Check for updates

Received: 2 March 2021 | Revised: 19 October 2021 | Accepted: 25 October 2021

DOI: 10.1002/qre.3025

WILEY

REVIEW ARTICLE

Design of Experiments and machine learning for product innovation: A systematic literature review

Rosa Arboretti[1] | **Riccardo Ceccato**[2] | **Luca Pegoraro**[2] | **Luigi Salmaso**[2]

[1] Department of Civil, Environmental and Architectural Engineering, Università degli Studi di Padova, Padua, Italy

[2] Department of Management and Engineering, Università degli Studi di Padova, Vicenza, Italy

Correspondence
Luigi Salmaso, Department of Management and Engineering, Università degli Studi di Padova, Vicenza, Italy.
Email: luigi.salmaso@unipd.it

Abstract

The recent increase in digitalization of industrial systems has resulted in a boost in data availability in the industrial environment. This has favored the adoption of machine learning (ML) methodologies for the analysis of data, but not all contexts boast data abundance. When data are scarce or costly to collect, Design of Experiments (DOE) can be used to provide an informative dataset for analysis using ML techniques. This article aims to provide a systematic overview of the literature on the joint application of DOE and ML in product innovation (PI) settings. To this end, a systematic literature review (SLR) of two major scientific databases is conducted, retrieving 388 papers, of which 86 are selected for careful analysis. The results of this review delineate the state of the art and identify the main trends in terms of experimental designs and ML algorithms selected for joint application on PI. The gaps, open problems, and research opportunities are identified, and directions for future research are provided.

KEYWORDS

causality, experimental design, product development, research and development, uncertainty

1 | INTRODUCTION

In recent years, industry has been undergoing a fourth industrial revolution, also referred to as "Industry 4.0." One of the main drivers of this phenomenon is the creation of cyber-physical systems which, by integration of physical equipment with digital systems, enables the generation of large and potentially continuous streams of data that are produced by various sources and characterized by diverse formats. These new types of data, often referred to as "big data," require the adoption of specific approaches for their analysis. Methods from traditional parametric statistics rely on assumptions about the shape of data that are not always met in "big data" settings in which the analyst typically has only scant control over the data collection process. Furthermore, traditional statistical methods can perform badly in situations where the volume of data is extremely large, or in situations of extreme high dimensionality ($p >> n$), which are typical in "big data" contexts. Other problems may be encountered when the data are not in numerical format, but rather image, video, or audio files. Machine learning (ML) refers to a family of techniques, which enable computer programs to learn specific patterns or trends from historical data in order to perform complex tasks such as prediction, classification, and clustering. As a consequence of the increasing availability of large and complex sets of data, ML has rapidly spread in industry as the

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *Quality and Reliability Engineering International* published by John Wiley & Sons Ltd.

main analytical tool. However, not all industries, nor all business functions are equally invested in the big data revolution, mainly because data availability is not the same throughout the varying contexts.

Consider a company in the steel industry. At a production level, steel slabs are produced continuously using automated equipment, which can be integrated with hundreds of sensors positioned at the various stages of the continuous casting process, producing terabytes of data.[1] The research and development (R&D) function of the same company may be interested in the development of a new alloy steel with enhanced properties. In this case, data will very likely be scarce because data in materials science usually consists of a small number of observations compared to the enormous search space of unexplored materials.[2,3]

This is a problem common to many industries in which product improvement and innovation relies on generating knowledge of the phenomenon, which drives product performance. In these industries, product performance depends upon one or more objective indicators, which can be measured and for the most part or entirely define the quality of the final product itself. Some examples of such indicators include fatigue strength for mechanical transmission system applications, tensile strength, and resistance to corrosion for generic metal applications, stain removal capacity for detergent products, solubility for pharmaceutical tablets, and many others. In such cases, strong and robust conclusions for product improvement and innovation can be achieved only as a result of analysis of the data collected on the phenomenon of interest. Due to the peculiarities of each case which limit the field of application, and the reluctance of companies to share sensitive data, very rarely are large datasets of this kind publicly available, and organizations are motivated to collect the data they need autonomously.

Experimentation based on a statistical plan is an effective approach to data collection, which has been applied in product and process improvement for decades. Design of Experiments (DOE) is a statistical method, which guides the execution of experiments, which in turn are analyzed to detect the relevant variables and optimize the process or phenomenon under investigation.[4] The use of DOE in product innovation (PI) can result in products that are easier and cheaper to manufacture, that have enhanced performance and reliability, and require a shorter product design and development time.[4]

Only recently have the first attempts to assess the implications of a joint application of DOE and ML been carried out, but these refer mainly to production settings in which large amounts of data are generated and only need to be captured by means of a monitoring infrastructure.[5] In their work,[5] the authors conclude that ML methods could potentially supplant DOE in this situation because, in the normal production process, the inputs are varied to an amount sufficient to capture the relevant effect of each factor, meaning that data abundance by means of continuous process monitoring could make experimentation unnecessary. The scope of this paper is different. It focuses on the application of DOE and ML as a means to increase the investigative power in a PI setting, which is typically characterized by data scarcity. Some work has been carried out on this topic, but it only presented a general background of the literature and a case study application.[6] Therefore, the aim of this article is to present a systematic literature review (SLR) covering the publications on the joint application of DOE and ML in a PI setting. To the best of our knowledge, this is the first comprehensive literature review of these topics, therefore our aim is to identify the state of the art of this branch of literature for the first time, identify opportunities, and propose future research directions.

The paper is organized as follows: Section 2 describes the background to the topics of DOE and ML; Section 3 presents the methodology and execution of the SLR; Section 4 shows the result of the SLR; and Section 5 discusses the findings. Finally, in Section 6, the contributions of this work are highlighted and some final conclusions are presented.

2 | BACKGROUND

In this section, we briefly introduce the topics of DOE in Section 2.1 and ML in Section 2.2. Further, we present a simplified framework, which is typically adopted in studies that apply ML on DOE data in Section 2.3. In Section 2.4, we briefly discuss a case study to better illustrate the framework described in Section 2.3.

2.1 | Design of Experiments

Germinal literature on DOE dates back to the 1920s, when Sir Ronald A. Fisher developed a statistical methodology for planning experiments in the field of agricultural research.[7] Soon the principles of DOE extended into industry, in some

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

cases even supplanting the inadequate and expensive "one factor at a time" experimentation, with many applications dating back to the 1950s or earlier.[8]

The basis of DOE is the identification of a set of factors, which can potentially drive process performance, the selection of reasonable levels for each of these factors, the definition of a set of combinations of factor levels and the execution of experiments according to the defined experimental design. Clearly, in order to assess the impact of each process variable, the factors should be organized at least on two levels and augmented if interest also lies in the detection of curvature effects. Consistent with this description, and given k factors and l levels, a full factorial design including all combinations of factors and factor levels will produce a total of l^k experimental runs. It is evident that as k or l increase, the number of experiments can quickly become unfeasible. To this end, fractional factorial designs that require only a fraction of the runs of a full factorial design have been introduced, nonetheless able to estimate a sufficient number of effects. Other classic designs used for the detection of interactions and quadratic effects, which are built upon 2^k factorial or fractional factorial designs, include the central composite design (CCD) and Box-Behnken Design (BBD). CCDs augment a 2^k factorial or fractional factorial design including $2k$ axial runs and c center points, while BBDs are constructed by combining 2^k factorials with incomplete block designs, leading to solutions, which are often more efficient than CCDs in terms of experimental runs for practical applications.[4] Full factorials, fractional factorials, CCDs, and BBDs are all designs used in the context of response surface methodology (RSM). Other designs used in this context are those with specific peculiarities such as mixture designs and designs with randomization restrictions.[4]

RSM is an experimental method based on a collection of statistical designs and procedures that was first introduced by Box and Wilson[9] and belongs to the family of DOE methods. The RSM has the dual objective of understanding the influence of factors on the response(s) and identifying the optimal configuration of the parameters, which leads to an optimization, typically by a maximization or minimization, of the response variable(s). As briefly pointed out, several families of designs have been proposed over the years and each contributed to making RSM the core of DOE in industrial experimentation. For a comprehensive review of RSM, please refer to the work of Myers et al.[10].

