






Beyond Traditional Computer-Aided Design Parameterization, Feature Engineering for Improved Surrogate Modeling in Engineering Design

Mohammad Arjomandi Rad¹ , Massimo Panarotto² , Ola Isaksson³ 

¹Chalmers University of Technology, radmo@chalmers.se

²Chalmers University of Technology, maspan@chalmers.se

³Chalmers University of Technology, iola@chalmers.se

Corresponding author: Mohammad Arjomandi Rad, radmo@chalmers.se

Abstract. Surrogate modeling in engineering design uses Computer-Aided Design (CAD) to create input features. To this end, CAD models are parameterized and have traditionally assisted design changes, automation, and standardization. However, this process leads to low flexibility, limited design space exploration, a high-dimensional design space, and ultimately extended design cycles. This paper builds on an existing methodology of correlation-based feature extraction in CAD to prevent dimensionality excess and improve the flexibility of surrogate models. We extend the 'sleeping parameters' concept from extraction to engineered features and position it in the overall machine modeling learning process. To count for efficacy validation as part of the process of training a prediction model, several correlation matrices are suggested to rank and select these new features, which complete the feature engineering loop. Utilizing a new case study on Thin-Walled Beams (TWBs) crashworthiness, we showcase how to construct the medial axis of a beam cross-section and extract numerous features in several categories. The results show meaningful relationships between the sleeping parameters and their resulting crashworthiness outputs. The implications of the findings suggest the possibility of achieving better predictions with fewer parameters and reduced dependency on CAD parameterization, potentially leading to accelerated design iterations in the development of TWBs.

Keywords: Surrogate modeling, Feature engineering, Data-driven design, CAD/CAE, Thin-walled tubes, Crashworthiness

DOI: <https://doi.org/10.14733/cadaps.2025.536-554>

1 INTRODUCTION

Building predictive surrogate models for assessing the performance of a design concept is a standard practice in engineering design. Traditional Computer-Aided Design (CAD) parameterization, while instrumental in

design changes, automation, and standardization, presents inherent challenges when faced with drastic design changes. A simple geometric shape can necessitate many parameters for definition, leading to high dimensional [27] or unnecessarily complex prediction models further down in the process. With the increase in the problem size and geometrical intricacies, prediction models become cumbersome or potentially lose their accuracy. As a result, it will be incredibly expensive to explore design space with more complex systems [7]. On the other hand, with another AI boom in the industry, more data availability, and increased complexity of modern products, information overload can easily overwhelm any design process [4, 17] and escalate the problem.

CAD features have historically played a pivotal role in building meta/surrogate models in engineering design [39, 43], and parameterization has been associated with CAD models from the beginning. Another issue is that conventional CAD parameterization tends to be deterministic and might not easily accommodate the variability and uncertainty inherent in real-world applications, which is crucial for robust prediction models [20]. For example, in the case of the introduction of a new design change (such as the removal or addition of a sketch dimension such as the one depicted in Figure 1), the already trained surrogate model can easily be useless or require maintenance. Because the dimensions that were being utilized as a feature in the Machine Learning (ML) training, such as a and b , might not be available anymore in the revised geometry or replaced by another dimension, such as e , as a result of the change in the shape.

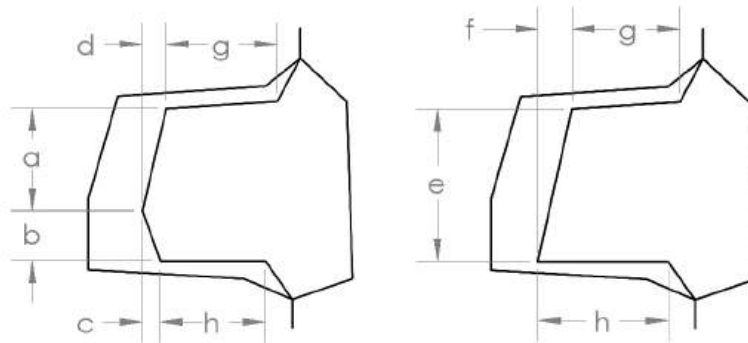


Figure 1: CAD model parameterization reduces the flexibility of surrogate models to design change.

Moreover, CAD parameters are frequently derived based on design intent [19], meaning designers deal with them early in the design process. Yet, the prediction models are often brought into the picture late during design iterations. So naturally, designers can unintentionally make decisions that include irrelevant or redundant features that are not revealed as an issue until the later phases of product development [29]. In design space exploration, this increases the computational demands and can also obscure meaningful relationships (between input and output), hindering the model's generalizability and performance. All these problems echo a pressing need for refined and purpose-driven parameterization techniques or alternative solutions that align more closely with the requirements of predictive surrogate models.

As a solution to reduce the complexity of managing CAD parameters in complex shapes and making predictive surrogate models less dependent on them, the sleeping parameters convention has been introduced recently by the authors [27]. Sleeping parameters are defined in contrast to conventional CAD parameterization as engineered features that are coupled to the geometry of the design but are independent of the geometry creation process. The term 'sleeping' emphasizes that these features, while not immediately visible or conventional, have a potential utility that can be awakened through appropriate processes. Unlike 'latent'

or 'hidden' in data science, this name suggests an inherent obscurity or underlying useful nature. Sleeping parameters can be constructed, extracted, selected, and then processed even if the geometry undergoes drastic changes. This process justifies borrowing and using the feature engineering terminology from the data science field. As illustrated in Figure 2, the machine learning process starts with raw data and continues in an iterative data preparation, which goes through two processes for pre-processing and feature engineering before the start of learning the data and ending with derived model [8].

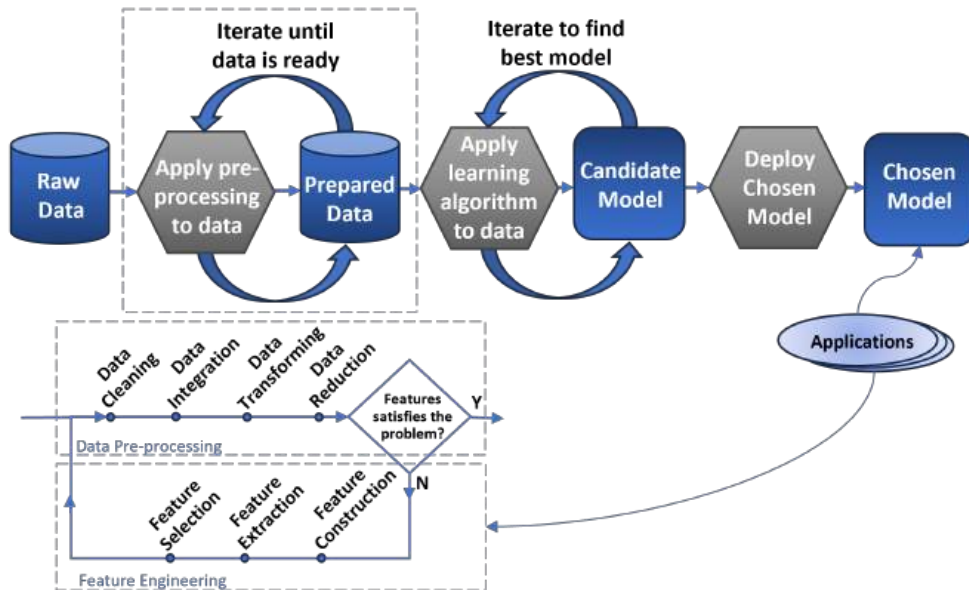


Figure 2: Feature engineering process nested in a machine learning application loop.

As illustrated in the figure, the first iterative process (also known as data preparation) includes data pre-processing and may or may not include feature engineering according to the definition (depending on the acquired data quality). Thus, feature engineering can be considered a specialized part of the broader data preparation process and encompasses a wider range of activities, including creating, transforming, and optimizing features for use in machine learning models.

