Journal of Building Engineering 86 (2024) 108766

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Full length article

Machine learning techniques for diagrid building design: Architectural–Structural correlations with feature selection and data augmentation

Pooyan Kazemi *, Alireza Entezami, Aldo Ghisi

Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

ARTICLE INFO

Keywords:
AI in building design
Architectural feature selection
Advanced data augmentation
High-rise buildings
Generative architectural forms
Design informed by structural insights

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) techniques are transforming building engineering. This work goes through the critical role of architectural parameters in influencing the structural responses of tall buildings, with a special focus on diagrid structures. The main aim of this study is to demonstrate how ML can improve the early design phase of diagrid buildings. Using a small, initially collected data set, enhanced through data augmentation, the classification of diagrid buildings in terms of design feasibility is investigated. This study identifies key architectural and structural parameters, employing various filter and wrapper methods for feature selection. The results show that our methods are effective in producing high-quality synthetic data, maintaining stable learning accuracies, and establishing accurate and robust relationships between architectural parameters and structural responses in diagrid buildings. These insights are crucial for facilitating more effective design processes in the realm of high-rise diagrid building design.

1. Introduction

The rapid rise of AI in recent years has opened new possibilities for architectural and structural designs of tall buildings. In the early phase of building design, conflicts can arise because visually appealing forms for architects may not align with structurally effective design choices. The key challenge lies in accurately estimating the structural behavior. This difficulty is particularly pronounced in the case of seismic loading, even moderate, and represents a significant challenge in the design process, which is generally unfavorable for all parties involved, including architects, engineers and clients. With the current availability of computing power, integrated design, and advanced tools capable of analyzing large data sets [1], it is feasible to propose an improved design approach [2,3].

Tools like parametric design software [4] allow specialists to create numerous geometries for high-rise building using computer-aided design (CAD). This information can then be passed onto structural codes, which can generate extensive data to describe the response, even under complex loading conditions. AI tools, based on ML [5,6] or even deep learning (DL) [7–9], can process these results, identifying correlations between input and output variables with an efficiency that surpasses human capabilities.

While this AI and ML-driven approach is promising, it is important to recognize that numerous practical details, both in the design process and in training the ML models, still necessitate human intervention for optimal results. Only through the correct application and interpretation of these procedures it is really possible to gain an advantage.

* Corresponding author.
E-mail addresses: seyedpooyan.kazemi@polimi.it (P. Kazemi), alireza.entezami@polimi.it (A. Entezami), aldo.ghisi@polimi.it (A. Ghisi).

https://doi.org/10.1016/j.jobe.2024.108766
Received 14 August 2023; Received in revised form 4 January 2024; Accepted 6 February 2024
Available online 15 February 2024

This study focuses on the class of tall buildings with outer diagrids [10,11]: for this category, the elegance of transferring the lateral-load bearing function to elements on the façade (the diagrids), thus freeing the interior from most of the columns, is combined with the need to ensure structural efficiency. Historically, numerous studies have explored diagrid structures, focusing on optimization (e.g. [12,13]), and decision-making models (e.g. [14,15]). Further literature [16–18] provides a comprehensive view on the assessment of key variables in diagrid structures, also in the nonlinear regime. However, when the attention is directed towards the field of building structural design, ML or DL have only recently found widespread application, even if their adoption is increasing rapidly. ML tools are utilized to both predict and assess the structural performance, exploiting numerical (e.g. [19]) and experimental data (e.g. [20]), often in combination with observations from structural health monitoring [21]. In particular, the use of ML tools to evaluate seismic behavior in various types of standard buildings has received attention, because the success of the ML approach to recognize hidden patterns in other nonlinear problems in several fields of physics and engineering [22–24] has been of inspiration to deal with the difficulties of the structural evaluation of the building behavior.

Studies [25,26] have investigated an extensive database for reinforced concrete structures subject to dozens of seismic excitations using various ML algorithms, demonstrating the significant role of the algorithms' parameters, or hyperparameters, and their influence on the classification of buildings in terms of seismic damage. In [27] a set of tens of cross laminated timber structures has been subjected to horizontal acceleration records and the outcomes of the numerical simulations have been analyzed through several classical ML regression algorithms to predict the building drift ratios. In the study [28], a ML approach has been utilized to identify the physical versus non-physical modes in a modal analysis of a 195-meter tall building, yielding the first five modes and damping ratios and recognizing their stochastic distribution. DL techniques, such as those outlined in [29], which exploit combinations of neural networks (NNs), have proven capable of predict the capacity curves of an eight-floor industrial building subjected to ground motion. Study [30] has determined the features influencing the economic losses due to earthquakes in a steel building using an extreme gradient boosting model. Several ML techniques have been employed in [31] to assess seismic damage in a 2D five-story, five-bay moment frame with a nonlinear behavior. After creating a database comprising 468 moment steel frames subjected to 240 ground motions, researchers have managed to rank eight ML methods and found good performance.

In the present work, instead, the support for the designers' decisions in the early design phase, possibly including also human-driven architectural sensitivities, is pursued, because the ML procedure aims to categorize the buildings into feasible choices combining both form and structural efficiency. This approach has been less investigated in the past, see e.g. the authors' work in [32]; in particular, it is evidently important to understand which geometrical and architectural choices have the most relevant effect on the structural efficiency, since these choices are also the typical object of the architectural expertise. This specific issue is also considered here within the sub-problem of feature selection (f.s.), a standard step in any ML procedure: in the following, several options are considered for this task and their effectiveness is discussed.

In recent years, surrogate models (SMs) have emerged as a pivotal tool in computational engineering, offering a pragmatic solution for approximating complex systems with a high degree of accuracy while significantly reducing computational costs [33]. These models are essentially simplified representations or emulations of more intricate systems and relate, *in a statistical sense*, input to output data, obtained from complex system simulations. SMs have found widespread application across various domains, particularly in scenarios where direct numerical simulations are either too time-consuming or computationally expensive. Within the realm of building engineering, SMs stand as a bridge between detailed, data-intensive simulations and efficient, real-time decision-making. The use of SMs in the context of diagrid building design efficiently approximates complex structural behaviors and interactions, which are otherwise challenging to capture in full-scale simulations [34]. These models facilitate the extraction of meaningful insights from large data sets, enabling a deeper understanding of the intricate architectural–structural correlations in tall buildings. By incorporating surrogate modeling approaches, our study aligns with contemporary computational strategies, striving to optimize the design process through intelligent data synthesis and reduction techniques. The application of these models in our research not only underscores the practicality of SMs in architectural and structural design but also showcases their potential in enhancing the accuracy and efficiency of ML algorithms deployed for building design optimization.

In our previous work [32], we carried out the groundwork by generating data, conducting structural analysis, and implementing classification for structural responses. This current study is indeed based on these foundations but significantly extends our earlier efforts by going deeper into the difference between f.s. and data augmentation (d.a.). Unlike [32], where the focus was primarily on data generation and basic classification, here we have introduced advanced methodologies for f.s. using filter, wrapper, and exhaustive methods, and augment the data set using the Gaussian copula approach. These enhancements not only provide a more distinctive understanding of the structural responses of diagrid buildings but also offer a novel approach to handling limited data sets in structural engineering research.

While this study offers valuable insights into f.s. and d.a. in the context of diagrid building design, it is important to note that the analysis considered in this study is based on a linear elastic model assuming a weak seismic event. This approach limits the direct applicability of the findings of this study to buildings in regions with high seismicity. The primary aim of this paper is to explore the methodology of f.s. and d.a. rather than exhaustive structural analysis under varied seismic conditions. Furthermore, it is vital to acknowledge that the focus on early-stage design necessitates linking architectural parameters to structural behavior through a simplified structural simulation. It should be noted that while a nonlinear approach is imperative for the structural design of a single tall building, the objective of assessing the structural behavior of multiple tall buildings in the initial design phase justifies the adoption of a simplified methodology. However, it is acknowledged that this limitation may affect the generalization of our insights regarding the efficiency of a *specific* diagrid building.

In Section 2 the engineering problem is summarized; the main geometrical input variables and the structural output quantities are listed as possible features or responses useful for the ML procedure. The aforementioned important step of feature/response

Table 1
Acronyms used in the paper.

AI	Artificial intelligence
AR	Aspect ratio
CAD	Computer aided design
CDF	Cumulative distribution function
d.a.	Data augmentation
DL	Deep learning
EDW	Expected design weight
FE	Finite element
f.s.	Feature selection
KS	Kolmogorov–Smirnov
LOOCV	Leave-one-out cross-validation
ML	Machine learning
NN	Neural network
PCC	Pearson correlation coefficient
RF	Random forest
SM	Surrogate model
TGA	Total gross area

selection is described in Section 3. Then, in Section 4, since the data set associated to numerically simulated complex buildings can be often relatively small for the selected features/responses because of the computational burden, a d.a. technique is investigated, aiming to provide a reasonable number of form options respectful of the stochastic properties of the initial data set. To validate the procedure and specifically to verify the quality of the augmented data, we discuss the results of a classification problem in Section 5. This problem uses a machine learning algorithm to determine the most feasible building designs based on previously identified responses. Table 1 provides a collection of acronyms used in this paper.

2. Methodology

2.1. The architectural and engineering problem

In this work, a tall building with an outer diagrid is considered as the base model, providing data that is subsequently fed into the ML procedure.

A typical exemplary model is depicted in Fig. 1 and includes: (i) an outer network of inclined beams working mainly axially in tension or in compression, (ii) an internal core obtained by adding a given a specific number of columns, and (iii) flat slabs connecting the diagrids and the core columns at each floor. The building form is characterized by its top and bottom floor geometry and a curvilinear transformation developing along the building height. Other geometrical variables include the aspect ratio, i.e. the ratio between the height and the maximum dimension in the plan, the building height, and the inclination angle for the diagrids. These geometrical variables constitute the essential input data for the parametric design phase and enable designers to create a significant number of alternative building forms. After discretization through beam and plate finite elements (FEs), loads and boundary conditions are imposed: namely, the nodes at the building base are totally restrained, a dead load of 4.50 kN/m^2 and a live load of 2.0 kN/m^2 are applied to all floors, a seismic force correspondent to a moderate earthquake is applied along a prescribed direction as a set of distributed horizontal forces at each floor. Details about the statically equivalent seismic load can be found in [32]. Also in the same reference, the choices for the beam and column cross-sections and slab thicknesses are described, as well as other details related to the structural modeling of the considered tall buildings.

As a weak seismic event is assumed [32], for each model a linear elastic analysis is carried out and its results are collected as responses for the ML data set.

2.2. Features and responses; workflow

In what follows, under the framework of supervised learning [5], distinctions between features and responses are made.

The former refer to geometrical or mass properties defining a building and they are typically used by parametric design tools, such as Grasshopper™, to create three-dimensional CAD models. From there, after discretization, the FE analysis is conducted. Among the features, architectural choices are identified, as the majority of the geometric properties of the buildings correspond to architectural parameters, such as the top and bottom plan geometry, the building height, and the total gross area of the cross sections. Additionally, certain features are directly associated with the structural modeling aspect, such as the diagrid degree at the top and bottom of the building, total mass (weight), and the position of the center of gravity. Within the ML framework features serve as the input of the training algorithm, while responses (also known as labels) act as the output from training.

Responses encompass instead variables derived from the numerical analyses, offering insights into the structural behavior of the models, including the top story displacement, the maximum utilization ratio across all structural members, the expected design weight, and their combination.

The comprehensive initial list of features and responses considered in this work is reported in Tables 2–3, respectively, together with the correspondent data type and range of values; their exact meaning is described in the Tables' last columns.

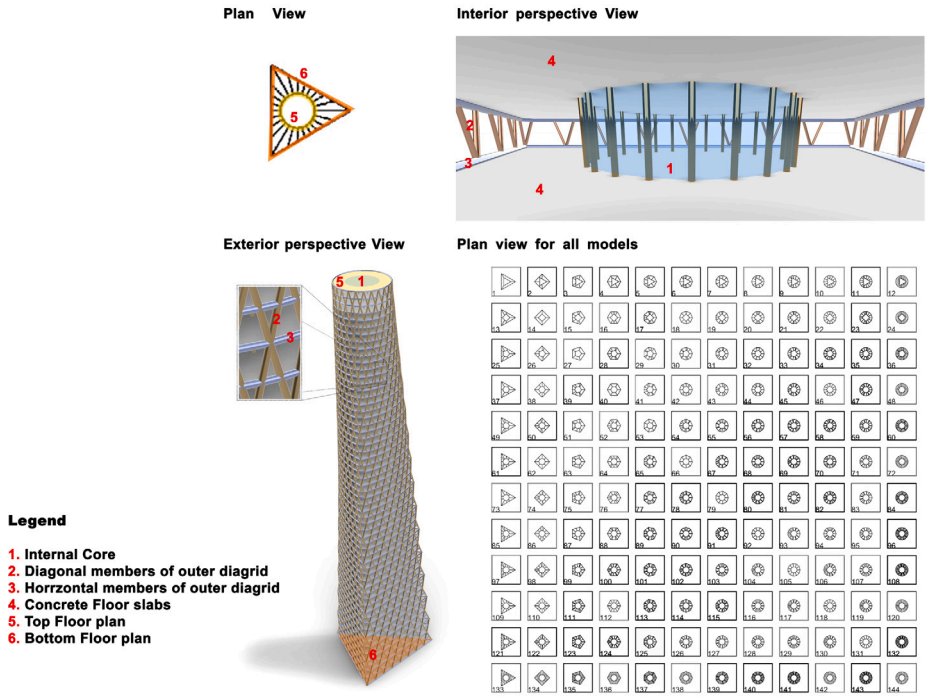


Fig. 1. Building models. (Top) plan and interior view with the outer diagrid members, the internal core columns, and the concrete slabs. (Bottom left) perspective view of an exemplary model. (Bottom right) top and bottom plan geometries for each of the 144 generated model, with corresponding model numbers.

It should be mentioned that all continuous data have been normalized to facilitate ML numerical convergence and to enhance the comprehension of the results. This normalization has been conducted using the Z-score method, translating features into a common scale with a mean of zero and standard deviation of one. Outliers have been identified using the inter-quartile range method and have been addressed accordingly to ensure they do not bias the learning process. In addition, no missing values have been reported in the data set.

The discreteness of data such as the number of plan sides inherently impacts the f.s. process, as these features exhibit non-linear relationships with the responses. The supervised learning models have been hence tuned to accommodate the discrete nature of such variables, ensuring that the predictive accuracy is not compromised.

Given the abundance of characteristics available as input during model generation and the extensive data gathered from numerical simulations, selecting the variables actually useful for the ML procedure is vital in the whole procedure. The f.s. step in a ML process aims to maximize relevance and minimize redundancy. Reducing the number of features has several benefits: simplification enhances interpretation, reduces misleading complexities, saves training time, and avoids the inclusion of non-essential features that might adversely affect learning metrics, such as accuracy. In summary, f.s. counters the curse of dimensionality, i.e. the difficulties in creating, managing and understanding an excessively large data set. Lastly, feature and response selection enhances generalization by reducing overfitting.

The entire procedure proposed in this work is outlined in the following steps.