In parallel with RSM, robust parameter designs (RPD) were developed mainly by Taguchi.[11–13] The idea of RPD is that processes are governed by two sets of factors: control factors, that are under complete control of the experimenter, and noise factors, that may be controlled solely during the experimentation process but not in production. The objective of RPD is to find the configuration of factors that are insensitive to noise, thus minimizing the transmission of variability from uncontrolled noise factors. This is achieved by putting control factors in an inner array design, and noise factors in an outer array design, as such obtaining a design that can guide experiments with the aim of increasing the quality of a process or a product in a cost-effective manner. For a comprehensive review of RPDs and critiques of Taguchi's contribution to DOE, please refer to Robinson et al.[14] and Box et al.[15].

Another relevant family of DOE designs is space-filling. The idea behind the construction of space-filling designs is to spread out experimental points in an even manner in order to collect as much information as possible on the design space. One of the most widespread space-filling designs is the Latin hypercube design (LHD),[16] which is based on a multidimensional expansion of the concept of Latin squares, that are constructed in a bidimensional space positioning one and only one sample in each row and column. Space-filling designs are becoming increasingly popular in industry, as they are adopted in the context of computer simulations of physical systems; for an extensive review of space-filling designs, please refer to Joseph.[17]

An additional concept that can be applied across all different types of DOE is sequential experimentation. The principle of sequential experimentation is to perform a series of experiments in a sequential manner, using the results of the previous experiments to drive the next steps. Box and Liu[18] assert that industrial innovation comes as a result of a dynamic and adaptive process of learning, thus a sequential approach to DOE is advocated, particularly in those settings in which preexisting knowledge is limited, such as in product or process development and innovation. Several authors stressed the importance of adopting a sequential approach for DOE,[19,20] however, the rather seldom adoption of this paradigm makes it an open issue in practical DOE studies.[21]

2.2 | Machine learning

ML refers to a vast set of methods and algorithms that enable computer programs to learn from data. One common categorization of ML is supervised versus unsupervised learning.[22] Supervised learning involves building a model for predicting one or more outputs based on one or more inputs. When the outputs are not available, only the patterns among the input variables can be studied; in such cases, the ML type is unsupervised learning. Other ML types exist, such as

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneitalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

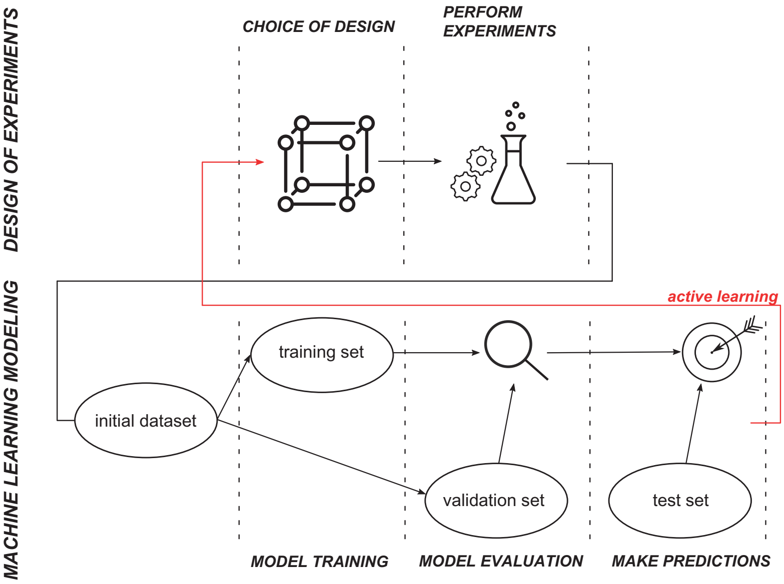


FIGURE 1 A simplified framework, which describes the application of ML as a tool to analyze Design of Experiments (DOE) data (DOE + ML framework). If the results of the machine learning (ML) modeling step are used to suggest the next experimental configurations, the specific framework becomes an active learning (AL) one (in red)

semi-supervised learning (the output is available only for few data points) and reinforcement learning, which is typically used for technologies such as self-driving cars and robotics.[23] In this article, given the type of application, we mainly refer to the supervised learning setting.

The task of developing supervised ML models from a set of training data is rather articulated, and specific applications may require different and specialized efforts. Nevertheless, a simplified framework includes the two macro-phases of data management and data analytics.[24] Data management mainly deals with the tasks of data acquisition, filtering, and cleaning in order to prepare data for the subsequent analytical phase. In data analytics, the main assignments are model training and evaluation; finally the model is put into production and used to make predictions on new and previously unseen data. A plethora of different ML models exists, and depending on the needs and peculiarities of different applications, one may be preferable over the others. For an excellent introduction to the main ML algorithms, please refer to the work of Hastie et al.[25]

2.3 | ML on DOE's data

As will be clear in the following sections of this paper, the application of ML models as a method to analyze DOE data has become more and more common in recent years. A simplified framework of this approach is summarized in Figure 1. In its most elementary implementation, a ML model substitutes the typical methods used for analysis of DOE data, as such performing prediction and optimization. We call this framework "DOE + ML." However, a recent trend in the application of DOE and ML in PI is based on an iterative application of DOE and supervised ML in which the learner proposes the next experimental configurations. Therefore, the learner is in control of the data from which the machine learns; this approach is often referred to as "active learning" (AL).[26] A concept strongly linked with AL and often used for adaptive experimental designs is Bayesian optimization (BO). BO incorporates a Gaussian process (GP) model and an acquisition function, which drives the adaptive sampling procedure, using the mean prediction and uncertainty of the GP model. For a recent review of BO for adaptive DOE, refer to the work of Greenhill et al.[27]

In both the nonsequential DOE + ML and AL frameworks, a crucial step concerns model assessment to evaluate the performance of the predictive model. This is also essential for tuning the hyperparameters of the ML algorithm, a phase that heavily impacts the ability of the method to provide accurate and reliable predictions. The typical approach consists in holding out a subset of the initial training data to assess the performance of the algorithms on these unseen data points.[22] This can either be done by randomly dividing the initial dataset into a training and validation set (VS approach) or by partitioning the initial data into k folds of (approximately) equal size, using one of these as a validation set and the remaining k−1 groups as training data (k-fold cross-validation – k-fold CV), repeating the procedure k times.[22] One special case of k-fold CV is when k equals the number of training data, identifying the so-called "leave-one-out cross-validation" (LOOCV).[22] Then, when the algorithm is put into production, it is used to make predictions on test data that

1091638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

was previously unseen and did not belong to the initial dataset. For simplicity, in Figure 1, we report the VS approach, but k-fold CV is largely employed in ML tasks as it gives a more reliable estimate of the test error.[22]

Some concerns have been raised about the appropriateness of partitioning DOE data for the purpose of training ML models. The main issue is that, since one of the objectives of DOE is the minimization of the number of experiments required to estimate a given number of parameters, the removal of even a few experimental configurations can potentially destroy the design structure leading to the development of models with inadequate performance or even to the impossibility of fitting a model capable of estimating the desired effects. In this context, it should be emphasized that while the latter risk is certainly valid for the typical use of quadratic regression with interactions to analyze DOE data, ML algorithms generally require less strict assumptions about the data size or structure, meaning that undesigned data can be effectively modeled, leading to satisfactory predictions on unseen data.