Our prior work introduced a correlation-based feature extraction method on an airbag design case [27]. However, any method's true potential and versatility are revealed when challenged in diverse environments [18]. By extending methodology from one context to another (crashworthiness case here), we hope to create a more rigorous foundation for the method and contribute to its generalizability. This paper extends the sleeping parameters concepts from only an extraction to cover the whole feature engineering loop. This is done by introducing more varied ways of extracting features from the constructed data from the medial axis. Additionally, more correlation metrics are suggested to be used for evaluating the extracted features. Borrowing the data science terminology for the feature engineering process that includes the construction, extraction, and selection [46], it has shown how we can construct the medial axis of different variants of TWBs, extract meaningful features from this data, and evaluate and select the best ones among them for further use in the machine learning process.

In the realm of computational modeling and design, the accurate prediction of system behaviors and outcomes is of paramount importance. Selecting the right features for analysis can make all the difference to achieve accuracy. The aim is to make the end prediction model (in the surrogate modeling process) more flexible

by crafting (or engineering) high-quality features that represent the form but are acquired independently of it. Unlike conventional feature selection techniques that may overlook certain intrinsic parameters with significant predictive power, this technique tries to introduce a new way of looking at the process of selecting features for surrogate modeling in engineering design. Thus, the research question for this paper will be how to extract features from Computer-Aided Design (CAD) and, in particular, geometrical descriptions of the concepts to enhance prediction accuracy.

To answer this question, the previously introduced medial axis concept will be employed over a new case. We aim to validate, refine, and potentially expand our original findings by applying a previously established method to a new scenario. The remainder of the paper is organized as follows. The next section lays down a background and literature review of features and data mining in engineering design. The third chapter introduces Thin-Walled Beams (TWB) as the case study in this paper and explains the finite element simulations performed to attain the final outputs. Next, seven features extracted from the medial axis process are presented and ranked based on different correlation metrics. Finally, we discuss the results and suggest a categorization of the features together with some suggestions for future studies.

2 BACKGROUND

2.1 Features in Design Science vs. Features in Data Science

Although data mining and feature engineering are not new topics, unlike data science, the literature on this topic in design society is scattered and has not gained much attention. Partially, this can be the result of the semantic confusion around the word features.

Features in computer-aided design and features in data science are completely referred to as different things. In the data science and machine learning community, a feature is defined as the numerical encapsulation of the raw data [49], serving as characteristics or property of the entities being analyzed [12]. Features serve as structures in a dataset and are meaningful within a scientific or engineering context [24]. Data scientists use algorithms to discover patterns and relationships within mined data to identify patterns, make predictions, or derive insights. Features in data science can be represented as columns in an Excel spreadsheet or as attributes within a dataset in various formats, such as CSV files, SQL databases, or data frames. These features are crucial for training machine learning models as they provide the necessary information to predict or classify outcomes based on the learned patterns from the data.

On the other hand, there is the notion of *CAD features* or *form features* in the design literature that refer to the fundamental building blocks of a design's form [35]. These features are essential geometric or functional components of a product, such as holes, slots, bosses, and other standardized shapes that can be combined to create the overall geometry of a product. In addition to the form, features are also defined as entities describing the function [13], encapsulating specific engineering significance used to represent attributes and relationships within a part or assembly. Features can also refer to the connection between two parts [23], also known as assembly features like mating and constraints, as well as parametric controllers like dimensions and angles that designers use as relationships that exist between different parts within an assembly. CAD features in parametric modeling allow for easier modifications and optimization of designs, where the relationships among features can be defined to update the entire design when changes are made automatically. Moreover, the terms kinematic features, manufacturing features, and functional features are also introduced in design literature [9], which are self-explanatory. User-defined features (UDF) enable users to build their own features in CAD [42, 5], allowing for greater flexibility and customization in the design process. These features can be tailored to specific engineering requirements or to optimize the manufacturing process, thereby enhancing the functionality and efficiency of the designed products. The definitions of features and the distinctions between Design Science and Data Science are succinctly summarized in Table 1. This table highlights how each field approaches the concept of 'features' from its unique perspective.

Data Science	Design Science
Numerical encapsulation of the raw data [49]	Form features are fundamental building blocks of a design's form [35]
Characteristics or property of the entities being analyzed [12]	Entity describing both the form and function of a design [13]
Structures in a dataset and meaningful in context [24]	A part feature a shape with specific geometric and topological characteristics and similarly, assembly feature is as a connection between two parts [23]

Table 1: Definitions for features in data science and design science literature

An overlap between data engineering in product design and process is the knowledge discovery field that enables understanding a large body of textual datasets. Initially, this included building cyber agents such as web tracers and web organizers to extract needed information for product development [11]. Such textual features in engineering studies aim to enhance knowledge reuse by introducing computation knowledge extraction in text format from design documents, testing reports, life cycle assessments, customer reviews, sales returns, and so on [28, 31]. Semantic literature constructs a tag similarity measure to emulate how humans recall tags from memory. This line of research aims to design information retrieval by utilizing a network of similar semantics [37]. Features that are being mined are structured information from written historical records [34]. Another semantics branch is to analyze sentiments for mining customer requirements in the conceptual design process [40, 44]. Features identified here are the emotional tone (positive, negative, neutral) expressed within the text. Moreover, there are other kinds of mining-related topics concerning product design, such as reasoning about designs through frequent pattern mining, product design using association rule mining, and text features for mining design rationale, which are all along the same lines. Text mining research in product development continues until today [26, 47] with advancements in large language models, but since it goes beyond the scope of this paper, we refer readers to a recent review paper [38] for a comprehensive review on the topic.

2.2 Data Mining in Engineering

Regardless of any possible input data type in machine learning models - scalar or binary, vector or time series, and matrices or images [2] - the data is translated to Real (R) numbers to be used in mathematical computation. Because of their rich data types, simulations are a natural choice for data mining in engineering. In continuous simulations (e.x. Finite Elements), node and shell section information stored in a meshed finite element model can yield input data for analyzing a part's performance after a geometric change [21]. Zhao et al. present a framework for data preparation on crash simulation data for studying occupant restraint systems parameters on crashworthiness properties based on attribute importance and decision trees [48]. This was reported to reduce the size of the data sets and delete irrelevant features from the data sets, especially in full vehicle model type geometries that have more than hundreds of parameters.

On the other hand, discrete event simulation models are used as cost projectors for estimations in life cycle assessments. Data mining on the history and cost-based features are used in the aerospace industry as tools to characterize cost drivers such as over-performing repair activities [25]. Such clustering-based simulation mining methods instantiate a vast design space offline. Given new design variants, most similar designs are looked up with a similarity index, and from the simulation results of its 'design neighbors', a behavior valuation for a given simulation is stated without a Finite Element Analysis (FEA) [6]. Generally, simulation data is exploited to learn heuristic connections between the design space and the simulation space, but the effectiveness of this method depends on how well your simulations represent real-world behavior. Bad simulations can lead to bad

heuristics. More recently, simulation data mining that uses mesh models has been shown to be effective in assisting designers in tracking design change [36].

Graening and Sendhoff suggest several methods for shape mining to enable data mining techniques in engineering design to integrate data across design teams dealing with different simulations, as they argue these techniques are restricted to single design processes and individual design teams [14]. Data mining for such research is more of knowledge discovery by looking at associations (finding dependencies in an analyzed data set), clustering (creating clusters of objects in a way to ensure the highest possible similarity between group members), classification/regression (creating a dependency model between independent variables describing given objects), or description (concise summarizing of analyzed data) [30].

3 STUDIED CASE

Automakers prioritize efforts to reduce the overall weight of their vehicles as heavier automobiles have greater inertia and rolling resistance, contributing to different issues such as higher fuel consumption [22]. The main structure of a car, without mounting the motor, seats, electronics, etc., is known as the Body-In-White (BIW). This is where the main energy-absorbing structures are located. The BIW accounts for 30%-40% of a car's total weight. Therefore, the need to develop a low-weight frame with high Specific Energy Absorbing (SEA) capacity is bigger than ever. The BIW of an automobile consists of multiple TWBs, and each one of these beams is designed to absorb the highest energy (in the moment of crash) and deform with the lowest Peak Crushing Force (PCF). The maximum force needs to be lower to minimize the amount of acceleration that affects the passengers [3]. This makes the automobile frame a complex system of different TWBs that must work together to fulfill the function of stiffness, crashworthiness, and low weight.