1. Redundant and irrelevant data are identified through various techniques (Section 3).
2. Additional synthetic data is generated to expand the original data set (Section 4). Obtaining quality data to feed the ML procedure is a significant aspect of ML, and, although parametric design followed by numerical simulation is now convenient, it still involves complicated computations in terms of resources and data collection. For practical reasons, the number of actual computations or “direct” analyses is often restricted; instead, d.a. techniques can be efficiently utilized.
3. Both the original and augmented data sets are employed for a classification task (Section 5).

3. Feature and response selection

Feature and response selection is a pivotal process in ML, involving the identification of the most pertinent variables for model training. This step enhances model performance, mitigates complexity, and helps in avoiding overfitting. In the AI context, the various methods adopted for feature/response selection have been traditionally collected into broader families, as reported in the existing literature [35].

Table 2
List of features for the studied data set.

ID	Features	Data type	Range/values	Meaning
F1	Number of top plan sides	Discrete, integer	[3,13]	Integer number of the edges of the polygon used for the tall building top plan
F2	Number of bottom plan sides	Discrete, integer	[3,13]	Integer number of the edges of the polygon used for the tall building bottom plan
F3	Total gross area (TGA)	Continuous, float	$[6.95{\cdot}10^4, 7.06{\cdot}10^4]$ m^2	Sum of floor plan area for all stories of the tall building
F4	Height	Discrete, integer	[232, 236] m	Net height of the building, equal to the number of floors times 4 meters
F5	Aspect ratio (AR)	Continuous, float	[3.92, 5.24]	Ratio between the building height and the maximum width of bottom floor plan
F6	Diagrid degree at top	Continuous, float	[41.45, 61.18]°	Inclination angle, with respect to the horizontal plane, of diagrids at the top plan of the building
F7	Diagrid degree at bottom	Continuous, float	[61.01, 75.70]°	Inclination angle, with respect to the horizontal plane, of diagrids at the top plan of the building
F8	Diagrid degree, average	Continuous, float	[51.90, 67.25]°	Average of the inclination angles, with respect to the horizontal plane, of all diagrids in the model
F9	Total façade area	Continuous, float	$[2.84{\cdot}10^4, 3.63{\cdot}10^4]$ m^2	Side area of the building, including diagrid area and holes between diagrids
F10	Total amount of diagrids	Discrete, integer	[4522, 8568]	Number of vertical, horizontal and inclined diagrids
F11	Total length of diagrid members	Continuous, float	$[1.87{\cdot}10^4, 2.59{\cdot}10^4]$ m	Sum of the lengths of all diagrids
F12	Total mass	Continuous, float	$[7.74{\cdot}10^3, 1.05{\cdot}10^4]$ tons	Total mass in the model, accounting for all the finite elements included
F13	Height of the center of gravity	Continuous, float	[104.55, 110.31] m	Position of the center of gravity with respect to the bottom plan

Table 3
List of responses for the studied data set.

ID	Responses	Data type	Range/values	Meaning
R1	Displacement of the top story	Continuous, float	[0.87, 1.14] m	Maximum horizontal displacement of the top story
R2	Maximum utilization, compression	Continuous, float	[−0.89, 0.35]	Ratio between current and allowable stress in compression
R3	Maximum utilization, tension	Continuous, float	[0.41, 1.13]	Ratio between current and allowable stress in tension
R4	Maximum normal force, compression	Continuous, float	[−11.7, −4.62] kN	Highest compressive axial force for the finite elements in the model
R5	Maximum normal force, tension	Continuous, float	[4.67, 10.3] kN	Highest tensile axial force for the finite elements in the model
R6	Maximum normal force, absolute	Continuous, float	[4.67, 10.3] kN	Highest axial force, compressive or tensile, for the finite elements in the model, in absolute value
R7	Expected design weight (EDW)	Continuous, float	$[1.49{\cdot}10^3, 1.78{\cdot}10^3]$ tons	The total weight of a structure, calculated based on the initially assigned mass for each component as per their given utilization ratio
R8	Elastic energy	Continuous, float	[2.39, 9.95] kJ	Total elastic energy for all the finite elements in the model
R9	Displacement/total mass	Continuous, float	$[8.53{\cdot}10^{-5}, 1.22{\cdot}10^{-4}]$ m/ton	Ratio between the displacement of the top story (R1) and the total mass of the model (F12)
R10	Displacement/EDW	Continuous, float	$[5.56{\cdot}10^{-4}, 6.48{\cdot}10^{-4}]$ m/ton	Ratio between R1 and R7
R11	EDW/AR	Continuous, float	$[2.95{\cdot}10^2, 4.46{\cdot}10^2]$ ton	Ratio between R7 and the aspect ratio (F5)
R12	EDW/TGA	Continuous, float	$[2.12{\cdot}10^{-2}, 2.53{\cdot}10^{-2}]$ ton/m^2	Ratio between R7 and the total gross area (F3)
R13	Total mass/TGA	Continuous, float	$[1.10{\cdot}10^{-1}, 1.49{\cdot}10^{-1}]$ ton/m^2	Ratio between F12 and the total gross area (F3)

Within the family of filter methods, selections are made based on purely statistical measures, independent of the algorithm adopted in the subsequent ML step: in this study, the Pearson Correlation Coefficient (PCC) [36] has been utilized. The PCC is a statistical measure that identifies linear correlations between variables, offering insights into their direct relationships.

In the wrapper method family, on the other hand, the selection is carried out by comparing the performances of combinations of features/responses during the learning process. Wrapper methods, including forward selection, backward elimination [37,38], and the exhaustive method [35], adopt a sequential approach. They iteratively add or remove features, assessing their impact on the model's performance. This allows for a distinctive selection process, considering the complex interplay between variables. These methods are particularly crucial in this study, where understanding the complex relationships between architectural features and structural responses is key.

A third family, the embedded methods, which combine characteristics of the previous two families, is not considered in this work.

3.1. Response selection

To ascertain the optimal responses, it is advisable to first explore their statistical correlation. In fact, it might not be necessary to train the model using two responses that exhibit a high correlation.

Various metrics can be employed to examine the statistical correlation between two sampled variables; in this study, as previously mentioned, the PCC metric is considered. This coefficient ranges between −1 and +1; a value close to one indicates a strong positive correlation, with negative values representing an inverse correlation (where one variable increases as the other decreases or vice versa); a value near zero signifies there is a minimal correlation. More precisely, the PCC between two random variables X and Y is defined as $\rho_{XY} = \mathrm{cov}(X,Y)/(\sigma_X\,\sigma_Y)$, that is, the ratio of the covariance of the two variables to the product of their standard deviations. The advantage of using this method resides mostly in its simplicity, because it is based on the empirical mean and covariance; the PCC is often considered an index of a linear correlation between the two random variables. An analysis of the correlation across all responses is conducted, with the results being visualized as a heatmap, shown in Fig. 2: pairs of responses exhibiting a correlation higher than 85% are identified, and one of the correlated responses is excluded [39]. The optimal responses identified through this process, and detailed in Table 4 (top line), encompass key structural parameters, including the displacement of the top story, maximum utilization (compression), defined as the ratio between maximum and allowable compressive stress in the FEs; EDW, comprising the sum of the product of each member's utilization and weight [32]; and the ratios displacement/EDW and EDW/AR. Therefore, only five out of thirteen responses are retained. An arbitrary threshold of 85% has been arbitrarily chosen to balance the number of responses that are labeled as irrelevant with those that are considered. There is no specific correlation threshold provided in existing literature that should be adopted, especially in studies using the PCC method. Different studies have utilized various threshold values such as 0.40, 0.50, 0.70, 0.80, 0.85, 0.90, and 0.95. Consequently, the threshold selection process appears to be somewhat arbitrary. In other research works [40,41], a set of thresholds ranging from 0.50 to 0.95 is defined, and the threshold leading to the highest accuracy is selected based on the training performance. This approach bears resemblance to a sequential f.s. method, as described in the next Section 3.2.3, where the primary criterion is achieving the highest accuracy. The rationale here is to balance the number of included responses against the excluded ones without dealing with significant computational costs, as 13 responses could be theoretically combined with 13 features for all the data set. The need for this human intervention, which is clearly necessary and relevant, is carried out for the responses, but it is handled differently later in case of f.s., as described in Section 3.2.3.

It is acknowledged that the essence of supervised learning is the reflection of the task's objectives through inherently defined responses. Therefore, an in-depth analysis has been carried out to evaluate the trade-offs between model simplification and potential accuracy loss when omitting responses with high correlation coefficients. This analysis aims to ascertain the acceptability of the error induced and its implications for the predictive accuracy of the model. Numerical simulations, comparing the model's performance with both the full set and the reduced set of responses, have been conducted. The outcomes, presented in Appendix B, confirm that while the error introduced by exclusion is non-negligible, the efficiency gains in terms of computational complexity and model interpretability justify the approach. The results demonstrate that a good degree of accuracy is retained in the reduced model, thereby confirming the suitability of the adopted response selection methodology for this specific application domain.

3.2. Feature selection

F.s. is a well-established topic in machine learning and is applied across various fields [35,42–45]. However, comparative studies focusing on its application to building structural design remain rare up to date.

In this Section, three different approaches are considered with an increasing level of complexity. First, the features are studied as an isolated set, similar to what has been done for the responses, by using the PCC. Secondly, the features are considered together with the responses, and the set is again evaluated through the PCC. Third, the relationships between features and responses are investigated through wrapper methods, namely the forward selection, the backward elimination and the exhaustive method [37,38]. In applying a wrapper method, the influence of each feature combined with every response is tested. The rationale behind the proposed approach, which gradually exploits more complex methods, is to check the robustness of the selection process.

3.2.1. Feature selection through the PCC

As outlined earlier, the first step to investigate the correlation between features adopts the same methodology outlined in Section 3.1 for the responses. The PCC is utilized to quantify the independence of each variable, with a threshold of 85% chosen to determine highly correlated features; consequently, one of the features exceeding this threshold will be eliminated from further consideration.

A heatmap, presented in Fig. 3, visually illustrates the correlation patterns among the features. Utilizing this approach, the final selection comprises eight critical structural features out of an initial thirteen: number of top or bottom plan sides; total gross area; building height; aspect ratio (AR); diagrid degree at the top; diagrid degree at the bottom; total façade area.

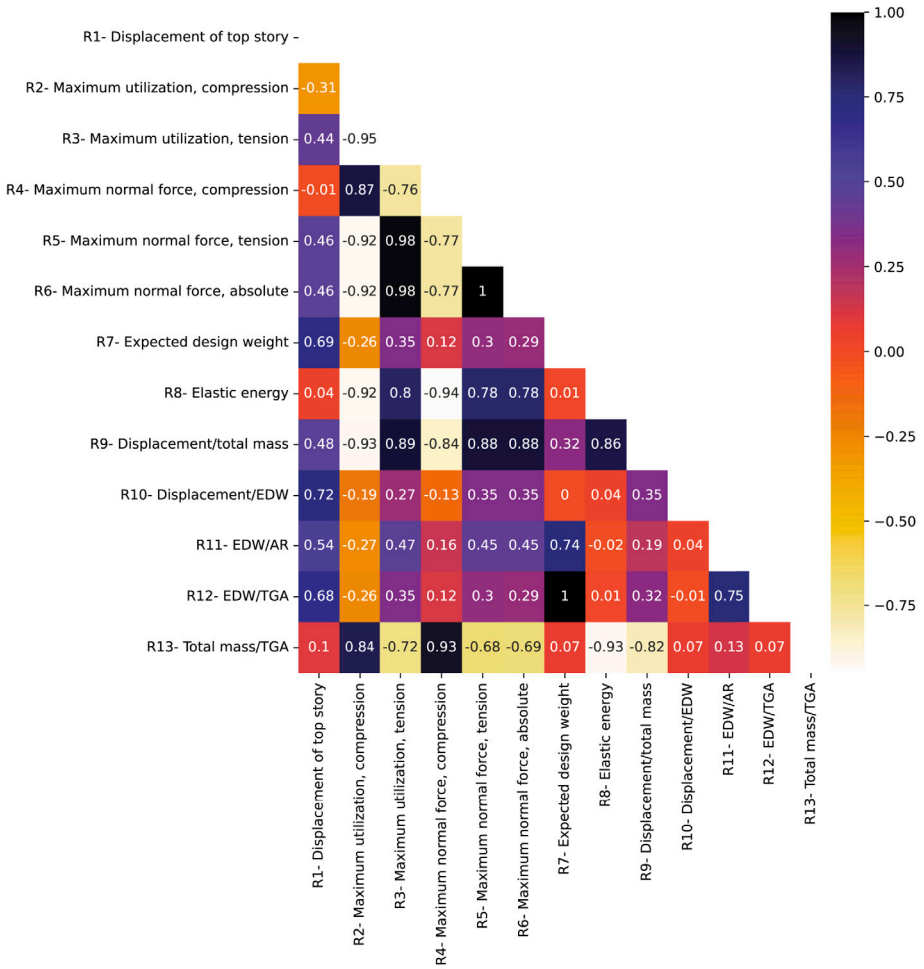


Fig. 2. Response correlation heatmap, computed using the PCC. Darker colors indicate a high positive correlation between the responses; lighter colors indicate a high inverse correlation. The gradient of red colors represents very low correlations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.2. Features selection by studying the relationship between features and responses

In this second scenario, the PCC is used to measure the correlation between the features and the five selected responses. The principle is that, if a feature demonstrates a weak correlation with a response, it is considered irrelevant for the learning process. Consequently, for each of the five responses selected in Section 3.1, the complete set of N=13 features is considered, and the correlations between the 14 variables (13 plus one response) are computed. Subsequently, the average correlation for each feature is calculated. Features with average correlations exceeding 25% are considered; they are presented in Table 4, first column. Once again, the determination of this threshold is conducted in an arbitrary manner. The underlying rationale is to retain features with high average correlation across all responses while eliminating those with low correlation, which indicates their lack of relevance to the responses. Moreover, this step is preliminary with respect to a more advanced f.s. shown in Section 3.2.3.

Through this streamlined approach, a total of seven features would be selected for further investigation, ranked 1–7 in the last column appearing in Table 4.

The presence of a correlation between the number of bottom plan sides and the responses, while the number of top plan sides exhibits no significant correlation, poses an intriguing question. In simpler terms, variations in the number of bottom plan sides have a direct impact on the responses. In contrast, increasing the number of top plan sides does not lead to either a corresponding increase or decrease in the responses. This disparity arises from the structural design of the models, even more because the plan cross section is tapered along the building height: the bottom plans hold greater influence for a cantilever-like structure, to which the global behavior of the tall building can be approximated. The building base plan section in fact entails the higher stress, induced by a bending moment and by the shear force, while the top plan section sustains a very low bending moment.