Nonetheless, we recognize that the partition of DOE data into training and validation sets is a controversial matter that needs more attention in the literature as, to the best of our knowledge, it has never been directly addressed, either analytically or empirically, by means of a dedicated study. Consequently, the real impact of splitting DOE data for model training and validation has not yet been quantified, and it is difficult to come to a definitive conclusion. One recent advancement that can potentially overcome this issue was proposed by Lemkus et al.[28] In their article, Lemkus et al.[28] propose an innovative method that uses a modification of the fractionally weighted bootstrap[29] in order to include all the available DOE observations in both the training and validation sets. A sufficient level of independence between the two partitions is obtained by weighting the same observation in the training and validation sets with exponentially distributed weights (a large weight in the training data corresponds to a small weight in the validation set and vice versa) so that maximum anticorrelation is obtained.

2.4 | Motivating example

In the remainder of the paper, we will provide a systematic and comprehensive overview of the literature on the application of DOE and ML for PI. In this section, as a motivating example, we present an industrial application that successfully employs the DOE + ML framework. The case study is taken from the work of Arboretti et al.,[6] and consists in the application of DOE + ML as a way to collect and analyze data for the development of a new detergent. The main objective was to develop models that can reliably predict washing performance on three different types of stains, based on the components of the detergent's formula. Moreover, although the goal is predictive in nature, the authors also wanted to develop fairly interpretable models in order to provide chemists with all the elements required for the generation of new knowledge of the phenomenon. Accordingly, together with prediction, inference was also one of the study's objectives.

To this end, the DOE + ML technique was selected because it was considered capable of providing both sufficient accuracy in prediction and a causal basis, due to the experimental nature of data. A BBD was chosen as the experimental design, primarily because it avoided extreme combinations of the factor levels that are known to undermine the stability of the chemical formulae. Furthermore, at the time, the company was not keen to explore factorial designs with more than three levels due to the expected increase in experimental complexity and cost. Nonetheless, a priori knowledge indicated that the phenomenon was characterized by high and rather unpredictable variability, thus a replicated three-level design was considered the best option.

After data collection, careful analysis of the data structure was performed, leading to the identification of both fixed and random effect terms influencing the responses. Accordingly, the ANOVA estimation method[30] was employed, leading to a quantification of the relevant variance components in terms of variability *between* and *within* replicates. This confirmed the presence of large and input-dependent noise (heteroscedasticity). The DOE data were then randomly split into training and test sets (75%–25%), and five-fold CV was further employed on the training set to appropriately set the hyperparameters of the predictive models. After careful tuning, three models were fitted for each response, namely: quadratic regression with interactions, artificial neural network ensemble (ANN ensemble) and Random Forest.[31] The ML algorithms were selected since they were expected to provide more accurate predictions than the traditional statistical model. However, since the objectives included inference, the authors tried to shed some light on the black box of ML models by providing a measure of variable importance together with a quantification of prediction uncertainty for each model. The authors believe, therefore, that the developed models represent a reasonable trade-off between model accuracy and interpretability.[22] This is confirmed by the final results of the study: the application of the ML models reduces the predictive error by about $\sim 20\% - 30\%$ depending on the response variable, if compared to the traditional statistical approach. At the same time, the estimation of variable importance and uncertainty quantification, together with an extensive

1091638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

analysis and discussion of the results, prove that ML models can be applied on DOE data without necessarily losing much in terms of interpretability with respect to parametric approaches. The authors[6] state that the developed algorithms can be included in a semi-automatic system for prediction of the washing performance, thus reducing the need for future experiments and facilitating the process of new product development.

In the following sections, we will observe that, in addition to the advantages of the DOE + ML framework highlighted in the industrial application described by Arboretti et al.,[6] several other advantages have been found in the literature.

3 | METHODOLOGY

In this section, we describe the adopted SLR approach, focusing on the literature research questions, search queries, the search process, exclusion criteria, database construction and analysis. We follow the general principles of SLR as detailed by Xiao and Watson[32] and Petersen et al.,[33] adapting them to this specific research, with the aim of achieving high-quality and reproducible results.

3.1 | Planning the review

In this phase, the main tasks include the formulation of literature research questions (LRQs) and definition of search queries, which should guide the search in online scientific databases.

3.1.1 | LRQs

This research deals with the application of DOE and ML in the field of PI. Even if not clearly stated, all the LRQs listed below are in the context of PI:

LRQ1: What are the advantages and challenges of the application of ML methods compared to traditional parametric statistical approaches used for the analysis of DOE data?
LRQ2: Considering a DOE + ML framework, what DOE strategies and DOE types are adopted?
LRQ3: Considering a DOE + ML framework, what ML algorithms are adopted?
LRQ4: Considering a DOE + ML framework, what data partitioning methods are adopted for ML model assessment?
LRQ5: What are the implications of adopting a DOE + ML framework?
LRQ6: What gaps, open questions and research opportunities are identified?

3.1.2 | Search queries

The next step consists in the definition of strings and selection of databases. The keywords used to compose the string are divided into three categories: ML, DOE, and PI. After some iterations and discussion among the authors of the paper, the final query formulation was defined and is reported in Figure 2. One thing to point out is that while the first two categories were searched in the title–abstract–keywords of the paper, the third category was searched within the whole document. This is the result of a careful investigation, which highlighted that the DOE + ML framework is often adopted in PI settings even if this is not clearly stated in the abstract of the article. In fact, it is often stated in the introduction, mainly because the adoption of DOE in PI is so widespread that it may not be specified, by means of the selected keywords, in the abstract. Nevertheless, application in PI may already be clear from the abstract if keywords specific to the subject area of application are adopted. The final query was inputted within Scopus and Web of Science (WoS) scientific databases on October 28, 2020.