Within the automotive industry, the analysis of structural components like TWBs is crucial, given their role in vehicle safety and performance. Many automakers use a repetitive design process to evaluate the performance of these designed beams. As the final design needs to be integrated with other systems, these design iterations can take up to many years [27]. Therefore, being able to predict the performance of these beams is of utmost importance. To show such ability in this paper, we use cross-section geometries for the existing beams in the literature [50]. Figure 3 shows 46 geometries that are a simplified representation of a Toyota RAV4's frame [51]. Using these images of the geometries, a similar scaled curve is extracted in SolidWorks for each shown cross-section. The cross-sections were systematically cataloged in STEP files using an image to CAD capabilities of SolidWorks software.

Later, to be able to show the predictability, simulation results of the SEA and PCF of each one of these tubes that are acquired [15]. A dynamic explicit simulation with semi-automatic mass scaling is used, and tests are carried out under lateral load. The STEP files are imported as a sketch to the FEA software ABAQUS and then extruded as shell elements; the process is visualized in Figure 4 (from geometry definition to discretization). A filter is used to stop the simulation at the correct time; this filter monitors the speed of the analytical wall and stops the simulation when the wall velocity reaches 0. It is worth mentioning that the FEA simulations in this paper are performed for only one thickness. However, testing for three different thicknesses (1, 2, and 3 mm) is carried out in a separate study, which shows that such superiority is not limited to the specific scale of the geometry.

Several output variables are requested from the simulation every 5E-5 seconds. The reaction force is the first variable requested, measured at every node in the tube's boundary condition. The other two variables are the velocity of the wall displacement. Since the wall touches the tube at the beginning of the simulation, the measured amount of wall displacement is the same as the tube displacement. In this way, all cross-sections are simulated to read out their crashworthiness characteristics, i.e., PCF and SEA. The verification of the FEA model was done with the results of published literature [1]. Many of the geometries that are simulated have spot welds between the different plates. To simplify the FEA process, they are considered rigid nodes as suggested in the literature [45]. Since the performance of the welds is not of interest in this study, this choice

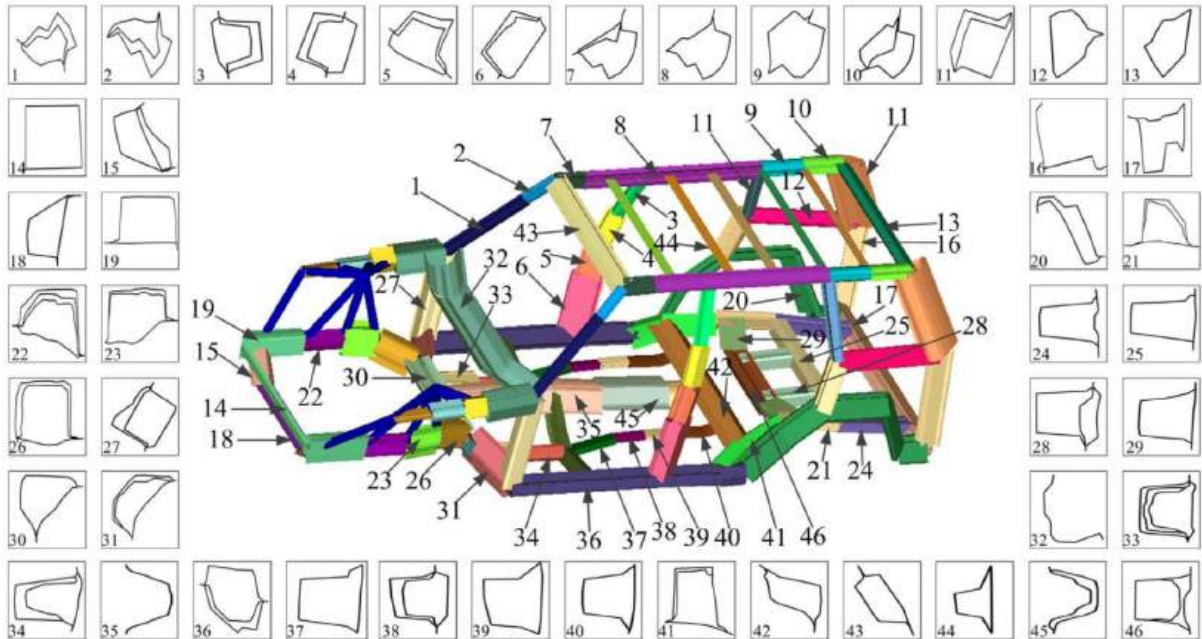


Figure 3: An example of frames in BIW with many different TWBs geometries.

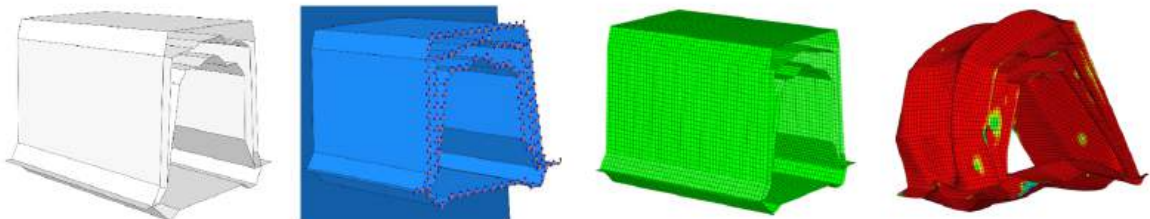


Figure 4: The finite element process of one example cross-section.

does not affect the final conclusions.

4 CONSTRUCTING FEATURES

This paper suggests using an alternative geometric representation of a shape as a means for data mining before starting feature engineering in the process of training machine learning models. The medial axis or the 'skeleton' of a shape is the set of all points inside the shape that have an equal distance to two or more points on the shape's boundary. The medial axis of a simple polygon is closely related to the Voronoi diagram constructed from its edges [27]. The shape of the medial axis is unique for each polygon and is reversible, so the medial axis can be used to restore a shape based on its medial axis. In general, one can consider the medial axis of a shape as an alternative geometric representation of the shape but in a lower dimension. This

is because the medial axis of a 3d shape is a 2d surface, and the medial axis of a 2d shape reduces down to a curve. The shape's convex or concaveness is not important for our use case aim because it just determines the direction and position of the medial axis. However, the shaper must be closed, and indeed, three of the cross-sections that are not closed polygons (numbers 16, 32, and 35) are removed from our analyses.

There are many ways to construct a medial axis of shape. In this paper, we use the Rhino Grasshopper. Rhino is a specialized 3D design software employed extensively in industrial design applications. Grasshopper is a complementary plug-in to Rhino, which offers visual programming capabilities and can aid in the creation of transparent design automation assets for designers [16]. The saved STEP files (from the previous section) were imported into Rhino Grasshopper using the yellow part of the script shown in Figure 5. The figure shows that after importing the geometry, the data goes through components to create a boundary surface, and then the surface splits into several segments in another component depending on how many sections exist in the geometry. The blue section in the figure reconstructs the medial axis of each section and then combines them with a series of components for later analysis. This is shown with two output components in the figure: the 'Radius of circles' and 'Medial axis' segments.