The correlation between the number of top/bottom plan sides and a representative response, namely the displacement of the top floor, is depicted in Fig. 4. While in Fig. 4a a correlation can be inferred, specifically the displacement decreases as the number

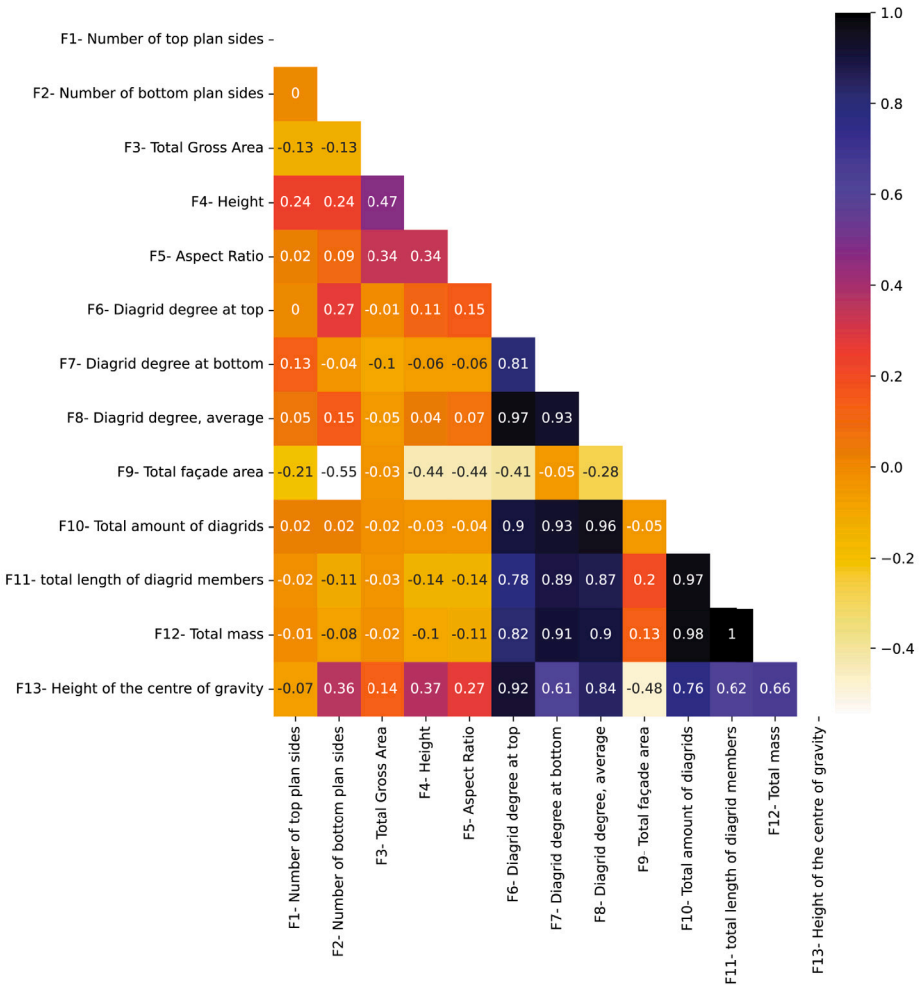


Fig. 3. Feature correlation heatmap based on PCC. High positive correlations are indicated by dark colors, high negative correlations by light colors, and low correlations by values in the red gradient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Correlation between features and responses with the Pearson metric.

	Displacement of top story	Maximum utilization, compression	EDW	Displacement / EDW	EDW/AR	Average	Rank
Total façade area	0.88	0.30	0.65	0.60	0.68	0.62	1
Center of gravity	0.48	0.87	0.35	0.33	0.40	0.49	2
Diagrid degree at top	0.43	0.96	0.30	0.30	0.28	0.46	3
Number of bottom plan sides	0.61	0.19	0.06	0.80	0.14	0.36	4
Diagrid degree, average	0.27	0.95	0.19	0.18	0.16	0.35	5
Aspect ratio	0.22	0.18	0.35	0.04	0.89	0.33	6
Total length of diagrid members	0.17	0.81	0.11	0.12	0.17	0.28	7
Total mass	0.10	0.84	0.07	0.08	0.12	0.24	8
Total amount of diagrids	0.06	0.91	0.06	0.03	0.00	0.21	9
Diagrid degree at bottom	0.00	0.83	0.01	0.01	0.04	0.18	10
Total gross area	0.19	0.00	0.02	0.24	0.23	0.14	11
Height	0.12	0.04	0.10	0.07	0.33	0.13	12
Number of top plan sides	0.02	0.03	0.03	0.00	0.00	0.02	13

of the bottom plan sides increases, for the number of top plan sides, as shown in Fig. 4b, there is no recognizable pattern and thus there is no influence of this feature on the displacement of the top story.

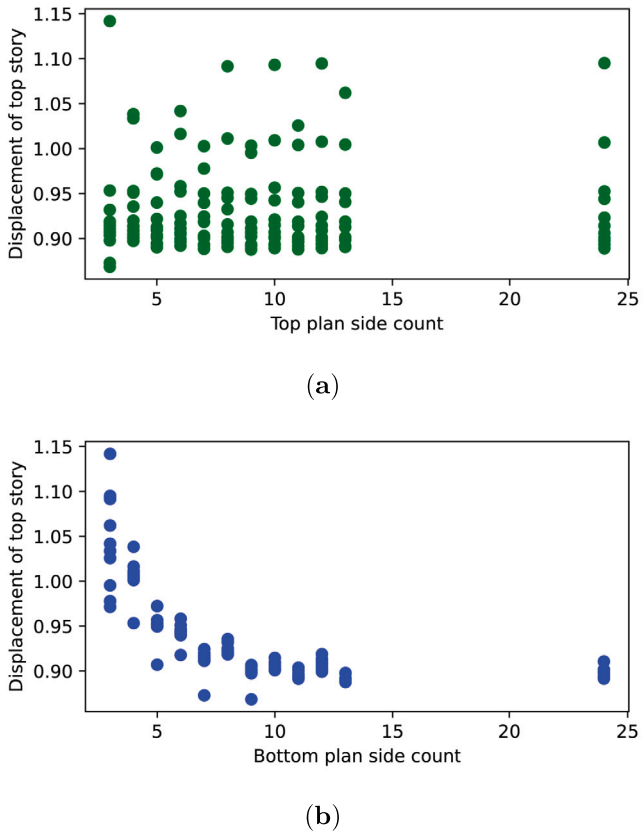


Fig. 4. Relationship between the number of (a) top and (b) bottom plan sides with respect to the displacement of the top story for the considered building models.

3.2.3. Feature selection using wrapper methods

A sequential f.s. procedure, adopted for the family of wrapper methods, is regarded as one of the most effective approaches [46–49]. It assesses the suitability of each feature during the learning process and determines whether to retain or discard it on the basis of its impact on the learning performance metrics, such as the accuracy or the mean squared error. Unlike other methods, such as the previously shown filter methods, which rely only on correlations, a sequential method directly affects the learning process, resulting in an enhanced effectiveness.

Subsequently, the forward selection, the backward elimination and the exhaustive selection methods are compared.

The forward selection method begins with an empty model; then, the most beneficial features are added iteratively one by one, to identify which ones yield the best performance according to a given learning metric.

The backward elimination method starts instead with the complete set of features and iteratively drops the ones leading to worst performances, until the desired number of features is attained.

3.2.4. Machine learning for feature selection within the framework of wrapper methods

In this work, ML is used both for f.s. and as a classification tool. In this Section, the methodology employed in the former case is described. The latter usage is instead shown further ahead, in Section 5.

The f.s. process begins with the PCC to eliminate redundant features, followed by more advanced wrapper methods using the random forest (RF) classifier. The main aim of this approach is to identify the most important architectural and structural features for diagrid buildings. The 4-fold cross-validation strategy is employed to ensure model robustness, particularly important given the data set's limited size. Then, the RF classifier is chosen for its effectiveness in high-dimensional data management and resistance to overfitting. Its use in our study is based on its proven superiority in comparative analyses and consistent performance with default hyperparameters.

The Mlxtends Python library [50] is utilized for these methods, commonly referred to as sequential approaches. These techniques aim to balance model performance and complexity by selecting the features that most significantly impact the metric used in the learning phase.

Since these f.s. techniques employ ML algorithms, some details of the ML implementation will be discussed here. The selection of the desired features involves an arbitrary parameter, designated as n_f, initially set to five in a first exploratory phase which

identifies potentially suitable features. In the subsequent phase, rather than being constrained to a fixed number of features, an optimal feature subset, without any predetermined number, is searched. Furthermore, to mitigate the risk of overfitting, a 4-fold cross-validation strategy [51] is implemented. It is worthwhile to note that k-fold cross-validation involves dividing the data into k subsets (folds) and then repeating the learning process k times. Each iteration entails applying the learning algorithm to the training set, which includes k-1 folds, while the last, remaining fold poses as the testing set. The ultimate evaluation metric is obtained by averaging the results across the k folds. A 4-fold cross-validation has been pursued because of the limited size of the data set: a low k ensures that each fold contains a sufficient number of data points. Moreover, a smaller k results in a larger training set for every fold, which could be beneficial while working with limited data. The RF classifier, using default hyperparameters, is utilized in this study due to its reputed superiority over other state-of-the-art ML methods [52,53]. As an ensemble learning method, RF is recognized for its ability to handle high-dimensional data sets and typically excels in classification and regression tasks. Additionally, compared to other methods such as NNs, RFs are less prone to overfitting, and their output is easier to interpret, an essential aspect for practical applications in building engineering. Moreover, a compelling rationale for adopting the RF algorithm stems from the inherent data imbalance in the original data set. RF adeptly handles imbalanced data by generating multiple models on diverse subsets and subsequently aggregating their predictions through averaging. This approach effectively mitigates the risk of model bias and adeptly addresses the challenge posed by the imbalanced nature of the data [54–59]. Finally, unlike other algorithms, such as the support vector machine, this classifier does not require extensive parameter tuning and is more computationally efficient, making it a practical choice for this research study. The selection of RF as the classifier for this study has also been justified by previous research findings: in [32], the authors presented a comparative evaluation of six classifiers, namely k-nearest neighbor, support vector machine, naive Bayes, an ensemble method, decision tree, and discriminant analysis, concluding that, for this database, the ensemble family of classifiers exhibited very good performance. To ensure reproducibility, a fixed random state has been chosen due to the stochastic nature of the algorithm. Ultimately, the decision to exclude a NN classifier is motivated by the data set limited size, a circumstance that presents significant challenges in terms of training and generalization capabilities for NNs [60,61].

Five labels (responses), which qualify the model's performance (refer to [32]), are assigned using two different approaches: (i) an equal number of data points is present in each class within the range of the response values (equal number of observations per class), and (ii) the range of the response values is divided into equal bins for each class (equal length per class). It is observed that the second approach yields superior learning performance, and thus it is selected for further analysis, see Table 9.

To enhance the performance of the ML algorithm, all data values are normalized, as it is widely recommended for improving algorithm understanding and performance [62–64].

The features identified by both the sequential methods are highlighted in blue in Table 5; the results for the forward/backward methods are collected in columns 2–5.

Considering that five responses are selected (refer to 3.1), the f.s. process is repeated for each response individually by the forward selection and backward elimination. In all cases, at least three common features are consistently identified by both the forward selection and backward elimination methods, with an accuracy exceeding 80%, see Table 5, columns 2–5. The accuracies achieved by both approaches are comparable.

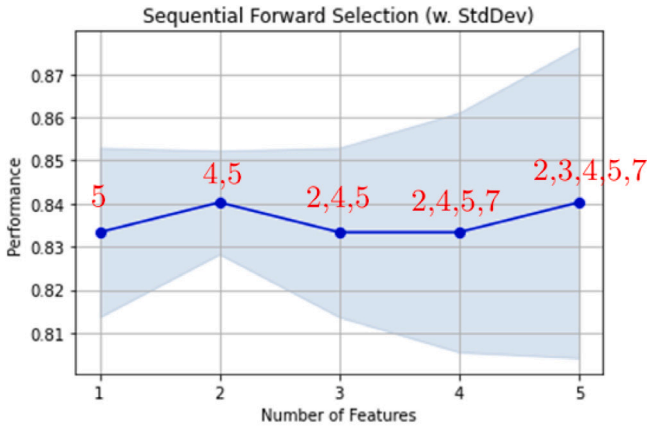
An important question arises regarding the rationale behind selecting five features. To address this question, as anticipated, a new procedure, adopting the same selection methods but with an unspecified number of desired features, is carried out; the algorithm is executed aiming to determine the optimal number of features, corresponding to the maximum accuracy, and for this reason it is labeled as n_f=best in the following. The results are presented in Table 5, columns 6–7. The n_f=best method explores various possibilities, thus demonstrating that the previous forward or backward approach is sufficient: the number of selected features ranges from 3 to 5 across different responses and it is the same for the forward and the backward methods; for this reason, only one column (column 6, labeled "B/F sel. n_f=best") is provided in Table 5.

The process of f.s. for a sample response, EDW/displacement, is visualized in Fig. 5, showcasing both the forward selection and backward elimination approach. In the same figure, the features that are added or eliminated (in this latter case their ID is crossed) at every step are also indicated in red color. In both methods, the performance of the learning process demonstrates improvement with respect to the starting point. The backward elimination, Fig. 5b, starting with the complete set of 13 features, explores a greater range of possibilities compared to the forward selection, Fig. 5a. The accuracy of the considered response, EDW/displacement, along with a 95% confidence interval, is represented as a light blue range in the graphs. Notably, the forward selection method, Fig. 5a, exhibits accuracy ranging from 83% with a single selected feature to 84% with five selected features. Conversely, the backward selection method, Fig. 5b, starts at 73% with all features and reaches 84% with five selected features.

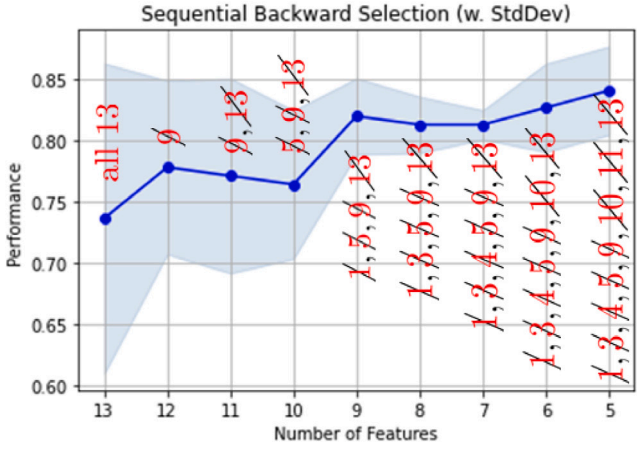
3.2.5. Exhaustive selection with best-selected features

Another alternative is the exhaustive sequential method, which explores all possible feature combinations [35]. Although computationally expensive, this method provides a more comprehensive analysis as it considers every scenario: in our study, with 13 features there is a total of $\sum_{i=1}^{13} \binom{13}{i} = 8191$ possible combinations.

One can specify a desired range for the minimum and the maximum number of features to explore; however, a minimum of 1 and a maximum of 13 have been chosen in this study by purpose, to explore all possibilities. For each response, the selected features, along with their corresponding accuracies, are presented in Table 5 in the last columns 8–9. When the algorithm selects five features, the accuracy is similar to what found for the previous sequential f.s. methods.