3.2 | Conducting the review

At this point, the review process starts. The first task is to refine the search to reduce the body of work to only relevant papers. The second task regards the definition of significant dimensions for descriptive and content analyses that will be used to organize the final results in a systematic and structured manner.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

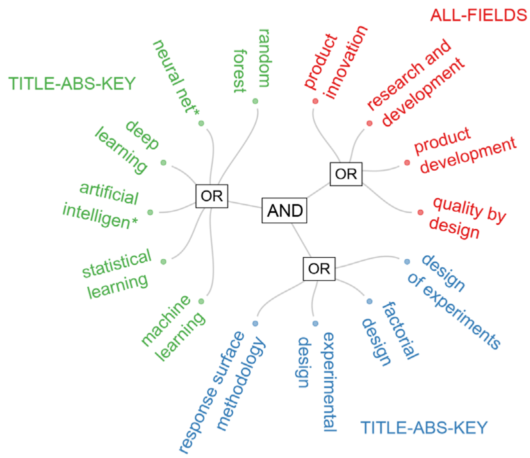


FIGURE 2 The final query inputted within Scopus and Web of Science (WoS) databases. The first category concerns machine learning (ML) (in green), the second concerns Design of Experiments (DOE) (in blue), and the third product innovation (PI) (in red). The first two categories were searched within title–abstract–keywords, while the third within the whole document

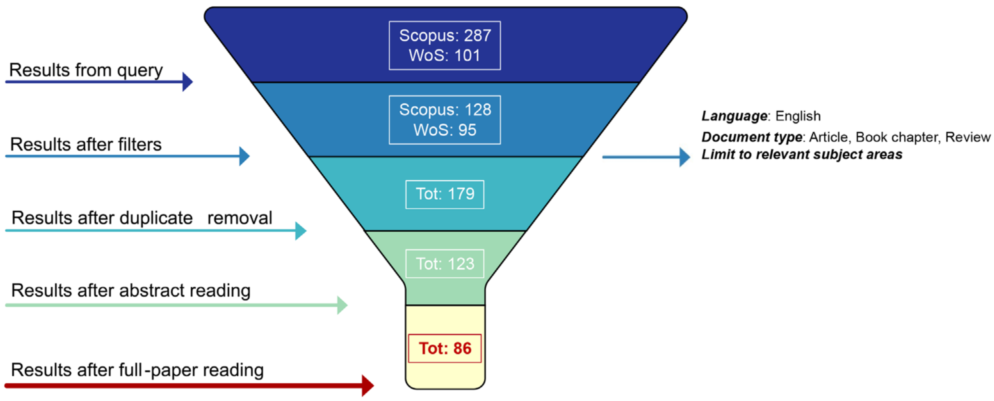


FIGURE 3 Systematic literature search and refinement process

TABLE 1 The quality criteria used for exclusion of papers based on full text reading

ID	Exclusion criteria
C1.	The terms of the search query are solely mentioned in the general introduction of the paper.
C2.	The general description of the application of DOE and/or ML is very unclear.
C3.	DOE and ML are not jointly applied to the same case study or discussed in the same context.
C4.	The field of application is distant from a PI setting.

Abbreviations: DOE, Design of Experiments; ML, machine learning; PI, product innovation.

3.2.1 | Search refinement

The initial database of papers undergoes a screening process with the aim of selecting only those works relevant to the scope of the SLR (Figure 3). The initial query found 287 results in Scopus and 101 results in WoS. These were initially filtered by language (only papers in English are considered), article type, and subject area, with the exclusion of fields such as medicine, social sciences, chemistry, and physics. The complete search queries are provided as supplemental material. A second filter was applied on duplicates, narrowing the body of work to 179 papers. All abstracts were then read and 56 papers were excluded as it was clear that the focus of the paper was far from the scope and LRQs of the present SLR. The remaining papers were carefully read, and further screened in line with the exclusion quality criteria listed in Table 1. A final database of 86 papers was used for subsequent analysis.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

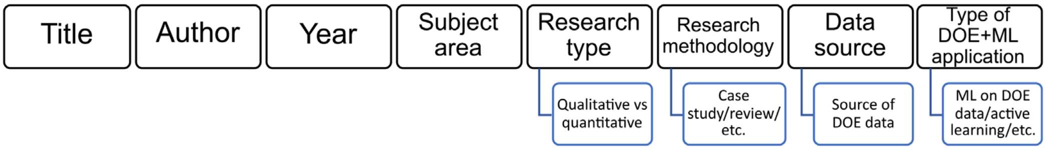


FIGURE 4 Framework for descriptive analysis

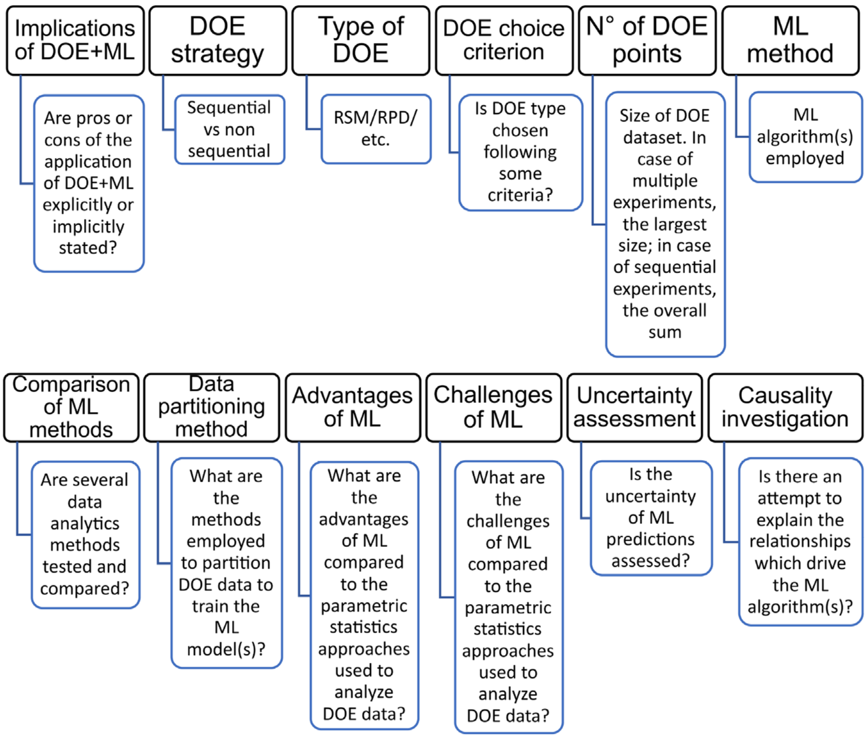


FIGURE 5 Framework for content analysis (CA)

3.2.2 | Descriptive and content analyses

In order to analyze the final database of papers from the search refinement process, descriptive and content analyses were conducted.

Descriptive analysis serves to gain some initial insights into the papers by providing basic information as detailed in Figure 4.

One definition of content analysis (CA) is "a research technique for the objective, systematic, and quantitative description of the manifest content of communication"[34]; CA is applied as a primary tool to synthesize the research topics and uncover essential matters in publications, with the objective of identifying open questions and research opportunities in a specific body of knowledge. The core step in CA is the identification, by an inductive or deductive approach, of analytic dimensions or categories that constitute the basic structure of the body of literature and are used to classify the reviewed material according to the LRQs, which drive the study.[35] Such dimensions constitute the framework for CA, as reported in Figure 5.

