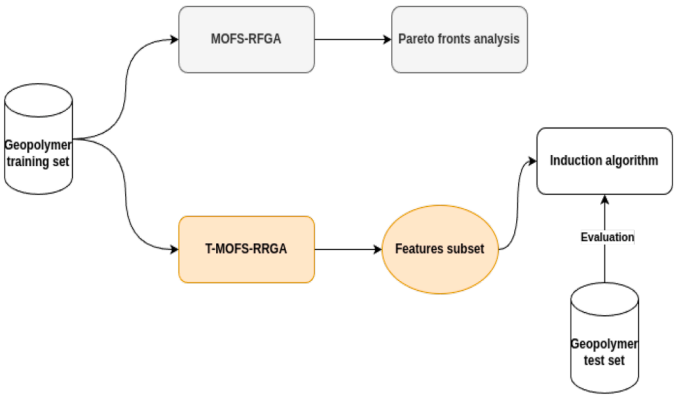


Fig. 1. Methodology.

$$W[A] = W[A] + \frac{P_{dY\&dA}[A]}{P_{dY}} - \frac{(P_{dA}[A] - P_{dY\&dA}[A])}{(m - P_{dY})} \tag{4}$$

where m is the number of iterations, A is an attribute name,

$$\gamma_y(O_i, O_j) = \frac{\left|y_i - y_j\right|}{Max(y) - Min(y)} \tag{5}$$

γ_y is the difference in the value of the continuous response y between observation O_i and O_j and d_{ij} takes into account the distance between the two instances O_i and O_j.

MOFS-RFGA uses ReliefF for algorithm weighting. However, this method was designed exclusively for classification problem, and we propose to adapt it for regression contexts.

3. Proposed method

3.1. Overview

The proposed method, T-Multi-Objective Feature Selection with Regressional Relief and Genetic Algorithm (T-MOFS-RRGA), is a multi-objective feature selection algorithm that modifies the weight initialization algorithm proposed by Xue et al. [8] and adapts the objective function for regression problems. This approach is based on MOFS-RFGA [8], which was originally developed for classification problems. The workflow used to find relevant features is presented in Fig. 1. After dividing the dataset into training and test sets, experiments were carried out. The MOFS-RFGA and NSGA-II algorithms were applied to the dataset, and the Pareto fronts were analyzed to evaluate the performance of the approach on this dataset. Subsequently, the T-MOFS-RRGA and NSGA-II algorithms were applied, and the Pareto fronts obtained, as well as the subsets obtained, were evaluated. After induction using the learning algorithms, they were evaluated against existing methods. The use of MOFS-RFGA and the T-MOFS-RRGA approaches are presented in the remainder of this section.

3.2. Apply MOFS-RFGA for classification

Given that the method proposed by Xue et al. [8] is applicable in a classification context, to ensure that it can also be used in the regression context, it was applied to the geopolymer dataset adapted for a classification task. "Pore type" is used as the categorical attribute of business significance. The results obtained were promising for consolidating Xue et al.'s approach [8]. The results allowed us to conclude that this multi-objective genetic algorithm is applicable to the regression context investigated in this study. An adapted version of this algorithm was then proposed for solving regression problems by modifying the attribute weighting algorithm using RreliefF [34] based on a filter approach and using a linear regression model instead of KNN [35] for prediction error

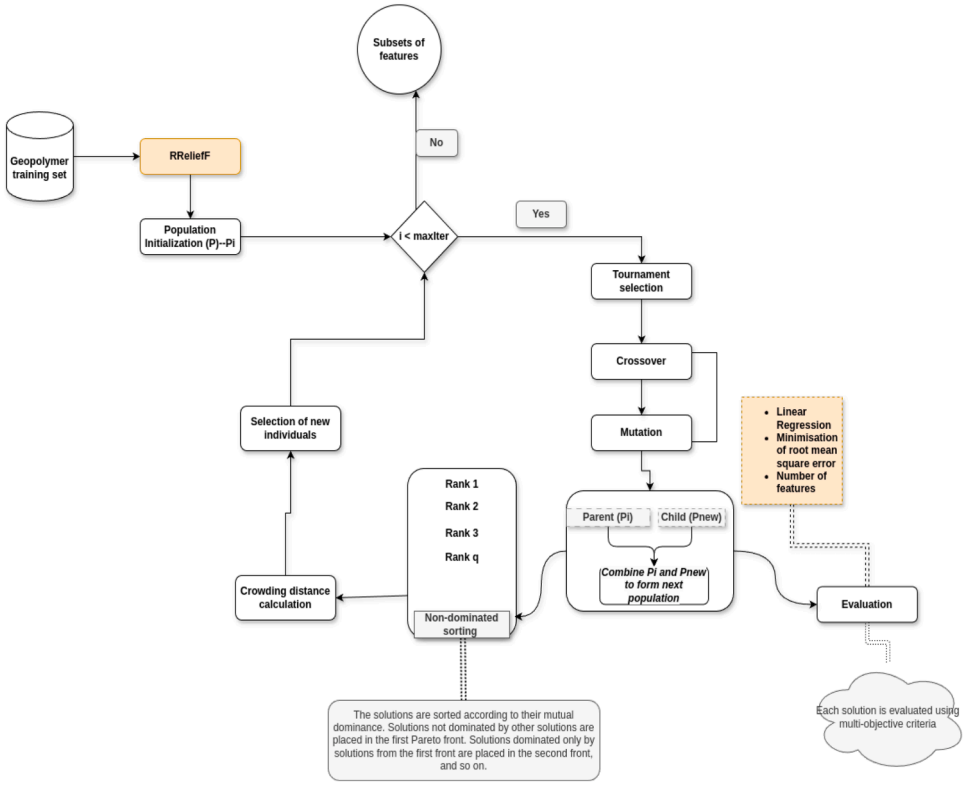


Fig. 2. T-MOFS-RRGA.

Table 1
Feature score after RRelief initialization.

Index	Features	Score RReliefF
1	Humidity	0.02
2	Water	0.18
3	Temperature	0.45
4	Tortutosity	0.10
5	Metakaolin	0.30
6	NaOh	0.12

minimization.

3.3. T-Multi-Objective feature selection with regressional relief and genetic algorithm(T-MOFS-RRGA)

The framework of the T-MOFS-RRGA method is shown in Fig. 2. TMOFS-RRGA stands for T-Multi-Objective Feature Selection with Regressional Relief and Genetic Algorithm is an adaptation of MOFS-RFGA approach for regression originally designed for classification. The motivation behind this work is the observation that most existing multi-objective feature selection methods have primarily focused on classification tasks and only later adapted to regression problems, often without properly addressing regression-specific challenges. As mentioned above, the process is the same as that proposed by Xue et al. [8], with the difference that the attribute weighting algorithm and the error minimization function are modified.

1. We adapt Relief to RRelief because the original Relief algorithm is designed for classification problems, where the objective is to minimize classification error. However, our work focuses on regression tasks, where the goal is to minimize prediction error. RReliefF is a regression-oriented extension of Relief that evaluates the relevance of features based on continuous target values, making it more suitable for our problem setting.
2. We chose the Mean Squared Error as the objective function for the regression task because it provides a strong mathematical incentive to penalize large deviations between predicted and actual values. This property is particularly relevant in our case, where accurate predictions are critical and large errors can lead to significant misinterpretation of the model's output.

After running the RReliefF algorithm, which produces relevance scores for each attribute, these scores are used to initialize the population. The population here constitutes a list of subsets of solutions. The next step is to evaluate this population, i.e., to calculate the value of each objective function for each individual in the population. The individuals best suited to reproduce in this population are selected to have a population of individuals that converges most towards the global optimum. These individuals are the parents who help to build the offspring. To mix the genetic heritage of the population already collected and avoid premature convergence of the algorithm towards the local extremum, crossover and mutation operations are carried out. The new descendants obtained are combined with the current population. Non-dominated sorting and crowding distance calculation are then used to determine the new population and preserve elitism. The process is iterative until the end condition is verified. The problem is therefore formalized as finding X such that we have:

Minimize $(F_1(X), F_2(X))$

Where $F_1(X) = \frac{1}{K}\sum_{l=1}^{K}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\left(y_i - \widehat{y_i}\right)^2\right)$ and

$$F_2(X) = \sum_{i=1}^{D} x_i \tag{6}$$

and $F_2(X)$ is the minimization of feature. X representing the solution, K the total number of groups or datasets, N_l is the size of group l, N_l the number of instances of the l^{th} dataset, D the number of features, x_i is the value of i^{th} value of the individual X, y_i is the actual value for the i^{th} data point and $\widehat{y_i}$ is the predicted value for the i^{th} data point.

Table 2
Initial population P_i based on feature score in Table 1.

	Humidity	Water	Temperature	Tortuosity	Metakaolin	NaOh
1	0	**1**	**1**	0	**1**	0
2	0	1	0	0	1	1
3	0	1	1	0	0	1
4	0	1	1	1	1	0

Table 3
Data dictionnary.

