Engineering Applications of Artificial Intelligence 136 (2024) 108957

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A cost modelling methodology based on machine learning for engineered-to-order products

Marco Mandolini [a,*], Luca Manuguerra [a], Mikhailo Sartini [a], Giulio Marcello Lo Presti [b], Francesco Pescatori [b]

[a] Department of Industrial Engineering and Mathematical Sciences, Università Politecnica Delle Marche, Via Brecce Bianche 12, 60131, Ancona, Italy
[b] Baker Hughes, Via Felice Matteucci 2, 50127, Firenze, Italy

ARTICLE INFO

Keywords:
Design to cost
Cost estimation
Machine learning
Cost engineering
Conceptual design
Conceptual costing

ABSTRACT

Recent scientific studies are targeted at applying and assessing the effectiveness of Machine Learning (ML) approaches for cost estimation during the preliminary design phases. To train ML prediction models, comprehensive and structured datasets of historical data are required. This solution is inapplicable when such information is unavailable or sparse due to the lack of structured datasets. For engineered-to-order products, the number of historical records is often limited and strongly influenced by different purchasing or manufacturing strategies, thus requiring complex normalisation of such data.

This method overcomes the above limitations by presenting an ML-based cost modelling methodology for the conceptual design that is applicable even when historical data are insufficient to train the prediction algorithms. The training dataset is generated through an analytical and automatic software tool for manufacturing cost estimation. Such a tool, starting from a 3D model of a product, can quickly and autonomously assess the related cost in different scenarios. An extensive and structured training dataset can be easily generated. The proposed methodology was based on CRISP-DM (Cross Industry Standard Process for Data Mining).

Cost engineers of an Oil & Gas company used the method to develop parametric cost models for discs and spacers of an axial compressor. The solution guarantees lower error (7% vs 9%) and significant time-saving (minutes instead of hours) than estimations based on other approaches. Cost models are more comprehensive (capable of analysing different scenarios), explainable (not conceived as a black box), and self-learning (can be updated by extending the training dataset).

1. Introduction and literature review

Nowadays, the cost of a product is a design driver as important as performance, environmental sustainability, and quality. Cost estimation demands significant manufacturing information to coordinate with many different areas, from design to production (Kadir et al., 2020). Knowledge-based methods are required to formalise and collect the information to be incorporated into software tools. Determining the production cost of a product during the preliminary design phases (e.g., conceptual design) is essential for a company's competitiveness (Lukić et al., 2016). Conceptual cost estimation should be feasible and fast, requesting only the information known during this design step (Ning et al., 2020a).

Cost-estimating methods based on parametric approaches (top-down) are the most suitable in these phases of product development (Masel et al., 2010). Parametric cost estimation methods work well when relationships between design variables (namely, cost drivers) and the cost are easily identifiable. Typically, costs are computed as the sum of elementary units representing the various resources used throughout the entire manufacturing cycle of a specific product, or they are calculated using an analytical function of variables reflecting multiple product attributes (Niazi et al., 2006).

1.1. Cost estimation relationship methods

From the industrial standpoint, engineers rely on linear regression approaches (cost estimation relationship – CER) to create simple parametric functions that relate cost to design features (e.g., mass, size, material) (Masel et al., 2010). Many linear regression-based parametric cost modelling applications have been published in the scientific

* Corresponding author.
E-mail address: m.mandolini@staff.univpm.it (M. Mandolini).

https://doi.org/10.1016/j.engappai.2024.108957
Received 11 December 2023; Received in revised form 12 June 2024; Accepted 8 July 2024

literature. Boothroyd and Reynolds (1989) used a parametric costing technique to estimate the cost early in the design process, using part volume as a parameter. Bertoni et al. (Bertoni and Bertoni, 2020) proposed a method using data parameters from concept design simulation through Computer Aided Engineering to estimate the life cycle costs of a Product Service System in the conceptual design phase of a collection of design alternatives. T. Kamps et al. (2018) proposed two integrated models for cost and life cycle assessment of producing low-volume or high-variant gear wheels through manufacturing parameters. Langmaak et al. (2013) presented a scalable cost model that predicts the unit cost of a gas turbine compressor's bladed disc based on an approach that uses process-based CERs to evaluate the cost of jet engine parts. COSYSMO 3.0 was a parametric cost estimation system that allowed users to understand numerically how its parameters affect cost estimates (Alstad, 2019).

1.2. Machine learning methods

The recent innovations introduced by Industry 4.0 (e.g., data mining and the Internet of Things) provide new tools and opportunities to overcome the issues of CER methods (Hammann, 2024; Van Nguyen et al., 2023). ML applications have drawn interest because they make industrial processes more efficient and make complex parameter causations easier to understand (Maier et al., 2022). For example, deep-learning techniques can automatically learn complex relationships between design features and manufacturing costs (Ning et al., 2020b). ML models for manufacturing cost estimation can be employed during the conceptual design phase to get precise predictions (Hennebold et al., 2022).

Recent scientific studies aim at applying and assessing the effectiveness of ML approaches for cost estimation during the preliminary design phases (Campi et al., 2021)(Kanyilmaz et al., 2022). ML is typically more effective in manufacturing cost estimation than conventional statistical and mathematical models (Yeh and Deng, 2012). Traditional techniques are still unable to forecast unknown feature values for a new piece and comprehend the relationships between the features of data samples (Dogan and Birant, 2021). Campi et al. also concluded that ML techniques are a better solution to linear regression techniques when the complexity of the problem increases (Campi et al., 2021). The same conclusion was drafted by Cavalieri et al. (2004), who demonstrated the excellent results of ML-based cost modelling techniques.

In the scientific literature, several approaches aim at creating cost models where cost formulas are made using regression analysis and neural networks (Verlinden et al., 2008). Loyer J. et al. (Loyer et al., 2016) demonstrated that ML appears to be an effective, affordable, accurate and scalable technique to estimate the cost of mechanical parts of a jet engine in the early stage of the design process. Chen et al. (Chen et al., 2021) and Bertoni et al. (Bertoni and Bertoni, 2020) give other ML applications for cost estimation in aviation. Campi et al. (2021) presented a cost estimation methodology based on ML (artificial neural networks, deep learning, random forest and linear regression) for manufacturing cost estimation of axial compressors. Cavalieri et al. (2004) demonstrated the effectiveness of ML techniques for evaluating the cost of a novel type of brake discs considering the weight, the unit cost of raw material, and the number of cores. Wang et al. (2013) built an ML model to estimate the cost of parts made by injection moulding. ML cost models are often combined with 3D computer-aided design (CAD) systems (Yoo and Kang, 2021; Ning et al., 2020a).

1.3. Challenges from industry and academia

CER methods used by the industry have several limitations. The cost prediction accuracy could be inadequate for the conceptual design phase (around 85%). The sparse historical data are often insufficient to precisely estimate the cost of new products (with different dimensions). For example, the models could not permit a comprehensive assessment when cost estimates are required in multiple production scenarios (e.g., country) (Martinelli et al., 2019). Models built on older data may need to be updated, or historical data may need to be cleaned (Weichert et al., 2019). The activities required to retrieve and normalise historical data could require an effort and a time that often is not congruent with the rapidity requested during the conceptual design. The recent challenges concerning the rapid growth of energy and raw material prices push cost engineers to update their analyses continuously. Cost models should be able to update over time through self-learning capabilities.

On the other hand, also ML methods have limitations. First, ML-based approaches suit repetitive products manufactured in medium to high volumes (Hammann, 2024). In these scenarios, comprehensive sets of manufacturing and financial information are available within databases of corporate software tools (e.g., ERP – Enterprise Resource Planning or MES – Manufacturing Execution System). Historical data are well structured and consistent. It is unnecessary to proceed with complex data cleaning and augmentation operations or normalisation (Rapaccini et al., 2023). Conversely, historical data may be sparse and poorly organised for engineered-to-order products. The number of records is often lower and strongly influenced by different purchasing (e.g., single sourcing) or manufacturing strategies (e.g., full-buy, farm-out) (Hammann, 2024). Here, the state-of-the-art solutions are not applicable.

Although ML techniques are a valid alternative widely used in many case studies, enterprises are still hesitant to adopt these techniques. Cost engineers consider such models black boxes (Hihn and Menzies, 2015). Thus, they often cannot interpret the results (Cavalieri et al., 2004)(Hihn and Menzies, 2015). Understanding the relationship between the main cost drivers and project costs is mandatory to adequately explain the cost model. Statistical learning approaches (e.g., feature selection and feature importance) help engineers meet this requirement (Elmousalami, 2021). In this way, it is possible to guarantee an objective selection of features without the subjectivity of cost engineers (Rapaccini et al., 2023).

1.4. Proposed methodology

The proposed method stems from the business's need to quickly and accurately estimate the cost of components during the conceptual design through a limited set of product parameters. The goal is to provide a cost modelling method that allows companies to develop models for conceptual cost estimation without using historical (sparse/unavailable) data.

This paper presents a systematic method for developing ML-based parametric models to estimate the cost of engineered-to-order products (and related components) during the preliminary design phase. Based on CRISP-DM (Cross Industry Standard Process for Data Mining), the approach overcomes the problem of sparse historical data by employing an analytical and automatic software tool to generate the dataset for training the cost model. The cost estimation tool can estimate the production cost by considering the complete manufacturing cost through its database of complex cost models, rules and parameters. Multiple production scenarios can be quickly evaluated (e.g., different production countries, machine hourly rates, raw material costs, energy costs) to construct a comprehensive training dataset. The originating ML-based model will allow design engineers to quickly estimate the manufacturing cost in different conditions (e.g., variation of energy cost, localisation/delocalisation, raw material shortage) during the conceptual design phase.

The proposed approach expects a step where cost drivers are sorted according to their importance in predicting costs. In this way, the cost model is explainable. Design engineers can use the prioritised list of design features to reduce the manufacturing cost of products effectively (Xie et al., 2023). Cost models developed through the proposed approach consider product and process parameters (e.g., dimensions, mass, material, production batch, country). The prediction models are

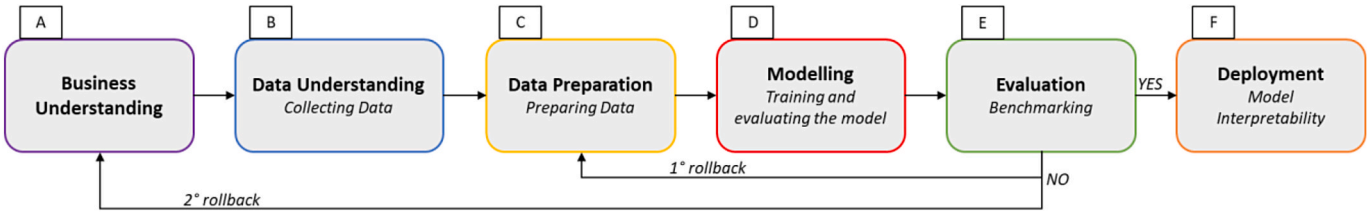

Fig. 1. Methodology workflow overview.

suitable for estimating the cost of parts similar to those used for training the ML algorithms (e.g., similar shape, dimensions and production process). For parts manufactured through different technologies and significantly different shapes and sizes, the prediction accuracy is lower than that computed by the cost model on the test dataset.