(a)



(b)

Fig. 5. A priori five f.s. (n_f=5). Performance of the algorithm for a sample response (EDW/displacement) in terms of accuracy and confidence interval. (a) forward selection, (b) backward elimination. In red color: features IDs (a) added or (b) eliminated at every step (IDs are listed in Tables 2–3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Comparison between the proposed wrapper feature selection methods

A comparison has been drawn between the aforementioned feature wrapper methods, namely the sequential f.s. methods and the exhaustive method, see Table 5. While the accuracy is mostly the same if compared to the backward elimination method with n_f=best, the exhaustive method guarantees the highest achievable accuracy. Notably, for the "displacement of the top story" response, the backward selection with n_f=best yields the same accuracy as the exhaustive method. However, the selected features vary, with the first method selecting five features and the second picking only three. This is an exception: despite this difference, the accuracy of both approaches only varies at the 5th decimal place. Moreover, if the accuracy of the backward method with n_f=best matches that of the exhaustive method for other responses, the selected features remain the same.

For instance, the learning accuracy for all 8191 combinations related to the EDW/AR response is depicted in Fig. 6. The y-axis of the graph represents the accuracy, while the x-axis corresponds to different scenarios with varying feature combinations. The 95% confidence interval is depicted using a light blue shading. A vertical red line is used to demarcate the boundaries that indicate the number of selected features within each scenario. In fact, scenarios 0 to 12 include the initial 13 scenarios, illustrating the performance of the ML algorithm with the selection of a single feature. The subsequent region, comprising scenarios 13 to 90, demonstrates the ML algorithm's performance when two features are chosen from the total pool of features. This pattern extends to other regions, each showcasing the ML algorithm's performance for selecting an increasing number of features. The most extensive range pertains to selecting 6 or 7 features out of the total 13, resulting in a substantial 1716 scenarios. The best performance,

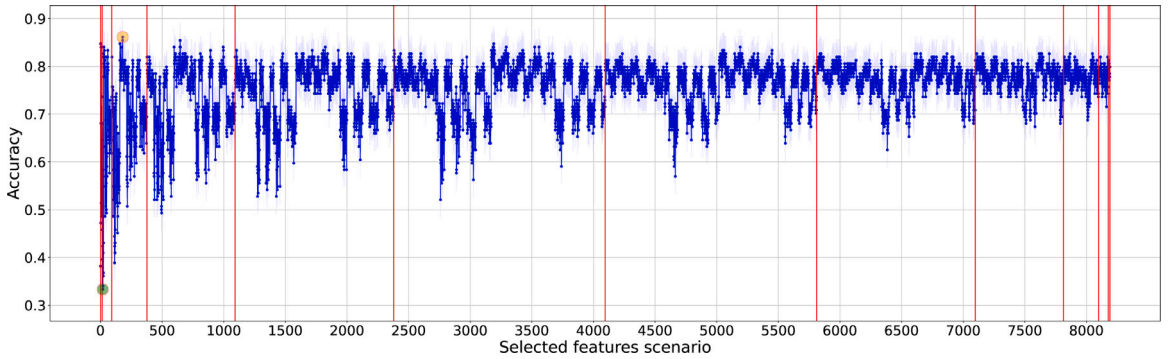


Fig. 6. The achieved accuracy of all scenarios with the exhaustive selection method for all 8191 possible feature combinations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

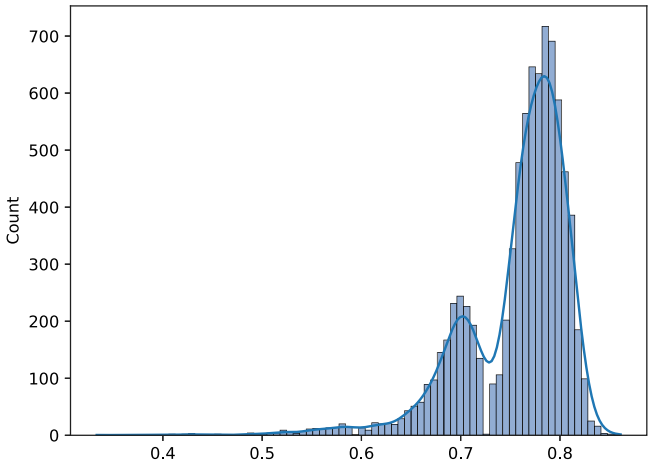


Fig. 7. The histogram plot of the kernel Gaussian probability distribution for the accuracy values among all scenarios. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

highlighted with an orange circle, occurs at scenario 180 within region 3, where the ML algorithm selects three features, namely "total amount of diagrids", "number of bottom plan sides" and "aspect ratio", achieving an accuracy value of 0.861. Conversely, the worst performance, represented by a green circle, takes place at scenario 18 of region 2, where the ML algorithm selects two features, namely "diagrid degree, bottom" and "number of top plan sides", yielding an accuracy value of 0.333. Overall, the graph illustrates that, as the number of features increases, the minimum accuracy within each region also increases, and the variation in accuracy range generally decreases. The mean and mode accuracy across all scenarios are 0.756 and 0.785, respectively. Furthermore, the histogram plot of all accuracy values is depicted in Fig. 7, where the kernel Gaussian probability distribution is also shown with a superimposed blue line.

In conclusion, while the exhaustive method unsurprisingly requires the highest computational effort, it yields the best features for the learning process. To decide on the features to retain, those common across all responses should be considered: the "number of bottom plan sides" is actually found in all cases, while the "diagrid degree at the bottom" is repeated four out of five times, and the remaining features are repeated a few times. Since these features are not highly similar, the optimal would be to use different sets of features for each response. The final list of selected features adopted in the rest of the paper corresponds to the exhaustive method (column 8 in Table 5). It is noteworthy that the "number of bottom plan sides" is selected by both the exhaustive method and the filter method (guided by PCC metric), which confirms the significance of this feature, while the "number of top plan sides" is deemed irrelevant by all wrapper methods. Even if the filter methods, based on the PCC metric and described in the Sections Section 3.2, point out some of the features recognized also by the most effective wrapper f.s. methods, they do not account for the best combinations.

4. Data augmentation

Given the limited size of the data set studied, which originally included 144 data points, it is imperative to use d.a. techniques to expand the data set, increase the stability of the learning process, and ensure the validity of the results.

Table 5
Comparison between the wrapper f.s. selection methods adopted in this work. Common features selected in the forward/backward methods are in blue color.

Response	Forward sel.(n_f=5)	Score	Backward elim. (n_f=5)	Score	B/F sel. (n_f=best)	Score	Exhaustive sel.	Score
Displacement of the top story	Number of top plan sides N. of bottom plan sides Height Diagrid degree average Total amount of diagrids	0.896	Number of top plan sides N. of bottom plan sides Height Diagrid degree, top Diagrid degree, bottom	0.910	Number of top plan sides N. of bottom plan sides Height Diagrid degree at the top Diagrid degree, bottom	0.910	N. of bottom plan sides Height Diagrid degree, bottom Total length of diagrids	0.910
Max utilization, compression	Height Aspect ratio Diagrid degree at the top Total amount of diagrids Height of center of gravity	0.861	N. of bottom plan sides Height Diagrid degree at the top Total amount of diagrids Height of center of gravity	0.868	N. of bottom plan sides Diagrid degree at the top Diagrid degree, bottom Total amount of diagrids	0.875	N. of bottom plan sides Diagrid degree at the top Diagrid degree, bottom Total amount of diagrids	0.875
EDW	N. of bottom plan sides Aspect ratio Diagrid degree, bottom Total façade area Height of center of gravity	0.875	N. of bottom plan sides Aspect ratio Diagrid degree at the top Total façade area Total length of diagrids	0.854	N. of bottom plan sides Aspect ratio Diagrid degree at the top Diagrid degree, bottom Total length of diagrids	0.868	N. of bottom plan sides Aspect ratio Diagrid degree at the top Diagrid degree, bottom Total façade area Total amount of diagrids	0.882
Displacement/EDW	N. of bottom plan sides Total gross area Height Aspect ratio Diagrid degree, bottom	0.840	N. of bottom plan sides Aspect ratio Diagrid degree, bottom Diagrid degree average Total mass	0.840	N. of bottom plan sides Aspect ratio Diagrid degree, bottom Total length of diagrids	0.875	N. of bottom plan sides Aspect ratio Diagrid degree, bottom Total length of diagrids	0.875
EDW/AR	N. of bottom plan sides Total gross area Height Aspect ratio Total amount of diagrids	0.819	N. of bottom plan sides Total gross area Height Aspect ratio Diagrid degree, bottom	0.826	N. of bottom plan sides Aspect ratio Total amount of diagrids	0.861	N. of bottom plan sides Aspect ratio Total amount of diagrids	0.861

D.a. in machine learning refers to the technique of generating new, synthetic data points from the existing dataset to artificially expand it. This approach is particularly beneficial for small datasets, as it enhances the model's ability to generalize [65,66].

While numerous augmentation algorithms exist for tabular data, the majority cater to continuous domains. However, when dealing with tabular data sets that include integers or labeled data, as in the present case, the choice of algorithms becomes more restrictive. The Gaussian copula algorithm from the synthetic data vault Python library [67] is exploited in this work. It constructs a multivariate Gaussian distribution to simulate the joint distribution of various features. It transforms the marginal distribution of each feature into a Gaussian distribution, followed by sampling from this multivariate distribution to generate new data points [68]. This method ensures that the synthetic data closely mirrors the statistical properties and correlations present in the original dataset, making it an effective tool for expanding data sets in structural design studies.

To implement the d.a. algorithm, first it is necessary to define the metadata of the tabular data set, including the list of the variables, as well as the type and sub-type of each variable. Subsequently, the desired constraints should be specified, such as the minimum values for the number of top/bottom plan sides, set to a minimum of three, and the requirement for the total height of the building, set to be an increment of the floor-to-floor height (4 m). Additionally, the algorithm enforces the generated data to remain within the boundary of the original data set.

It is worth noting that the values for the number of top/bottom plan sides exhibit a hiatus: they increase regularly from 3 to 12, then jump to 24 (a high number of sides originally intended to approximate a circular shape). Since the augmentation algorithm generates values that do not exist in the original data set in the large hiatus 12–24, a possible source of misbehavior would be generated. To deal with this issue, a "cleaned" data set (11×11 models instead of 12×12) is considered, by excluding the 24-sided polygon as the top or bottom plan.

The cleaned data set is then used for augmentation. The augmented data set is chosen to be ten times larger than the original one, resulting in 1200 data points. Different types of distributions, such as Gaussian, gamma, beta, student-T, Gaussian-kernel density estimation, and truncated-Gaussian, are tested for each column of the data set, by trial and error, evaluating the quality of the augmented data for each parameter and in the learning process.

Subsequently, the Gaussian copula algorithm follows a process where the values of each column are transformed, first by converting them into their respective cumulative distribution function (CDF) values based on their marginal distribution. These transformed values are then subjected to an inverse CDF transformation using a standard normal distribution. The algorithm proceeds to learn the correlations among the newly generated random variables; next, sampling is performed from a multivariate standard normal distribution, taking into account the learned correlations. Finally, the sampled values undergo a reversal process, where their standard normal CDF is computed, followed by the application of the inverse CDF corresponding to their respective marginal distributions [68,69].

The augmented data demonstrates a very high quality: the column shape score reaches 90.57% and the column pair trend score achieves 97.87%; the overall quality score, defined as the average of these two score, is equal to 94.22%. The column shape score is based on the Kolmogorov–Smirnov (KS) complement metric, which evaluates the similarity between the original and synthesized data CDFs, each represented by a column of values. The resulting disparity is confined to a numerical range from 0 to 1, since the KS statistic is subtracted from 1.

The column pair trend score is instead based on correlation similarity. Firstly, the correlation between each pair of columns in the original data is computed; then, the correlations are determined for each column pair in the augmented data. Correlation

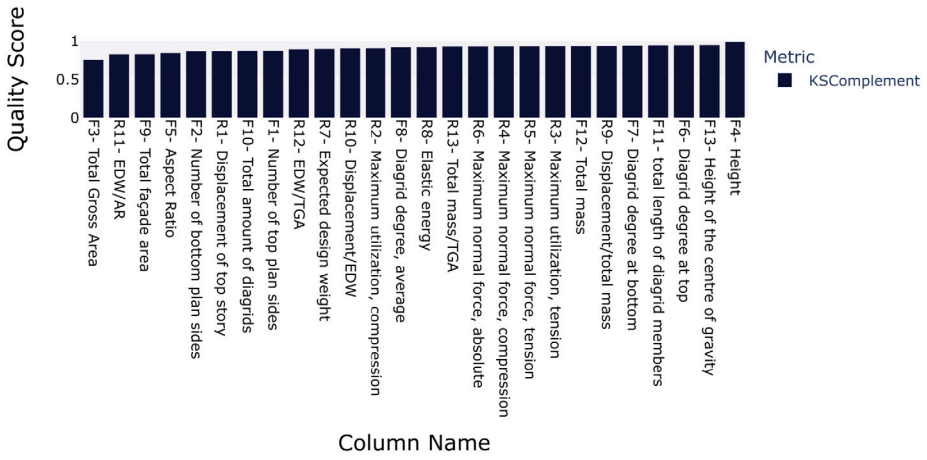


Fig. 8. Bar chart depicting the column shape score of the augmented data using the KS complement metric.

calculations can be based on either the PCC or other indices. The correlation similarity is finally defined as half of the absolute difference between the two correlations (augmented and original data) of the same column pair; this value is then subtracted from 1. The resulting score ranges between 0 and 1; a value close to one signifies that the correlation between column pairs in the augmented data is accurately preserved.

A bar chart representing the column shapes score is shown in Fig. 8. To assess the quality of column pairs in both the augmented and original data sets, the correlation similarity metric based on the PCC is employed, similar to what described in Section 3.1. In this case, a high correlation is positive, because it would indicate that the augmented data preserves the information derived from the original data set: therefore, the results indicating an average score of 98% for the column pairs, see Fig. 9, emphasize the process soundness. The presented heatmap exhibits symmetry, and it visualizes highly correlated column pairs using a green gradient, which indicates correlations near 1. Conversely, poorly correlated column pairs are depicted with a red gradient, suggesting correlations close to zero. The most substantial correlation between column pairs is observed between “maximum normal force, tension“ and “maximum utilization, tension”, with a correlation coefficient of 1. On the other hand, the weakest correlation among column pairs is found between “EDW to AR“ and “total façade area”, with a correlation coefficient of 0.786. Another confirmation of the preservation of column correlations between the original and augmented data sets is visible when a side-by-side comparison of the correlation matrices is conducted, as shown in the Fig. 10, for the original and augmented data set, respectively. Also this heatmap exhibits obviously symmetry; nearly all column pair correlations demonstrate a high degree of similarity. Remarkably, the correlation matrices appear very similar, reaffirming again the high quality of the synthetic data.

Finally, the comparative analysis of the augmented vs original CDFs can be observed for all the features and responses in the Figs. B.12–B.15.

It is important noting that the d.a. algorithm performs optimally when the complete set of features and responses is provided; a reduced data set, considering only the selected features, shows instead a lower accuracy. After the augmentation process, the selected responses and features used in the following ML step are therefore extracted from the augmented data set, resulting in a refined data set for further analysis.