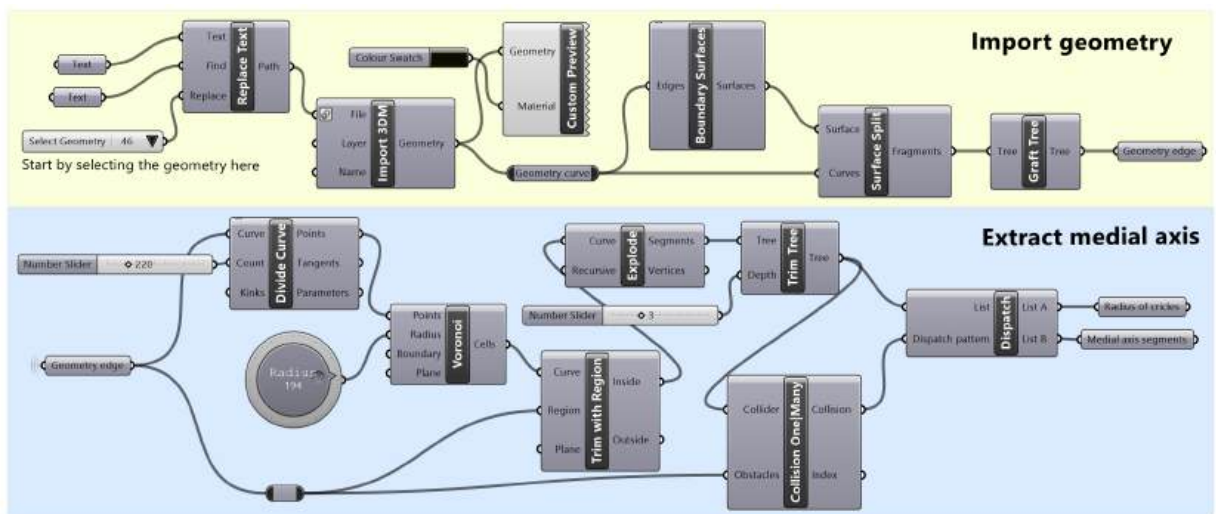


Figure 5: The visual representation of the Rhino grasshopper code.

The steps for constructing the medial axis of a shape are as follows. First, the boundary is divided into equally distanced points (the number can be adjusted with a slider component), and then a circle is grown on each cell to create Voronoi cells. Here, every point in the circumference acts as a seed for these Voronoi cells. When these cells reach each other (due to increasing the size of cells), they create the desired medial axis. It is important to note that not all the edges of the Voronoi diagram belong to the medial axis, and the constructed lines need to be pruned, which is done by trim components shown in Figure 5. Depending on the pruning process, some noise will be introduced to the gathered data [41]. However, since the nature of the use here is comparative, the introduced noise will not affect our conclusion. The assumption is if the gathered (noisy) data is shown to be effective for the end purpose (prediction model), the healthier data can even show more promising results and not worse. On the other hand, the hyperparameters for creating such a medial axis were selected similarly for all construction; therefore, the noise should affect all geometries equally.

Four steps of creating a medial axis from dividing the curve from a STEP file to edge points for each geometry, growing the seeds of Voronoi, and pruning are shown in Figure 6. From left to right, the seeds and

small circles growing from them are shown in the first figure, and then the next two figures show how circles behave when they start to grow more and reach each other. In the next figure, the blue line represents the shape's medial axis or skeleton, and the final figure shows the fully developed medial axis shown in the blue lines. Here, all the gray-colored lines, which are radii of the circles that have developed the medial axis, need to be pruned.

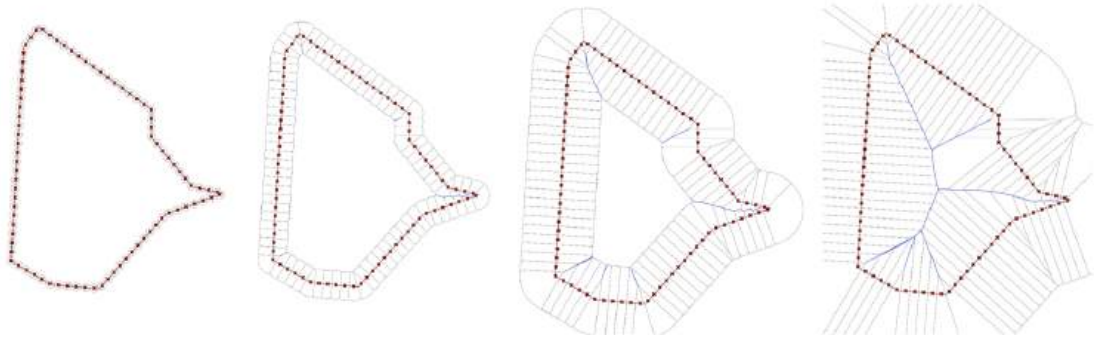


Figure 6: Voronoi process to get the medial axis of a cross-section.

5 EXTRACTING FEATURES

Construction of the medial axis from Voronoi diagrams of the shape's boundary is an advantageous data mining process that then needs to be continued further by extracting useful information according to the features engineering process shown earlier. Medial axis construction not only gives back the skeleton of the shape but also the radius of the circles (which resembles the fractal's width), which contributes to different kinds of information about geometry. These radii can reveal other useful information for the predictive task at hand. Figure 7 illustrates the results of the applied process and the acquired medial (in blue line) axis and radius of the circles (in gray lines) for three cross sections. The black curves correspond to the geometry's outer circumference. It can be argued that while the medial axis is coupled to the shapes' intricate geometry, the constructing radii are related to the regions of the shape and thus can hold regional-based information for end prediction models.

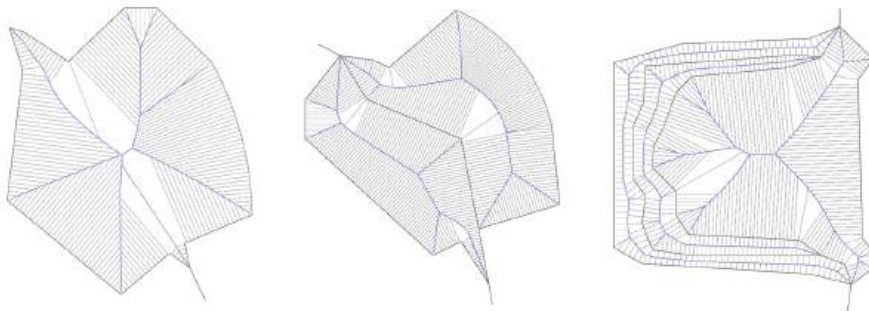


Figure 7: Several geometries (3 out of 46) after the Voronoi operation.

5.1 Region-Based Features

The region-based features are informative about those characteristics of the shape that are related to the area and regions. The gray lines in Figure 7, which are the radius of the circles and cover the shape area, can be used to extract two examples here by averaging and summing the length of all the gray lines in the shape, as shown in Figure 8. The two features in the dataset are called *Avg. circle radius* and *Width information*. The average circle radius acts as a measure to show how regular or symmetrical the shape is. Shapes with a more consistent average circle radius tend to be more regular shapes. This can be important in classifying objects or detecting anomalies in a dataset. The average circle radius is indirectly related to the shape's overall size and scale. While it doesn't give the exact area, it provides a sense of whether shapes are predominantly large or small compared to the other shapes in the dataset. If the size of the shapes matters for the prediction model, this feature could be a useful input.

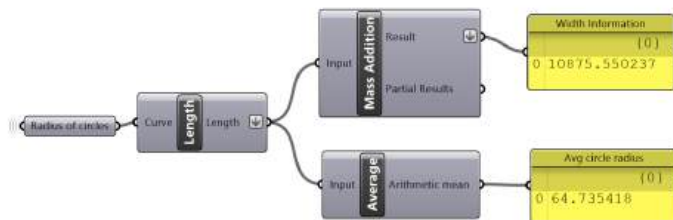


Figure 8: Extracting 'avg. circle radius' and 'Width information' features.