Feature name		Unit	Values
1	Humidity	%	R
2	Temperature	degree	R
3	Raw material (Metakaolin)	gram	80, 60, 50
4	RHA	gram	20, 40, 50
5	NaOh	gram	33.6, 29.6
6	Sodium silicate(Na_2SiO_2)	gram	67.2, 37.6
7	Si/Al	//	6.4793, 1.93, 7.98, 4.01, 2.71, 2.73, 2.45
8	Water	gram	0, 20
9	Hydrogen peroxide (H_2O_2)	gram	0, 0.06, 0.07, 0.08, 0.09
10	Aluminium(Al)	gram	0, 0.06, 0.07, 0.08, 0.09
11	Porosity	%	R
12	Tortuosity	//	R
13	Pore diameter(D)	µm	R
14	The ratio metakaolin/ RHA	//	1.5, 4, 1
15	Thermal conductivity	//	R
16	The type of pore	//	Mesopore, millipore
17	Density	//	R
18	m	gram	30, 40, 50, 20
19	Inter-pore space	//	0.057, 0.07, 0.067, 0.2, 0.051
20	**Hygroscopy capacity**	%	R

3.4. Example

Suppose we have a dataset of geopolymer matrices with the following features: Humidity, Water, Temperature, Tortutosity, Metakaolin, NaOh. After applying RReliefF algorithm, the relevance weights obtained are represented in Table 1.

Suppose $R = 3$ (number of features to be selected). Through a repeated binary tournament based on the above scores, the indices {2, 3, 5} are selected. These arrays represent a set of features selected for formulations.

0	**1**	**1**	**0**	**1**	0
||	||	||	||	||	||
Humidity	Water	Temperature	Tortuosity	Metakaolin	NaOh

The algorithms start by initializing a number a population of size N who represent a set of features relevance for geopolymer matrices. In this example, $N = 4$ is represented in Table 2. After initializing the initial population, the crossover operator favors the transmission of common chemical characteristics between three parental formulations (e.g., Temperature, NaOh), while the mutation operator, guided by the importance of each characteristic (according to RReliefF), dynamically adjusts the composition by activating or deactivating less or more relevant variables to optimize mechanical performance. Once these new compositions are obtained, we will combine this with the initial population P_i. Each individual in the population is evaluated according to the multi-objective criteria. Once this is completed, the nondominated sorting is performed. A solution dominates another if it offers a more

Table 4
Descriptive statistics of the dataset.

Feature names	Mean	Std	Min	25 %	50 %	75 %	Max
Humidity	83.26	10.90	51.00	81.00	87.50	90.00	95.00
Temperature	24.50	2.00	20.00	23.00	24.00	26.00	29.00
Metakaolin	64.10	12.86	50.00	50.00	60.00	80.00	80.00
RHA	36.67	12.48	20.00	20.00	40.00	50.00	50.00
NaOh	31.76	2.00	29.60	29.60	33.60	33.60	33.60
$Si0_2$	53.58	14.76	37.60	37.60	67.20	67.20	67.20
Water	9.20	9.98	0.00	0.00	0.00	20.00	20.00
Al	0.03	0.04	0.00	0.00	0.00	0.07	0.09
H2O2	0.04	0.04	0.00	0.00	0.06	0.08	0.09
porosity	66.31	8.59	56.00	60.00	64.00	70.50	87.00
tortuosity	1.34	0.15	1.13	1.25	1.32	1.37	1.70
Density	0.84	0.11	0.62	0.80	0.89	0.92	0.98
interpore	0.07	0.04	0.05	0.06	0.07	0.07	0.20
mk rha	2.17	1.31	1.00	1.00	1.50	4.00	4.00
Thermal_conductivity	0.15	0.03	0.12	0.13	0.14	0.17	0.20
m	35.83	11.16	20.00	27.50	40.00	42.50	50.00
SiO2 Al2O3	4.41	2.25	1.93	2.45	4.01	6.48	7.98
D	317.58	195.05	72.49	147.71	312.68	415.15	697.75
ab des	−5.01	5.68	−21.06	−8.44	−3.46	−0.32	3.14

efficient formulation in terms of strength, durability, or cost without being inferior in any of these aspects, which allows the selection of optimal formulations called non-dominated forming the Pareto front. For more information on mutation crossover, non-dominated sorting, and crowding distance calculation, refer to [36,37].

4. Experimental design

4.1. Collected datasets and description

According to [1], the design of interconnected thin layers of porous geopolymer materials was based on the idea that effective absorption and desorption will occur in a thin layer of geopolymer. Since the intrinsic properties of geopolymer are nano and mesoporosity [38], a porogen agent is gradually added to the new pastes at different concentrations over time, as shown in Table 3, where the volume of larger pores grows while preserving a homogeneous microstructure. There are distinct S_iO_2/Al_2O_3 used to make the geopolymer pastes. During this time, the lengthening of polysialate chains

(H—N-A-S) and modifications to the distribution of pore sizes enable the creation of porous matrices with hierarchical porosity and physicochemical stability and strength. All the mixture values are presenting in Table 3. During this time, the lengthening of polysi-alate chains (H—N-A-S) and the alteration of the distribution of pore sizes enable the creation of porous matrices with hierarchical porosity and physicochemical stability and strength. Rice Husk Ash (RHA) was used to modify the silica concentration in order to determine the SiO2/Al2O3 ratio values. Alumino-silicate precursor was a typical metakaolin that contained silica and alumina [39,1]. A combination of sodium hydroxide (8 M) and sodium silicate (SiO2/NaO2) was utilized as the alkaline solution. The right amounts of RHA and metakaolin are ballmilled until all of the particles are smaller than 200 μm. After adding the appropriate amount of solution, the paste is ball-milled once more for five minutes. The paste and porogen agent are combined quickly (a few seconds), and the resulting slurries are then poured into plastic molds. Hydrogen peroxide was employed as the porogen agent, and its concentrations are shown in Table 3 vol. %. To prevent contact with the surrounding air, the porous matrices are covered with plastic for the first 72 h. After that, the curing process is carried out in the lab at 23 ± 2 °C and 54 ± 2 % till the mass remains constant.

After building porous materials [1,2], the moisture absorption and desorption data were collected for each sample. Temperature and relative humidity were influenced by weather conditions such as rain, wind, and sun. Measurements of relative humidity, temperature, and sample mass were taken before sunrise and at 6 PM when there was no more sun. The mass of the sample was measured using an ELECTRONIC SCALE (W. SCALE MOD. BC4100C-4200) with an accuracy of 0.01 g Table 3 presents all the mixture design features of the geopolymer matrices as well as the quantities. A total of 600 observations for experimental absorption/desorption tests were collected for each formulation. Factors investigated in this study as potential features that affect the hygroscopy of porous geopolymers, and by extension the capacity of the matrices to ensure thermal comfort, include (i) direct factors such as humidity, temperature, porosity (pore volume, interpore space, pore size, etc.), and tortuosity; (ii) indirect factors such as metakaolin, RHA, alkaline solution, porogen agent, $S_iO_2/Al_2/O_3$ ratio, etc. Indirect factors are parameters governing the optimization of direct factors. They affect the formation of pores, the stability of the interpore space, the pore size, the pore size distribution, and the pore network. Temperature and humidity appear as direct factors affecting hygroscopy but are considered external factors.

The statistical description of the adopted dataset can be observed in Table 4. It contains the range of minimum and maximum values, standard deviation (Std), and average values of the adopted input and output variables. To develop the best predictive model, the proposed expression in the current study should be used within the specified range of parameters. The statistical analysis illustrates that the adopted dataset covers a large range of ingredients, and Std depicts the distribution of the dataset along the mean values. Fig. 3 illustrates the distributions of the input variables using histograms and kernel density estimates (KDE). During the preprocessing step, no missing values were detected in the dataset. In addition, all duplicated entries related to humidity, temperature, and adsorption-desorption measurements were removed during data collection to ensure the quality and integrity of the data. The features were analyzed in their raw form, without prior transformation or normalization, to better capture their original characteristics and behavior. This approach allows for an unbiased assessment of each variable's intrinsic nature, highlighting potential asymmetries, multimodalities, or outliers. It also promotes scientific transparency while leaving the door open for appropriate preprocessing or transformations in future work.

All the matrices have been optimized (Fig. 4a) to have a matrix with intrinsic nanopores and capillary pores homogeneously dispersed. The capillary pores have diameters under 1000 $^{\mu m}$ (Fig. 4b) for the most part. These classes of pores are responsible for the aeration of the porous network and are applied to control the hygroscopic behavior of the porous matrices by limiting the condensation of humidity while controlling the absorption desorption process. The role of S_iO_2/AlO_2 is to ensure the durability and stability of the matrix under temperature and humidity fluctuations.