2. Cost modelling methodology

The cost modelling methodology employing supervised ML is based on the CRISP-DM (Cross Industry Standard Process for Data Mining) method (Fig. 1). CRISP-DM is a process model for data science and representation. It provides an overview of the data mining life cycle. Its flexibility and easy customisation allow for the creation of a data mining model that fits the goal of this work. Fig. 1 provides an overview of the proposed methodology using a UML Activity Diagram. The steps of the CRISP-DM and ML modelling workflows are reported respectively in bold and italic. The following sections describe each phase.

2.1. Business understanding

The Business Understanding phase focuses on conceptualising the project's objectives and requirements. This phase identifies business opportunities or customer needs and assesses if the available resources suit them. The present work proposes a cost modelling approach based on ML techniques. The parametric cost models should be employed to estimate the cost of mechanical parts at the conceptual design stage. So, for this methodology, the business understanding phase defines two essential requirements (*activity A1*).

- *The acceptance performance of the cost model*: At this phase, defining the cost model accuracy (prediction error) is crucial. For example, a cost estimate with an error of −15%/+20% may be acceptable at the conceptual design stage (ASTM E2516 – 11 Standard Classification for Cost Estimate Classification System).
- *Cost breakdown*: The total manufacturing cost is divided into set-up, labour, energy, investment and material. Often, it is requested to break down the cost to know better how to reduce it. Furthermore, the total manufacturing cost can be predicted more precisely by adopting specific cost models (one for each cost element). In other words, this point clarifies how many dependent parameters should be predicted (cost breakdown), so it is possible to define the number (N) of parametric cost models to create.

2.2. Data Understanding

The Data Understanding phase is usually used to identify, collect, analyse and verify the datasets to reach the project goals. In this case, this stage represents the initial step to collect all the data and information linked to parts of the family to estimate the cost, and it is essential for applying the methodology. The architecture of a machine (e.g., a cross-section of turbomachinery) consists of many components (e.g., disc, spacer, shaft, blade, nozzle, shroud), called items, often configured according to master models. So, the first step involves identifying the parts (of the same family) from existing machines for which the parametric cost model is to be created. For each identified part, two types of data are collected (*activity B1*, Fig. 2).

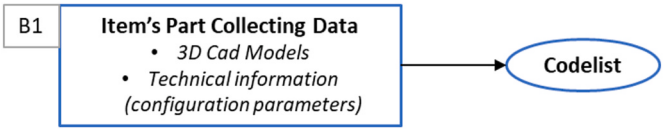

Fig. 2. Data Understanding workflow.

a. *3D CAD models*: the 3D CAD model for each piece is required for *C2 activity*. Parts without a 3D model are excluded from the following steps.
b. *Technical information*: a part family is described by several geometrical (e.g., dimensions) and non-geometrical parameters (e.g., material, part code, description, stage, production batch, unitary material cost) called configuration parameters. Such information is directly retrieved from PDM/EDM (e.g., material, mass, bounding box dimensions). For identifying other parameters (e.g., dovetail type of a blade, rotor type of a shaft), feature recognition algorithms can be employed when available. At last, manual identification from 3D models or drawings could be required.

The data are grouped and ordered in a single first dataset (called "Codelist"). All the associated configuration parameters describe each part. So, this dataset consists of as many records as the number of components and fields as the configuration parameters.

2.3. Data preparation

Data preparation is one of the crucial phases of the methodology (Fig. 3). It prepares the final dataset(s) for modelling. Hence, a well-processed and structured dataset enables the method to obtain accurate algorithms (i.e., parametric cost model). Data quality assessment in machine learning involves evaluating the suitability of a dataset for specific tasks, considering several factors (Mazurek and Wielgosz, 2023). Data can be compromised by errors or irregularities introduced during collection, aggregation, or annotation, necessitating thorough profiling and assessment (Gupta et al., 2021). Several monitoring activities, defined through metrics, can be adopted to assess the quality of a dataset (Budach et al., 2022).

- *Completeness*: many missing values in datasets can affect the prediction accuracy of the ML algorithms. This metric aims to identify and eliminate empty values within the dataset.
- *Features Accuracy*: machine learning models identify correlations within datasets; hence, ensuring error-free values is crucial. Data can have incorrect values due to various factors, such as user input errors. Feature accuracy measures how closely the feature values in a dataset match their ground truth values.
- *Target Accuracy*: refers to the precision and correctness of the labels or values of the target variable (the variable the model is trying to predict). It is the deviation of its target feature values from their ground truth values.

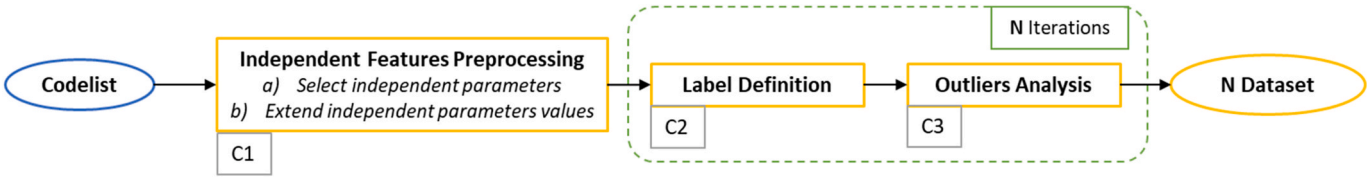

Fig. 3. Data Preparation workflow.

- *Uniqueness*: refers to the degree to which each record in a dataset is distinct and not duplicated. It ensures that every entry in the dataset represents a unique entity or observation.
- *Class Balance*: homogeneous distribution of feature and target values. A check ensures no unbalanced numbers of records between the feature value groups. Similarly, it is verified that the target variable (label) has a homogeneous distribution for training the model.
- *Consistency*: data consistency refers to the uniformity and coherence of data across a dataset. It ensures the data follows a consistent format, structure, and value range, maintaining logical integrity throughout the dataset.

The proposed metrics are general and can be used in different scenarios in which ML algorithms are developed (i.e. classification, regression and clustering). Nevertheless, depending on the scenario, some metrics are more important than others. In detail, this paper is defined as a regression study. In this scenario, completeness, feature accuracy and target accuracy have an essential influence on the dataset's quality. Uniqueness and balance have moderate importance, while consistency has low significance, and it is neglected in this study (Budach et al., 2022).

The data collected within the Codelist groups the available components and all their configuration parameters. These components can be limited in number, and not all configuration parameters are necessary to create the cost model. Therefore, starting with the Codelist, a first preprocessing feature action is needed. This phase enhances the suitability of these features (independent parameters) for modelling.

The first activity defines all the geometrical and non-geometrical parameters that affect the cost (*activity C1-a*). These parameters are called independent parameters (or cost drivers). Cost drivers are extracted from parameters of the technical information collected in *activity B1*. The number of geometrical cost drivers could be less or equal to the configuration parameters (e.g., some design variables are unrelated to the manufacturing cost). Non-geometrical cost drivers (e.g., material and its unitary cost, production batch quantity) are independent of the geometry but still affect the cost. In general, several non-geometric parameters influence cost. However, some parameters (e.g., tolerance, roughness, supplier, heat treatments, and post-processing operations) could not vary within the same family of parts. Therefore, they are constants and may not be considered in the dataset creation.

A*ctivity C1-b* represents a crucial step in the Data Preparation phase. The accuracy of a parametric cost model can be influenced by several factors, including the characteristics of the dataset used for training the ML algorithm. The quality of a dataset can be assessed in terms of the following.

a. The number of records: a dataset with a significant number of records allows the algorithm to have better training and, hence, better accuracy of the results.
b. Differentiation between the independent parameter values of each record. Having different values guarantees the risk reduction of the overfitting.

The dataset obtained from *activity B1* may not meet this requirement. For example, parts may not be evenly distributed from the minimum to the maximum mass. Besides, the minimum or maximum mass of identified components may not be adequate for a robust cost model. Thus, this activity (A*ctivity C1-b,* data augmentation) aims at extending the number of records with different values of independent geometric and non-geometric parameters, obtaining a new extended dataset. Regarding the size, parts can be scaled (up and down) to extend the dimensional variability. At the same time, other non-geometric parameters (e.g., material, production facility, unitary material cost, production batch) can also be varied to simulate different scenarios. The records extension is carried out considering the different metrics guaranteeing the dataset's quality. Scaling geometric parameters and variation of non-geometric parameters is carried out to avoid creating records with identical values. This procedure guarantees the uniqueness of the dataset (Uniqueness). In addition, the scaling allows the distribution of geometric parameter values, ensuring the dataset's homogeneity (i.e., features Class Balance). The dataset's creation is then evaluated and controlled by a second annotator. This check makes it possible to detect errors and guarantee the accuracy of independent parameter values (Features Accuracy).

Activity C1 involves several techniques to define and prepare the independent parameters of the dataset. Since this methodology is based on supervised machine learning, preparing the dataset for the modelling phase also requires the definition of the dependent parameter/s. *Activity A1-b* clarifies how many dependent parameters should be predicted (cost breakdown) so the number (N) of parametric cost models to generate. In other words, for each dependent parameter (output) to be predicted, a dedicated dataset should be established.

Activity C2 calculates the cost breakdown for all the parts of the extended datasets obtained in *C1*. Manufacturing cost is assessed through a validated analytical cost estimation software tool for mechanical components. Commercially available software tools can automatically estimate the manufacturing cost (with related breakdown) of parts archives from 3D CAD models designed according to the model-based definition (MBD) paradigm. Indeed, such models embed product manufacturing information (PMI), which is required to assess the manufacturing process and cost correctly.

The extension of the non-geometric parameters (activity C1) also affects the homogeneity of the target variable value distribution (target Class Balance). For example, parameters such as batch quantity must be chosen to ensure that the corresponding cost values are evenly distributed within a given cost range. In detail, the cost has a hyperbolic trend with batch quantity. The cost difference is small with high production batch values, resulting in similar costs despite different batch values. In contrast, high-cost differences are obtained between cost estimates based on low production batch values. Therefore, a higher density of low-production batch values than high-production batch values is necessary.

Before finishing the Data Preparation phase, it is necessary to analyse the cost values obtained from the software analysis. The costing tool can make errors (i.e., outliers) during the automatic estimation, which must be detected and eliminated (*activity C3*). For example, missing information from 3D CAD models, incorrect data entry by humans or issues during feature recognition can generate non-coherent datasets. Analysing the values of the dependent parameters obtained makes it possible to assess the accuracy of the target variable (Target Accuracy). In this case, comparing the values obtained through automatic costing with the ground truth values (i.e. manual costing by an experienced cost

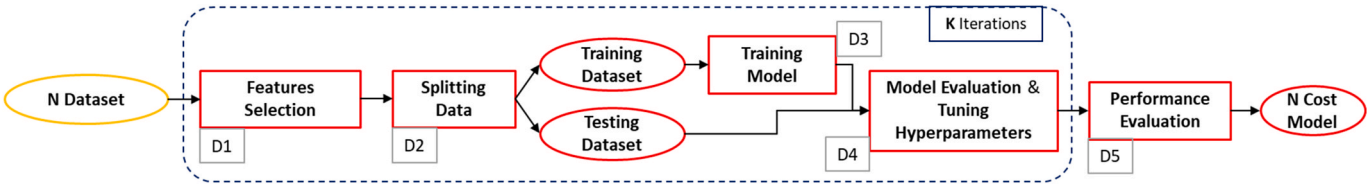

Fig. 4. Modelling workflow.

engineer) is not feasible. It is, however, possible to perform automatic evaluation analyses (i.e. outlier analysis) to identify and eliminate incorrect target variable values. Outlier costs must be removed for every dataset obtained in *C2*. Outliers can be identified through various mathematical methods and operators. For the present study, the Numeric Outlier Quartile method is applied. The identification of outliers is carried out via the interquartile range. It can be defined using the expression:

$$x_i > Q_3 + k(IQR) \quad V \quad x_i < Q_1 - k(IQR) \tag{1}$$

with $IQR = Q_3 - Q_1$ and $k \geq 0$.