5. Machine learning applications in original and synthetic data sets

The architectural design process often involves categorizing design options into feasible and non-feasible categories, or according to efficiency levels. Hence, from the decision-making viewpoint, classification provides a structured way to make these distinctions clear and actionable. On the other hand, taking the nature of responses/data into account, many of the responses in our study, such as structural responses and efficiency metrics, are inherently continuous, so alternatively a regression analysis could be carried out. In this study a categorical approach has been purposely chosen to simplify the complex decision-making process for designers and left the regression analysis for a future work. By categorizing options, we aim to present clear-cut choices that are easier to interpret and apply in practical scenarios. However, while binning continuous data into categories can lead to a loss of detailed information, it is authors’ opinion that for the scope of this study the benefits of simplification outweigh this loss. One of the secondary objectives of this research is to underscore the collaborative dynamics between architects and engineers. To facilitate this, the decision-making process must be comprehensible and align with the thought processes of both disciplines. Given that architectural design often relies on categorical and qualitative metrics, the rationale for categorizing labels based on structural results stems from this consideration. Hence, in this Section the ML methods are all used to produce a classification of the tall buildings.

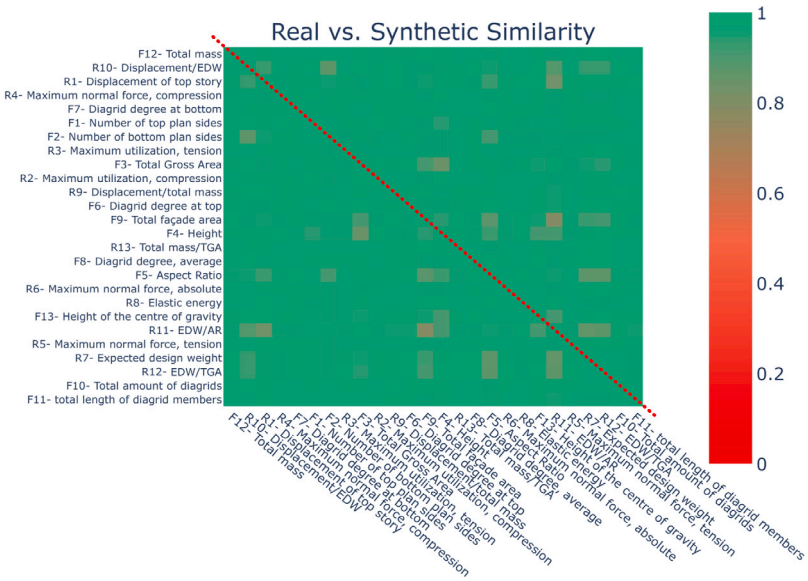


Fig. 9. Quality of the synthetic vs the original data; correlation between the column pairs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Average accuracy for selected responses in the original and augmented data set.

	Original data		Augmented data	
	Equal length	Equal no. of obs.	Equal length	Equal no. of obs
Displacement of the top story	0.653	0.849	0.386	0.676
Max utilization, compr.	0.723	0.663	0.621	0.649
EDW	0.473	0.790	0.372	0.509
Displacement/EDW	0.545	0.826	0.289	1.000
EDW/AR	0.612	0.779	0.564	0.713

5.1. Decision tree analysis on the original and augmented data sets

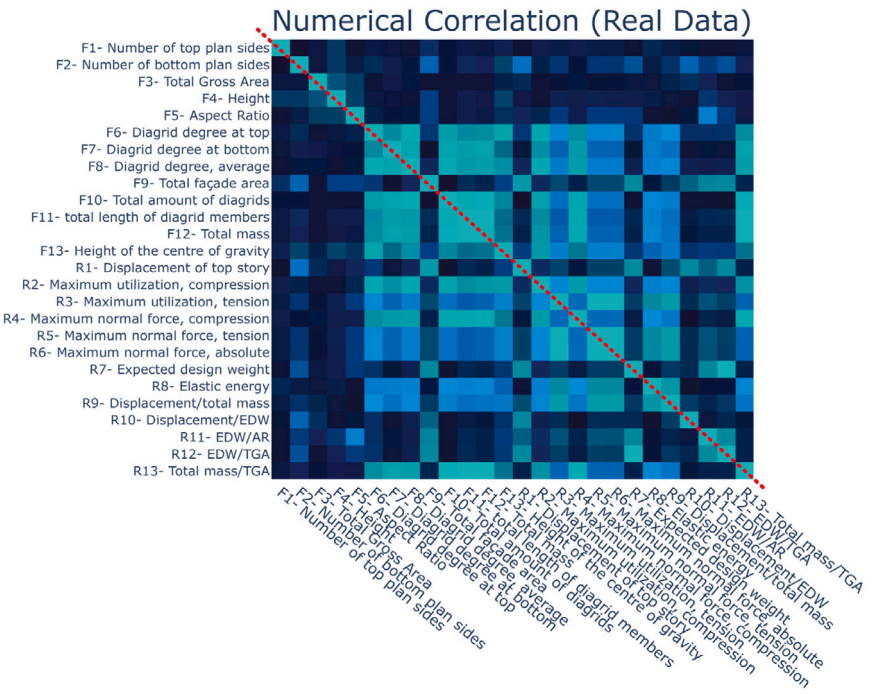
DTs are a non-parametric supervised learning method used for classification (and regression) tasks. They are particularly beneficial for interpretability, as they visually represent decisions and the decision-making process. For this reason, before addressing more complex ML methods in the following Sections 5.2–5.3, a DT approach is followed here to give insights into the learning algorithm.

The original data set, comprising 121 data points, has been first subjected to DT analysis. The process has involved creating a tree-like model of decisions. For each building design in the data set, the DT algorithm identifies the most critical features that affect the structural responses, following a hierarchical decision-making process. This analysis aids in understanding how different architectural and structural parameters interact to influence the overall design efficiency. As in the following Sections 5.2–5.3 for the RF approach, here a DT classifier with 4-fold cross-validation and 10 different random state parameters is defined, and the results are presented in Table 6. The results, visualized as a tree diagram in Fig. 11, show the paths of decision-making and the importance of each feature in the final decision. Following the DT analysis of the original data set, the same methodology to the augmented data set, which contains 1200 data points, has been applied. Again, a DT classifier with 4-fold cross-validation and 10 different random state parameters is defined, and the results are presented in the Table 6. This larger and more diverse data set provides a more robust platform for DT analysis, allowing for a deeper exploration of the intricate relationships between features and responses; its (far larger) tree diagram is shown as a file in the Supplementary material to this paper.

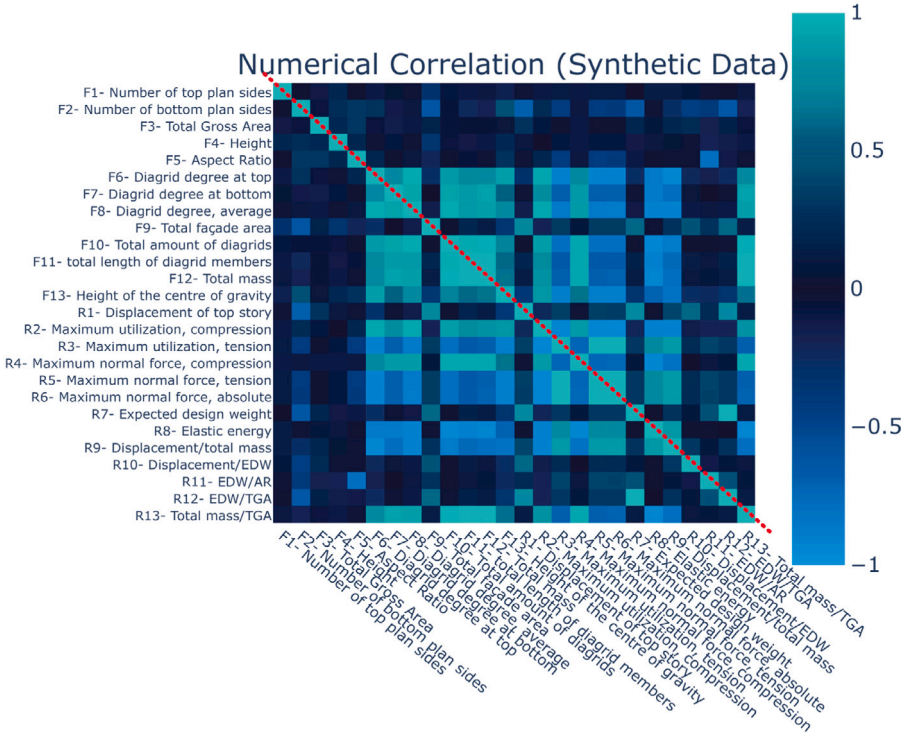
The comparative analysis between the original and augmented data sets utilizing Decision Trees (DT), demonstrates that while some key features remain consistently significant, the augmented data set offers a more distinctive understanding of the feature-response relationship. This comparison highlights the necessity and effectiveness of data augmentation in revealing understated, yet critical, aspects of building design that might not be apparent in smaller data sets.

5.2. Classification exploiting the original data set

Following the identification of the (five) optimal responses in Section 3.1 and the determination of the best features in Section 3.3, the learning process for the original data set is explored. The same algorithm utilized in the sequential f.s. is adopted for consistency



(a)



(b)

Fig. 10. Quality of the synthetic vs the original data; comparison of the feature and response reciprocal correlations for the (left) original vs (right) the augmented data set.

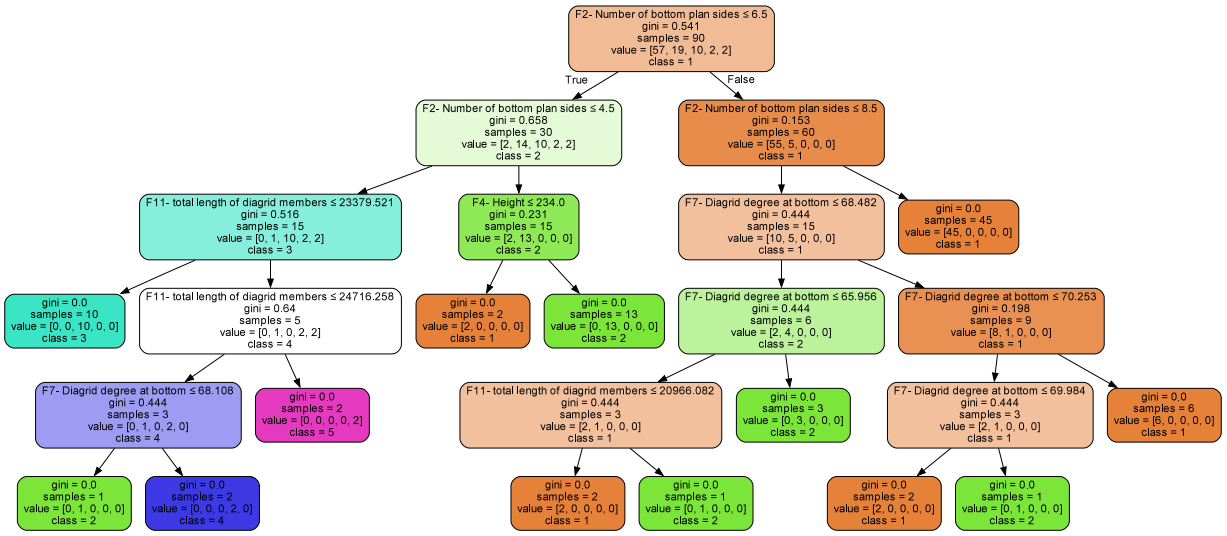


Fig. 11. DT path visualization for the original data set.

Table 7
Original data set. Accuracy of the top story displacement response of a specific random state, average and range.

Fold 1	Fold 2	Fold 3	Fold 4	Average	Max–min
0.806	0.900	0.900	0.767	0.843	0.133

Table 8
Original data set. Average cross-validation accuracy for top story displacement with 10 random state parameters.

1	2	3	4	5	6	7	8	9	10	Average	Max–min
0.843	0.843	0.860	0.810	0.827	0.802	0.843	0.835	0.818	0.810	0.829	0.05

Table 9
Original data set. Average accuracy for selected responses.

	Equal length	Equal number of observations
Displacement of the top story	0.829	0.636
Maximum utilization, compression	0.790	0.793
EDW	0.798	0.538
Displacement/EDW	0.861	0.630
EDW/AR	0.816	0.589

also for the learning process: thus, a RF algorithm with 4-fold cross-validation is employed. Due to the limited size of the original (cleaned) data set, which comprises 121 data points, a high number of folds for cross-validation might lead to an absence of data points for specific classes in some folds. Delving into further details, 10 different random states are chosen to account for the stochastic nature of the RF algorithm. Each RF is composed of 100 DTs, resulting in a total of 40 RFs created for each response, and the average of their results is taken into consideration. All values are normalized to maintain consistency.

The learning accuracy for the "displacement of the top story" response is presented in Table 7, showcasing the average accuracy and range across different folds. Due to the small size of the data set, the accuracies exhibit considerable variation. This variability is evidently magnified by the limited number of observations, particularly when certain classes have fewer data points than the number of the folds.

Table 8 reports the average cross-validation score across all ten different random states. Given the limited size of the data set, the learning accuracy is unstable but variable within 5% of each other.

The evaluation of the accuracy for all five responses considers the best features for each response, utilizing the two aforementioned strategies for defining class labels in classification (equal length or equal number of observations). As illustrated in Table 8, in all cases improved performance is obtained compared to other strategies when equal length bins per class are defined. To explore the role of each algorithmic choice, four distinct cross-validation methods have been employed, namely k-fold cross validation, stratified cross validation, repeated random subsampling, and leave-one-out cross-validation (LOOCV). The selection of an appropriate cross-validation method becomes in fact relevant because of the limited number of data points.

k-fold cross-validation was previously mentioned in Section 3.2.3, see also [51], while the stratified k-fold method mirrors the aforementioned approach and guarantees that the distribution of each class is preserved in the training and testing sets in proportion to their occurrence in the entire data set [70].

Table 10
Original data set. Learning accuracy for alternative cross-validation approaches.

	k-fold	Stratified *k*-fold	Shuffle	LOOCV
Displacement of the top story	0.829	0.847	0.823	0.836
Maximum utilization, compression	0.790	0.806	0.756	0.821
EDW	0.798	0.794	0.779	0.778
Displacement/EDW	0.861	0.838	0.849	0.823
EDW/AR	0.816	0.856	0.841	0.835

Table 11
Augmented data set. Accuracy of the displacement of the top story.

Fold 1	Fold 2	Fold 3	Fold 4	Average	Max–min
0.713	0.726	0.730	0.750	0.730	0.037

Table 12
Augmented data set. Cross-validation accuracy for the displacement of the top story, averaged across 10 random state parameters, displaying average and range.

1	2	3	4	5	6	7	8	9	10	Average	Max–min
0.730	0.723	0.714	0.718	0.715	0.711	0.716	0.720	0.715	0.716	0.718	0.019

Table 13
Augmented data set. Average accuracy for selected responses using two label definitions.

	Equal length	Equal number of observations
Displacement of the top story	0.718	0.436
Maximum utilization, compression	0.718	0.683
EDW	0.592	0.416
Displacement/EDW	1.000	0.318
EDW/AR	0.771	0.606

Repeated random subsampling, alternatively known as shuffle split cross-validation or Monte Carlo cross-validation, differs from the *k*-fold method by not confining each sample to a single fold; it rather draws random samples from the entire data set in each iteration, creating distinct training and test sets [71]. Since sampling is carried out from the entire data set at each iteration, certain values selected in an iteration could be chosen again in subsequent iterations; the only requirement is to maintain a different train–test split percentage.