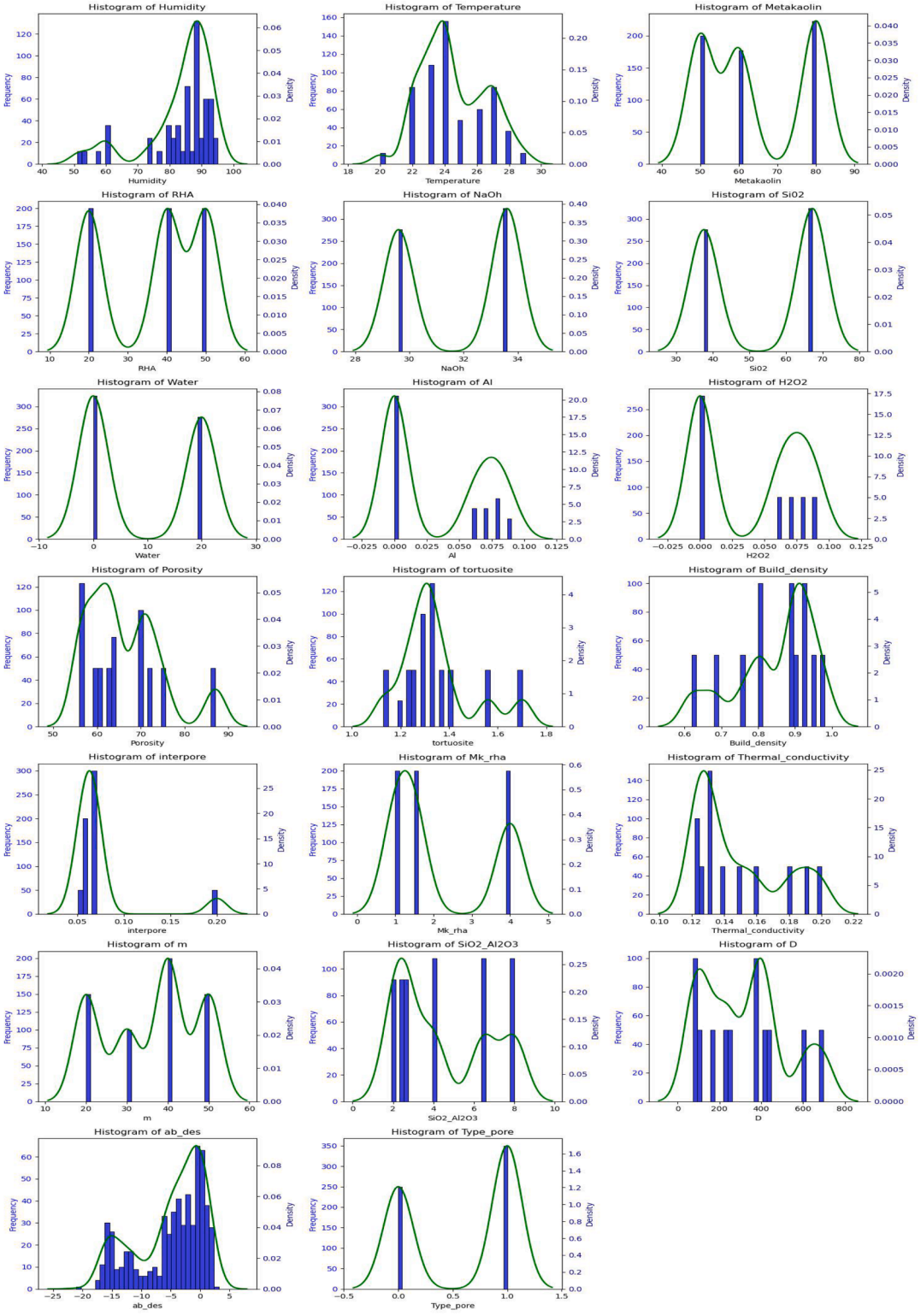
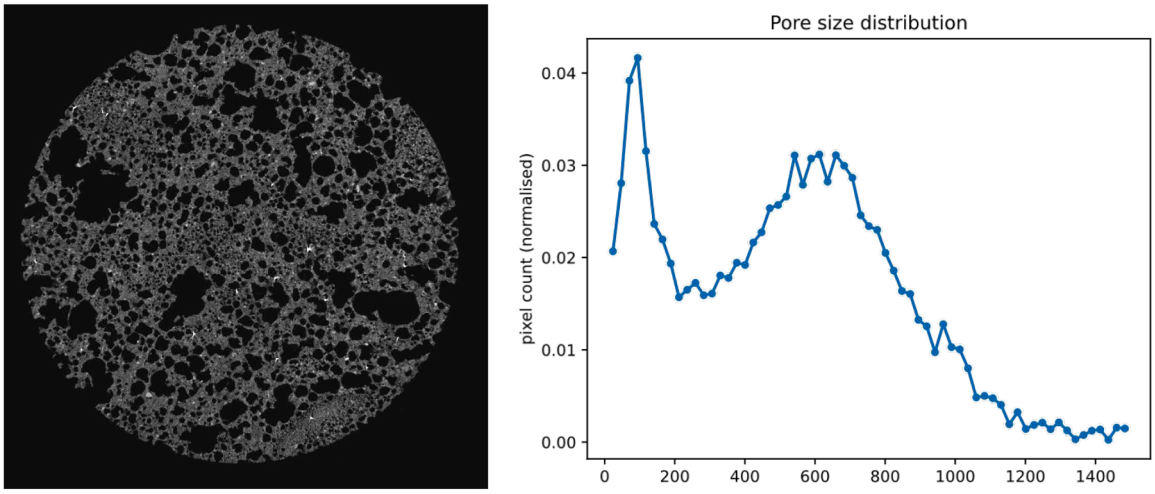


Fig. 3. Distribution of variables used in the study. Each subgraph shows a histogram enriched with a density curve.

Results of the macrotomography analysis showed that pore networks can be easily designed with desired pore volume, pore size distribution, and pore connectivity using controlled processing parameters such as S_iO_2/AlO_2 (RHA content), the porogen agent, and the rheology of the fresh pastes [1,2].

4.2. *Experimental tools*

We performed experiments on two levels to assess the proposed approach. To evaluate the effectiveness of their approach under identical circumstances, we first replicated the experimental configuration suggested by Xue et al. [8] using NSGA-II as the comparative algorithm to benchmark MOFS-RFGA.



(a) 2D microtomograph of a typical porous geopolymer

(b) Pore size distribution

Fig. 4. Materials and pore size distribution.

Table 5
Experimental setting MOFS-RFGA.

Parameters	NSGA-2	MOFS-RFGA
Dataset size	Train set: 420, Test set: 180	Train set: 420, Test set: 180
Croissover probability	0.9	//
Mutation probability	1/19	//
Initialization of the population	Random	RelieFF algorithm
Initial population size	100	100
Run number	10	10
Class number	2	2
Class name	Type of pore	Type of pore
Classifier	KNN ($k = 3$)	KNN($k = 3$)

Table 6
Algorithms.

Algorithms	Parameters
LASSO [40]	{Alpha: 1}
SVM-Recursive feature elimination [41]	{kernel: linear, kbest feature: 14}
Mutual information regression [42]	{kbest feature: 14}

Second, we adapted this experimental framework to suit the specifics of our regression problem and evaluated the effectiveness of the proposed TMOFS-RRGA(T-Multi-Objective Feature Selection with Regressional Relief and Genetic Algorithm) approach within this revised setting. For consistency and comparability, we retained the experimental procedure of Xue et al. when implementing our method, as described in Section 3.2. The experimental setup and dataset features are described in detail in Tables 5 and 7. The state-of-the-art algorithms used for comparison are compiled in Table 6, along with the parameter settings after that correspond to them. All multi-objective algorithms were executed over 2000 iterations.

4.3. Induction algorithm and metrics evaluation

Nowdays, ensemble methods are widely recognized in various domains to improve prediction accuracy [43]. They stand out notably for their ability to highlight the importance of each attribute in predicting the final model. Once subsets of features are obtained, they are evaluated using five artificial learning algorithms such as linear regression,

Table 7
Experimental setting T-MOFS-RRGA.

Parameters	NSGA-2	T-MOFS-RRGA
Initialization of the population	Random	RReliefF algorithm
Regression algorithm	Linear regression	Linear regression
Label name	Absorption/desorption	Absorption/desorption
Validation	$K = 3$ fold cross validation	$K = 3$ fold cross validation

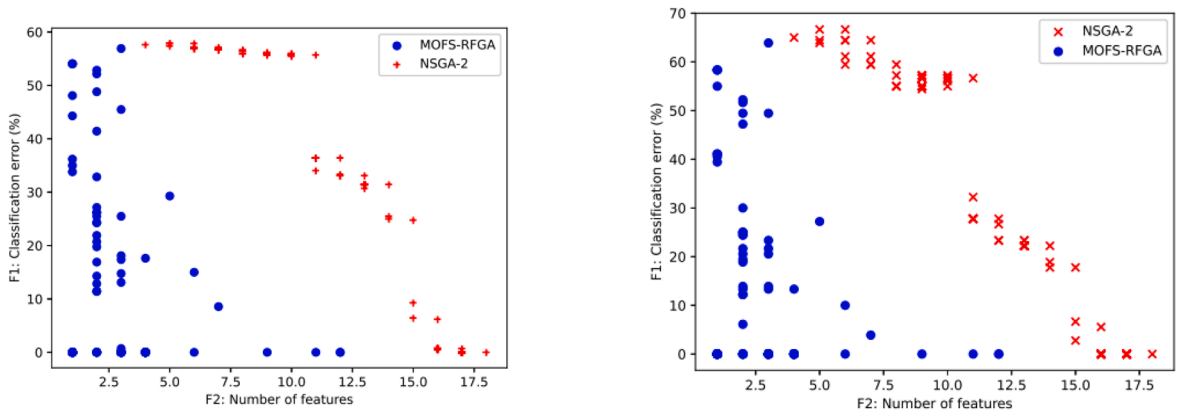
regression tree, Random forest, Adaboost and Gradient boosting, which serve as induction algorithms. Due to the computational cost of others algorithms, focus is made on linear regression. The metrics such as mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) are used to evaluate prediction results [44]. To evaluate the multiobjective algorithm, some metrics are used. The initial criterion is the Inverted Generational Distance (IGD) which primarily evaluates the algorithm's convergence and distribution performance. This is achieved calculating the total of the smallest distances between each point on the true Pareto front and the individuals acquired through the algorithm [45]. Furthermore, the solution set coverage (sc) [46] is used to assess the convergence discrepancy between the two algorithms. The third metric, the number of feature subsets, is used to evaluate the distributional disparities between the algorithms.