5.2 Fractal-Based Features

The medial axis-based features can play a vital role in more intricate geometries that allow for no parameterization or require enormous parameters to control every degree of freedom. The medial axis (blue curve) shown in Figure 7 can be used to extract several features such as the *length of the medial axis* and *number of branch points* as called in the dataset. Figure 9 shows the process of measuring these features. The length of the medial axis can offer insight into the overall size and extent of the shape. Longer medial axes generally correspond to larger, more elongated shapes. This information can be crucial in differentiating between shapes of varying sizes and proportions. On the other hand, the number of branch points reflects the complexity of the shape's internal structure. Shapes with many branch points often exhibit intricate branching patterns or protrusions.

5.3 Boundary-Based Features

Perimeter offers a basic size estimator. While it doesn't directly capture intricate details, it provides a general size reference for comparison between shapes. This can be valuable for a prediction model involving any kind of size normalization or adjustment. In topology, 'handles' refer to the number of holes in the shape. A higher number of handles correlates with a more complex boundary, potentially indicating cavities or indentations. This metric can be important for classification or feature analysis when surface complexity is significant. Since it uses the edge curve of the original shape, as shown in Figure 10, it is placed among the boundary-based features.

The list of features one can extract from an alternative representation of geometry, like the medial axis, can be long and may vary based on the use case. Some features that show good quality for one use case

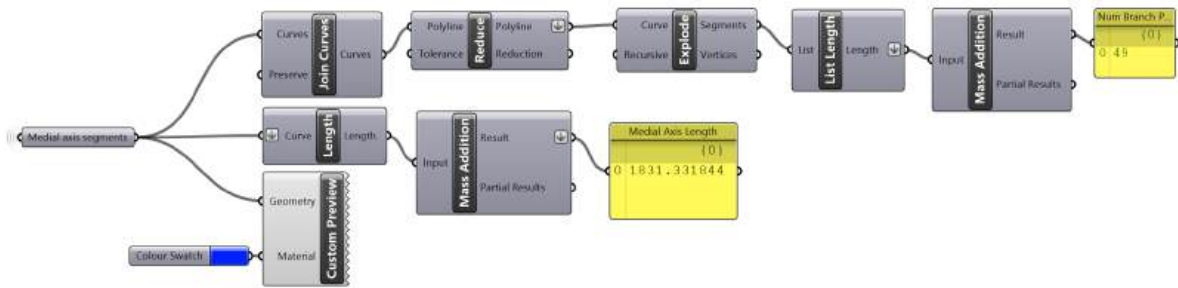


Figure 9: Extracting 'Number of branch points' and 'Medial axis length' features.

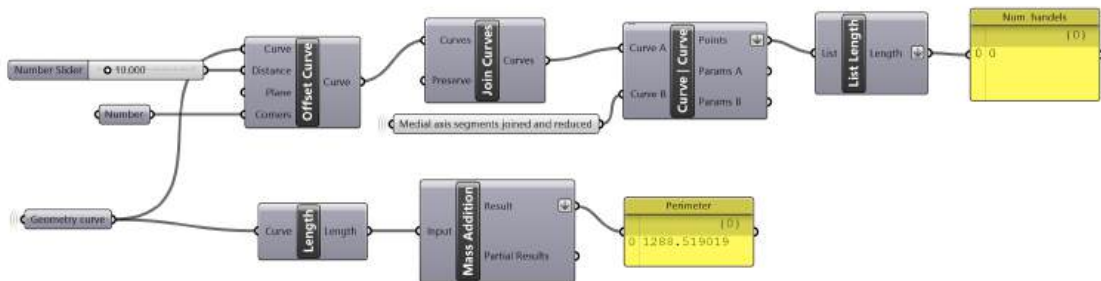


Figure 10: Extracting 'Number of handles' and 'perimeter' features.

might not show the same importance for others. Here, we only give examples of different possible scenarios. One other possible scenario in this case study is to use a mix of the above sources to extract a feature. For example, the compactness ratio of a polygon shape is a well-known metric for 2d and 3d shapes and is an intrinsic property of objects [32]. This is calculated as the ratio of a shape's area to its bounding circle's area (shown in Figure 10). For the same area, shapes that deviate significantly from being circular will have lower compactness ratios. This feature can offer insights into how well a shape fills its space. Some other features can help differentiate shapes with similar areas but drastically different contours.

5.4 Other Features

The list of features one can extract from an alternative representation of geometry, like the medial axis, can be long and may vary based on the use case. Some features that show good quality for one use case might not show the same importance for others. Here, we only give examples of different possible scenarios. One other possible scenario in this case study is to use a mix of the above sources to extract a feature. For example, the compactness ratio of a polygon shape is a well-known metric for 2d and 3d shapes and is an intrinsic property of objects [32]. This is calculated as the ratio of the area of a shape to the area of its bounding circle (shown in Figure 11). For the same area, shapes that deviate significantly from being circular will have lower compactness ratios. This feature can offer insights into how well a shape fills its space. Some other features can help differentiate shapes with similar areas but drastically different contours.

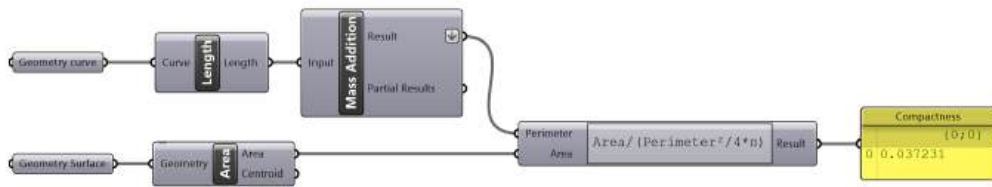


Figure 11: Extracting 'compactness ratio' as a feature.

After analyzing all 46 cross-sections that are introduced in Figure 3 with the shown procedures. Finally, all extracted features are recorded in a spreadsheet to be further analyzed in the feature selection process. This database is shown partially in Table 2. It is worth mentioning that the order of magnitude and unit are of no interest due to the nature of the analysis being performed, and if necessary, this can be done later by a normalizing method as a subtask under the matching learning process.

Num	Length of Medial Axis	Width Information	Number of Handles	Branching Points	Shape Perimeter	Avg Circle Radius	Compactness	SEA (KJ/Kg)	PCF (N)
1	3104.1	26054.3	15	96	1830.8	29.6	0.0268	15.16	92033
2	3706.2	23187.1	13	106	2223.7	26.3	0.0211	12.52	108135
3	3446.5	33735.4	14	83	1922.9	38.3	0.0314	14.51	94598
...
41	3451.4	30293.5	11	83	2209.1	34.4	0.0282	13.11	106251
42	1746.4	25227.4	7	30	1316.8	57.3	0.0527	21.65	64662
43	1614.5	25573.9	6	29	1213.7	58.1	0.0553	22.87	61466
44	1502.3	17552.7	4	29	1221.6	39.9	0.0433	23.26	59824
45	1860.9	8991.6	11	62	1671.4	20.4	0.0155	17.03	80601
46	3279.2	41802.3	13	70	1826.2	47.5	0.0386	14.47	95184

Table 2: Dataset with extracted features and the two CAE outputs as target values.

6 SELECTING THE EXTRACTED FEATURES

After the feature extraction process (Figure 2), it is crucial to employ feature selection techniques to determine which one of the extracted features has the strongest predictive power. Correlation measures the strength and direction of a linear relationship between two variables and thus is a suitable metric to measure the quality of extracted features in the context of prediction modeling [27]. Unlike previous studies, we propose to use different correlation techniques to study the quality of the extracted features to account for both linear relations in the data as well as nonlinear relations. A high absolute correlation coefficient (close to 1 or -1) indicates a strong relationship, suggesting that the feature significantly impacts the target variable and can be

a good predictor. Table 3 shows three studied correlations for all extracted features. For example, the Pearson correlation between the *Length of Medial Axis* and *SEA* is -0.91, and between the same feature and *PCF* is 0.94, which shows this feature is highly correlating with both defined target values.