The quartiles divide the analysed data into four equal parts. The first quartile (Q_1) is the value at the 25th percentile (meaning 25% of the data points are below it). The second quartile (Q_2) is at the 50th percentile, and the third quartile (Q_3) is at the 75th percentile. The interquartile range (IQR) is the range between the first and third quartiles ($Q_3 - Q_1$). So, it allows understanding how the data points are spread in the middle of the dataset. The multiplier value k is to specify the stringency of the boundaries to define which values are outliers. Based on different tests, k -values of 1 or 1.5 are considered suitable for this study. Analysing equation (1), the outlier value x_i is the one that lies outside the interquartile range.

Outliers must be analysed through indicators that link the cost to the most critical independent parameters (e.g., mass). General analyses consider material cost vs mass (i.e., €/kg) and machining cost vs machined mass (i.e., €/kg). Furthermore, specific indicators can be considered (e.g., cost vs blade height). The indicators mentioned above have a slight standard deviation for parts that belong to the same family and thus have a similar shape. This procedure does not allow for empty values within the dataset, guaranteeing its completeness (Completeness).

Once *C2* and *C3* are completed, the different costs are added to the extended datasets, resulting in N datasets: one for each estimated cost.

2.4. Modelling

Modelling represents the phase in which various models are built and evaluated using multiple techniques (Fig. 4). The main goal is to find the best ML algorithms (one for each parametric cost model) to predict the manufacturing cost. To do that, K different ML algorithms (e.g., linear regression (Su et al., 2012), random forest (Liu et al., 2012), neural network (Bishop, 1994), deep learning (LeCun et al., 2015) and gradient boosting (Natekin and Knoll, 2013)) must be compared.

Examining Fig. 4 from a broader perspective, the Modelling phase starts by processing one dataset at a time. For each dataset, all K-selected algorithms are trained and tested individually. The algorithm with the best performance will represent the cost model for that specific dataset. This activity ends with N cost models for each dataset based on their best-performing algorithms.

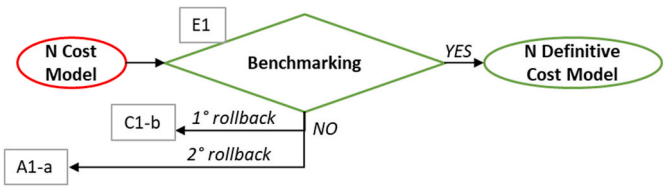

Fig. 5. Evaluation workflow.

The Modelling phase defines the evaluated algorithm and proceeds with the feature selection (*activity D1*). The main goals of feature selection are to avoid overfitting, enhance model performance (better cluster detection for clustering and prediction performance for supervised classification), provide faster and more cost-effective models, and gain a deeper understanding of the underlying processes that produced the data (Saeys et al., 2007). This activity evaluates and defines cost drivers (independent parameters) with a high-cost sensitivity. It allows obtaining a cost model with only the parameters that drive the cost for the chosen algorithm.

After the feature selection activity, the dataset is ready to be processed and used to create the model. Therefore, the dataset is divided into two groups (*activity D2*): the training and testing sets. The first group is used to train the model. The second group is for testing the developed model. In this step, choosing the proportion in which the initial dataset is divided is essential. 80% and 20% are commonly used for training and testing sets.

During the training model phase (*activity D3*), the algorithm iteratively adjusts its parameters based on the input data and corresponding outputs to minimise prediction errors, thereby enhancing its predictive accuracy. The training phase concludes with the model achieving a state of convergence, where further training does not significantly alter the parameters, indicating readiness to evaluate test data.

In the model evaluation phase (*activity D4*), the trained algorithm is assessed using a separate test (testing dataset) to gauge its predictive performance and generalisation capability to unseen data. To do this, several indicators are used to evaluate the performance of the algorithms. These indicators allow a cost engineer to understand how the model performs on new data and, on the other hand, to improve the predictive capabilities by tuning hyperparameters. The random forest algorithm, for example, is based on two primary hyperparameters: the number of trees and maximum depth. Optimising the values of these parameters improves the algorithm's predictive performance. The *activities D2*, *D3*, and *D4* are carried out using a data science software tool.

According to the requirements defined in *activity A1*, it is crucial to establish the cost models' overall performance and trends. So, two performance indicators were chosen: Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE). MPE represents the average percentage errors by which model forecasts differ from actual values. This indicator makes it possible to assess whether the average trend of the model underestimates or overestimates the cost. In cost prediction, underestimating is more dangerous than overestimating. The limitation is that the negative and positive values cancel each other out when averaged. Thus, MAPE is introduced to indicate the model's accuracy performance. MAPE takes the absolute deviation between the actual and forecast values. This indicator represents the accuracy in percentage terms without considering the direction of errors. These are the main errors that can be used to evaluate a cost model. Then, the most appropriate one depends on the context and targets defined in the business understanding phase.

Activity D4 calculates the performance metrics for K*N cost models. The performance evaluation (*activity D5*) directly compares these metrics to identify the most accurate model for each dataset. The model

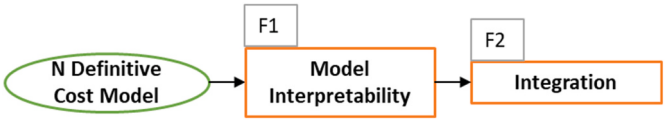

Fig. 6. Deployment workflow.

demonstrating the lowest error is then chosen for the evaluation phase.

2.5. Evaluation

The cost models from the Modelling phase represent the best prediction result for specific dependent parameters (Fig. 5). The benchmarking (*activity E1*) involves comparing the newly developed cost model's performance against established baselines (*activity A1-a*) or industry standards to gauge its relative effectiveness and efficiency. So, this activity evaluates if the cost model performance meets the requirements of the Business Understanding phase.

If the model conforms to the requirements, it can be used to move on to the next deployment phase. If the comparison shows non-conformity with the requirements, the model cannot be used, so re-assessing the conditions is needed. In this case, two steps can be taken. With the first rollback, it is possible to increase the number of records in the dataset (*active C1-b*) further to improve the model's prediction performance. If expanding the dataset is inadequate to meet the cost model's conformity, reviewing the requirements established in the business understanding phase (*activity A1-a*) is necessary.

2.6. Deployment

The deployment phase refers to integrating a trained ML model into an existing production environment to make it operational and accessible for end-users or systems (Fig. 6). This study allows users to manage the parametric cost model and evaluate how each independent parameter affects the cost.

Model interpretability (*activity F1*) intends to provide a cost model that is not a black box but perfectly interpretable. The feature importance study allows design engineers to know the weight of every cost driver (i.e., design variable) and how it affects the cost. Different algorithms can evaluate the feature's importance (Molnar, 2022). Feature Permutation Importance (FPI) makes it possible to independently identify the product/process variables that influence cost on the ML algorithm.

Finally (*activity F2*), models are incorporated into the existing IT infrastructure, which could involve embedding the model into a web service, a cloud-based application, an enterprise system, or an IoT (Internet of Things) device, depending on the use case. This solution will allow design team members to use the models during their design decisions.

3. Case study

The objective of the case study is to apply the proposed cost modelling method for discs and spacers of an axial compressor, an engineered-to-order product. The design and cost engineering teams of Baker Hughes Company were involved in this case study and validation. The company needs a method for estimating costs during the conceptual design of its products. It should ensure the main benefits listed in the introduction (e.g., accuracy higher than the empirical models already developed internally based on linear regressions, time-saving, more comprehensive analysis).

During the conceptual design phase, design engineers are responsible for defining the cross-section of the turbomachine. Engineers perform this task using a configuration software tool, which sketches the 2D cross-section view of all the turbomachine components by considering configuration parameters (29 for discs and spacers). These parameters are the same as those used to create the training datasets.

The company followed the entire cost modelling procedure. It provided the required documentation (i.e., 3D CAD models and drawings) and technical know-how (i.e., bill of materials, configurations parameters, manufacturing information). The authors supervised the project by working alongside designers and cost engineers. Initially, a set of reference turbomachines with related discs and spacers was identified to create the training dataset. The cost information was obtained by using the LeanCOST cost estimation software. This software utilises an analytical approach to determine the costs of the parts. Costs are calculated starting from the 3D CAD models of the components. The company cost engineers previously validated this tool. LeanCOST can automatically estimate the costs of 3D CAD models without user interaction (batch mode). It can also update previous cost estimations to simulate multiple scenarios (e.g., inflation).

3.1. Business understanding

In *activity A1*, the cost estimation error was set within the range −15%/+20% as indicated by E2516 − 11, Standard Classification for

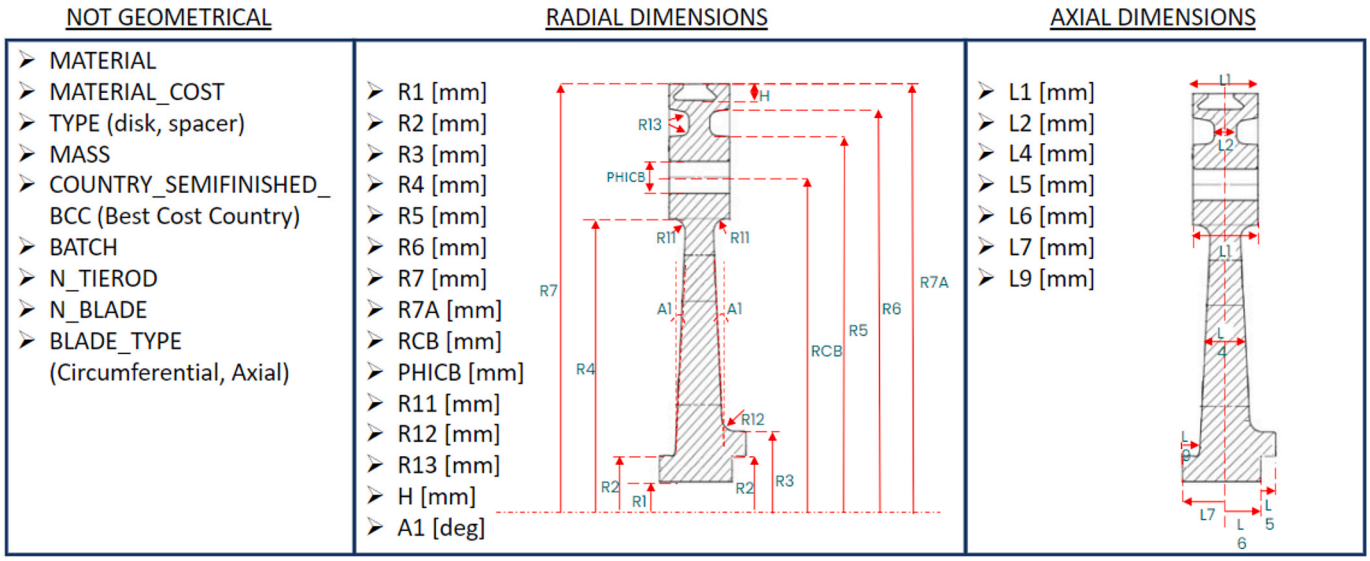

Fig. 7. A half-section view of the disc/spacer with related configuration parameters.