5. Results and discussions

Tables 5,6,7 show the parameters used to predict the hygroscopic capacity of porous geopolymers. The data have been divided into training and test set using holdout for selection and training. T-MOFS-RRGA(T-MultiObjective Feature Selection with Regressional Relief and Genetic Algorithm) is compared with several variable selection algorithms, including LASSO [40], SVM-RFE [41], Mutual Information [42] and NSGA-2 [9].

5.1. Analysis of MOFS-RFGA to porous geopolymer dataset

After running the MOFS-RFGA and NSGA-2 algorithms on the geopolymer dataset, a set of individuals is obtained. Each individual is characterized by a rank, the set of selected features, and the objective value, which is the fitness. This fitness is a vector containing the



(a) MOFS-RFGA vs NSGA-II training set (b) MOFS-RFGA vs NSGA-2 test set

Fig. 5. Best pareto fronts on classification.

Table 8
IGD, SC and Number of solutions analysis.

	IDG		Coverage(sc)		Number of solution	
	MOFS-RFGA	**NSGA-2**	**SC(MOFS-RFGA, NSGA-2)**	**SC(NSGA-2, MOFS-RFGA)**	**MOFS-RFGA**	**NSGA-2**
average	**0.135**	0.282	**0.943/ 0.14**	0	88	**100**

classification error and the number of features selected.

5.1.1. Analysis of MOFS-RFGA pareto front

In Fig. 5, each point represents an individual in the population and visually represents the classification error as a function of the number of features. The curve shows the Pareto front, which contains a set of solutions that represent the best compromise between the two objective functions. It can be observed that the MOFS-RFGA solutions dominate the NSGA-2 solutions, as demonstrated by the convergence results. A minimum error of 0 is observed on the training data and test set for each algorithm. The difference between the two lies in the number of variables. NSGA-2 returns a subset constituted by variables 15, 16, 17, and 18 for a training and test error of 0, whereas with MOFS-RFGA, a subset is constituted by variables 1, 2, 3, 4, 6, 9, 11, and 12 for an error of 0.

5.1.2. Analysis of IGD, convergence and diversity

Calculating IGD presupposes knowledge of the optimal Pareto front. To obtain it, NSGA-2 and MOFS-RFGA are executed. After the first execution, the two solutions are combined, and non-dominated sorting is used to obtain the theoretical optimal solution. After 10 independent executions for each algorithm, it is noted that MOFS-RFGA converges better with an average value of 0.135 compared to NSGA-2, which has a value of 0.282. Column 3 of Table 8 represents the coverage of MOFS-RFGA and NSGA-2. An average for sc(MOFS-RFGA, NSGA-2) equal to 0.943, a standard deviation of 0.14, and 0 for NSGA-2 were obtained. This confirms that NSGA-2 solutions do not dominate the MOFS-RFGA solutions. The column "Number of solutions" in Table 8 shows the average number of solutions obtained. The population is initialized with a size of 100 individuals for each algorithm. It is noted that NSGA-2 retains its size after each run but does not cover the whole search

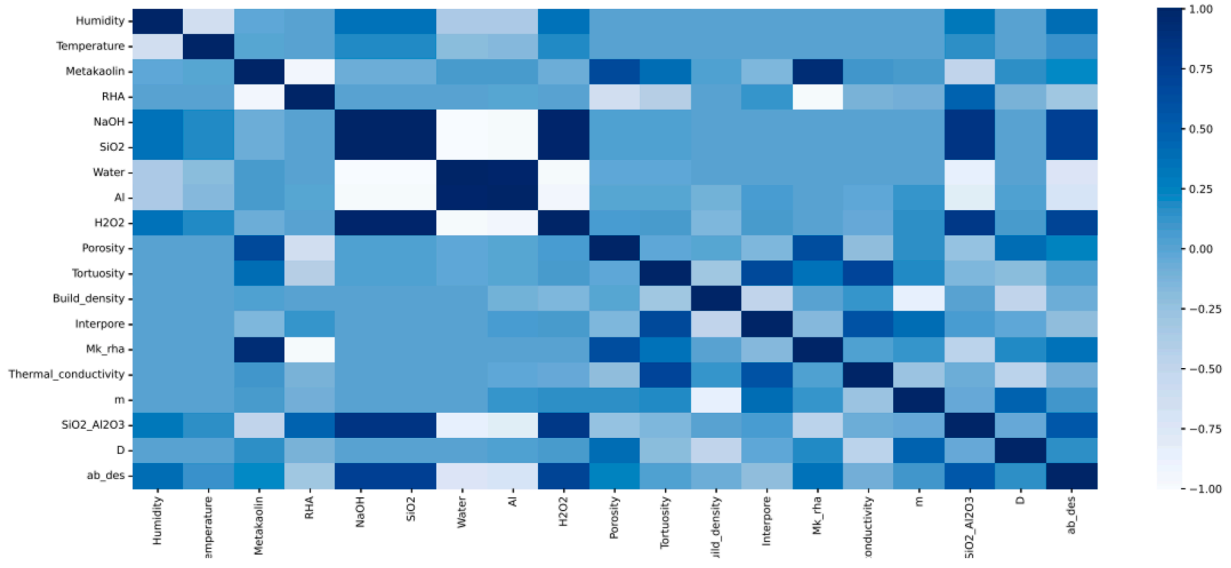


Fig. 6. Correlation matrix between features.

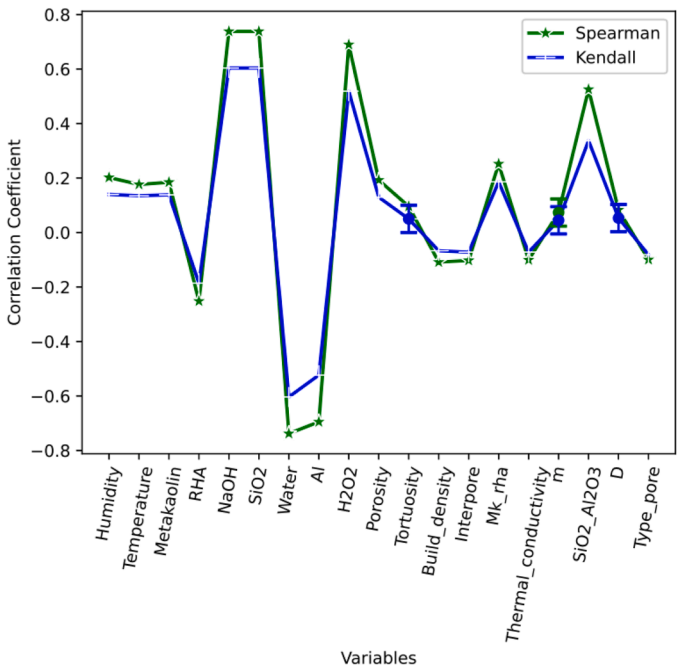


Fig. 7. Correlation Kendall and Spearman.

space, whereas MOFS-RFGA covers the search space but obtains an average number of individuals of 94. The analysis results obtained demonstrate that MOFS-RFGA [8] is appropriate in the context of materials chemistry and engineering and can be adapted to solve the problem.

5.2. Analysis of correlations between attributes

5.2.1. Relationship between features

To check the correlations between the data, the correlation between pairs of attributes was evaluated. Fig. 6 shows the correlation between each pair of attributes. Several correlations between variables have been observed.

Firstly, humidity and temperature display a negative correlation. Additionally, metakaolin (MK) and rice husk ash (RHA) exhibit a strong negative correlation, indicating that as one quantity increases, the other decreases within the geopolymer matrix setup. Silicate and soda solutions show a strong correlation, reflecting their combined use in forming the alkaline solution. Furthermore, aluminum, water, and hydrogen peroxide are correlated, with the exclusionary relationship between aluminum and hydrogen peroxide being highlighted, as well as the role they play as pore-forming agents. Water is also correlated with aluminum. Lastly, thermal conductivity is correlated with tortuosity and inter-pore space. These correlations provide insights into the intricate relationships within the studied thematic field.

It is important to underline Pearson's speciality in detecting linearity between data and target, which does not present direct multiple correlations in Fig. 6. It is also essential to note that although Al, NaOH, water and S_iO_2/Al_2O_3, are strongly correlated with hygroscopy, they are not sufficient to determine the absorption/desorption capacity, as climatic conditions must be taken into account. Spearman's coefficient and Kendall's rank measure the monotonicity of the relationship and the ordinal association between two or more variables, as illustrated in Fig. 7, highlighting the relationship between these variables and the target similar to that observed with Pearson's coefficient. After a significance test of the *p values* at $n - 2$ degrees of freedom, It is found that the correlation between most of these variables is statistically significant. A variable taken individually may not be correlated to the variable of interest but taken together, they can significantly impact the target.

Nonlinear analysis using mutual information present in Fig. 8 using pair of features highlights the central role of pore diameter (D), which is present in the majority of combinations most correlated with the target variable. In particular, the Tortuosity + D association is the most decisive, reflecting the joint importance of pore size and the complexity of diffusion pathways in transport mechanisms. The SiO_2/Al $2O_3$ + D and Build density + D interactions highlight that the porous microstructure acts synergistically with chemical composition and bulk density. The silica/alumina ratio modulates phase formation and thermal stability, while density and overall porosity influence mechanical strength and permeability. Finally, combinations involving H_2O_2, thermal conductivity, or interpore parameters confirm that material properties depend not only on D, but also on its interaction with synthesis conditions and other microstructural factors. Thus, the final performance results from a balance between pore geometry, chemical composition, and organization of the solid network.