4 | RESULTS

In this section, we present the results of the descriptive and content analyses of the selected literature with the aim of responding to the LRQs of this research.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

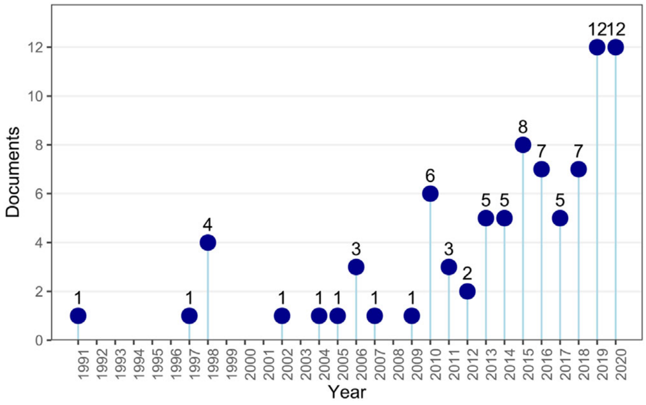
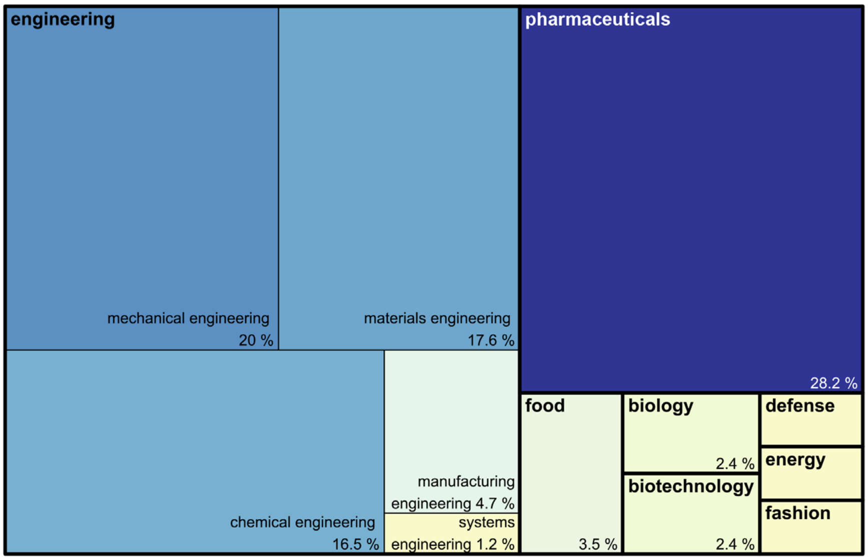


FIGURE 6 Number of papers per year

FIGURE 7 Distribution of papers (total = 86) per subject area



4.1 | Descriptive analysis

Figure 6 shows the distribution of publications over the years. The results confirm a growing interest in this research area since half of the papers have been published in the last 5 years, with a peak in 2019 and 2020.

Engineering is the most influential field of application of the methodology, with 60% of publications similarly divided between mechanical, materials, and chemical engineering (Figure 7). Another very relevant topic is pharmaceutical applications, accounting for 28.2%. This distribution is in line with the application of DOE in the fields of engineering and pharmaceutical product development, two of the major fields of application of this statistical technique.[36,37]

Among the 86 papers considered in the database, 92% adopted a quantitative research approach, the majority of which being case study applications discussing the adoption of the DOE + ML framework for the innovation of specific products. The second category is conceptual papers, in which there is a degree of discussion of the DOE + ML approach adopted and not simply a straight application of the method. Less relevant are the reviews and simulation studies. For papers adopting a mixed research methodology, the same paper was counted in each pertinent category (Figure 8).

Among the quantitative papers that employ DOE data, the majority used real data from physical experiments rather than data sourced from computer experiments using software that simulates a specific process. Fewer works consider data taken from published literature papers, or synthetic data artificially generated in a simulation study, mainly to prove robustness of a methodology in different scenarios (Figure 9).

The type of DOE + ML application is typically a nonsequential or sequential (only two papers) application of ML algorithms to analyze DOE data, here defined as the "ML on DOE data" strategy (Figure 10). On the other hand, if the ML model suggests the data from which it learns, the application is of the AL type. Such an approach has recently been

1091638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

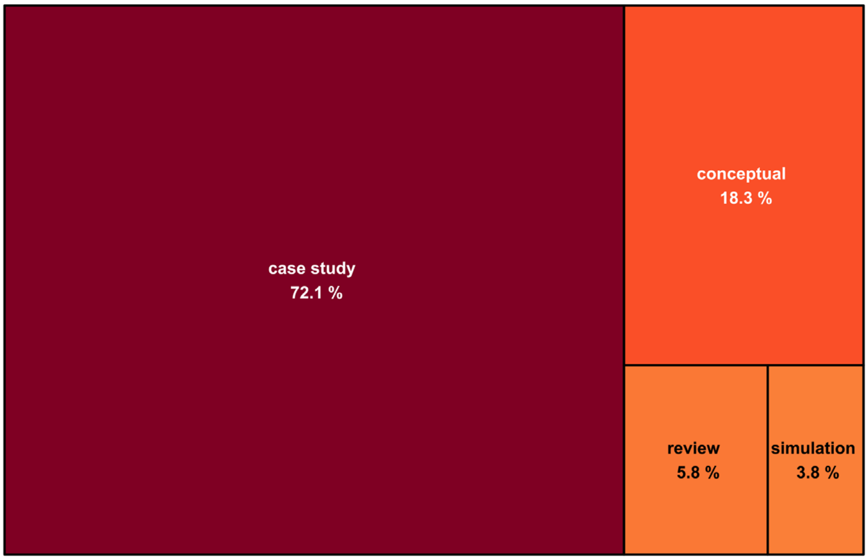


FIGURE 8 Distribution of papers (total = 86) per research methodology

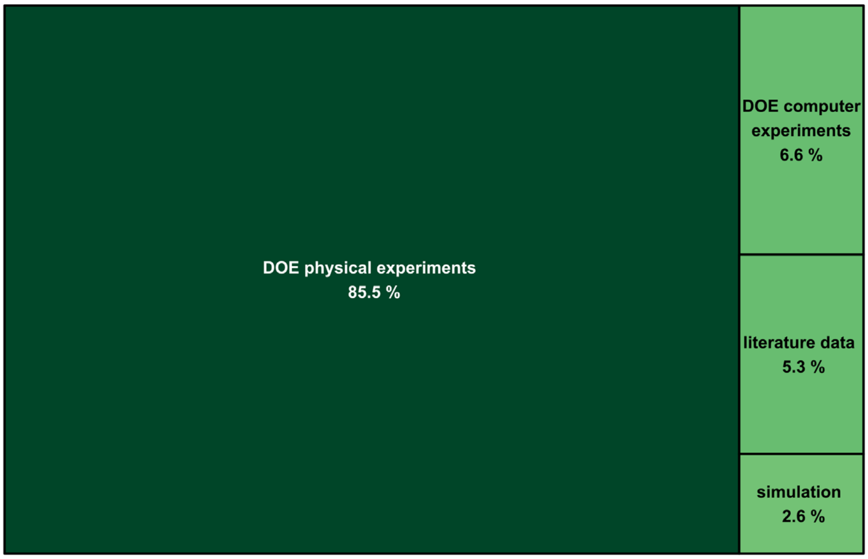


FIGURE 9 Distribution of papers (total = 76) per data source. Note that the number of articles considered is less than the systematic literature review (SLR) total, since in some papers (e.g., some reviews and conceptual papers), Design of Experiments (DOE) data are not used. Similar considerations apply in the next figures