	Linear regression score		Pearson correlation		Spearman correlation	
	SEA	PCF	SEA	PCF	SEA	PCF
Length of Medial Axis	0.83	0.89	-0.91	0.94	-0.93	0.92
Width Information	0.26	0.23	-0.50	0.48	-0.57	0.58
Num. of Handels	0.68	0.77	-0.82	0.88	-0.81	0.81
Branching Points	0.70	0.79	-0.83	0.89	-0.86	0.85
Shape Perimeter	0.87	0.98	-0.93	0.99	-0.99	0.98
Avg. Circle Radius	0.34	0.30	0.58	0.54	0.58	-0.58
Shape Compactness	0.57	0.52	0.75	0.72	0.74	-0.74

Table 3: Correlation between mined features and two FEA outputs.

In this table, a simple linear relationship derived from fitting a linear line to the data is used to describe the complexity of the relation. This parameter can be extracted from the score of the linear regression model in Keras [10]. This score mainly tells us how well a simple linear model explains the variability in one variable based on changes in another variable.

Parametric Correlations are used when data is assumed to follow a normal distribution. For example, the Pearson Correlation Coefficient is the most common measure that captures the strength and direction of a linear relationship between two continuous variables. There are also nonparametric Correlations, such as Spearman, which is rank-based and is used when data is not normally distributed or contains outliers. Spearman Correlation (Spearman's rho) measures the strength and direction of a monotonic relationship between two variables (continuous or ordinal). A monotonic relationship means the variables consistently increase or decrease together, but not necessarily at a constant rate [33].

Using three types of correlation - Linear Regression Score, Pearson Correlation, and Spearman Correlation - provides a more comprehensive evaluation of the relationships between features and the target variable in your data. Each of these metrics offers a unique perspective on the data, allowing for a more nuanced understanding of how different features might influence the model's predictions.

7 DISCUSSION

This paper demonstrates the value of utilizing feature engineering in CAD for surrogate modeling. To this end, the medial axis, derived from Voronoi diagrams of a shape's boundary, is used for constructing, extracting, and selecting features. The medial axis provides a skeletal representation of the shape and can potentially hold more valuable information for prediction tasks. One of these valuable features that was identified was the radii of the contributing circles, which offer insights into regional variations and can be interpreted as fractal width measurements. Other features identified were the *number of handles*, *number of branching points*, *shape perimeter*, *average circle radius*, and *shape compactness*. These extracted features could be integrated with machine learning models to develop robust predictive tools for various engineering applications.

Table 3 shows the length of the medial axis has a good correlation with the crashworthiness characteristics of the geometry (calculated from finite elements as mentioned earlier). Specifically, an average correlation of 0,8613 and 0,8104 for peak crushing force and specific energy absorption, respectively. This is highly advantageous since such parameters that are hidden in CAD can represent the objective function and thus

can be used as part of the selected features for further data analysis in predicting the outputs. This table also reveals which features do not show good correlation and thus should be removed in the selection process (*average circle radius*).

Developing thin-walled tubes is a resource-intensive process; thus, surrogate models are heavily used in this domain to predict TWBs' crashworthiness characteristics. The methodology shown in this paper for feature engineering on CAD geometries can free surrogate models from parameterization and reduce their dimensionality. More efficient surrogate models can accelerate the design loops. Considering that one FEA simulation of a TWB takes, on average, 30 minutes while calculating the sleeping parameter only takes a few seconds, shows how such feature engineering can be beneficial for accelerating the development process of TWBs. Especially for more advanced models that take hours and days of simulation time over the development process iteratively.

It should be noted that the number of features that need to be constructed in a feature engineering process depends on the complexity of the response surface at hand. And it is not always for granted that most extracted features will be of high quality. This puts the burden on the domain engineer to understand the extracted features and suggest a way forward. However, to give an idea about what possibilities exist with our methodology, we present a categorization based on the nature of the information that these features hold. Also, to inspire designers for more possibilities, several other possible features that are not studied for this use case are also added to Table 4.

Feature Type	Examples	Description
Boundary-based	Perimeter, convex hull, radius of gyration, Euler number, profile, bounding box parameters	Extracted directly from the shape's boundary
Region-based	Area, mean intensity, Eccentricity (elongated or stretched) variance, entropy, texture, Central moments, Hu moments, or Zernike moments can capture shape properties, compactness	Extracted from the interior of the shape
Fractal-based	Fractal dimension, Skeletonization Features (length, branches, loop handles), eigenvalues and eigenvectors of the covariance matrix of shape points, Tortuosity (Measuring the "wiggleness" of the medial axis), Angles between Branches	Describes the self-similar structure of the shape

Table 4: Categorization of features beyond traditional feature selection in CAD.

Boundary based features are valuable measures of the shape's overall size or material usage. Among the three types of features, boundary-based features are the most susceptible to minor imperfections or noise along the boundary, so pre-processing steps should be noted to smooth the boundary or reduce the impact of noise. As it can be inferred from the results in Table 4 for our use case, the correlation of perimeter is the best among features. This means shape descriptors derived from boundary and medial axis data can potentially aid in quality control, identifying deviations from the intended design and catching defects for other prediction models in the industry.

Region based features hold information about local variations in thickness. A shape with a narrow geometric distribution suggests consistent thickness, while a wide distribution implies significant changes in curvature. This can potentially correlate with areas susceptible to high stress. For example, in an optimization process of an aircraft wing spar, it can be used for thickening areas along the medial axis where stress is highest, and conversely, in areas experiencing lower stress, the thickness can be reduced to save weight without compromising structural integrity. As another use case, maintaining consistent thickness is crucial for analyzing flow paths within a fluidic system. By monitoring the average circle radius, engineers can ensure specific regions

within the system are optimized to prevent turbulence and maintain uniform pressure. This consistency helps in achieving efficient flow and system reliability.

Compactness is a unique feature that combines regional and boundary information. Table 3 shows it is moderately important concerning correlating with the FEA output. However, the importance of such hybrid features can only be revealed when studying if they correlate with the functional performance of engineered components. Since compactness is defined as the ratio of area to its bounding box, it can exhibit useful information when used in specific domains. For example, in aerodynamic analysis, the shapes with higher compactness exhibit lower drag, making them more aerodynamically efficient, or in heat transfer analysis, more compact shapes have a higher surface-area-to-volume ratio, which can mean faster heat dissipation.

Fractal-based features can hint at a structure's strength, how well it can distribute stress, and its potential deformation under a load. For instance, objects with a higher fractal dimension tend to have more surface area, potentially increasing stress distribution capacity. In engineering design, fractal attributes extracted from shapes offer insights into their complexity and their behavior under different conditions, and therefore, it is no surprise that the length of the medial axis holds a very good rank among the extracted features. On the other hand, the way a fractal structure branches out and the angles between branches hold information about the underlying generative process of the shape. The power of fractal-based features is their ability to capture intricate geometric details that traditional parametric models often miss. They provide a powerful feature for engineers and designers to quickly analyze and optimize different design cases.

In CAD models, complex shapes are often represented by a multitude of parameters (dimensions, angles, curvatures, etc.). Extracting features from alternative geometric representation can offer features with a highly descriptive metric that encapsulates significant shape information. This dimensionality reduction is essential for making the machine learning surrogate model more accurate and efficient as well as interpretable.

Another advantage of the introduced feature engineering process is that having such a reductionist approach in finding new features provides a holistic measure that is relatively insensitive to minor boundary irregularities, enhancing the reliability of machine learning models to design changes. This is because one problem with prediction models is that when applying design iterations to the shape at hand, designers often are forced to add or remove from the geometric shape (a simple example is adding a curve or removing a line shown in Figure 1). This small change drastically affects how CAD parameterization has been done and will make the trained model useless. In other words, using the approach introduced in this paper will prevent the design change from propagating to the trained machine-learning model and allow for further analyses without the maintenance of such models.