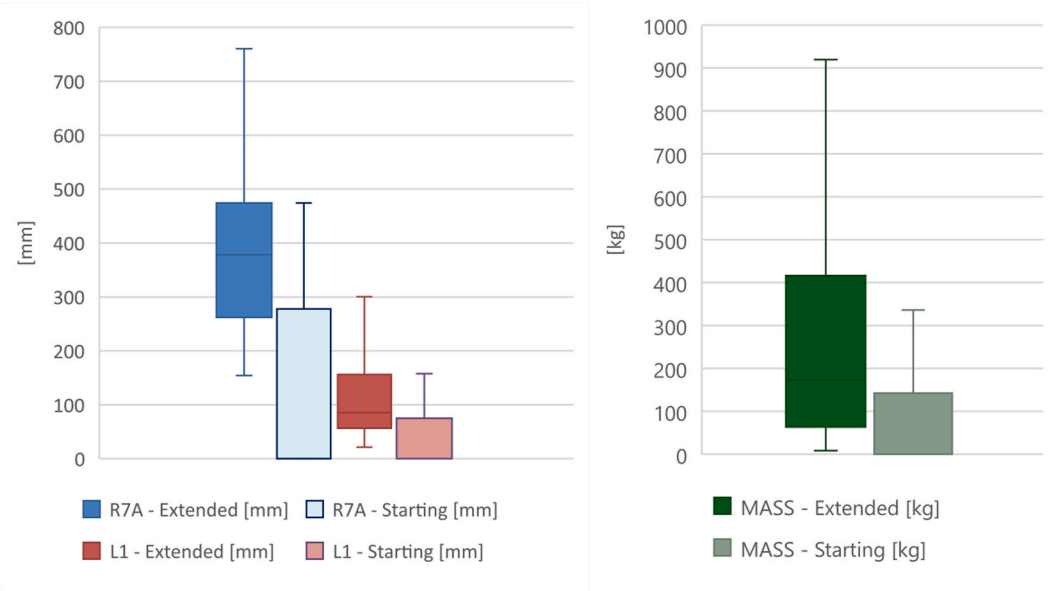

Fig. 8. Excerpt of parameters distribution before and after data augmentation.

Cost Estimate Classification System, Class 4 (i.e., concept study or feasibility). The output desired from the cost modelling method consists of two cost models, one for the semi-finishing and the other for the finishing phases of discs or spacers. Cost estimates obtained with this breakdown allow the company to compare predicted with actual costs (those incurred when manufacturing or purchasing the parts).

3.2. Data Understanding

The cost modelling method is used for turbomachines' discs or spacers family. Five turbomachines were analysed, and 50 discs and spacers were identified (*activity B1*). The 3D CAD models of the finished and semi-finished parts were downloaded from the company's PDM system. The turbomachines were chosen considering shape, size (power), and material variability. Geometric information was extracted from the 3D CAD model. Still, non-geometric information such as batch, production country, material, and material cost has also been collected. Thirty-four parameters were collected, including the names of the machines and the corresponding semi-finished and finished part codes. These last three parameters are identification (ID) parameters and are insignificant for cost prediction. All this collected information was arranged in a table to create the Codelist.

3.3. Data preparation

Through the analysis of drawings and 3D models (*activity C1-a*), the main parameters of standard designs among the various discs/spacers were identified. A general simplified shape type of disc/spacer was defined (Fig. 7). The support of the company's cost managers made it possible to ascertain that the features with the most significant impact on cost were included within the simplified representation. Non-geometric parameters of interest in the conceptual design phase were also defined. A total of 31 configuration parameters were selected after excluding the ID parameters. Hereunder, there are the geometrical and non-geometrical parameters. The complete list is available in Fig. 7.

- NON-GEOMETRIC PARAMETERS:

  - BATCH: this parameter represents the overall number of components produced in a single batch. The manufacturing cost directly relates to this parameter since it affects fixed costs (e.g., set-up cost). Increasing the batch quantity provides a hyperbolic trend in unitary manufacturing costs.
  - COUNTRY: the manufacturing cost depends on the country where the components are produced. Different countries allow for different hourly rates. It is to be noted that analytical cost models estimate the manufacturing cost of components. Overhead costs (e.g., transportation, taxes, duties) are not considered.
  - MATERIAL: the material affects the manufacturing cost. On the one hand, it defines the baseline unitary cost of the material used for producing the components. On the other hand, all the mechanical operation technological parameters depend on the material being processed.
  - MATERIAL_COST: with this parameter, it is possible to define different unitary costs for the same material. This cost range on material costs allows accounting for inflation phenomena related to the raw material market. Consequently, the proposed method enables the actualisation of the cost according to inflation. The considerations were made based on the experience gained by the purchasing department.
  - MASS: defines the mass of the component. Together with the material cost, it contributes to determining the total manufacturing cost.
  - N_TIEROD: represents the total number of holes drilled in the disc/spacer. The holes are those defined by the geometric parameter PHICB.
  - N_BLADE: number of slots where the blades are fixed onto the disc/spacer. Millings are identified in correspondence with the parameter H
  - BLADE_TYPE: different blades can be clamped to the disc/spacer. This difference influences the machining on the disc/spacer to enable subsequent assembly.

- GEOMETRIC PARAMETERS: The most relevant geometric parameters affecting the manufacturing cost can be identified as radial and axial parameters (e.g., R7, L1). As a family of parts, the geometric parameters can be related. In this case, only one is chosen. Regarding tolerances, they are constant for all the discs and spacers. The general surface roughness is 25 [μm] and 3.2 [μm] for semi-finished and finished parts. Tolerances and roughness are included as PMI in the 3D models.

Starting from the 50 discs and spacers identified during the Data

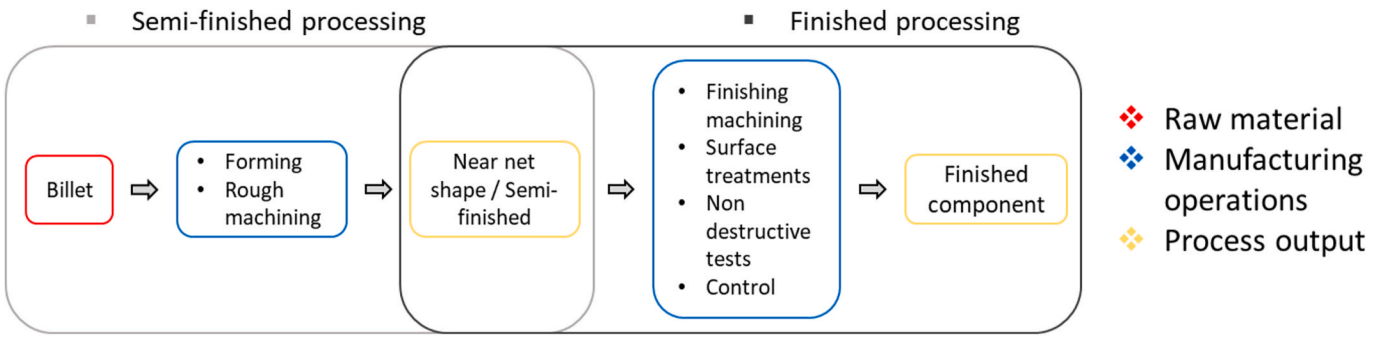

Fig. 9. The manufacturing process for discs and spacers.

Understanding phase, different configuration parameters were chosen to extend the Codelist (*activity C1-b*), enabling different scenarios and variations that had a significant impact on costs.

- GEOMETRIC PARAMETERS: first, some discs and spacers were selected among those available. The selection was made considering the different types and morphologies. To choose the scaling factors, first, a dimensional parameter was identified as representative of the disc's and spacer's size, which could be an index to the geometric variability. The chosen parameter is R7A, which represents the maximum radius of the disc/spacer. The scaling factors were selected to consider discs and spacers engineers can design in the future. Also, dimensional jumps were deemed to be acceptable and not excessive. In this way, the training dataset considers different geometries, which are large and heterogeneous. Part scaling contributed by adding 58 parts to the initial baseline, thus obtaining 108 3D CAD models.
- BATCH (1, 2, 3, 4, 5, 6, 8, 12, 25, 50–10 values): the following production batches were chosen: 1, 2, 3, 4, 5, 6, 8, 12, 25, and 50. The percentage cost deviation between two consecutive production batches is 1%. Starting from 108 records, this second step allows the Codelist to be extended up to 1080.
- COUNTRY (BCC, WCC – 2 values): manufacturing costs will be simulated in two countries, the best cost country (BCC, e.g., Far East) and the worst cost country (WCC, e.g., Europe) only for semi-finished parts. Starting from 1080 records, this second step allows the Codelist to be extended up to 2160.
- MATERIAL_COST (xx, yy, zz, – 3 values): for this case study, only one material was considered. So, starting from the baseline cost related to the selected material, a range from −50% to +100% was used. Beginning from 2160 records, this third step extends the Codelist to 6480.

Finally, a dataset with 31 configuration parameters and 6480 records was obtained. The Completeness and Uniqueness of the dataset are assessed at the end of the scaling procedure by verifying missing data and the absence of duplicates. Choosing proper scaling factors leads to high-quality datasets regarding Class Balance (for features and targets). Feature accuracy was checked as discussed in the method.

Fig. 8 shows the distribution of values for three identification parameters (R7A, L1 and MASS), comparing the starting dataset and the extended one. R7A is the maximum radial external dimension, L1 is the maximum axial external dimension, and MASS is the finished mass of the component. The comparison makes it possible to see that the distribution considers a broader range.

The cost models to be developed concern semi-finished and finished products (*activity C2)*. Two datasets are obtained: the one for the semi-finished product comprises 6480 rows. In contrast, the one for the finished product is made up of 1080 rows because it does not involve the multiplication of the parameters related to the material cost and the country of the production site. The reason for dividing the total cost into semi-finished and finished lies in the supply strategy. Farm-out (the suppliers make the semi-finished product while the finished product is produced internally) is the strategy for this case study.

Table 1
Statistics of the main configuration parameters for discs and spacers.