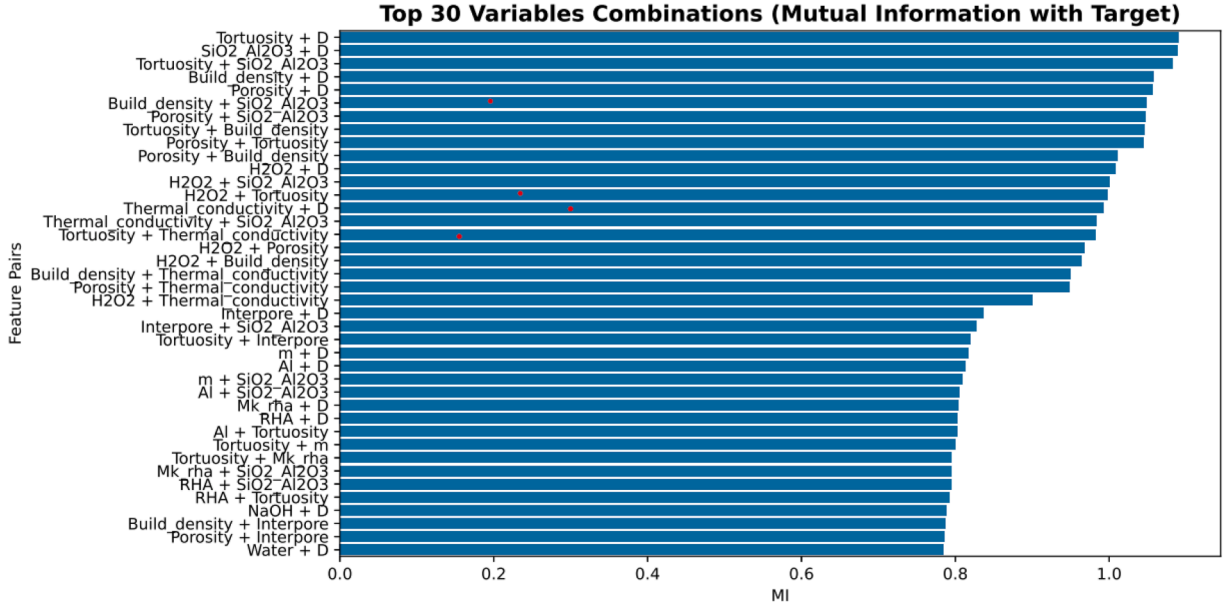
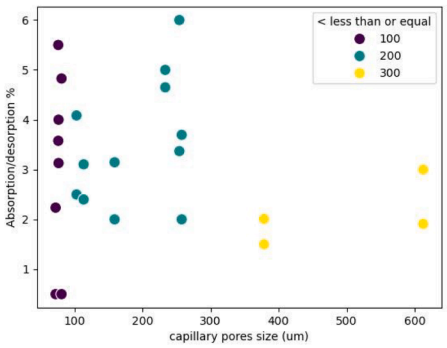
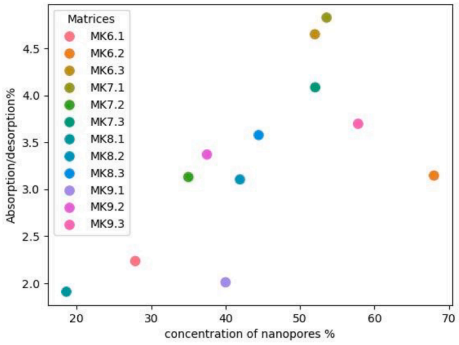


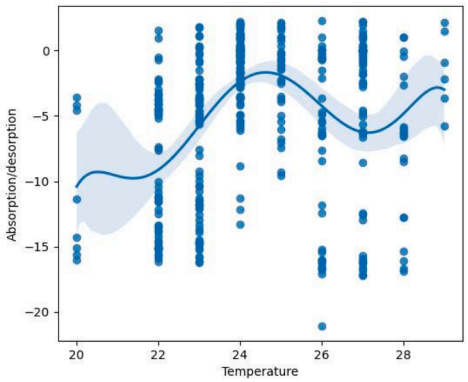
Fig. 8. Combinations (Mutual Information with Target).



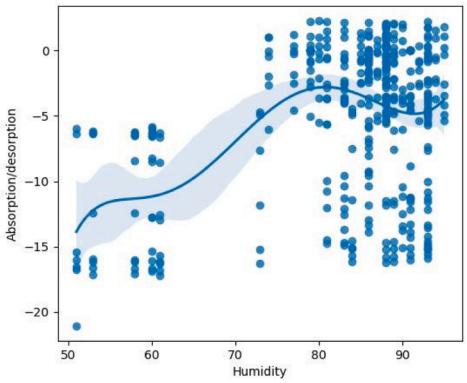
(a) Absorption/desorption vs capillary pores size



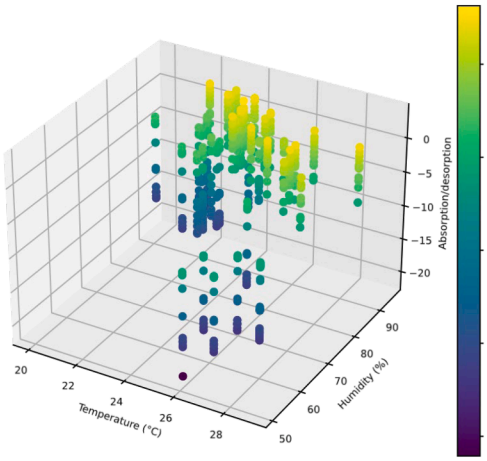
(b) Absorption/desorption vs concentration of nanopores



(c) Absorption/desorption vs Temperature



(d) Absorption/desorption vs Humidity

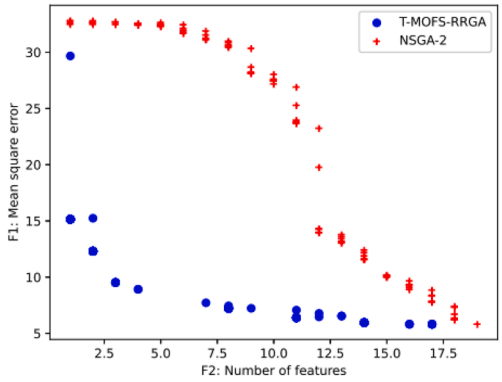


(e) Absorption/desorption vs Humidity and Temperature

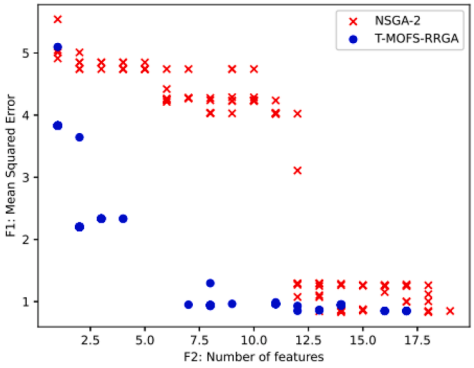
Fig. 9. Absorption/Desorption capacity of porous geopolymer.

5.2.2. Relationship between absorption/desorption, humidity and temperature

Data summarized into the Fig. 9 show that the hygroscopic behavior of the porous geopolymer is function of the relative humidity and the temperature (Figs. 9c and 9d). In fact in the normal room conditions ($T = 23\pm2\,^{\circ}C$), (RH=55±5 °C) porous geopolymers are having humidity in equilibrium in their pores. Fig. 9 evidences the role of direct factors affecting the hygroscopic behavior of porous geopolymers. In the Figs. 9a and 9b, it is observed that, the concentration of small porosity (nanometric and mesometric) allows high absorption of humidity. Making the hypothesis that micrometric pores should be as low as possible in concentration with role focused to the aeration of the pore network and improvement of the interconnectivity, important for absorption/desorption cycles. In the Figs. 9c and 9d, the role of direct



(a) T-MOFS-RRGA vs NSGA-II training set



(b) T-MOFS-RRGA vs NSGA-II test set

Fig. 10. Best pareto fronts on regression.

Table 9
IGD, SC and Number of solutions analysis for T-MOFS-RRGA.

IDG		Coverage(sc)		Number of solution	
T-MOFS-RRGA	**NSGA-2**	**sc(T-MOFS-RRGA, NSGA-2)**	**sc(NSGA-2, T-MOFS-RRGA)**	**T-MOFS-RRGA**	**NSGA-2**
0.1393	0.2121	**1/0**	0/0	**100**	**100**

factors indicates temperature and humidity govern the absorption/desorption. The absorption that is enhanced when temperature is low and humidity high. This explains the negative correlation observed between both factors in Fig. 6. Desorption is important with high RH (≥ 70 °C) and the raise of temperature. Situation perfectly comparable with the regulation of the thermal comfort in building environment. In a 3D projection (Fig. 9e), it is observed that low temperature and relatively low humidity correspond to situation where porous geopolymer remain with poor activity in term of absorption and desorption. In practice, corresponding to the condition in with the thermal comfort is insured. With the increase in temperature, the thermal comfort is better insured with the relative high humidity (Fig. 9e). Porous geopolymers appeared hence as ideal matrices for the control of the thermal comfort with efficient capacity even in extreme environment ($22 \leq T \leq 35$ °C).

5.3. Prediction of absorption/desorption capacity using T-MOFS-RRGA: pareto front analysis and study of convergence/divergence

Fig. 10 shows the best Pareto fronts obtained for NSGA-2 regression and T-MOFS-RRGA. It can be observed that there is a very good balance between intensification and diversification. Table 9 shows the IGD, coverage, and average number of solutions for 10 independent runs of T-MOFS-RRGA and NSGA-2. With a value of **0.1393** for the IGD of T-MOFS-RRGA, it can be seen that it converges better than NSGA-2 and covers the whole search space, since coverage (T-MOFS-RRGA, NSGA-2-R) = **1** means that all the solutions of T-MOFS-RRGA dominate the solutions of NSGA-2. Both algorithms used the same initial population size.