LOOCV represents an extreme variant of *k*-fold cross-validation and involves repeating the learning process for each data point [72]. During each iteration, a single observation is designated as the test sample, while the remaining data comprises the training set; the overall learning performance is assessed by taking the average of the results across all iterations. This approach is particularly recommended for cases involving a small data set.

Table 10 illustrates the accuracy outcomes obtained using each of the aforementioned cross-validation methods. All other hyperparameters have been kept constant throughout the process. *k*-fold cross-validation demonstrates superiority over shuffle cross-validation, since a higher accuracy is observed in four out of five analyzed responses. Stratified *k*-fold cross-validation exhibits superiority over *k*-fold cross-validation in three out of five response accuracy measurements due to the preservation of class proportions in both the training and testing sets. Although the accuracy of responses in the LOOCV approach does not consistently surpass other methodologies, since it considers a significantly larger number of cross-validation possibilities, it provides in general higher accuracy.

5.3. Classification exploiting the synthetic data set

The learning accuracy for the "displacement of the top story" response is evaluated in each fold for the augmented data set, with results displayed in Table 11; these results show a consistent performance across all folds, attributable to an adequate number of data points. The average cross-validation score for various random states is presented in Table 12, demonstrating stable accuracy due to a sufficient amount of data and a mitigated impact of RF's stochasticity on the learning process. Table 13 reports the accuracy of all five responses, for each of them considering the best features under the two different label definition strategies: the performance improves when the equal length bins approach for every class is used. Three separate cross-validation approaches have been implemented for the augmented data set. In the original data set, the accuracy achieved by each cross-validation method has varied significantly due to the limited size of the data set. However, with the augmented data set, which comprises more than 1000 observations, the disparities in accuracy across different cross-validation methods approximate to 2%. As a result, the impact of the choice of cross-validation method in this case is not particularly critical. Additionally, it should be noted that the LOOCV approach is actually computationally intensive in this scenario, requiring 12000 computations for each response when considering 10 different random states for RFs. Consequently, the LOOCV approach has been disregarded for the augmented data. The corresponding outcomes are collected in Table 14. While it is apparent that the accuracy is slightly lower, given the previously mentioned low disparity during cross-validation, the results of the augmented data are deemed more reliable than that of the smaller, original data set.

Table 14
Augmented data set. Comparison among three different cross-validation approaches.

	k-fold	Stratified k-fold	Shuffle
Displacement of the top story	0.718	0.714	0.728
Maximum utilization, compression	0.718	0.806	0.756
EDW	0.592	0.606	0.612
Displacement/EDW	1.000	1.000	1.000
EDW/AR	0.771	0.761	0.758

Table 15
Performance analysis with 100% original 121 data points.

Responses	Accuracy	Precision	Recall	F1-score	Training time (s)	Data Prep. time (s)
Displacement of the top story	0.83	0.72	0.75	0.71		
Max utilization, compression	0.79	0.79	0.75	0.74		
EDW	0.80	0.49	0.54	0.50	24.14	15.28
Displacement/EDW	0.86	0.84	0.87	0.84		
EDW/AR	0.82	0.79	0.76	0.77		

Table 16
Performance analysis with 75% original, 25% augmented 391 data points.

Responses	Accuracy	Precision	Recall	F1-score
Displacement of the top story	0.71	0.45	0.45	0.44
Max utilization, compression	0.67	0.68	0.65	0.65
EDW	0.59	0.43	0.39	0.39
Displacement/EDW	0.93	0.81	0.70	0.73
EDW/AR	0.74	0.76	0.72	0.73

Table 17
Performance analysis with 50% original, 50% augmented 660 data points.

Responses	Accuracy	Precision	Recall	F1-score
Displacement of the top story	0.70	0.33	0.32	0.32
Max utilization, compression	0.70	0.72	0.68	0.69
EDW	0.61	0.41	0.39	0.39
Displacement/EDW	0.96	0.67	0.59	0.62
EDW/AR	0.75	0.79	0.73	0.74

It is also worth mentioning that, with respect to original data, the training and validation time increases from 24 s to 39 s with augmented data. Moreover, a data preparation time of about 78 s, including data collection, cleaning, and augmentation processes, must be added. Therefore, while d.a. enhances the robustness of the model in terms of accuracy and precision, it also imposes additional computational demands, as evidenced by increased training and preparation times. These findings align with the observations in [32], where similar trade-offs have been noted.

5.4. Comparative analysis of learning effectiveness with varying ratios of augmented/original data and with alternative learning metrics

In this Subsection, data sets with varying ratios of original and augmented data are evaluated, including 100%/ 75%/ 50%/ 25%/ no original data, no/ 25%/ 50%/ 75%/ 100% augmented data. This gradation allows for a detailed examination of how the introduction of augmented data affects model accuracy and generalization. The following Tables 15–19 show the effect of increasing the number of data points by adding augmented data to the original ones on four metrics: accuracy (i.e. the ratio of correctly predicted instances to the total number of instances), precision (i.e. the proportion of correctly predicted positive instances out of all actual positive instances, no matter whether originally true or false), recall (i.e. the proportion of negatives that are truly called negatives), and F1-score (i.e. the harmonic average of precision and recall), see [73] for more detail. 10 random state variables with 4-fold cross-validation for an RF classifier is considered: therefore, averaged values are included in the tables. While augmented data show lower metrics, the results are less variable within cross-validations and are therefore deemed more reliable than the original, small data set.

6. Conclusion

In this work, major practical and technical challenges in the application of AI to tall building structural design have been addressed. These challenges include the complex task of selecting the optimal features and responses, handling data set sizes, imbalanced data distribution, and difficulties in achieving stable learning accuracy. A thorough investigation has been conducted on the influence of architectural parameters on structural responses in tall buildings with outer diagrids using robust statistical

Table 18
Performance analysis with 25% original, 75% augmented 930 data points.

Responses	Accuracy	Precision	Recall	F1-score
Displacement of the top story	0.72	0.28	0.28	0.27
Max utilization, compression	0.72	0.74	0.72	0.72
EDW	0.60	0.40	0.38	0.38
Displacement/EDW	0.98	0.27	0.27	0.27
EDW/AR	0.78	0.79	0.73	0.74

Table 19
Performance analysis with all augmented 1200 data points.

Responses	Accuracy	Precision	Recall	F1-score	Training time (s)	Data Prep. time (s)
Displacement of the top story	0.72	0.31	0.30	0.30		
Max utilization, compression	0.72	0.74	0.72	0.73		
EDW	0.59	0.35	0.35	0.35	38.92	76.25
Displacement/EDW	1.00	1.00	1.00	1.00		
EDW/AR	0.77	0.62	0.59	0.60		

Table A.20
Training performance on selected responses on original data (with feature selection).

Responses	Accuracy	Precision	Recall	F1-score
Displacement of the top story	0.83	0.72	0.75	0.71
Max utilization, compression	0.79	0.79	0.75	0.74
EDW	0.80	0.49	0.54	0.50
Displacement/EDW	0.86	0.84	0.87	0.84
EDW/AR	0.82	0.79	0.76	0.77
Average	0.82	0.73	0.74	0.71

Table A.21
Training performance on all responses on original data (with all features).

Responses	Accuracy	Precision	Recall	F1-score
Displacement of top story	0.83	0.72	0.75	0.71
Max utilization, compression	0.79	0.79	0.75	0.74
Max utilization, tension	0.78	0.80	0.74	0.73
Max normal forces, compression	0.85	0.81	0.78	0.77
Max normal forces, tension	0.78	0.80	0.74	0.73
Max normal forces, absolute	0.78	0.80	0.74	0.73
EDW	0.80	0.49	0.54	0.50
Elastic energy	0.84	0.79	0.76	0.75
Displacement to total mass	0.72	0.80	0.74	0.74
Displacement to EDW	0.86	0.84	0.87	0.84
EDW to AR	0.82	0.79	0.76	0.77
EDW to TGA	0.73	0.51	0.53	0.50
Total mass to TGA	0.96	0.97	0.97	0.96
Average	0.81	0.76	0.75	0.73

techniques. A d.a. algorithm has been then employed to mitigate the limitations imposed by the small size of the original data set, resulting in an augmented data set that maintains fidelity to the original data set. The quality of the synthetic (i.e. augmented) data has been critically assessed, demonstrating a high overall quality score. Further, various class labeling strategies have been evaluated, and the enhanced performance obtained using equal-length class bins has been underscored. A strong emphasis is placed on ensuring the preservation of column shapes and correlation trends, which are subsequently validated using appropriate metrics. Correlations between architectural parameters and structural responses have been established using classification techniques.

The investigation into f.s. has evolved, moving from the application of a simple statistical correlation index to more sophisticated approaches, such as the forward selection, backward elimination, and the exhaustive method. These methods have been utilized in view of the results derived from the learning process of different feature subsets. It has been found that to maximize the accuracy for a given response, it is better to extract different features from the original pool, rather than using a unique set for all the responses. This result holds substantial significance for architectural design. It indicates that each decision should consider specific, varying aspects among the alternative choices for every objective.

From the perspective of data derived from numerical simulations, the procedure often starts with a data set of limited size, relative to the entire design space; therefore, d.a. is deemed necessary to arrive at sound conclusions. By exploiting Gaussian copula for this objective and augmenting the data set size tenfold, it has been shown that this step must be viewed in conjunction with the f.s. process, as d.a. tends to work more effectively for the entire data set after outlier removal.

Therefore, f.s. must be implemented after d.a. to achieve superior results. Furthermore, the ML algorithm, RF, has demonstrated more stable results under different cross-validation procedures for the augmented data set with respect to the original, smaller data

Table A.22
Training performance on selected responses on augmented data (with feature selection).

Responses	Accuracy	Precision	Recall	F1-score
Displacement of the top story	0.72	0.31	0.30	0.30
Max utilization, compression	0.72	0.74	0.72	0.73
EDW	0.59	0.35	0.35	0.35
Displacement/EDW	1.00	1.00	1.00	1.00
EDW/AR	0.77	0.62	0.59	0.60
Average	0.76	0.60	0.59	0.60

Table A.23
Training performance on all responses on augmented data (with all features).

Responses	Accuracy	Precision	Recall	F1-score
Displacement of top story	0.72	0.31	0.30	0.30
Max utilization, compression	0.72	0.74	0.72	0.73
Max utilization, tension	0.75	0.76	0.73	0.74
Max normal forces, compression	0.76	0.78	0.76	0.77
Max normal forces, tension	0.75	0.77	0.73	0.75
Max normal forces, absolute	0.76	0.77	0.73	0.74
EDW	0.59	0.35	0.35	0.35
Elastic energy	0.75	0.77	0.73	0.74
Displacement to total mass	0.77	0.78	0.77	0.78
Displacement to EDW	1.00	1.00	1.00	1.00
EDW to AR	0.77	0.62	0.59	0.60
EDW to TGA	0.60	0.38	0.33	0.33
Total mass to TGA	0.96	0.97	0.96	0.97
Average	0.76	0.69	0.67	0.68

Table A.24
Computing time analysis with the reduced response set.

Metric	Full set	Reduced set
Computing time, original data (s)	67.50	24.68
Computing time, augmented data (s)	1046.76	149.04
Exhaustive f.s. time (s)	4723.76	1576.27

set; however, the accuracy appears lower in case of the augmented data set. A cautious conclusion might be that the price to pay for a limited initial data set could be a slightly reduced, but still acceptable, accuracy following the proposed d.a. Nevertheless, given the significant benefits in terms of time efficiency and complexity reduction, such a trade-off should be carefully evaluated.

Further studies, some of which are already in progress, are necessary to confirm the trend and also to check whether alternative extractions of the values for the considered features, e.g. exploiting a random or Latin hypercube or orthogonal sampling method, would lead to different results. It would be also interesting to consider a regression analysis in turn of the classification and compare the outcome with the approach here proposed. Moreover, the structural analyses could transition towards dynamics to account for the influence of higher structural modes on the building response and, consequently, on the ML procedure.

CRediT authorship contribution statement

Pooyan Kazemi: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Alireza Entezami:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. **Aldo Ghisi:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgment

This research is not received any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

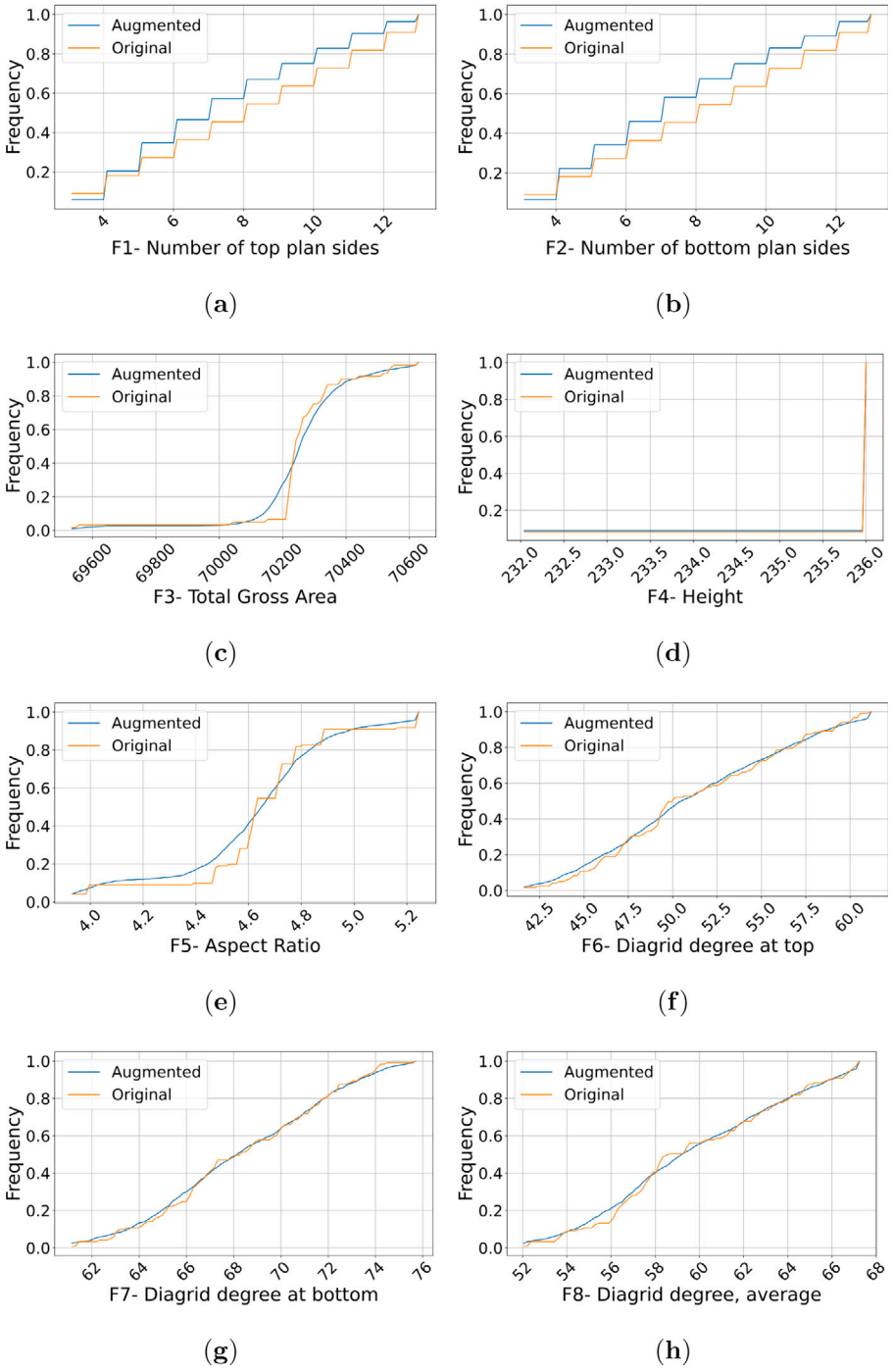


Fig. B.12. Comparison, in terms of CDFs, between the original and augmented data for features 1–8 (see Table 2).