Parameter	Max	Min	Mean	Standard deviation
R1 [mm]	53	0	16	14
R2 [mm]	324	19	65	59
R3 [mm]	346	28	86	62
R4 [mm]	611	88	256	133
R5 [mm]	811	0	300	182
R6 [mm]	870	0	350	207
R7 [mm]	898	151	398	193
R7A [mm]	900	154	402	196
RCB [mm]	703	100	293	153
PHICB [mm]	84	13	34	17
R11 [mm]	93	4	23	18
R12 [mm]	83	5	22	15
R13 [mm]	53	0	14	12
H_S [mm]	47	0	6	9
A1 [DEG]	4.5	0	2	1
L1 [mm]	1024	21	154	179
L2 [mm]	967	0	92	178
L4 [mm]	666	7	91	127
L5 [mm]	48	4	15	8
L6 [mm]	339	8	52	54
L7 [mm]	512	17	81	86
L9 [mm]	33	4	14	7
N_TR	26	16	22	5
N_BL	68	0	29	27
MASS [kg]	14463	9	859	2165

CAD models are analysed using LeanCOST (by Hyperlean srl, Italy), a company's cost-estimating tool. In this way, it was possible to generate a dataset of manufacturing costs by overcoming issues related to sparseness or unavailability of historical datasets.

Fig. 9 shows the entire manufacturing process estimated by the software to produce the finished disc or spacer in detail.

The outlier analysis of semi-finished and finished parts was performed for the Codelist (*activity C3*). The cost estimation carried out by LeanCOST is unsupervised. Thus, the manufacturing process (e.g., missing operations, improper stock selection) and related costs may be subject to errors. The goal is to identify parts with an inappropriate estimated cost. The ratio between raw material cost and stock mass was used for semi-finished parts to identify outliers. In contrast, the ratio between finishing cost and machined volume (stock mass minus finished mass) was used for finished parts. The Numeric Outlier Quartile is the method employed to identify outliers. For semi-finished parts, 16 geometries out of 108 were excluded, which led to the exclusion of 960 rows. A dataset of 5520 rows was obtained. For the finished, 12 geometries out of 108 (120 rows) were excluded, resulting in a dataset of 960 parts. Through this procedure, it was possible to achieve a good level of Target Accuracy.

Finally, two datasets were obtained (Table 1). They have 31

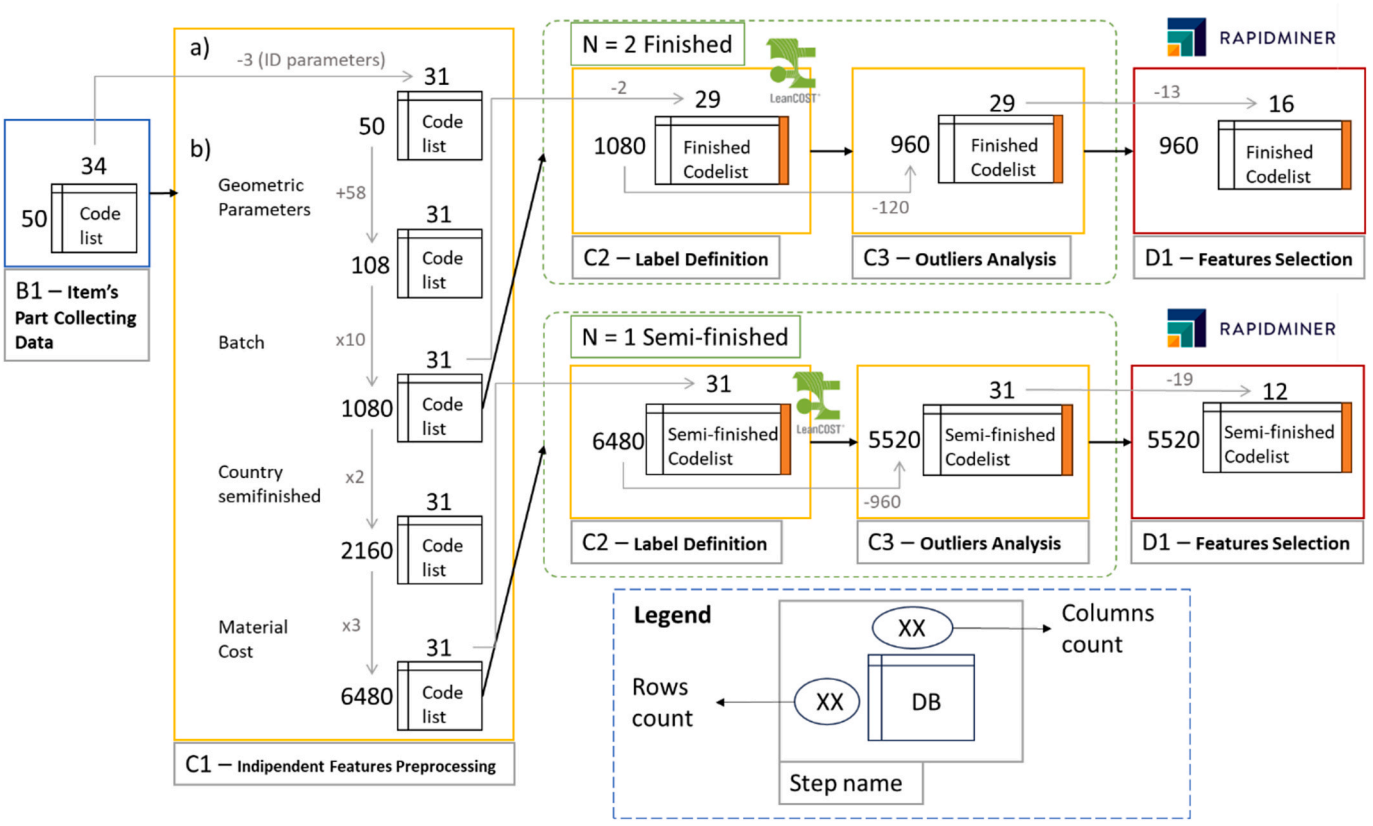

Fig. 10. ML modelling for the case study.

configuration parameters and 5520 records (semi-finished), 29 configuration parameters and 960 records (finished). The parameters "MATERIAL_COST" and "COUNTRY" were removed from the finished dataset. "MATERIAL_COST" was excluded because it does not influence the cost of the finishing operations (i.e., machining and quality controls). "COUNTRY" was removed because of the company's production strategy (in-house finishing). Indeed, the finishing operations are only performed in one country (their internal plant).

3.4. Modelling

The case study aims at developing two cost models, so the following steps must be performed twice. For convenience, the results for finished and semi-finished products are presented together.

ML algorithms tested for the case study were selected among those managed by RapidMiner Studio (By Altair Engineering, USA), a data science software.

- *Generalised Linear Models (GLMs)*: it offers interpretability and simplicity. The model's coefficients provide clear insights into the impact of each feature on the outcome, facilitating a better understanding of the relationships. Additionally, GLMs are suitable for statistical inference, providing p-values and confidence intervals. However, their limitation lies in their assumption of linear relationships, making them less effective in capturing complex, non-linear patterns in the data.
- *Deep Learning (DL)*: DL models, characterised by their neural networks, excel in capturing intricate patterns and non-linear relationships within data. They are highly expressive and capable of feature learning, eliminating the need for extensive manual feature engineering. Despite their power, they come with drawbacks, such as computational intensity requiring substantial resources for training. Moreover, the incomprehensible "black box" nature of DL models can pose challenges for interpretability, potentially impacting trust in the predictions.
- *Decision Trees (DTs)*: DTs are known for their interpretability, clearly visualising the decision-making process. They do not assume specific data distributions and can naturally handle non-linear relationships. However, DTs are prone to overfitting, capturing noise in the training data and exhibiting instability with minor changes. While adequate for simple tasks, individual DTs might lack the expressiveness to capture more complex patterns in the data.
- *Random Forest (RF)*: these models mitigate the overfitting issues of individual trees and provide more robust predictions. They excel in handling non-linear relationships and offer insights into the importance of features. However, their computational complexity increases with the number of trees, making them resource-intensive. Despite being less of a "black box" than DL models, Random Forests can still present challenges in interpretation compared to simpler models.
- *Gradient Boosted Trees (GBTs)*: these models build sequentially, correcting errors of previous trees and resulting in high predictive accuracy. They handle missing data effectively and exhibit robustness to outliers. However, the computational demands, especially with numerous trees, can be a drawback. There is also a risk of overfitting noisy data if not properly tuned. While more interpretable than some complex models, GBTs still pose challenges compared to simpler models like linear regression.

A multi-objective evolutionary algorithm (provided by RapidMiner) was used to find the best feature set (cost drivers) for each model (*activity D1*). It is noted that cost drivers decreased from the initial list of configuration parameters, namely 31 for semi-finished and 29 for finished. The number of selected cost drivers was 12 for semi-finished and 16 for finished (Fig. 12). Thus, each algorithm was tested using the best set.

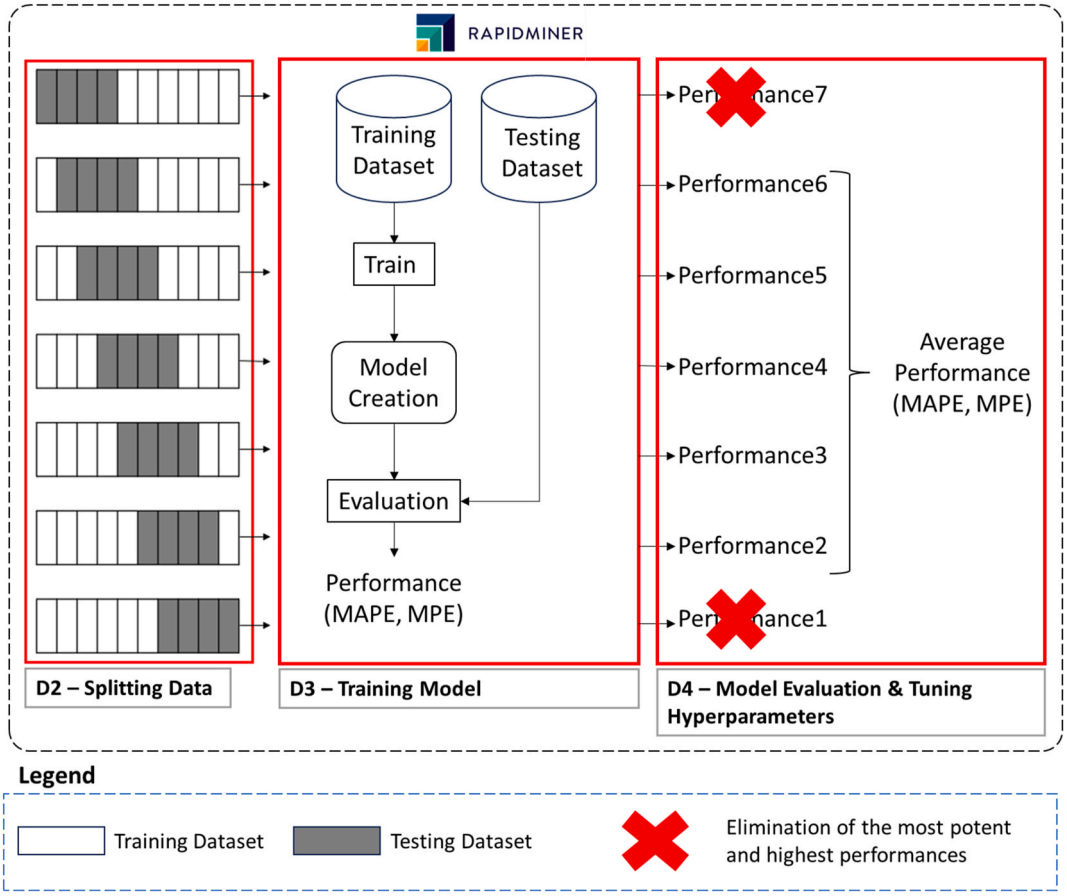

Fig. 11. ML modelling procedure in RapidMiner.