We found that at the level of 14, 16, and 17 variables, the errors become approximately constant and minimal (Fig. 10a). The three subsets of attributes having 14, 16, and 17 variables are used to evaluate the proposed method described in this work, under the different constraints defined initially (induction algorithm, error functions, etc.). It is observed that the minimal error is obtained for the subset of 16 variables using the gradient boosting regressor induction algorithm.

Fig. 11 illustrates the relevance of each feature using the gradient boosting regressor as the induction algorithm for T-MOFS-RRGA. Aluminum ranks first, followed by the Si_2O/Al_2O_3 ratio and hydrogen peroxide. It's noteworthy that aluminum and H_2O_2 serve the same role as pore-forming agents, and the use of one excludes the other. The Si_2O/Al_2O_3 ratio in the second position acts as the network former influencing the structure, microstructure and durability of porous geopolymer. The primary observation is that these parameters contribute to the matrix composition, influencing pore formation, tortuosity, thermal conductivity, pore diameter, and spacing. Although intrinsic parameters such as pore diameter, porosity, and tortuosity are important, they are influenced by the quantity of pore-forming agent. The set of features obtained from Tables in appendix section present the subsets of solutions for each of the methods described.

The importance of tortuosity, porosity, Si/Al ratio, pore space, thermal conductivity, and pore diameter indicates that intrinsic matrix properties are critical to hygroscopic efficiency. Fig. 12 shows the scatterplot between actual and predicted hygroscopic capacity. The predicted values were obtained by training the five algorithms on the 16 variables obtained with T-MOFSRRGA. It is noted that, gradient boosting regressor best approximates the difference between the predicted and experimental values.

5.4. Analyzing the effectiveness of classical feature selection methods

Starting from the idea that all the attributes found in the literature are important, and to test our hypothesis, several models were trained. Table A.10 and A.11 shows the test results for the various models. With LASSO regression, the coefficients of six features are different from zero. It is noted that it takes into account attributes related to the composition of the porous geopolymers, external attributes such as temperature and humidity, and the parameter governing the microstructure. Although LASSO doesn't perform better on induction algorithms, it does group one attribute per category and remove redundant features. Fig. 11b shows the relevance of these features to hygroscopic capacity. Selection using mutual information requires identifying the k best variables for the induction algorithm. In this work, k is unknown and need to be defined. For k ranging from one to the number of features, mutual information is run on the training set to obtain a subset, which is then evaluated by further training a RandomForestRegressor classifier that serves as the objective function. A value of k best is retained equivalent to the base case, meaning that all attributes are important. Fig. 11a shows the relevance of each feature to absorption/desorption capacity using mutual information. According to this graph, pore diameter is very important for predicting absorption/desorption capacity, while pore type is the least important. The drawback with this method for choosing

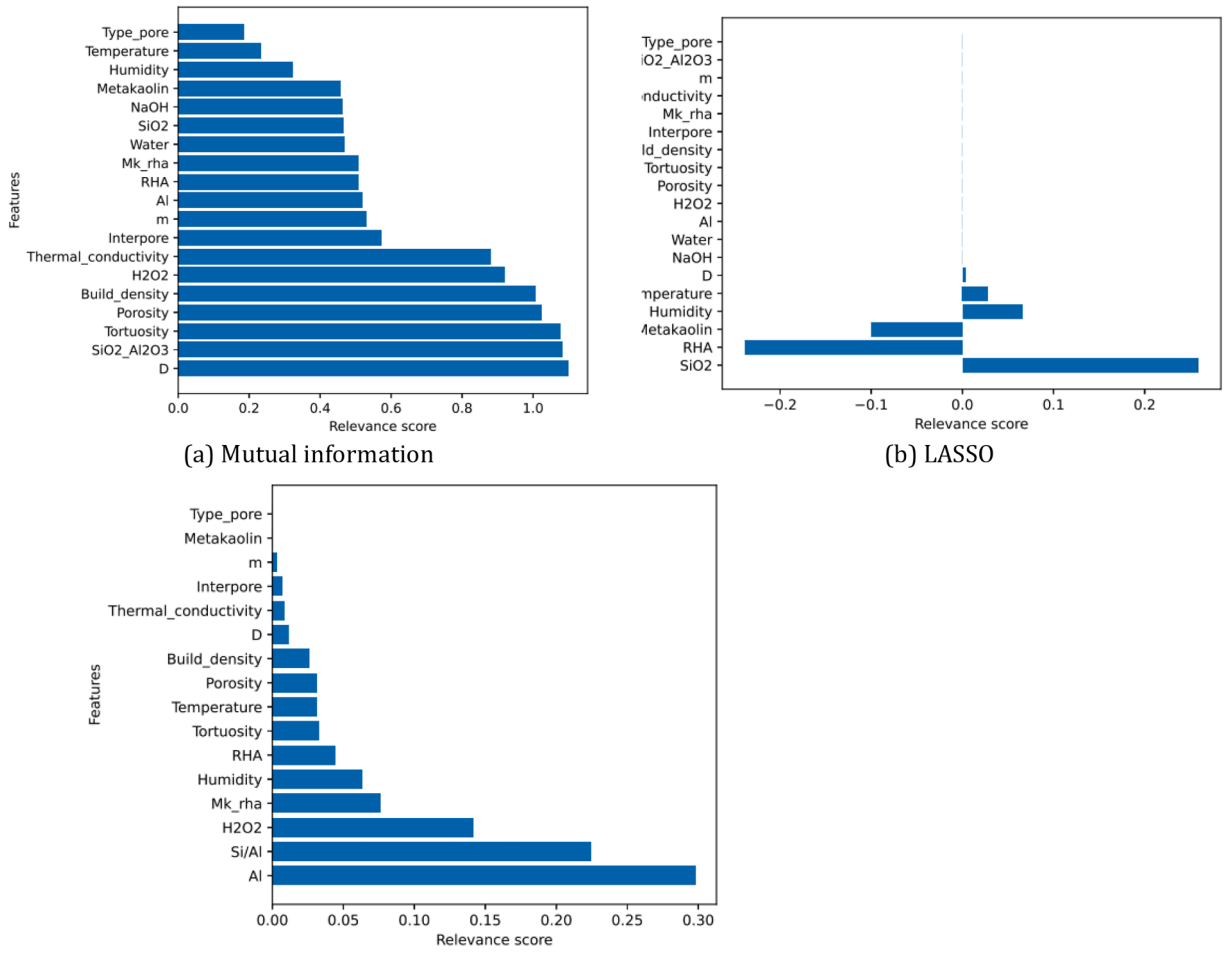


Fig. 11. Relevance of feature to hygroscopic capacity.

k variables is that mutual information returns output scores that are not normalized to quantify the degree of relevance on the same scale. With T-MOFS-RRGA, 14 is obtained as the size of the subset with the lowest learning error. We proposed to build all possible subsets of size 14 and evaluate them. We trained a RandomForestRegressor for each variable subset, and in the end, we retained the one with the minimum Mean Square Error, which is *0.8932*. The following subsets were selected: H_2O_2, porosity, density, interpore, m, S_iO_2/Al_2O_3, D(pore diameter), type pore, Humidity, Temperature, Metakaolin, RHA, *NaOh,* Na_2SiO_2 were selected.For Mutual Information (MI), the selected features are D (pore diameter), S_iO_2/Al_2O_3, tortuosity, porosity, density, H_2O_2, thermal conductivity, interpore, Al, m, RHA, mk rha, water, NaOh. From the first fourteen mutual information features, five have been replaced. Tortuosity, thermal conductivity, aluminum, water, and mk/rha ratio have been replaced by pore type, humidity, temperature, metakaolin, S_iO_2/Al_2O_3. After examining the diagram describing the correlation (Fig. 6), it is noticed that the inter-pore space is correlated with tortuosity and thermal conductivity. This justifies their absence. Secondly, the mk rha ratio is absent in the generated subset and replaced by metakaolin. Indeed, the mk/rha ratio is highly correlated according to Fig. 6. The addition of silicate in the generated subset and its absence in the mutual information subset indicate that this solution is needed to prepare the alkaline solution. So these 14 variables are equivalent. With SVM-RFE, the selected features are humidity, temperature, metakaolin, RHA, sodium silicate, Al, H_2O_2, tortuosity, density, interpore, mk rha, thermal conductivity, S_iO_2/Al_2O_3, type pore are obtained. It is noted that pore diameter, which is very important according to mutual information, is absent. Using GradientBoostingRegressor as a classifier to induce, the value of 0.921687 is obtained for RMSE, which is better than mutual information but still lower than the base case.