Moreover, specific CAD parameters might be highly domain-dependent. Feature engineering on CAD replaces them with universally relevant geometric properties. Machine learning models trained with such features have the potential to generalize better across different kinds of shapes and applications. However, the disadvantage is the lack of understanding of specific prediction tasks and domain knowledge, which is the biggest key to successful feature extraction and selection. As it is well known, correlation does not imply causation. Finding a correlation suggests an association, but further investigation is needed to determine if one variable actually causes changes in another. However, since the aim of feature engineering in CAD is to find features for prediction models that are not dependent on the parameterization of the geometry, the fact that it shows correlations or causation is not of importance. Because the selection process validates their predictive power. Thus, the nature of the relation should not matter for the machine learning task that is planned to be carried out further in the process.

The concepts of the medial axis, radius of constructing circles, and other derived features can be extended into 3D analysis for even richer geometric insights. However, this study and further investigation of the suggested features in Table 4 are planned to be pursued in future studies.

8 CONCLUSIONS

The paper seeks to generalize the concept of sleeping parameters as an alternative way of using latent information in CAD to construct, extract, and select features for data-driven approaches in engineering design. Feature engineering has been used in data science to improve the quality of the inputs as a preprocess for machine learning. By combining the medial axis representation with Voronoi-derived circle radii, we obtain a rich set of features that capture both the skeletal and regional properties of complex shapes. These features offer valuable insights for predictive tasks across diverse engineering domains. The methodology is showcased on a crashworthiness case and is an example of how the medial axis can create new features that correlate highly with the objective performance. We conclude that the application of this concept can also improve the application of data science in engineering design by enabling much more efficient mapping between input and output while it enables design loops to be independent of the CAD parameterization. By leveraging this approach, engineers and designers can enhance the efficiency of design processes, facilitate iterative loops, and improve the accuracy of regression models.

ACKNOWLEDGEMENTS

This work has been carried out at CHALMERS University of Technology. But the authors would like to sincerely thank researchers in Jönköping University Andreas Hedlund Daniel Blom for supporting this work and Professor Roland Stolt for his invaluable guidance.

Mohammad Arjomandi Rad, <https://orcid.org/0000-0002-7894-7734>

Massimo Panarotto, <https://orcid.org/0000-0001-5216-0944>

Ola Isaksson, <https://orcid.org/0000-0003-0373-3720>

REFERENCES

- [1] Al Galib, D.; Limam, A.: Experimental and numerical investigation of static and dynamic axial crushing of circular aluminum tubes. 42(8), 1103–1137. ISSN 0263-8231. <http://doi.org/10.1016/j.tws.2004.03.001>.
- [2] Arjomandi Rad, M.; Cenanovic, M.; Salomonsson, K.: Image regression-based digital qualification for simulation-driven design processes, case study on curtain airbag. 34(1), 1–22. ISSN 0954-4828. <http://doi.org/10.1080/09544828.2022.2164440>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/09544828.2022.2164440>.
- [3] Arjomandi Rad, M.; Khalkhali, A.: Crashworthiness multi-objective optimization of the thin-walled tubes under probabilistic 3d oblique load. 156, 538–557. ISSN 0264-1275. <http://doi.org/10.1016/j.matdes.2018.07.008>.
- [4] Bang, H.; Selva, D.: iFEED: Interactive feature extraction for engineering design. In Volume 7: 28th International Conference on Design Theory and Methodology, V007T06A037. American Society of Mechanical Engineers. ISBN 978-0-7918-5019-0. <http://doi.org/10.1115/DETC2016-60077>.
- [5] Bonde, J.M.; Brahma, A.; Panarotto, M.; Isaksson, O.; Wärmefjord, K.; Söderberg, R.; Kipouros, T.; Clarkson, P.J.; Kressin, J.; Andersson, P.: Assessment of weld manufacturability of alternative jet engine structural components through digital experiments. In 2022 Proceedings of ISABE. International Society for Air Breathing Engines (ISABE). <http://doi.org/http://dx.doi.org/10.5281/zenodo.7973381>.
- [6] Burrows, S.; Stein, B.; Frochte, J.; Wiesner, D.; Muller, K.: Simulation data mining for supporting bridge design. In Australasian Data Mining Conference. <https://dl.acm.org/doi/10.5555/2483628.2483647>.

- [7] Camba, J.; Contero, M.; Company, P.; Hartman, N.: The cost of change in parametric modeling: A roadmap. In CAD'20, 31–35. CAD Solutions LLC. <http://doi.org/10.14733/cadconfP.2020.31-35>.
- [8] Chappell, D.: Introducing azure machine learning. http://download.microsoft.com/download/3/B/9/3B9FBA69-8AAD-4707-830F-6C70A545C389/Introducing_Azure_Machine_Learning.pdf.
- [9] Cheng, Z.; Ma, Y.: Explicit function-based design modelling methodology with features. 28(3), 205–231. ISSN 0954-4828. <http://doi.org/10.1080/09544828.2017.1291920>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/09544828.2017.1291920>.
- [10] Chollet, F.; et al.: Keras. <https://keras.io>, 2015.
- [11] Dagli, C.H.; Lee, H.C.: Impacts of data mining technology on product design and process planning. In F. Plonka; G. Olling, eds., International Conference on Computer Applications in Production and Engineering (CAPE97) 5-7 November 1997, Detroit, Michigan, USA, IFIP - The ation for Information Processing, 58–70. Springer US. ISBN 978-0-387-35291-6. http://doi.org/10.1007/978-0-387-35291-6_6.
- [12] Dong, G.; Liu, H.: Feature engineering for machine learning and data analytics. CRC press. <https://doi.org/10.1201/9781315181080>.
- [13] Dong, X.; DeVries, W.R.; Wozny, M.J.: Feature-based reasoning in fixture design. 40(1), 111–114. ISSN 0007-8506. [http://doi.org/10.1016/S0007-8506\(07\)61946-5](http://doi.org/10.1016/S0007-8506(07)61946-5).
- [14] Graening, L.; Sendhoff, B.: Shape mining: A holistic data mining approach for engineering design. 28(2), 166–185. ISSN 1474-0346. <http://doi.org/10.1016/j.aei.2014.03.002>.
- [15] Hedlund, A.; Blom, D.: Correlation-based analysis on thin walled tubes. <https://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-58398>.
- [16] Heikkinen, T.: TRANSPARENCY OF DESIGN AUTOMATION SYSTEMS USING VISUAL PROGRAMMING - WITHIN THE MECHANICAL MANUFACTURING INDUSTRY. 1, 3249–3258. ISSN 2732-527X. <http://doi.org/10.1017/pds.2021.586>.
- [17] Isaksson, O.; Eckert, C.: Product development 2040. <http://doi.org/https://doi.org/10.35199/report.pd2040>.
- [18] Isaksson, O.; Eckert, C.; Panarotto, M.; Malmqvist, J.: YOU NEED TO FOCUS TO VALIDATE. 1, 31–40. ISSN 2633-7762. <http://doi.org/10.1017/dsd.2020.116>.
- [19] Kim, J.; Pratt, M.J.; Iyer, R.G.; Sriram, R.D.: Standardized data exchange of CAD models with design intent. 40(7), 760–777. ISSN 0010-4485. <http://doi.org/10.1016/j.cad.2007.06.014>.
- [20] Koch, P.; Allen, J.; Mistree, F.; Mavris, D.: The problem of size in robust design. <http://doi.org/10.1115/DETC97/DAC-3983>.
- [21] Kuhlmann, A.; Vetter, R.M.; Lübbling, C.; Thole, C.A.: Data mining on crash simulation data. In P. Perner; A. Imiya, eds., Machine Learning and Data Mining in Pattern Recognition, 558–569. Springer. ISBN 978-3-540-31891-0. http://doi.org/10.1007/11510888_55.
- [22] Malen, D.E.: Fundamentals of automobile body structure design. SAE Technical Paper. <https://www.sae.org/publications/technical-papers/content/R-394/>.
- [23] Murshed, S.M.M.; Dixon, A.; Shah, J.J.: Neutral definition and recognition of assembly features for legacy systems reverse engineering. 615–628. American Society of Mechanical Engineers Digital Collection. <http://doi.org/10.1115/DETC2009-86739>.
- [24] Obermaier, H.; Peikert, R.: Feature-based visualization of multifields. In C.D. Hansen; M. Chen; C.R. Johnson; A.E. Kaufman; H. Hagen, eds., Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization, 189–196. Springer. ISBN 978-1-4471-6497-5. http://doi.org/10.1007/978-1-4471-6497-5_17.
- [25] Painter, M.K.; Erraguntla, M.; Hogg, G.L.; Beachkofski, B.: Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management. In Proceedings of the 2006