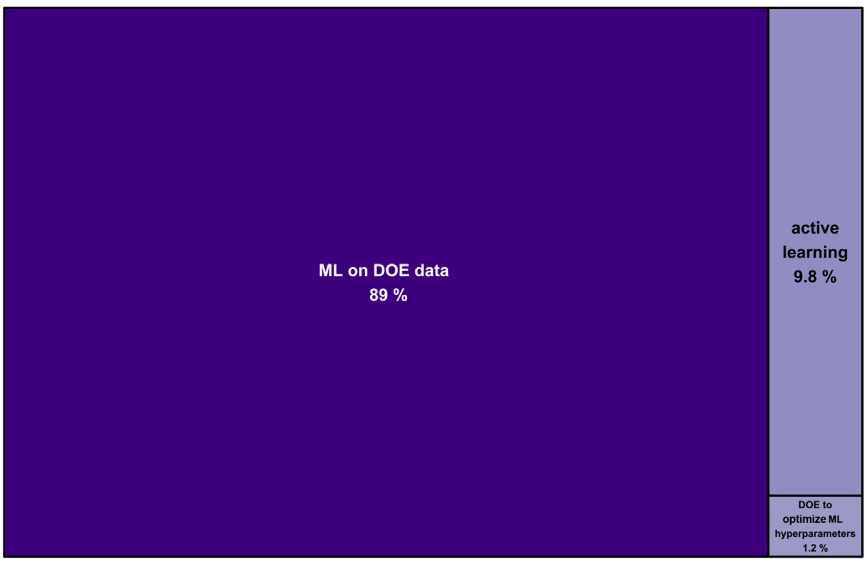


FIGURE 10 Distribution of papers (total = 82) per type of DOE + ML application

1091638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

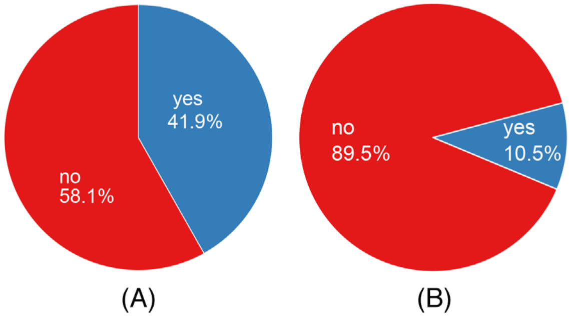


FIGURE 11 Distribution of papers (total = 86) that discuss the advantages (A) and challenges (B) of the application of machine learning (ML) methods compared to traditional parametric statistical approaches used for the analysis of Design of Experiments (DOE) data

TABLE 2 The advantages of the application of machine learning (ML) compared to traditional parametric statistical approaches used for the analysis of Design of Experiments (DOE) data in terms of percent value (%) and number of papers (**N**)

Description	%	*N*
ML provides more accurate predictions	58.3	21
ML can model highly complex nonlinear relationships	33.3	12
ML can find better optimal configurations than RSM strategy	25	9
ML can deal with undesigned data	19.4	7
ML does not require assumptions on distribution of data	16.7	6
ML can use production data, minimizing the need for production stops and enhancing automation	11.1	4
ML can deal with many predictors, even uncontrollable ones	8.3	3
ML has a plug-and-play approach	5.6	2

Abbreviations: ML, machine learning; RSM, response surface methodology.

gaining traction, with one paper in 2015, 2018, and 2020 and five papers in 2019. One work uses DOE techniques to optimize ML hyperparameters.

4.2 | Content analysis

In this section, we present the results of the CA and the answers to LRQs 1 to 5. Each answer is supported by a figure or table that shows the current state in a quantitative manner. For a complete list of the papers organized by dimensions of the CA, see the Appendix.

4.2.1 | LRQ1: What are the advantages and challenges of the application of ML methods compared to traditional parametric statistical approaches used for the analysis of DOE data?

Less than 42% of papers discuss the advantages of the application of ML compared to the traditional data analytics methods used to analyze DOE data, and this percentage is reduced to 10.5% when the focus shifts to the challenges (Figure 11).

In general, the advantages refer mainly to the ability of ML to provide accurate predictions even with complex nonlinear and undesigned data, also leading to a better optimal configuration of the factors while not requiring particular assumptions on the distribution of data (Table 2). However, such advantages are very often limited to the specific application, and cannot be generalized to a wider setting.

Challenges, on the other hand, deal mainly with well-known algorithmic problems that broadly affect ML methods[38] (Table 3).

TABLE 3 The challenges of the application of machine learning (ML) in terms of percent value (%) and number of papers (**N**)

Description	%	N
ML is sensitive to data quality	33.3	3
ML tends to overfit data	22.2	2
ML tends to be a black box	22.2	2
ML presents computational challenges	11.1	1
ML does not provide a quantification of uncertainty	11.1	1
There is a lack of expertise in ML	11.1	1
Traditional DOE is unsuitable to fit advanced ML models	11.1	1

DOE, Design of Experiments; ML, machine learning.

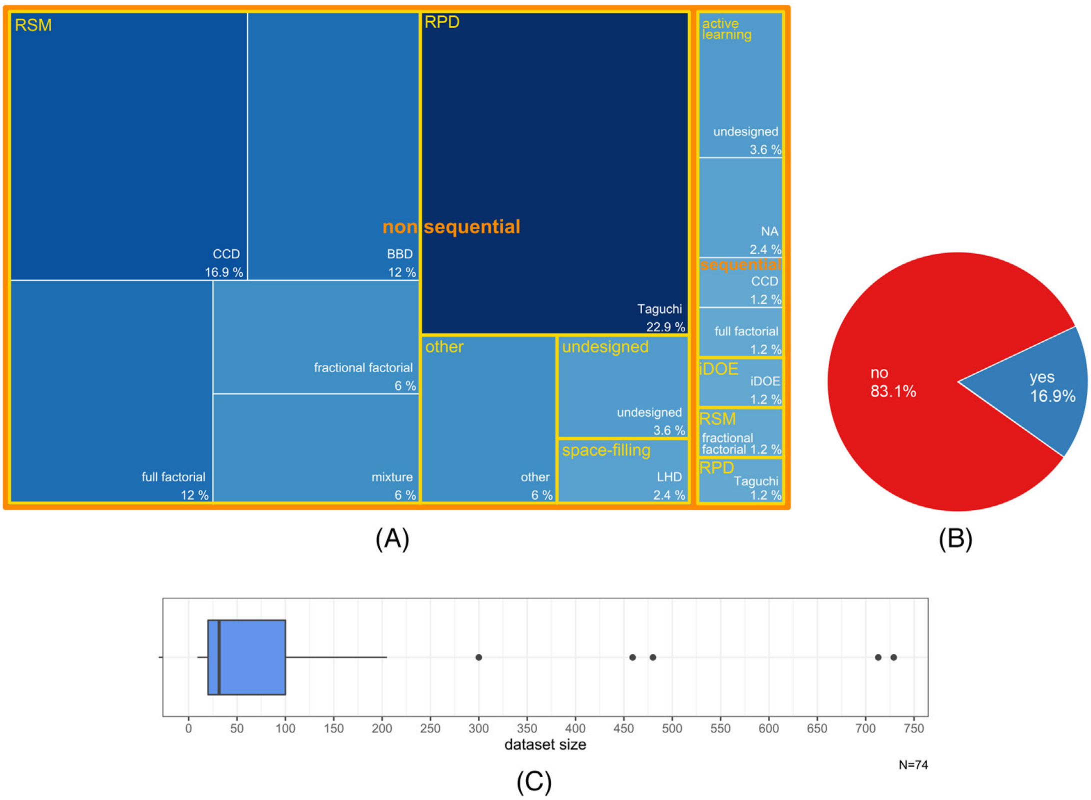


FIGURE 12 Distribution of studies (total = 83) by Design of Experiments (DOE) strategy and DOE type (A), and by the use of some criteria for the selection of the design (B). The size of DOE datasets (total = 74) (C) is also shown

4.2.2 | LRQ2: Considering a DOE + ML framework, what DOE strategies and DOE types are adopted?