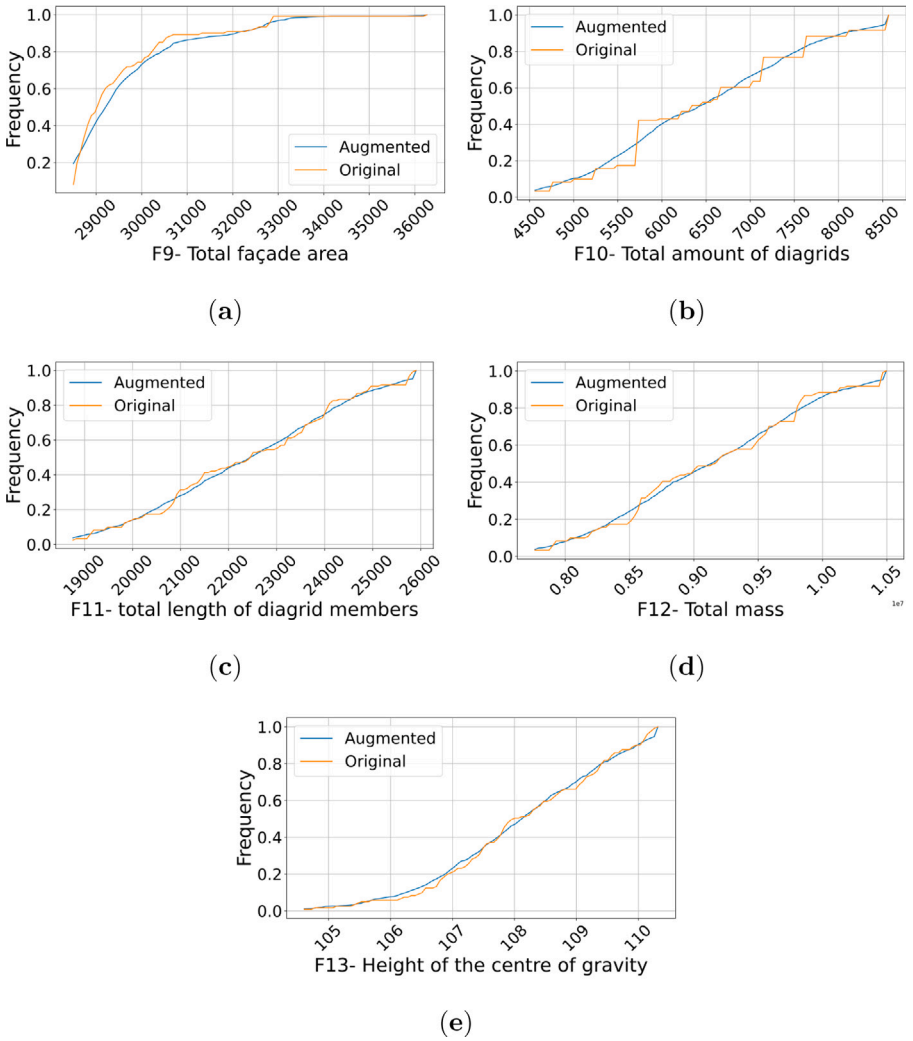


Fig. B.13. Comparison, in terms of CDFs, between the original and augmented data for features 9–13 (see Table 2).

Appendix A. About reducing the response set

Data analysis on the implications of response selection based on feature correlation is shown in this Appendix A. Tables A.20–A.24 present a comparative analysis of model performance using the full versus the reduced response sets across multiple validation frameworks on original and augmented data. A range of performance metrics, including accuracy, precision, recall, and F1 score, has been employed. The results indicate that the reduced model, while inherently more efficient, shows an acceptable decrease in accuracy, supporting the methodological choices made. In other words, the reduction in complexity gain is considered an advantage compensating for the performance reduction. These findings align with the task-specific nature of the study, where the goal is to establish a reliable yet efficient predictive model for diagrid building design.

Table A.24 provides a comparison of the computational efficiency between using the full set and the reduced set of responses with the original and the augmented data set. It highlights the benefits in terms of computational time, resource usage, and model complexity.

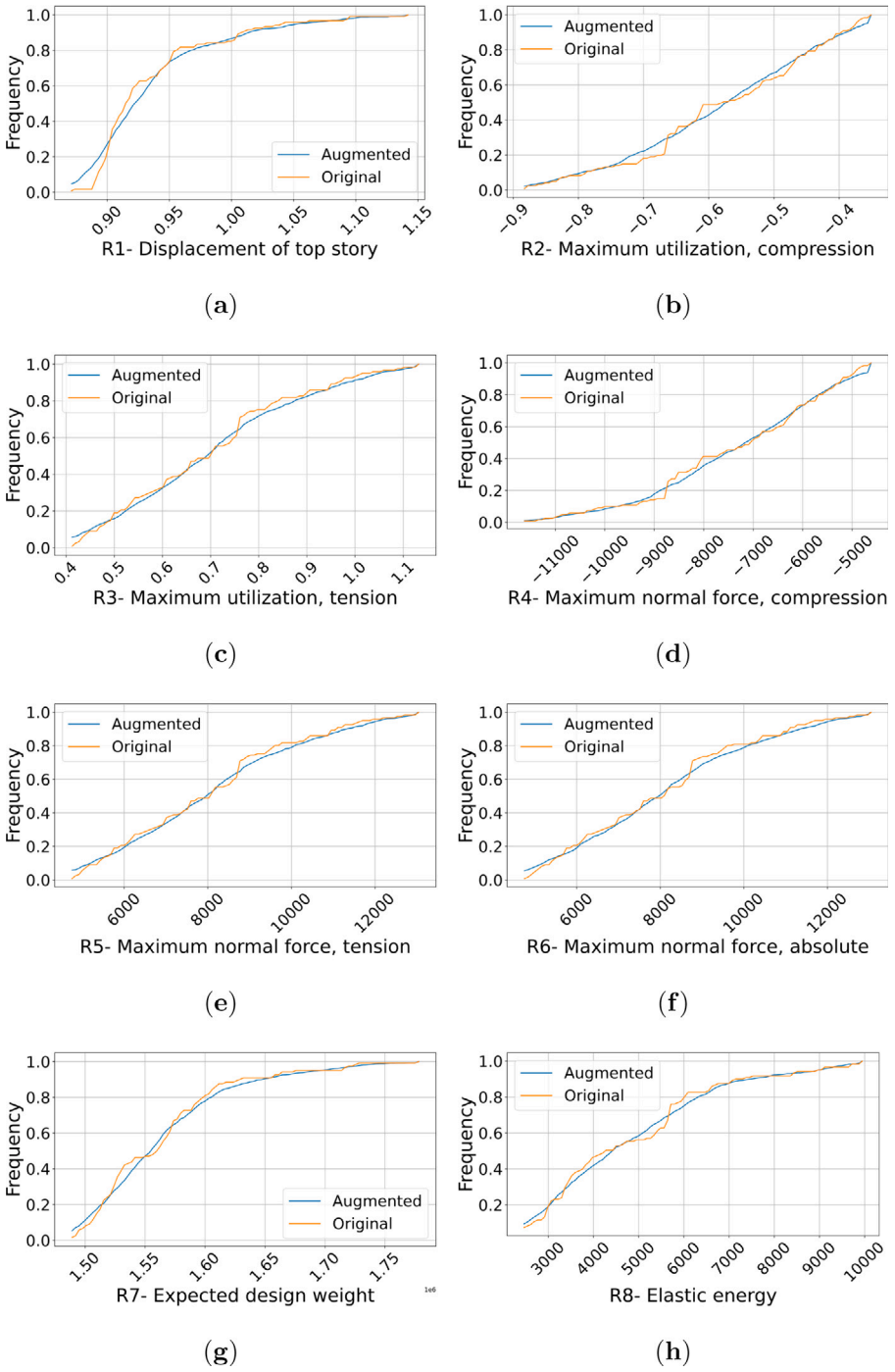


Fig. B.14. Comparison, in terms of CDFs, between the original and augmented data for responses 1–8 (see Table 3).

Appendix B. Comparison of the CDFs for original and synthetic data

The CDFs for the 13 features, Figs. B.12–B.13, and for the 13 responses, Figs. B.14–B.15, are collected in this Appendix A. For ease of reference, the exact meaning of the features and responses considered in this work is reported in the Tables 2 and 3.

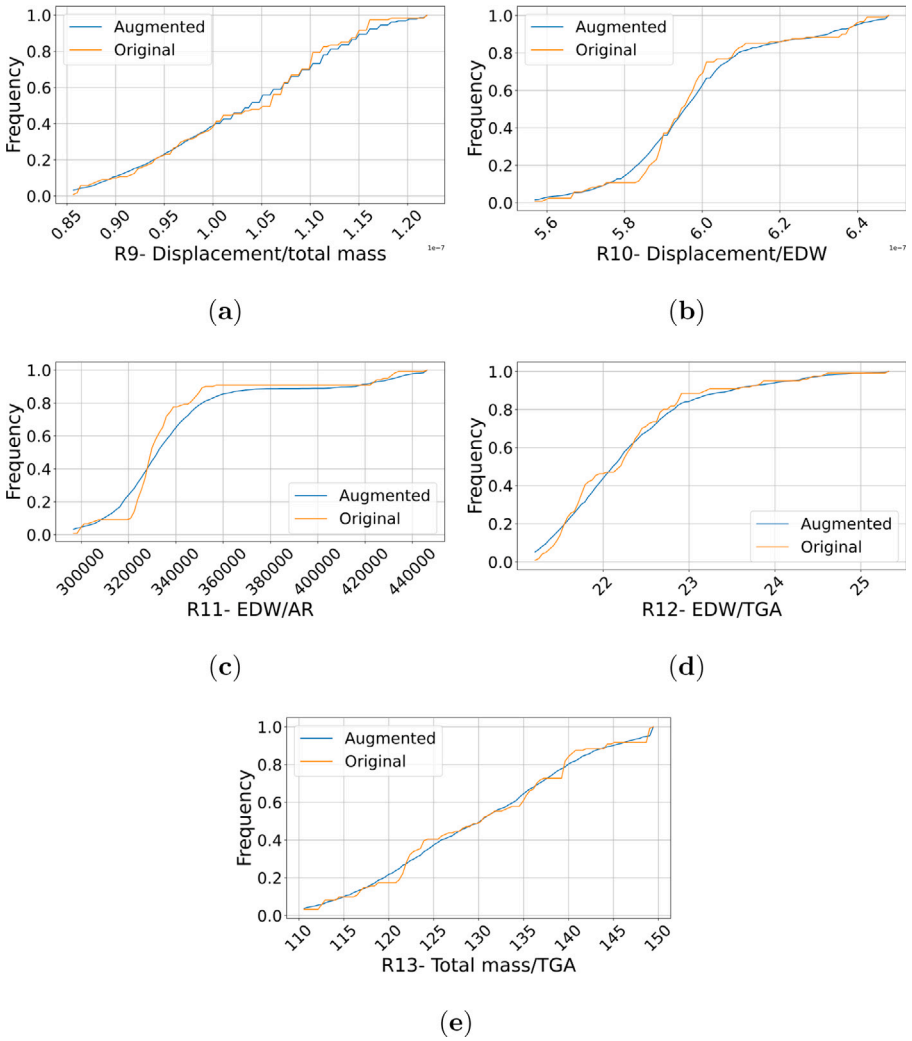


Fig. B.15. Comparison, in terms of CDFs, between the original and augmented data for responses 9–13 (see Table 3).

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jobe.2024.108766.

References

[1] S. Zhu, T. Yu, T. Xu, H. Chen, et al., Intelligent computing: the latest advances, challenges, and future, Intelligent Computing 2 (2023) 0006, http://dx.doi.org/10.34133/icomputing.0006.
[2] D. Fang, N. Brown, C. De Wolf, C. Mueller, Reducing embodied carbon in structural systems: A review of early-stage design strategies, J. Build. Eng. 76 (2023) 107054, http://dx.doi.org/10.1016/j.jobe.2023.107054.
[3] I. Hens, R. Solnosky, N. Brown, Parametric framework for early evaluation of prescriptive fire design and structural feasibility in tall timber, J. Archit. Eng. 29 (1) (2023) 04022040, http://dx.doi.org/10.1061/JAEIED.AEENG-145.
[4] C. Preisinger, Linking structure and parametric geometry, Archit. Des. 83 (2) (2013) 110–113, http://dx.doi.org/10.1002/ad.1564.
[5] H. Sun, H.V. Burton, H. Huang, Machine learning applications for building structural design and performance assessment: State-of-the-art review, J. Build. Eng. 33 (2021) 101816, http://dx.doi.org/10.1016/j.jobe.2020.101816.
[6] K.S. Ochoa, P.O. Ohlbrock, P. D'Acunto, V. Moosavi, Beyond typologies, beyond optimization: Exploring novel structural forms at the interface of human and machine intelligence, Int. J. Archit. Comput. 19 (3) (2021) 466–490, http://dx.doi.org/10.1177/1478077120943062.
[7] T.D. Akinosho, L.O. Oyedele, M. Bilal, A.O. Ajayi, M.D. Delgado, O.O. Akinade, A.A. Ahmed, Deep learning in the construction industry: A review of present status and future innovations, J. Build. Eng. 32 (2020) 101827, http://dx.doi.org/10.1016/j.jobe.2020.101827.
[8] L. Bleker, R. Pastrana, P.O. Ohlbrock, P. D'Acunto, Structural form-finding enhanced by graph neural networks, in: C. Gengnagel, O. Baverel, G. Betti, M. Popescu, M.R. Thomsen, J. Wurm (Eds.), Towards Radical Regeneration, Springer International Publishing, Cham, 2023, pp. 24–35.
[9] Z. Guo, K. Saldana Ochoa, P. D'Acunto, Enhancing structural form-finding through a text-based AI engine coupled with computational graphic statics, in: Proceedings of IASS Annual Symposia, vol. 2022, (no. 8) International Association for Shell and Spatial Structures (IASS), 2022, pp. 1–11.