6. Conclusion

Porous geopolymers have been investigated for their effectiveness as hygroscopic materials for enhancing thermal comfort in building environments. Feature selection methods were used for predicting hygroscopic properties. After data collection and analysis, a set of algorithms were used to investigate the correlations between parameters governing the design and production of porous networks and the direct aspects of hygroscopy (temperature, relative humidity, etc.). From the results, the following conclusions were drawn:

1. Validation of the hybrid feature selection approach: The MOFS-RFGA method proved applicable in this context and was efficiently adapted into T-MOFS-RRGA, a multi-objective algorithm suited for regression focused hybrid methods.
2. Correlation of synthesis parameters with pore structure: A significant correlation exists between the SiO_2/Al_2O_3 ratio, porogen agent, and

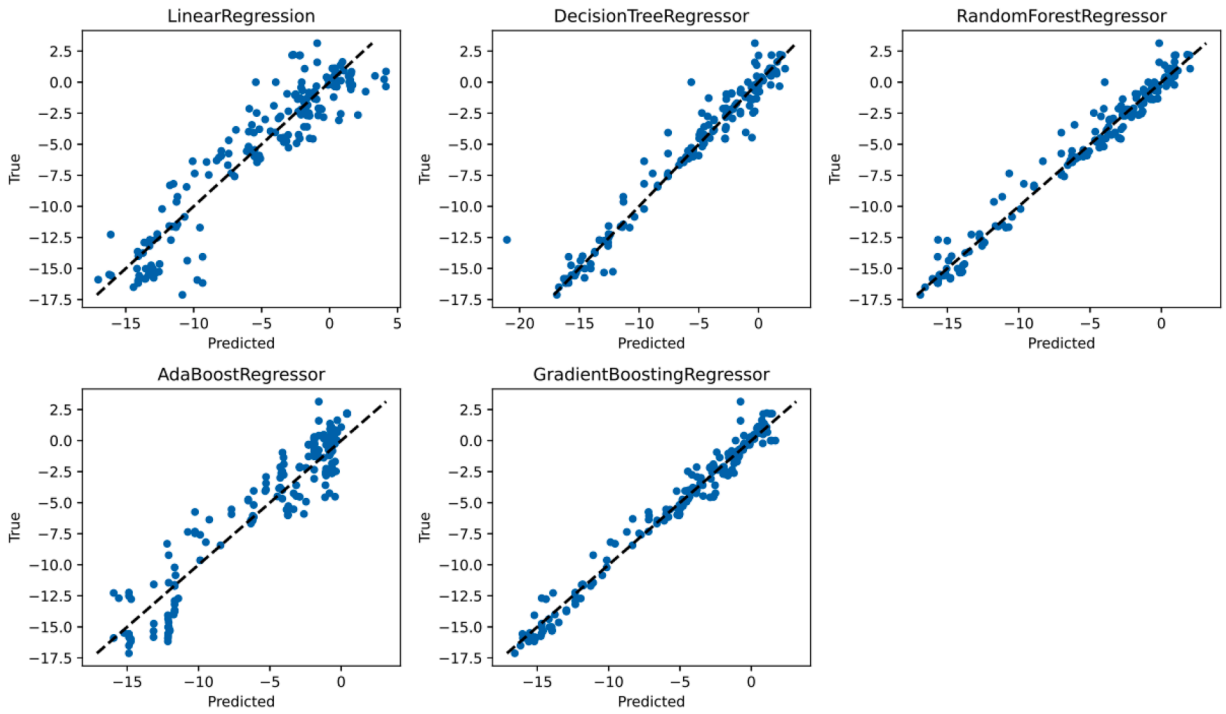


Fig. 12. Scatter plot between actual values and predicted values.

alkaline solution with the characteristics of the pore network (pore volume, size distribution, and connectivity).

3. Key features governing hygroscopic behavior: Sixteen features were identified as relevant to controlling the absorption/desorption capacity. Indirect features such as SiO_2/Al_2O_3 ratio, alkaline solution, and porogen agent affect hygroscopicity by governing direct pore features (volume, distribution, size, connectivity). External variables such as temperature and relative humidity also exert a direct influence on hygroscopic processes.
4. Performance of prediction algorithms: Among the machine learning algorithms tested using these features, GradientBoostingRegressor performed best, achieving RMSE = 0.8491, MSE = 0.72, and MAE = 0.6179. The adapted T-MOFS-RRGA algorithm further minimized error, confirming its robustness for this application.

To further advance this research, more robust learning algorithms beyond linear regression as objective function will be explored. Additionally, experimental validation under real laboratory conditions will be carried out to confirm the relevance and applicability of the selected features and predicted properties. This step is crucial to bridge the gap between data-driven predictions and their practical implementation in material engineering and building design.

CRediT authorship contribution statement

Lauraine Tiogning-Djiogue: Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Herman Tcheneghon Motcheyo:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elie Kamseu:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sylvie Rossignol:** Writing – review & editing, Supervision. **Cristina Leonelli:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors of this manuscript wishe to acknowledge the support of the European Union-OEACP (PRICNAC-EEPER: MD 2022).

The EEPER: MD team wishes to acknowledge Dr. Paulin Melatagia for his assistance and support.

Appendix A. Regression Performance of T-MOFS-RRGA Com-pared to Baseline and Traditional Feature Selection Techniques

Table A.10
Feature score for 16 variables.

Variable	score
Humidity	0.06296740781743355
Temperature	0.03140586666288773
Metakaolin	4.4444183943051475e-05
RHA	0.04405489757699458
Al	0.29789813988192454
H2O2	0.141502511987174
Porosity	0.03138688601492719
Tortuosity	0.03311687796847146
Build density	0.025847388861991328
Interpore	0.007314271847208835
Mk rha	0.07607028471285278
Thermal conductivity	0.008533885318715755
m	0.0034456137464866734
Si/Al	0.22472750096625452
D	0.011684022452734108
Type of pore	0.00001

Table A.11
Regression Performance of T-MOFS-RRGA Compared to Baseline and Traditional Feature Selection Techniques.

Methods	Induction algorithms	RMSE	MSE	MAE	Number of fea tures	Set obtained
Base case	LinearRegression	2.131170	4.541888	1.631190	19	{1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19 18,19}
	DecisionTreeRegressor	1.344463	1.807581	0.741664		
	RandomForestRegressor	0.931776	0.868206	0.649974		
	AdaBoostRegressor	1.935119	3.744686	1.649085		
	GradientBoostingRegress	jr0.849332	0.721364	0.617997		
LASSO	LinearRegression	3.126063	9.772271	2.502698	6	{1 2 3 4 6 1 3}
	DecisionTreeRegressor	1.499914	2.249742	0.826440		
	RandomForestRegressor	1.006200	1.012439	0.667336		
	AdaBoostRegressor	1.933699	3.739194	1.572794		
	GradientBoostingRegressor	1.117693	1.249238	0.832147		
Information Mutuelle	LinearRegression	2.441699	5.961892	1.841945	14	{13 7 12 11 17 9 15 19 10 18 4 14 8 15}
	DecisionTreeRegressor	1.290558	1.665540	0.748160		
	RandomForestRegressor	0.955922	0.913788	0.651702		
	AdaBoostRegressor	1.898723	3.605151	1.563066		
	GradientBoostingRegressor	0.926808	0.858973			
T-MOFS-RRGA	LinearRegression	2.131170	4.541888	1.631190	17	{1 2 3 4 8 10 9 11 12 17 19 14 15 18 7 13 16}
	DecisionTreeRegressor	1.253829	1.572086	0.714479		
	RandomForestRegressor	0.988373	0.976882	0.673321		
	AdaBoostRegressor	1.818240	3.305995	1.494418		
	GradientBoostingRegressor	0.849878	0.722293	0.619892		
T-MOFS-RRGA	LinearRegression	2.133768	4.552968	1.630897	16	{1 2 3 4 10 9 12 17 19 14 15 18 7 13 16}
	DecisionTreeRegressor	1.356159	1.839168	0.748779		
	RandomForestRegressor	0.979614	0.959644	0.668209		
	AdaBoostRegressor	1.928392	3.718697	1.615431		
	GradientBoostingRegressor	0.849102	0.720974	0.617997		
T-MOFS-RRGA	LinearRegression	2.244087	5.035925	1.702785	14	{1 2 3 4 8 11 17 19 14 15 7 13 16}
	DecisionTreeRegressor	1.377352	1.897097	0.783529		
	RandomForestRegressor	0.988214	0.976568	0.669736		
	AdaBoostRegressor	1.754176	3.077135	1.484259		
	GradientBoostingRegressor	0.936810	0.877614	0.686070		
RFE-SVM	LinearRegression	2.286517	5.228160	1.747265	14	{1 2 3 4 6 10 9 12 17 19 14 15 7 16}
	DecisionTreeRegressor	1.111506	1.235447	0.680107		
	RandomForestRegressor	0.948371	0.899408	0.652266		
	AdaBoostRegressor	1.950446	3.804238	1.621636		
	GradientBoostingRegressor	0.921687	0.849508	0.644594		
NSGA-II	LinearRegression	2.131170	4.541888	1.631190	19	{1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19 18,19}
	DecisionTreeRegressor	1.344463	1.807581	0.741664		
	RandomForestRegressor	0.931776	0.868206	0.649974		
	AdaBoostRegressor	1.935119	3.744686	1.649085		
	GradientBoostingRegressor	0.849332	0.721364	0.617997		

References

[1] Z.N. Ngouloure, E. Kamseu, L.M.B.A. Moungam, H. Tchakoute, L. Valentini, C. Leonelli, Design of porous geopolymers for hygrothermal applications: role of nano and meso porosity, Silicon 14 (15) (2022) 10045–10059.

[2] E. Kamseu, Z.N. Ngouloure, B. Nait-Ali, L. Valentini, S. Zekeng, S. Rossignol, C. Leonelli, Pore network and microstructure in the prediction of heat flux transport in sponge-like geopolymers for thermal insulation, J. Therm. Anal. Calorim. 147 (22) (2022) 12329–12344.

[3] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: introduction and review, J. Biomed. Inform. 85 (2018) 189–203.