In papers that directly describe the application of a DOE study or present a review of such applications, almost 88% of the occurrences use the strategy of nonsequential DOE experimentation (Figure 12(A)). RSM is the methodology applied in 52.9% of the cases, with CCDs, BBDs, and full factorial designs being the main DOE types adopted. RPD follows, as 22.9% of the applications use Taguchi's designs. Few papers also employ undesigned data, mainly sourced from literature or unstructured online databases, and space-filling designs. For clarity, designs employed in less than 2% of the papers are grouped into the "other" category.

The sequential strategy is adopted in around 12% of the cases. Among these, the AL framework is the main approach followed, often starting from an initial undesigned dataset that is then enlarged by suggestions of the learner. Rarely are designs from the RSM literature adopted while in other cases the initial DOE type is not specified (NAs). Other sequential approaches not within the AL framework are adopted.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

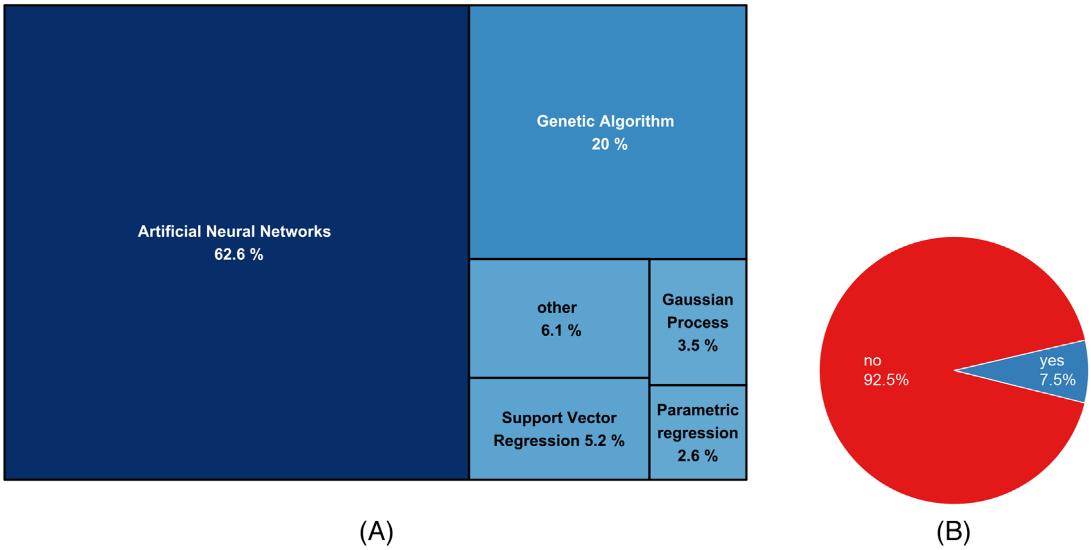


FIGURE 13 Distribution of studies (total = 80) by machine learning (ML) algorithm employed (A). The pie chart shows the proportion of studies, which carry out a comparison of different methods for data analysis (B)

This high fragmentation in type of DOE adopted is not justified by strong reasoning, since in only 16.9% of cases do the authors explain the rationale behind the choice of design or perform a comparison of concurrent designs before making the final choice (Figure 12(B)).

Another relevant result concerns the number of experimental configurations tested in the DOE study: about 50% of the time such a dataset has a size of 29 instances or smaller, and only about 25% of the experiments have 100 or more runs (Figure 12(C)).

4.2.3 | LRQ3: Considering a DOE + ML framework, what ML algorithms are adopted?

In the considered literature, ANNs are predominantly adopted as the ML method for the analysis of DOE data (Figure 13(A)). ANNs are mainly employed for regression tasks in which they are used for prediction purposes, and shallow networks with one or two hidden layers are preferred in the almost totality of the papers. In those studies in which an optimization of the factors is also enforced, ANNs are typically paired with a genetic algorithm (GA) that sequentially applies selection, crossover, and mutation for the generation of a configuration that can optimize the response(s).[39] Support vector regression (SVR) models are also used as an alternative to ANNs. Another algorithm that is often used in the AL and BO framework is the GP model. Parametric regression (PR), mainly based on quadratic regression with interactions, is also used principally to carry out a comparison of the ML methods with the traditional methodologies used for analysis of DOE data. However, this comparison between different algorithms or methodologies for data analysis is only seldom carried out, also within the ML domain: only 7.5% of papers test different strategies or algorithms for data analysis (Figure 13(B)). Note that if a prediction algorithm and an optimization algorithm (e.g., ANN and GA) are sequentially employed, this is not considered to be a comparison of different strategies since the purpose of the two methods is different. Methodologies used less than 2% of the time fall into the "other" category.

4.2.4 | LRQ4: Considering a DOE + ML framework, what data partitioning methods are adopted for ML model assessment?

Figure 14 shows the different choices made in relation to data partitioning and model assessment found in the literature. The most used method is the VS approach (57.1%), in which the DOE data are split into two subsets, one used for training and the other for validation of the developed models. Several authors decided not to divide the DOE data, preferring instead to carry out further experiments as confirmation of the goodness of predictions, without using these additional

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

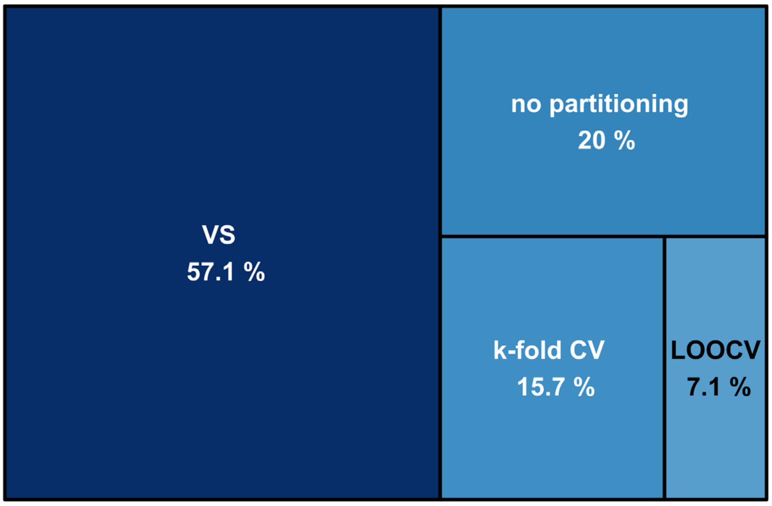


FIGURE 14 Distribution of papers (total = 70) per type of data partitioning method adopted

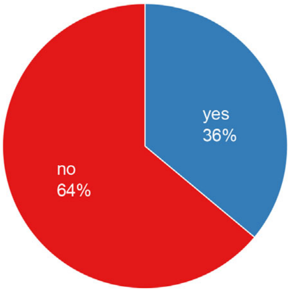


FIGURE 15 Distribution of papers (total = 86) that discuss the implications of the adoption of a DOE + ML framework

runs to build the model(s). Other resampling methods, such as k-fold CV (with k typically equal to 5 or 10) and LOOCV, are employed more seldomly, but still represent relevant options in the literature.

4.2.5 | LRQ5: What are the implications of adopting a DOE + ML framework?