- Winter Simulation Conference, 1253–1260. <http://doi.org/10.1109/WSC.2006.323221>. ISSN: 1558-4305.
- [26] Park, S.; Joung, J.; Kim, H.: Spec guidance for engineering design based on data mining and neural networks. 144, 103790. ISSN 0166-3615. <http://doi.org/10.1016/j.compind.2022.103790>.
- [27] Rad, M.A.; Salomonsson, K.; Cenanovic, M.; Balague, H.; Raudberget, D.; Stolt, R.: Correlation-based feature extraction from computer-aided design, case study on curtain airbags design. 138, 103634. ISSN 0166-3615. <http://doi.org/https://doi.org/10.1016/j.compind.2022.103634>.
- [28] Reich, Y.: Data mining of design products and processes. In O. Maimon; L. Rokach, eds., *Data Mining and Knowledge Discovery Handbook*, 1167–1187. Springer US. ISBN 978-0-387-25465-4. http://doi.org/10.1007/0-387-25465-X_55.
- [29] Robinson, T.; Friel, I.; Armstrong, C.G.; Murphy, A.; Butterfield, J.; Price, M.; Marzano, A.: Computer-aided design model parameterisation to derive knowledge useful for manufacturing design decisions. 232(4), 621–628. ISSN 0954-4054. <http://doi.org/10.1177/0954405417708218>. Publisher: IMECHE.
- [30] Rogalewicz, M.; Sika, R.: Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. 7(4), 97–108. <https://bibliotekanauki.pl/articles/407431.pdf>. Publisher: Polska Akademia Nauk. Czytelnia Czasopism PAN.
- [31] Romanowski, C.J.; Nagi, R.: A data mining-based engineering design support system: A research agenda. In D. Braha, ed., *Data Mining for Design and Manufacturing: Methods and Applications*, 161–178. Springer US. ISBN 978-1-4757-4911-3. http://doi.org/10.1007/978-1-4757-4911-3_7.
- [32] Santiago-Montero, R.; Bribiesca, E.; Santiago, R.: State of the art of compactness and circularity measures. 4, 1305–1335. <https://www.m-hikari.com/imf-password2009/25-28-2009/bribiescaIMF25-28-2009.pdf>.
- [33] Schober, P.; Boer, C.; Schwarte, L.A.: Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5), 1763–1768, 2018. <http://doi.org/https://doi.org/10.1213/ane.0000000000002864>.
- [34] Sexton, T.; Fuge, M.: Organizing tagged knowledge: Similarity measures and semantic fluency in structure mining. 142(31111). ISSN 1050-0472. <http://doi.org/10.1115/1.4045686>.
- [35] Shah, J.J.; Rogers, M.T.: Functional requirements and conceptual design of the feature-based modelling system. 5(1), 9–15. ISSN 2054-0353. <http://doi.org/10.1049/cae.1988.0004>. Publisher: IET Digital Library.
- [36] Shao, Y.; Liu, Y.; Ye, X.; Zhang, S.: A machine learning based global simulation data mining approach for efficient design changes. 124, 22–41. ISSN 0965-9978. <http://doi.org/10.1016/j.advengsoft.2018.07.002>.
- [37] Shi, F.; Chen, L.; Han, J.; Childs, P.: A data-driven text mining and semantic network analysis for design information retrieval. 139(111402). ISSN 1050-0472. <http://doi.org/10.1115/1.4037649>.
- [38] Siddharth, L.; Blessing, L.; Luo, J.: Natural language processing in-and-for design research. 8, e21. ISSN 2053-4701. <http://doi.org/10.1017/dsj.2022.16>. Publisher: Cambridge University Press.
- [39] Simpson, T.; Poplinski, J.; Koch, P.N.; Allen, J.: Metamodels for computer-based engineering design: Survey and recommendations. 17(2), 129–150. ISSN 1435-5663. <http://doi.org/10.1007/PL00007198>.
- [40] Sun, H.; Guo, W.; Shao, H.; Rong, B.: Dynamical mining of ever-changing user requirements: A product design and improvement perspective. 46, 101174. ISSN 1474-0346. <http://doi.org/10.1016/j.aei.2020.101174>.

- [41] Tam, R.; Heidrich, W.: Feature-preserving medial axis noise removal. In A. Heyden; G. Sparr; M. Nielsen; P. Johansen, eds., *Computer Vision - ECCV 2002*, vol. 2351, 672–686. Springer Berlin Heidelberg. ISBN 978-3-540-43744-4 978-3-540-47967-3. http://doi.org/10.1007/3-540-47967-8_45. Series Title: Lecture Notes in Computer Science.
- [42] Tang, M.; Wen, Y.; Mi, X.; Dong, J.: Parametric modeling with user-defined features. In *Proceedings of the Sixth International Conference on Computer Supported Cooperative Work in Design (IEEE Cat. No.01EX472)*, 207–211. <http://doi.org/10.1109/CSCWD.2001.942258>.
- [43] Wang, G.G.; Shan, S.: Review of metamodeling techniques in support of engineering design optimization. 129(4), 370–380. ISSN 1050-0472. <http://doi.org/10.1115/1.2429697>.
- [44] Wu, X.Y.; Hong, Z.X.; Feng, Y.X.; Li, M.D.; Lou, S.H.; Tan, J.R.: A semantic analysis-driven customer requirements mining method for product conceptual design. 12(1), 10139. ISSN 2045-2322. <http://doi.org/10.1038/s41598-022-14396-3>. Publisher: Nature Publishing Group.
- [45] Xiang, Y.; Wang, Q.; Fan, Z.; Fang, H.: Optimal crashworthiness design of a spot-welded thin-walled hat section. 42(10), 846–855. ISSN 0168-874X. <http://doi.org/10.1016/j.finel.2006.01.001>.
- [46] Xue, B.; Zhang, M.; Browne, W.N.; Yao, X.: A survey on evolutionary computation approaches to feature selection. 20(4), 606–626. ISSN 1941-0026. <http://doi.org/10.1109/TEVC.2015.2504420>. Conference Name: IEEE Transactions on Evolutionary Computation.
- [47] Yang, M.; Jiang, P.; Zang, T.; Liu, Y.: Data-driven intelligent computational design for products: method, techniques, and applications. 10(4), 1561–1578. ISSN 2288-5048. <http://doi.org/10.1093/jcde/qwad070>.
- [48] Zhao, Z.; Jin, X.; Cao, Y.; Wang, J.: Data mining application on crash simulation data of occupant restraint system. 37(8), 5788–5794. ISSN 0957-4174. <http://doi.org/10.1016/j.eswa.2010.02.029>.
- [49] Zheng, A.; Casari, A.: Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc.". <https://books.google.com/books?hl=en&lr=&id=sthSDwAAQBAJ&oi=fnd&pg=PT24&dq=features+engineering+in+data+science&ots=ZPXarX1jAZ&sig=UxB0G4-Ut1mZLQ01hSPVUz1TPeg>.
- [50] Zuo, W.; Lu, Y.; Zhao, X.; Bai, J.: Cross-sectional shape design of automobile structure considering rigidity and driver's field of view. 115, 161–167. ISSN 0965-9978. <http://doi.org/10.1016/j.advengsoft.2017.09.006>.
- [51] Zuo, W.J.; Bai, J.T.: Cross-sectional shape design and optimization of automotive body with stamping constraints. 17, 1003–1011. <https://doi.org/10.1007/s12239-016-0098-6>. Publisher: Springer.