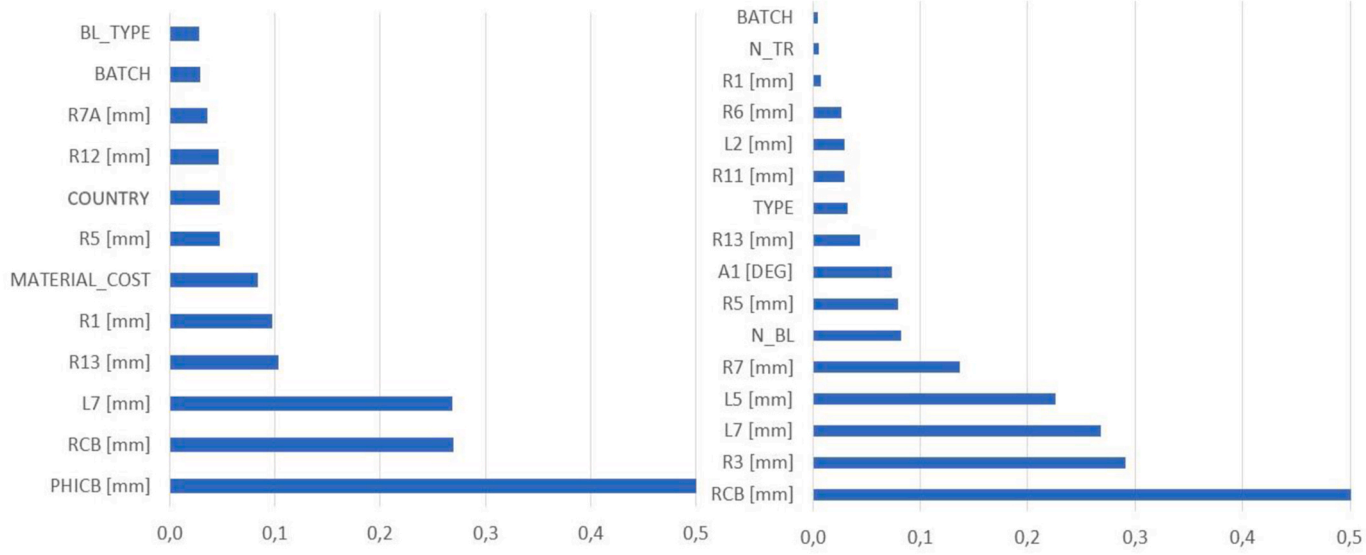

Fig. 12. Model Interpretation: feature importance for the semi-finished (left) and finished (right) cost models.

Fig. 10 shows the dataset evolution from the beginning (*activity B1*) to the final Codelists (*activity D1*).

In *activities D2 and D3*, cost models were generated using the Auto Model plug-in of RapidMiner Studio, which streamlines the creation and validation of prediction models. A 40% hold-out set has been used to calculate performance. The software starts with this hold-out set as input. Next, using a multi-hold-out-set validation, the software determines the performances for seven disjoint subsets. The most potent and highest performances are eliminated. The average of the remaining five performances is calculated (RapidMiner Documentation, 2023) (Fig. 11).

Table 2 contains MAPE and MPE indicators for the ten cost models (five for semi-finished and five for finished) developed by RapidMiner (*activity D4*) (Fig. 11). The best-performing algorithm for both models

Table 2
Cost model indicators of the semi-finished and finished discs and spacers.

Model	Semi-finished discs and spacers			Finished discs and spacers		
	MAPE	MPE Positive	MPE Negative	MAPE	MPE Positive	MPE Negative
Generalised Linear Models (GLMs)	113%	99%	125%	21%	25%	18%
Deep Learning (DL)	34%	24%	40%	16%	20%	10%
Decision Trees (DTs)	11%	13%	10%	6%	7%	5%
Random Forest (RM)	30%	40%	17%	10%	10%	8%
Gradient Boosted Trees (GBTs)	1%	1%	1%	3%	3%	2%

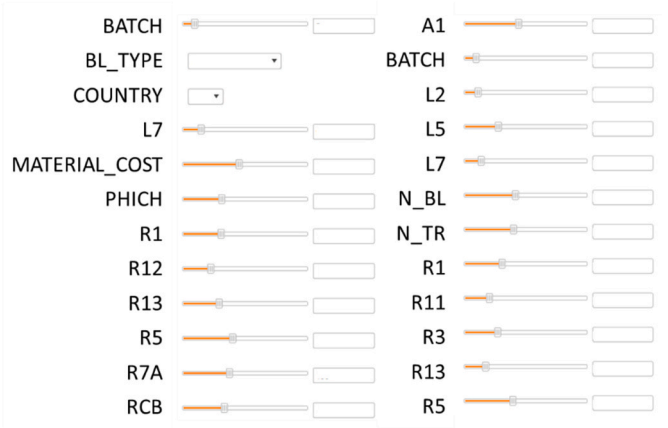

Fig. 13. Extract of the simulator environment of RapidMiner for semi-finished (left) and finished (right) cost models. Values are missing for confidentiality reasons.

was GBTs, which was thus selected (*activity D5*). RapidMiner automatically tuned the hyperparameters to achieve the following values (Optimise Parameters Quadratic Operator): number of trees: 150, maximal depth: 7 and learning rate: 0.1.

3.5. Evaluation

The MAPE of the GBTs algorithm is 1% for the model of semi-finished and 3% for finished discs. The models are valid since the MAPE and MPE indicators are lower than the thresholds set in business understanding (−15%/+20%). They can be deployed for cost estimation (*activity E1*).

3.6. Deployment

Feature importance analysis allows engineers to highlight the impact of cost parameters on the model (*activity F1*). In the case of the disc model of the semi-finished products, the algorithm identified the PHICB parameter (Fig. 12) as the most relevant (i.e., the hole through the tierods). For finished parts (Fig. 12), RCB (i.e., the distance of tierods from the disc axle) is the most important.

At last (*activity F2*), the cost models were deployed and stored in a dataset to be used by cost and design engineers through the simulator module of RapidMiner (Fig. 13).

4. Validation and results

Verifying the accuracy of the estimate is the process of cost validation. By confirming the data and estimating techniques, this procedure guarantees the precision and dependability of the outcomes. It is a systematic procedure for approving the data and techniques applied throughout the cost-estimating process.

In this study, a formal validation of the cost modelling methodology was conducted by considering actual cost data. The cost models obtained through the cost modelling methodology (§3) were used to evaluate the rotor discs and spacers of a six-stage gas turbine compressor that had never been designed before. The architecture of the new machine is the same as that of one already developed in the past, but with different dimensions. The parts analysed, therefore, have similar shapes and production processes but different dimensions. The dimensional parameters of the analysed parts were defined by a product configurator capable of determining the cross-section of the machine (§3). The system engineer responsible for the machine development defines the other parameters (e.g., production country, batch).

Three different cost-estimating methods have been implemented to evaluate the benefits of the proposed methodology. A cost engineer defined the actual cost through LeanCOST after 3D modelling the parts and identifying the PMIs. It is noted that the 3D modelling of the parts and the analytical cost evaluation were carried out only for the scope of this validation. Such activities are not performed during the conceptual design phase but during embodiment and detail. Thus, the economic values represent a benchmark since the parts to be analysed are made with production technologies characterised by validated analytical cost models and process routings. This information is formalised within the estimating software tool.

The estimated costs were defined using the cost models developed according to the proposed methodology (TO-BE approach in Fig. 14). The prediction models implemented in RapidMiner provide economic values based on the dimensional parameters. Finally, the company involved a cost engineer in defining the third economic information (AS-IS approach in Fig. 14). The value was determined based on the (little) historical data, integrated with evaluations from the technician's implicit experience. The AS-IS and TO-BE cost estimates were compared with the benchmark to calculate the MAPE (Table 3).

To evaluate the time-saving of the proposed methodology, the authors measured the time for cost estimation and the initial investment for cost modelling. The effort (time) for cost estimation was 4 h for AS-IS (cost engineer analysis) and 1 min for TO-BE (proposed approach). A junior cost engineer (less than three years of experience in cost estimation and modelling) was engaged to develop the ML-based model. The authors simulated two scenarios. The first implies the development of the new cost models (i.e., for semi-finishing and finishing processes) from scratch. The second refers to updating the cost models (i.e. adding more lines to the codelist) once a new gas turbine is designed and developed. The authors registered the effort for each activity (Table 4), splitting between time spent by the cost engineer and the computation time requested by the software (e.g., cost estimation, data augmentation, data analysis).

Data understanding is the most labour-intensive activity. To build the training dataset, a cost engineer must first identify and download the models and drawings from a PDM system. Then, he has to extract the complete geometrical information from 3D models and drawings to create a part of the training dataset. No CAD plug-ins (e.g., feature recognition tool) were used for the case study to automate the process. Data preparation, on the other side, is the most computationally intensive. Most of the effort is requested to run the cost estimation tool (batch mode) for the entire set of components. This effort is only marginally significant because it runs unattended, often overnight. Updating a cost model is much faster than creating a new one because it is assumed that the cost engineer should manage fewer parts.