[10] M.M. Ali, K. Al-Kodmany, Tall buildings and urban habitat of the 21st century: A global perspective, Buildings 2 (4) (2012) 384–423, http://dx.doi.org/10.3390/buildings2040384.
[11] M.M. Ali, K.S. Moon, Advances in structural systems for tall buildings: Emerging developments for contemporary urban giants, Buildings 8 (8) (2018) http://dx.doi.org/10.3390/buildings8080104.
[12] D. Scaramozzino, B. Albitos, G. Lacidogna, A. Carpinteri, Selection of the optimal diagrid patterns in tall buildings within a multi-response framework: Application of the desirability function, J. Build. Eng. 54 (2022) 104645, http://dx.doi.org/10.1016/j.jobe.2022.104645.
[13] B. Ekici, Z.T. Kazanasmaz, M. Turrin, M.F. Taşgetiren, I.S. Sariyildiz, Multi-zone optimisation of high-rise buildings using artificial intelligence for sustainable metropolises. Part 1: Background, methodology, setup, and machine learning results, Sol. Energy 224 (2021) 373–389, http://dx.doi.org/10.1016/j.solener.2021.05.083.
[14] E. Asadi, A.M. Salman, Y. Li, Multi-criteria decision-making for seismic resilience and sustainability assessment of diagrid buildings, Eng. Struct. 191 (2019) 229–246, http://dx.doi.org/10.1016/j.engstruct.2019.04.049.
[15] R.E. Weber, C. Mueller, C. Reinhart, Solar exoskeletons – An integrated building system combining solar gain control with structural efficiency, Sol. Energy 240 (2022) 301–314, http://dx.doi.org/10.1016/j.solener.2022.05.048.
[16] D. Fang, C. Liu, Mechanical characteristics and deformation calculation of steel diagrid structures in high-rise buildings, J. Build. Eng. 42 (2021) 103062, http://dx.doi.org/10.1016/j.jobe.2021.103062.
[17] C. Liu, D. Fang, L. Zhao, J. Zhou, Seismic fragility estimates of steel diagrid structure with performance-based tests for high-rise buildings, J. Build. Eng. 52 (2022) 104459, http://dx.doi.org/10.1016/j.jobe.2022.104459.
[18] C. Liu, D. Fang, Separation of long-period components of ground motion and its impact on seismic response of long-period diagrid structures, Soil Dyn. Earthq. Eng. 150 (2021) 106942, http://dx.doi.org/10.1016/j.soildyn.2021.106942.
[19] H. Luo, S.G. Paal, Artificial intelligence-enhanced seismic response prediction of reinforced concrete frames, Adv. Eng. Inform. 52 (2022) 101568, http://dx.doi.org/10.1016/j.aei.2022.101568.
[20] S. Paal, J.-S. Jeon, I. Brilakis, R. DesRoches, Automated damage index estimation of reinforced concrete columns for post-earthquake evaluations, J. Struct. Eng. 141 (9) (2015) 04014228, http://dx.doi.org/10.1061/(ASCE)ST.1943-541X.0001200.
[21] C.R. Farrar, K. Worden, Structural Health Monitoring: A Machine Learning Perspective, John Wiley & Sons Inc, 2012.
[22] S.R. Vadyala, S.N. Betgeri, J.C. Matthews, E. Matthews, A review of physics-based machine learning in civil engineering, Results Eng. 13 (2022) 100316, http://dx.doi.org/10.1016/j.rineng.2021.100316.
[23] F.N. Khan, Q. Fan, C. Lu, A.P.T. Lau, An optical communication's perspective on machine learning and its applications, J. Lightwave Technol. 37 (2) (2019) 493–516, http://dx.doi.org/10.1109/JLT.2019.2897313.
[24] J. Hegde, B. Rokseth, Applications of machine learning methods for engineering risk assessment – A review, Saf. Sci. 122 (2020) 104492, http://dx.doi.org/10.1016/j.ssci.2019.09.015.
[25] K. Morfidis, K. Kostinakis, Approaches to the rapid seismic damage prediction of R/C buildings using artificial neural networks, Eng. Struct. 165 (2018) 120–141, http://dx.doi.org/10.1016/j.engstruct.2018.03.028.
[26] K. Kostinakis, K. Morfidis, K. Demertzis, L. Iliadis, Classification of buildings' potential for seismic damage using a machine learning model with auto hyperparameter tuning, Eng. Struct. 290 (2023) 116359, http://dx.doi.org/10.1016/j.engstruct.2023.116359.
[27] E. Junda, C. Málaga-Chuquitaype, K. Chawgien, Interpretable machine learning models for the estimation of seismic drifts in CLT buildings, J. Build. Eng. 70 (2023) 106365, http://dx.doi.org/10.1016/j.jobe.2023.106365.
[28] K. Zhou, D.-L. Xie, K. Xu, L.-H. Zhi, F. Hu, Z.-R. Shu, A machine learning-based stochastic subspace approach for operational modal analysis of civil structures, J. Build. Eng. 76 (2023) 107187, http://dx.doi.org/10.1016/j.jobe.2023.107187.
[29] A.A. Torky, S. Ohno, Deep learning techniques for predicting nonlinear multi-component seismic responses of structural buildings, Comput. Struct. 252 (2021) 106570, http://dx.doi.org/10.1016/j.compstruc.2021.106570.
[30] S.-H. Hwang, S. Mangalathu, J. Shin, J.-S. Jeon, Estimation of economic seismic loss of steel moment-frame buildings using a machine learning algorithm, Eng. Struct. 254 (2022) 113877, http://dx.doi.org/10.1016/j.engstruct.2022.113877.
[31] H.D. Nguyen, J.M. LaFave, Y.-J. Lee, M. Shin, Rapid seismic damage-state assessment of steel moment frames using machine learning, Eng. Struct. 252 (2022) 113737, http://dx.doi.org/10.1016/j.engstruct.2021.113737.
[32] P. Kazemi, A. Ghisi, S. Mariani, Classification of the structural behavior of tall buildings with a diagrid structure: A machine learning-based approach, Algorithms 15 (10) (2022) http://dx.doi.org/10.3390/a15100349.
[33] R. Alizadeh, J.K. Allen, F. Mistree, Managing computational complexity using surrogate models: A critical review, Res. Eng. Des. 31 (2020) 285–298, http://dx.doi.org/10.1007/s00163-020-00336-7.
[34] K. Zhong, J.G. Navarro, S. Govindjee, G.G. Deierlein, Surrogate modeling of structural seismic response using probabilistic learning on manifolds, Earthq. Eng. Struct. Dynam. 52 (2023) 2407–2428, http://dx.doi.org/10.1002/eqe.3839.
[35] N. Pudjihartono, T. Fadason, A. Kempa-Liehr, J. O'Sullivan, A review of feature selection methods for machine learning-based disease risk prediction, Front. Bioinform. 2 (2022) 927312, http://dx.doi.org/10.3389/fbinf.2022.927312.
[36] D.G. Bonett, T.A. Wright, Sample size requirements for estimating pearson, kendall and spearman correlations, Psychometrika 65 (2000) 23–28, URL https://api.semanticscholar.org/CorpusID:120558581.
[37] F.J. Ferri, P. Pudil, M. Hatef, J. Kittler, Comparative study of techniques for large-scale feature selection, in: E.S. Gelsema, L.S. Kanal (Eds.), Pattern Recognition in Practice IV, in: Machine Intelligence and Pattern Recognition, vol. 16, North-Holland, 1994, pp. 403–413, http://dx.doi.org/10.1016/B978-0-444-81892-8.50040-7.
[38] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recognit. Lett. 15 (11) (1994) 1119–1125, http://dx.doi.org/10.1016/0167-8655(94)90127-9.
[39] M. Kuhn, K. Johnson, Applied Predictive Modeling, in: SpringerLink : Bücher, Springer, New York, 2013, URL https://books.google.nl/books?id=xYRDAAAAQBAJ.
[40] S. Sabila, R. Sarno, K. Triyana, Optimizing threshold using pearson correlation for selecting features of electronic nose signals, Int. J. Intell. Eng. Syst. 12 (2019) 81–90, http://dx.doi.org/10.22266/ijies2019.1231.08.
[41] A. Layeb, Two novel feature selection algorithms based on crowding distance, 2021, http://dx.doi.org/10.13140/RG.2.2.33509.93923.
[42] J. Miao, L. Niu, A survey on feature selection, Procedia Comput. Sci. 91 (2016) 919–926, http://dx.doi.org/10.1016/j.procs.2016.07.111.
[43] J. Xie, M. Sage, Y.F. Zhao, Feature selection and feature learning in machine learning applications for gas turbines: A review, Eng. Appl. Artif. Intell. 117 (2023) 105591, http://dx.doi.org/10.1016/j.engappai.2022.105591.
[44] H. Sarmadi, A. Entezami, B. Behkamal, C. De Michele, Partially online damage detection using long-term modal data under severe environmental effects by unsupervised feature selection and local metric learning, Journal of Civil Structural Health Monitoring 12 (5) (2022) 1043–1066, http://dx.doi.org/10.1007/s13349-022-00596-y.
[45] H. Sarmadi, A. Entezami, F. Magalhães, Unsupervised data normalization for continuous dynamic monitoring by an innovative hybrid feature weighting-selection algorithm and natural nearest neighbor searching, Structural Health Monitoring (2023) http://dx.doi.org/10.1177/14759217231166116, 14759217231166116.

[46] J. Suto, S. Oniga, P.P. Sitar, Comparison of wrapper and filter feature selection algorithms on human activity recognition, in: 2016 6th International Conference on Computers Communications and Control, ICCCC, 2016, pp. 124–129, http://dx.doi.org/10.1109/ICCCC.2016.7496749.
[47] D.M. Belete, D.H. Manjaiah, A comparative study of filter and wrapper methods on EDHS – HIV/AIDS dataset, in: 2020 Third International Conference on Smart Systems and Inventive Technology, ICSSIT, 2020, pp. 1264–1271, http://dx.doi.org/10.1109/ICSSIT48917.2020.9214212.
[48] N.D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella, A. Scotto di Freca, Comparing filter and wrapper approaches for feature selection in handwritten character recognition, Pattern Recognit. Lett. 168 (2023) 39–46, http://dx.doi.org/10.1016/j.patrec.2023.02.028.
[49] B. Kumari, T. Swarnkar, Filter versus wrapper feature subset selection in large dimensionality micro array: A review, Int. J. Comput. Sci. Inf. Technol. 2 (2011) 1048–1053.
[50] S. Raschka, Mlxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack, J. Open Source Softw. 3 (24) (2018) http://dx.doi.org/10.21105/joss.00638.
[51] S.J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, third ed., Pearson, 2009.
[52] A. Parmar, R. Katariya, V. Patel, A Review on Random Forest: An Ensemble Classifier, 2019, pp. 758–763, http://dx.doi.org/10.1007/978-3-030-03146-6_86.
[53] K. Fawagreh, M.M. Gaber, E. Elyan, Random forests: From early developments to recent advancements, Syst. Sci. Control Eng. 2 (1) (2014) 602–609, http://dx.doi.org/10.1080/21642583.2014.956265.
[54] C. Chen, L. Breiman, Using Random Forest to Learn Imbalanced Data, University of California, Berkeley, 2004.
[55] A. Karlhede, Tackling imbalanced data in random forest to predict free-to-fee transitions of a subscription saas-application, in: TRITA-EECS-EX, (no. 2020:479) KTH, School of Electrical Engineering and Computer Science (EECS), 2020, p. 98.
[56] J. Dong, Q. Qian, A density-based random forest for imbalanced data classification, Future Internet 14 (3) (2022) http://dx.doi.org/10.3390/fi14030090.
[57] Q. Chen, X. Zhang, Y. Wang, Z. Zhai, F. Yang, Applying a random forest approach to imbalanced dataset on network monitoring analysis, in: W. Lu, Y. Zhang, W. Wen, H. Yan, C. Li (Eds.), Cyber Security, Springer Nature Singapore, Singapore, 2022, pp. 28–37.
[58] R. O'Brien, H. Ishwaran, A random forests quantile classifier for class imbalanced data, Pattern Recognit. 90 (2019) 232–249, http://dx.doi.org/10.1016/j.patcog.2019.01.036.
[59] A. More, D. Rana, I. Agarwal, S. Vallabhbhai, Random forest classifier approach for imbalanced big data classification for smart city application domains, Int. J. Comput. Intell. Syst. 1 (2020) URL https://ssrn.com/abstract=3354727.
[60] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[61] I.H. Sarker, A. Kayes, P. Watters, Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage, J. Big Data 6 (1) (2019) 1–28, http://dx.doi.org/10.1186/s40537-019-0219-y.
[62] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Appl. Soft Comput. 97 (2020) 105524, http://dx.doi.org/10.1016/j.asoc.2019.105524.
[63] A. Tlebaldinova, N. Denissova, O. Baklanova, I. Krak, G. Györök, Normalization of vehicle license plate images based on analyzing of its specific features for improving the quality recognition, Acta Polytech. Hungarica 17 (6) (2020) 193–206, http://dx.doi.org/10.12700/APH.17.6.2020.6.11.
[64] Z. Hu, Y.V. Bodyanskiy, N.Y. Kulishova, O.K. Tyshchenko, A multidimensional extended neo-fuzzy neuron for facial expression recognition, Int. J. Intell. Syst. Appl. 9 (9) (2017) 29, http://dx.doi.org/10.5815/ijisa.2017.09.0.
[65] A. Entezami, A.N. Arslan, C. De Michele, B. Behkamal, Online hybrid learning methods for real-time structural health monitoring using remote sensing and small displacement data, Remote Sensing 14 (14) (2022) 3357, http://dx.doi.org/10.3390/rs14143357.
[66] B. Behkamal, A. Entezami, C. De Michele, A.N. Arslan, Elimination of thermal effects from limited structural displacements based on remote sensing by machine learning techniques, Remote Sensing 15 (12) (2023) 3095, http://dx.doi.org/10.3390/rs15123095.
[67] N. Patki, R. Wedge, K. Veeramachaneni, The synthetic data vault, in: IEEE International Conference on Data Science and Advanced Analytics, DSAA, 2016, pp. 399–410, http://dx.doi.org/10.1109/DSAA.2016.49.
[68] S.M. Ross, 6 - the multivariate normal distribution and copulas, in: S.M. Ross (Ed.), Simulation (Sixth Edition), sixth ed., Academic Press, Boston, 2023, pp. 99–110, http://dx.doi.org/10.1016/B978-0-32-385738-3.00011-0.
[69] D. Tjøstheim, H. Otneim, B. Støve, Chapter 5 - local Gaussian correlation and the copula, in: D. Tjøstheim, H. Otneim, B. Støve (Eds.), Statistical Modeling using Local Gaussian Approximation, Academic Press, 2022, pp. 135–159, http://dx.doi.org/10.1016/B978-0-12-815861-6.00012-2.
[70] J. Motl, P. Kordík, Stratified cross-validation on multiple columns, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence, ICTAI, 2021, pp. 26–31, http://dx.doi.org/10.1109/ICTAI52525.2021.00012.
[71] Q.-S. Xu, Y.-Z. Liang, Monte Carlo cross validation, Chemometr. Intell. Lab. Syst. 56 (1) (2001) 1–11, http://dx.doi.org/10.1016/S0169-7439(00)00122-2.
[72] C. Sammut, G.I. Webb (Eds.), Leave-one-out cross-validation, in: Encyclopedia of Machine Learning, Springer US, Boston, MA, 2010, pp. 600–601, http://dx.doi.org/10.1007/978-0-387-30164-8_469.
[73] R. Irizarry, Advanced Data Science. Statistics and Prediction Algorithms Through Case Studies, Chapman & Hall, 2020, Chapter 25.