[4] N. Dess`ı, B. Pes, L.M. Cannas, An Evolutionary Approach for Balancing Effectiveness and Representation Level in Gene Selection, 4, IGI Global, Hershey, PA, 2017, pp. 2557–2574, https://doi.org/10.4018/978-1-5225-17597.ch105.

[5] A. Faramarzi, M. Heidarinejad, B. Stephens, S. Mirjalili, Equilibrium optimizer: a novel optimization algorithm, Knowledge-Based Systems (Jan. 2019). doi:10.1016/j.knosys.2019.105190.
[6] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Appl. Intell. (2022) 1–39.
[7] K.-H. Chang, in: K.-H. Chang (Ed.), Chapter 19 - Multiobjective Optimization and Advanced Topics, Academic Press, Boston, 2015, pp. 1105–1173, https://doi.org/10.1016/B978-0-12-382038-9.000193. Ed.), e-Design.
[8] Y. Xue, H. Zhu, F. Neri, A feature selection approach based on nsga-ii with relieff, Appl. Soft. Comput. 134 (2023) 109987, https://doi.org/10.1016/j.asoc.2023.109987.
[9] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: nsga-ii, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.
[10] V. Cascione, D. Maskell, A. Shea, P. Walker, A review of moisture buffering capacity: from laboratory testing to full-scale measurement, Constr. Build. Mater. 200 (2019) 333–343, https://doi.org/10.1016/j.conbuildmat.2018.12.094.
[11] H. Gao, L. Liao, Y. Liang, X. Tang, H. Liu, L. Mei, G. Lv, L. Wang, Improvement of durability of porous perlite geopolymerbased thermal insulation material under hot and humid environment, Constr. Build. Mater. 313 (2021) 125417, https://doi.org/10.1016/j.conbuildmat.2021.125417.
[12] H. Gao, L. Liao, H. Liu, L. Mei, Z. Wang, D. Huang, G. Lv, G. Zhu, C. Wang, Optimization of thermal insulation performance of porous geopolymers under the guidance of thermal conductivity calculation, Ceram. Int. 46 (10) (2020) 16537–16547, https://doi.org/10.1016/j.ceramint.2020.03.221. Part B.
[13] H. Dai, H. Gao, B. Jiang, Q. Yang, X. Li, X. Guo, Z. Cheng, Y. Xiong, X. Li, X. Chen, J. Wu, L. Wang, Enhancement effect of basalt fiber on the foamy kaolinite-based composite thermal insulator, J. Build. Eng. 95 (2024) 110144, https://doi.org/10.1016/j.jobe.2024.110144.
[14] R. Feng, Q. Yang, H. Dai, M. Deng, H. Wang, C. Hou, Z. Cheng, H. Gao, Effect of glass fiber on the mechanical and thermal insulation performances of kaolinite-based thermal insulator, Case Stud. Constr. Mater. 21 (2024) e03879, https://doi.org/10.1016/j.cscm.2024.e03879.
[15] X. Lü, T. Lu, C. Kibert, Q. Zhang, M. Hughes, A novel methodology and new concept of structural dynamic moisture buffering for modeling building moisture dynamics, Build. Environ. 180 (2020) 106958, https://doi.org/10.1016/j.buildenv.2020.106958.
[16] Z.N. NGouloure, B. Nait-Ali, S. Zekeng, E. Kamseu, U. Melo, D. Smith, C. Leonelli, Recycled natural wastes in metakaolin based porous geopolymers for insulating applications, J. Build. Eng. 3 (2015) 58–69.
[17] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, IEEE, 2014, pp. 372–378.
[18] M. Verleysen, D. Fran¸cois, The curse of dimensionality in data mining and time series prediction, Eds., in: J. Cabestany, A. Prieto, F. Sandoval (Eds.), Computational Intelligence and Bioinspired Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 758–770.
[19] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, J. Saeed, A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction, Jo. Appl. Sci. Technol. Trends 1 (1) (2020) 56–70, https://doi.org/10.38094/jastt1224.
[20] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024, 40th-year commemorative issue.
[21] D. Theng, K.K. Bhoyar, Feature selection techniques for machine learning: a survey of more than two decades of research, Knowl. Inf. Syst. 66 (3) (2024) 1575–1637.
[22] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis, P. Pintelas, Feature selection for regression problems, in: Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece 2022, 2007.
[23] P. Langley, et al., Selection of relevant features in machine learning, in: Proceedings of the AAAI Fall symposium on relevance, California 184, 1994, pp. 245–271.
[24] D. Wolpert, W. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (1997) 67–82, https://doi.org/10.1109/4235.585893.
[25] K. Yu, W. Li, W. Xie, L. Wang, A hybrid feature-selection method based on mrmr and binary differential evolution for gene selection, Processes 12 (2) (2024), https://doi.org/10.3390/pr12020313.
[26] X. Deng, M. Li, S. Deng, L. Wang, Hybrid gene selection approach using xgboost and multi-objective genetic algorithm for cancer classification, Med. Biol. Eng. Comput. 60 (2022) 101751, https://doi.org/10.1007/s11517-021-02476-x.
[27] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, H.-S. Ye, A hybrid genetic algorithm with wrapper-embedded approaches for feature selection, IEEe Access. 6 (2018) 22863–22874, https://doi.org/10.1109/ACCESS.2018.2818682.
[28] M. Hammami, S. Bechikh, C.-C. Hung, L.B Said, A multi-objective hybrid filter-wrapper evolutionary approach for feature construction on high-dimensional data, 2018, IEEE Congr. Evol. Comput. (CEC) (2018) 1–8, https://doi.org/10.1109/CEC.2018.8477771.
[29] J. Li, T. Luo, B. Zhang, M. Chen, J. Zhou, Imoabc: an efficient multi-objective filter–wrapper hybrid approach for highdimensional feature selection, J. King Saud Univ. - Comput. Inf. Sci. 36 (9) (2024) 102205, https://doi.org/10.1016/j.jksuci.2024.102205.
[30] B. Zhao, D. Yang, H.R. Karimi, B. Zhou, S. Feng, G. Li, Filter-wrapper combined feature selection and adaboost-weighted broad learning system for transformer fault diagnosis under imbalanced samples, Neurocomputing 560 (2023) 126803, https://doi.org/10.1016/j.neucom.2023.126803.
[31] X. Wu, L. Wang, B. Chen, Z. Feng, Y. Qin, Q. Liu, Y. Liu, Multiobjective optimization of shield construction parameters based on random forests and nsga-ii, Adv. Eng. Inf. 54 (2022) 101751, https://doi.org/10.1016/j.aei.2022.101751.
[32] A.S. Ghareb, A.A. Bakar, A.R. Hamdan, Hybrid feature selection based on enhanced genetic algorithm for text categorization, Expert. Syst. Appl. 49 (C) (2016) 31–47, https://doi.org/10.1016/j.eswa.2015.12.004.
[33] Z. Li, Y. Wang, J. Zhang, H. Guan, Feature selection for indoor temperature prediction in large-space buildings based on transfer entropy and life cycle cost, Build. Environ. 243 (2023) 110722, https://doi.org/10.1016/j.buildenv.2023.110722.
[34] M. Robnik-Sikonja, I. Kononenko, Theoretical and empirical analysis of relieff and rrelieff, Mach. Learn. 53 (2003) 23–69.
[35] A. Kataria, M. Singh, A review of data classification using k-nearest neighbour algorithm, Int. J. Emerg. Technol. Adv. Eng. 3 (6) (2013) 354–360.
[36] B. Huang, B. Buckley, T.-M. Kechadi, Multi-objective feature selection by using nsga-ii for customer churn prediction in telecommunications, Expert. Syst. Appl. 37 (5) (2010) 3638–3646, https://doi.org/10.1016/j.eswa.2009.10.027. URL, https://www.sciencedirect.com/science/article/pii/S0957417409008938.
[37] Y. Xue, H. Zhu, F. Neri, A feature selection approach based on nsga-ii with relieff, Appl. Soft. Comput. 134 (2023) 109987, https://doi.org/10.1016/j.asoc.2023.109987.
[38] E. Landi, V. Medri, E. Papa, J. Dedecek, P. Klein, P. Benito, A. Vaccari, Alkali-bonded ceramics with hierarchical tailored porosity, Appl. Clay. Sci. 73 (2013) 56–64.
[39] C. Galle, Effect of drying on cement-based materials pore structure as identified by mercury intrusion porosimetry: a comparative study between oven-, vacuum-, and freeze-drying, Cem. Concr. Res. 31 (10) (2001) 1467–1477.
[40] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B: Stat. Methodol. 58 (1) (1996) 267–288.
[41] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[42] B.C. Ross, Mutual information between discrete and continuous data sets, PLoS. One 9 (2) (2014) e87357.
[43] Y. Zhang, J. Liu, W. Shen, A review of ensemble learning algorithms used in remote sensing applications, Appl. Sci. 12 (17) (2022), https://doi.org/10.3390/app12178654.
[44] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation, PeerJ Comput. Sci. 7 (2021) e623, https://doi.org/10.7717/peerj-cs.623.
[45] H. Ishibuchi, R. Imada, Y. Setoguchi, Y. Nojima, Reference point specification in inverted generational distance for triangular linear pareto front, IEEE Trans. Evol. Comput. 22 (6) (2018) 961–975, https://doi.org/10.1109/TEVC.2017.2776226.
[46] E. Zitzler, L. Thiele, M. Laumanns, C. Fonseca, V. da Fonseca, Performance assessment of multiobjective optimizers: an analysis and review, IEEE Trans. Evol. Comput. 7 (2) (2003) 117–132, https://doi.org/10.1109/TEVC.2003.810758.