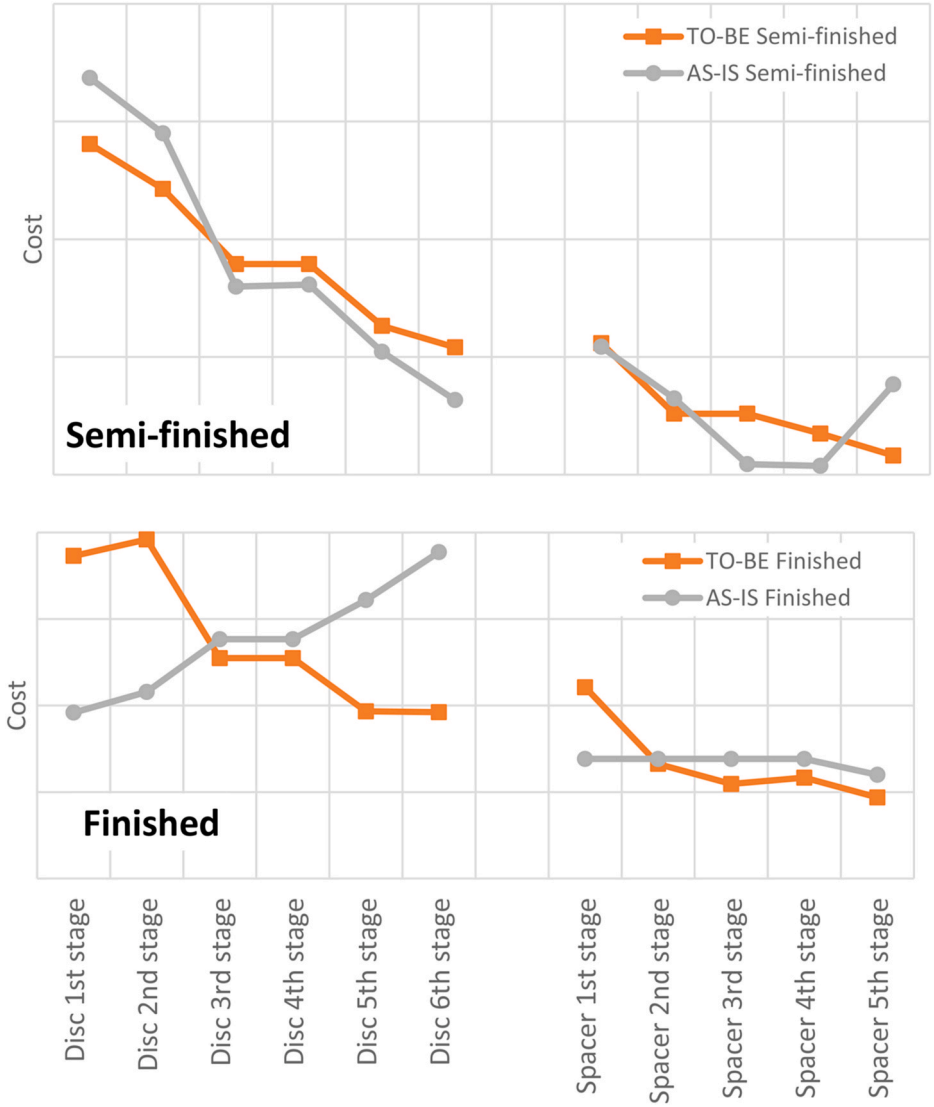

Fig. 14. Cost estimates through the AS-IS and TO-BE approaches. Values are missing for confidentiality reasons.

5. Results discussion

The results obtained from the case study presented in the previous section draft the strengths and weaknesses of the proposed methodology through qualitative and quantitative indicators. Weaknesses provide indicators for future work presented within the conclusions.

5.1. Strengths

5.1.1. Improved accuracy of cost estimations

The MAPE of the obtained cost models is 1% (semi-finished) and 3% (finished). The error is significantly lower than the threshold the business understanding requires (−15%/+20%: "Class I", as E2516 – 11, Standard Classification for Cost Estimate Classification System). Thus, the proposed method is promising for conceptual design and feasibility. The case study highlights MAPE values slightly higher than those obtained by the test dataset (5% vs 1% for semi-finished and 11% vs 3% for finished). These values are due to the slightly different shapes of discs and spacers considered in the case study and the limited number of records for the training dataset. Anyway, MAPE values are still lower than the requirements of business understanding.

The algorithm chosen for obtaining the model is the Gradient Boosted Trees, which was the best for realising a cost model for turbomachinery components. This result agrees with J.-L. Loyer et al. (2016), who examined the efficacy of five statistical models for estimating the manufacturing cost of jet engine components. The investigation demonstrates that modern methods, like GBTs and Support Vector Regression, are up to twice as effective as those frequently seen in the literature (Multiple Linear Regression and Artificial Neural Networks). However, it is not viable to generalise the approach to other industrial environments.

The estimated costs of employing the models obtained from the proposed cost modelling approach are much closer to the benchmark than the costs manually assessed by a cost engineer (Table 3). ML-based cost models can capture semi-finished and finished discs and spacers trends. On the other hand, the cost engineer could not precisely evaluate finishing operations for discs. Thus, the MAPE for the total manufacturing cost (semi-finishing and finishing processes) in the AS-IS approach (9%) is higher than TO-BE (7%). This result highlights the proposed approach's benefits in improved cost estimation accuracy.

5.1.2. Time-saving

Cost estimating generated using ML-based models with the proposed method would be significantly shorter (from hours to minutes) than traditional methods. Often, a cost engineer searches for economic values in the few and unstructured historical data. To make accurate cost

Table 3
MAPE of cost estimates was obtained through the AS-IS (estimation carried out manually by a cost engineer) and TO-BE (analysis based on ML-based cost models developed following the method proposed in this paper) approaches. The benchmark values were obtained by a cost engineer working with the analytical cost estimation tool.

	TO-BE			AS-IS		
	Semi-finished	Finished	Total	Semi-finished	Finished	Total
Disc 1st stage	7%	26%	17%	3%	39%	21%
Disc 2nd stage	5%	6%	5%	1%	21%	10%
Disc 3rd stage	8%	10%	9%	6%	12%	9%
Disc 4th stage	6%	10%	8%	5%	12%	8%
Disc 5th stage	2%	13%	7%	0%	29%	12%
Disc 6th stage	2%	24%	8%	6%	50%	16%
Spacer 1st stage	3%	4%	4%	4%	13%	8%
Spacer 2nd stage	1%	6%	3%	0%	5%	2%
Spacer 3rd stage	5%	9%	7%	1%	13%	6%
Spacer 4th stage	7%	8%	7%	4%	11%	7%
Spacer 5th stage	6%	5%	5%	0%	2%	1%
Mean	**5%**	**11%**	**7%**	**3%**	**19%**	**9%**

forecasts (e.g. in scenarios characterised by high inflation or changes in energy costs), the economic data (e.g. deriving from previous purchase orders) will have to be integrated with other information deriving from different sources (e.g. inflation rates, energy costs). These studies, as well as being time-consuming, undermine the estimate's accuracy and strongly depend on the experience of the cost engineer. Even if cost estimation through the proposed approach is much faster, a company must consider an initial investment for cost modelling (Table 4). Comparing the initial effort for cost modelling (56.4 h) and the time-saving obtained by employing the so-achieved cost models (43.8 h), it is possible to estimate that a company returns from the investment after analysing a second configuration or scenario. During the conceptual design phase of a gas turbine, a design team typically analyses many different configurations (i.e., shapes, dimensions, manufacturing processes) and different scenarios (e.g., country, batch, material cost, energy cost). Thus, the company returns from the investment in a couple of months (the product development process is about three years long).

It is worth underlining that a design team can reduce the time wasted due to process inefficiencies by employing the proposed cost modelling approach. For example, using regression techniques (e.g., CER) on a few historical datasets risks having inaccurate cost models. A wrong estimation could lead a company to make sub-optimal design decisions. A company could invest time and resources to engineer solutions that are subsequently discarded because they are not cost-effective. The time spent on redesign loops could significantly exceed that spent developing a parametric cost model, as indicated by the proposed methodology. Moreover, the cost wasted by such inefficiencies could be more important than the time spent on cost modelling. For ETO and complex products, design activities are carried out by large design teams.

5.1.3. Comprehensive cost estimation

Creating a reliable dataset for ML algorithm training was made possible by software for analytical cost assessment. Its use enabled circumventing the constraints of creating a parametric cost model by sparse and unnormalised historical cost data. By employing an analytical cost estimation software tool, it is possible to develop coherent (data deriving from a unique source and generated from validated analytical cost models) and comprehensive (that contains multiple scenarios, such as production countries, batch, and material cost) training datasets. The proposed approach is independent of any software application. Other data science and analytical cost estimation software can be used to implement the methodology.

The availability of a comprehensive and coherent dataset enables cost engineers to develop versatile cost models. Such estimation tools may accurately and rapidly assess the cost in different scenarios during the conceptual design phase. This feature allows engineers to quickly find the best design configuration and production scenario before investing in further design and engineering activities.

5.1.4. Explainability of cost models

The explainability of cost models is an essential feature for design and cost engineers. Both users must know the input-output relationship. The former can understand the most cost-effective design features during conceptual design. The latter can quickly assess the cost in different scenarios, being more confident about values estimated by the models.

Cost engineers recognised the cost drivers defined through the feature selection algorithm as the most cost-effective. Moreover, all of them are known during conceptual design. For the proposed case study, the problem's dimensionality was decreased using a multi-objective evolutionary algorithm as a feature selection method. For semi-finished and finished cost models, cost drivers decreased from 31 to 12 and 28 to 16, respectively. The reduction did not penalise the prediction accuracy of cost models.

5.1.5. Automatic cost estimation process

The proposed systematic method can be automated since it requires repetitive tasks. The analytical cost estimation tool (e.g., LeanCOST) and data science software (e.g., RapidMiner) can be integrated (e.g., through the relative application programming interfaces – APIs) to develop an orchestrator capable of automating the entire workflow.

5.1.6. Self-learning

Dataset maintainability (i.e., self-learning) is an essential feature to avoid the problem of making wrong and outdated estimates. The method guarantees such a requirement by leveraging the native benefits of ML. The technique can update and extend an original cost model by iterating the data preparation, modelling and evaluation phases. For example, the training dataset can be augmented by 3D scaling the original dataset of

Table 4
Cost modelling effort for generating and updating a cost model for discs and spacers of a gas turbine axial compressor.

	New development			Update		
	Computation time [h]	Cost engiener time [h]	Sub-Total time [h]	Computation time [h]	Cost engiener time [h]	Sub-Total time [h]
Business understanding	0.0	0.0	0.0	0.0	0.0	0.0
Data understanding	0.0	53.0	53.0	0.0	12.8	12.8
Data preparation	14.6	3.1	17.6	1.3	3.1	4.4
Modelling & Evaluation	0.7	0.2	0.8	0.7	0.2	0.8
Deployment	0.1	0.1	0.2	0.1	0.1	0.2
Total time [h]	**15.3**	**56.4**	**71.7**	**2.0**	**16.2**	**18.2**

CAD models or considering other prices or manufacturing processes. The effort required to update a cost model is much lower than needed for its first development (Table 4).

5.2. Weaknesses

The prediction accuracy worsens when a parametric model assesses the cost of parts with significantly different production technologies or shapes (compared with the parts of the training dataset). In this situation, the MAPE of a cost model is inconsistent with its actual accuracy. Depending on the distance between the estimated part and the training set, a cost engineer must consider a more significant value. The distance can be measured regarding how much the chosen part's production process, shape and dimension differs from the training dataset. The higher the distance, the higher the error. For example, cost engineers could consider a triangular function to correct the cost estimation. A typical range is −15%/+20% for parts with the same technology and design. Such a range can increase (e.g., −30%/+40%) for parts that belong to the same family but with an entirely new design and manufacturing process.

The number of records removed during the outlier analysis (around 12%, Fig. 10) is not negligible. It resulted from issues met by the software tool in estimating the cost for the 3D CAD models generated during the dataset augmentation (data preparation phase). Data augmentation (e.g., 3D scaling of CAD models) should be further investigated to avoid generating geometries that are too different from actual parts that the software tool cannot properly recognise and cost estimate.

The proposed methodology is a concept that deserves further development so that a company can effectively use it. So far, it consists of many manual operations to orchestrate the employed software tools and manage data.

6. Conclusions

The paper presented a cost modelling methodology for the cost estimation of products during the conceptual or preliminary design phases of engineered-to-order products. The proposed method uses CRISP-DM. The approach is grounded on ML algorithms trained using datasets generated through an automatic and analytical software tool for cost estimation. This solution eliminates those issues linked to the unavailability or inconsistency of historical data. Moreover, feature selection and importance algorithms are used to reduce system dimensionality and increase the understanding of models.

The method was used to develop two cost models (semi-finishing and finishing phases) of gas turbine parts (disc and spacers). GBTs turned out to be the best-performing prediction algorithm. The resulting cost models have 1% and 3% MAPE for the semi-finishing and finishing phases. Models originating from the proposed approach are suitable for estimating the cost of similar parts. However, the authors measured higher values (5% and 11%) through a case study.

Unlike conventional approaches, ML-based cost estimation will take much less time (from hours to minutes). Developing parametric cost models by employing the proposed approach takes some days for each parts family. Companies may shortly return from this initial investment because of its benefits (e.g., faster, improved and more comprehensive cost estimations).

Through a comprehensive and coherent training set generated as presented, cost engineers can create versatile cost models useable during the conceptual design phase for fast (from hours to minutes for each part) and precise assessments (the accuracy is much higher than required by AACE for feasibility study). Feature selection algorithms allow cost engineers to identify the most relevant drivers to improve the explainability of models.

Further research will examine the approach's applicability for larger product families by testing it on different items and industrial contexts. Clustering algorithms should be employed to determine the groups of components that need to be estimated with specific cost models. Clustering will also help engineers know which cost model to use for one part. Moreover, by computing the distance between the part to be estimated and the centre of gravity of each cluster, it will be possible to modulate the accuracy of the prediction (the higher the distance, the higher the prediction error).

Data augmentation should be further analysed to reduce the number of outliers during the data preparation phase of the methodology.

Last, the implementation of the approach should be investigated by considering solutions to speed up the process of generating, distributing, and using cost models. The authors recognised many manual and labour-intensive operations during the validation process when creating the training dataset. Retrieving metadata of parts from PDM systems or employing geometric feature recognition algorithms could speed up this phase, thus reducing the effort for cost modelling and improving the benefits.

Ethical approval

This study contains no studies with human or animal subjects performed by authors.

Funding

No funds, grants, or other support were received.

Data or code availability

The data that has been used is confidential.

CRediT authorship contribution statement

Marco Mandolini: Writing – original draft, Methodology, Conceptualization. **Luca Manuguerra:** Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Mikhailo Sartini:** Writing – original draft, Software, Methodology, Conceptualization. **Giulio Marcello Lo Presti:** Supervision, Funding acquisition. **Francesco Pescatori:** Writing – review & editing, Validation.

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

The data that has been used is confidential.

Acknowledgement

None.

References

Alstad, J.P., 2019. Development of COSYSMO 3.0: an extended, unified cost estimating model for systems engineering. Procedia Comput. Sci. 153, 55–62. https://doi.org/10.1016/j.procs.2019.05.055.

Bertoni, A., Bertoni, M., 2020. PSS cost engineering: a model-based approach for concept design. CIRP J Manuf Sci Technol 29, 176–190. https://doi.org/10.1016/j.cirpj.2018.08.001.

Bishop, C.M., 1994. Neural networks and their applications. Rev. Sci. Instrum. 65, 1803–1832. https://doi.org/10.1063/1.1144830.

Boothroyd, G., Reynolds, C., 1989. Approximate cost estimates for typical turned parts. J. Manuf. Syst. 8, 185–193. https://doi.org/10.1016/0278-6125(89)90040-X.

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., Harmouch, H., 2022. The Effects of Data Quality on Machine Learning Performance.

Campi, F., Mandolini, M., Santucci, F., Favi, C., Germani, M., 2021. Parametric cost modelling of components for turbomachines: preliminary study. Proceedings of the Design Society 1, 2379–2388. https://doi.org/10.1017/pds.2021.499.

Cavalieri, S., Maccarrone, P., Pinto, R., 2004. Parametric vs. neural network models for the estimation of production costs: a case study in the automotive industry. Int. J. Prod. Econ. 91, 165–177. https://doi.org/10.1016/j.ijpe.2003.08.005.

Chen, X., Huang, J., Yi, M., 2021. Development cost prediction of general aviation aircraft using combined estimation technique. Chinese J. Aeronautics. 34, 32–41. https://doi.org/10.1016/j.cja.2020.07.024.

Dogan, A., Birant, D., 2021. Machine learning and data mining in manufacturing. Expert Syst. Appl. 166, 114060 https://doi.org/10.1016/j.eswa.2020.114060.

Elmousalami, H.H., 2021. Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis. IEEE Trans. Eng. Manag. 68, 183–196. https://doi.org/10.1109/TEM.2020.2972078.

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R., Munigala, V., 2021. Data quality for machine learning tasks. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA, pp. 4040–4041. https://doi.org/10.1145/3447548.3470817.

Hammann, D., 2024. Big data and machine learning in cost estimation: an automotive case study. Int. J. Prod. Econ. 269, 109137 https://doi.org/10.1016/j.ijpe.2023.109137.

Hennebold, C., Klöpfer, K., Lettenbauer, P., Huber, M., 2022. Machine learning based cost prediction for product development in mechanical engineering. Procedia CIRP 107, 264–269. https://doi.org/10.1016/j.procir.2022.04.043.

Hihn, J., Menzies, T., 2015. Data mining methods and cost estimation models: why is it so hard to infuse new ideas?. In: 2015 30th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW). IEEE, pp. 5–9. https://doi.org/10.1109/ASEW.2015.27.

Kadir, A.Z.A., Yusof, Y., Wahab, M.S., 2020. Additive manufacturing cost estimation models—a classification review. Int. J. Adv. Des. Manuf. Technol. 107, 4033–4053. https://doi.org/10.1007/s00170-020-05262-5.

Kamps, T., Lutter-Guenther, M., Seidel, C., Gutowski, T., Reinhart, G., 2018. Cost- and energy-efficient manufacture of gears by laser beam melting. CIRP J Manuf Sci Technol 21, 47–60. https://doi.org/10.1016/j.cirpj.2018.01.002.

Kanyilmaz, A., Tichell, P.R.N., Loiacono, D., 2022. A genetic algorithm tool for conceptual structural design with cost and embodied carbon optimisation. Eng. Appl. Artif. Intell. 112, 104711 https://doi.org/10.1016/j.engappai.2022.104711.

Langmaak, S., Wiseall, S., Bru, C., Adkins, R., Scanlan, J., Sóbester, A., 2013. An activity-based-parametric hybrid cost model to estimate the unit cost of a novel gas turbine component. Int. J. Prod. Econ. 142, 74–88. https://doi.org/10.1016/j.ijpe.2012.09.020.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Liu, Y., Wang, Y., Zhang, J., 2012. New Machine Learning Algorithm: Random Forest, pp. 246–252. https://doi.org/10.1007/978-3-642-34062-8_32.

Loyer, J.L., Henriques, E., Fontul, M., Wiseall, S., 2016. Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components. Int. J. Prod. Econ. 178, 109–119. https://doi.org/10.1016/j.ijpe.2016.05.006.

Lukić, D., Borojevic, S., Đurđev, M., Milošević, M., Borojević, S., Vukman, J., Antić, A., 2016. MANUFACTURING COST ESTIMATION IN the CONCEPTUAL PROCESS PLANNING, Machine Design.

Maier, S., Zimmermann, P., Berger, J., 2022. MANU-ML: methodology for the application of machine learning in manufacturing processes. Procedia CIRP 107, 798–803. https://doi.org/10.1016/j.procir.2022.05.065.

Martinelli, I., Campi, F., Checcacci, E., Presti, G.M. Lo, Pescatori, F., Pumo, A., Germani, M., 2019. Cost estimation method for gas turbine in conceptual design phase. Procedia CIRP 84, 650–655. https://doi.org/10.1016/j.procir.2019.04.311.

Masel, D.T., Dowler, J.D., Judd, R.D., 2010. Adapting bottoms-up cost estimating relationships to new systems. In: SPA/SCEA Joint Annual Conference and Training Workshop.

Mazurek, S., Wielgosz, M., 2023. Assessing Dataset Quality through Decision Tree Characteristics in Autoencoder-Processed Spaces.

Molnar, Christoph, 2022. Interpretable Machine Learning A Guide for Making Black Box Models Explainable.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorob. 7 https://doi.org/10.3389/fnbot.2013.00021.

Niazi, A., Dai, J.S., Balabani, S., Seneviratne, L., 2006. Product cost estimation: technique classification and methodology review. J. Manuf. Sci. Eng. 128, 563–575. https://doi.org/10.1115/1.2137750.

Ning, F., Shi, Y., Cai, M., Xu, W., Zhang, X., 2020a. Manufacturing cost estimation based on a deep-learning method. J. Manuf. Syst. 54, 186–195. https://doi.org/10.1016/j.jmsy.2019.12.005.

Ning, F., Shi, Y., Cai, M., Xu, W., Zhang, X., 2020b. Manufacturing cost estimation based on the machining process and deep-learning method. J. Manuf. Syst. 56, 11–22. https://doi.org/10.1016/j.jmsy.2020.04.011.

Rapaccini, M., Cadonna, V.L., Leoni, L., De Carlo, F., 2023. Application of machine learning techniques for cost estimation of engineer to order products. Int. J. Prod. Res. 61, 6978–7000. https://doi.org/10.1080/00207543.2022.2141907.

RapidMiner Documentation, 2023. https://docs.rapidminer.com/9.3/studio/auto-model/ [WWW Document]).

Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517. https://doi.org/10.1093/bioinformatics/btm344.

Su, X., Yan, X., Tsai, C.-L., 2012. Linear regression. Wiley Interdiscip Rev Comput Stat 4, 275–294. https://doi.org/10.1002/wics.1198.

Van Nguyen, T.H., Huang, P.-M., Chien, C.-F., Chang, C.-K., 2023. Digital transformation for cost estimation system via meta-learning and an empirical study in aerospace industry. Comput. Ind. Eng. 184, 109558 https://doi.org/10.1016/j.cie.2023.109558.

Verlinden, B., Duflou, J.R., Collin, P., Cattrysse, D., 2008. Cost estimation for sheet metal parts using multiple regression and artificial neural networks: a case study. Int. J. Prod. Econ. 111, 484–492. https://doi.org/10.1016/j.ijpe.2007.02.004.

Wang, H.S., Wang, Y.N., Wang, Y.C., 2013. Cost estimation of plastic injection molding parts through integration of PSO and BP neural network. Expert Syst. Appl. 40, 418–428. https://doi.org/10.1016/j.eswa.2012.01.166.

Weichert, D., Link, P., Stoll, A., Rüping, S., Ihlenfeldt, S., Wrobel, S., 2019. A review of machine learning for the optimisation of production processes. Int. J. Adv. Des. Manuf. Technol. 104, 1889–1902. https://doi.org/10.1007/s00170-019-03988-5.

Xie, J., Sage, M., Zhao, Y.F., 2023. Feature selection and feature learning in machine learning applications for gas turbines: a review. Eng. Appl. Artif. Intell. 117, 105591 https://doi.org/10.1016/j.engappai.2022.105591.

Yeh, T.-H., Deng, S., 2012. Application of machine learning methods to cost estimation of product life cycle. Int. J. Comput. Integrated Manuf. 25, 340–352. https://doi.org/10.1080/0951192X.2011.645381.

Yoo, S., Kang, N., 2021. Explainable artificial intelligence for manufacturing cost estimation and machining feature visualisation. Expert Syst. Appl. 183 https://doi.org/10.1016/j.eswa.2021.115430.