The implications of the adoption of a DOE + ML framework are discussed in 31 of the 86 articles considered, that is a proportion of 36% (Figure 15). Table 4 displays the description of each implication that was discussed in the articles, also detailing if such implication has an immediate impact on the PI process, and if it is a distinguishing feature of DOE + ML in general or a hallmark of the AL framework. It is interesting to note that the top advantages of the DOE + ML framework in terms of percentage all qualify the AL framework, confirming that a sequential approach, preferably of the AL type, can have several advantages, mainly in the reduction of the experimental trials, which is often achieved by the suggestion of the optimal, most informative combinations of the factor levels. Other advantages that are typical of the AL framework include the possibility to explore experimental regions omitted by DOE alone, thus possibly leading to the discovery of regions with high and unexpected potential for optimization and also the opportunity to fully automate the experimental procedure if robotic equipment is available. Further analysis of the literature database revealed that while the advantages of a DOE + ML framework are evident in 25% of papers that adopt a nonsequential strategy, this is true for 90.9% of the works that adopt a sequential (including AL) approach, thus confirming the enormous potential of sequential experimentation, especially if driven by the ML algorithm.

Several of the implications of the DOE + ML framework reported in the literature have a direct and immediate impact on the PI process. For instance, several authors state that through such a framework, it was possible to achieve a minimization in the number of experiments, thus presenting clear advantages from the point of view of time and material costs. Another advantage includes the possibility to fully automate the experimentation process, making this phase more efficient and saving working hours. This framework also has a positive impact on the return on investments, since the final product quality appeared to be preferable to the one achieved using competitive approaches. An additional

1099163 8, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 The implications of the adoption of a DOE + ML framework in terms of percent value (%) and number of papers (**N**). Whether each implication has a direct impact on product innovation (PI) (**Impact on PI**) and if it is a distinguishing feature of the active learning (AL) framework (**Qualifying AL**) is also reported

Description	%	*N*	Impact on PI	Qualifying AL
In DOE + ML, a minimization of the number of experiments is achieved	35.5	11	x	x
In DOE + ML, ML can suggest optimal configurations for DOE trials	29	9		x
DOE + ML can lead to the exploration of experimental regions omitted by DOE alone	25.8	8		x
DOE + ML provides better final product quality	25.8	8	x	
In DOE + ML, DOE analysis can be used to explain the relationships, which drive the ML algorithms and select variables	25.8	8		
DOE + ML can lead to full automation in experimentation	22.6	7	x	x
DOE + ML provides a systematic, nonsubjective method for PI	16.1	5	x	
In DOE + ML, DOE can optimize ML hyperparameters	6.5	2		
In DOE + ML, DOE provides reasonable datasets in small data settings	6.5	2	x	
In DOE + ML, DOE controlled experiments provide a support for causal claims	3.2	1	x	

Abbreviations: AL, active learning; DOE, Design of Experiments; ML, machine learning; PI, product innovation.

organizational improvement consists in the definition of a systematic and objective method, which can drive PI by relying on robust conclusions based on data. The ability of DOE to generate highly informative datasets in small data settings and the possibility of supporting, at least to some extent, causal claims thanks to the controlled experiments are other implications of the framework, which have a collateral impact on the PI process.

5 | DISCUSSION

In this section, we discuss the results of the SLR and answer LRQ6 by highlighting the gaps and opportunities for further research.

The first thing to point out about the body of knowledge considered in the present study is the general lack of articles that comprehensively discuss the problem of the application of the DOE + ML framework in the context of PI. This is already highlighted by the distribution of articles per research methodology (Figure 8), which shows a predominance of case studies, the majority of which do not focus any attention on the methodology from a statistical perspective, but proceed with a straight application of DOE and ML. In our opinion, this is an indication that the research field is far from saturation, and many opportunities are available for new contributions. The relative immaturity of this branch of research is also noted when looking at the advantages and challenges of the application of ML over PR for the analysis of DOE data in PI as reported in Tables 2 and 3. The validity of some of the advantages is certainly questionable (e.g., the perception of a "plug-and-play" nature of ML), and the almost totality of these advantages are not extendable to a general setting, but true or false depending on the specific case. This is a common problem in ML, since the empirical nature of the methodology makes most of the conclusions problem specific. Similar considerations can be made on the challenges, given also how rare discussion of these aspects is (Figure 11). Nevertheless, some of the challenges identified are confirmed by the other results of the SLR and will be further discussed.

In the literature a nonsequential strategy for DOE is preferred (Figure 12(A)), although many advantages are shown to be linked with a sequential approach, especially of the AL type (Table 4). Elevated fragmentation is shown for the DOE type, although this does not appear to be justified since the rationale for the choice of one specific design is only rarely evident from reading the papers (Figure 12(B)). One possible explanation is that the majority of practitioners do not have enough experience in the statistical field, therefore may rely on the easiest or most widespread methods in their specific area. An example is the application of Taguchi's methodology, which is applied the most in the analyzed literature, mainly because of its popularity among engineers.[36]

It is the opinion of the authors that the present field of research would largely benefit from additional studies carried out on DOE because, as pointed out in Table 3, traditional DOE methodologies from RSM or RPD may not be suitable to take advantage of the high flexibility of advanced ML algorithms. These benefits would not be limited to the nonsequential DOE setting, since the sequential and AL frameworks also need an initial set of points to train the algorithms in the

1091638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

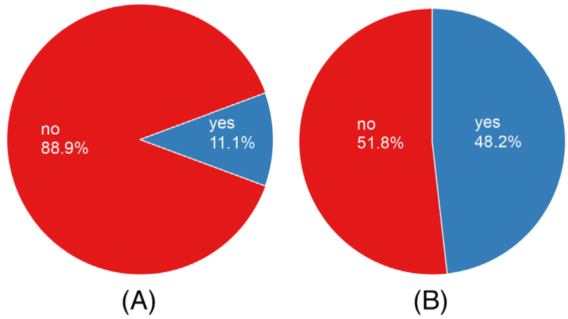


FIGURE 16 Distribution of papers that provide a quantification of predictions' uncertainty (A, total = 81) and that perform an investigation of causality (B, total = 83)

TABLE 5 The methodologies used for a quantification of prediction uncertainty in terms of percent value (%) and number of papers (**N**)

Description	%	*N*
Bayesian framework	44.4	4
Ensemble methods	22.2	2
Bootstrap methods	22.2	2
Using parametric statistics strategies	11.1	1

first iteration. Taking into consideration the scarcity of data, which characterizes the PI setting (Figure 12(C)), and the consequential need for highly informative datasets, further research on designs tailored to exploit the characteristics of ML models is recommended. One possible direction of research may concern space-filling designs, mainly because the homogeneous distribution of experimental configurations may favor flexible interpolating algorithms, as confirmed by the already widespread adoption of GP models for the analysis of space-filling experiments.[17]

Similar problems to those discussed for the DOE phase are present for the choice of ML algorithm, with an even smaller percentage of papers that seem to apply any statistical-based criteria, at least by a direct comparison of competitive models, for the selection of the final predictive algorithm (Figure 13(B)). Therefore, the preponderance of ANN models (Figure 13(A)) is not to be attributed to a clear algorithmic advantage, but rather to incidental factors probably connected with software availability or personal skills of the practitioners. This is also reinforced by the fact that 11 models, considering both prediction and optimization tasks, are applied in the examined literature, with five of them being applied only once. Acknowledging the wealth of ML algorithms that have been proposed over the years, this is a surprisingly low number. Furthermore, some very well-established algorithms such as Random Forests[31] are never employed.

Two additional matters that deserve discussion are the problems of uncertainty quantification and causal investigation that affect ML methodologies, as also pointed out in Table 3. The decisions taken in the early stages of a product development process may greatly impact the success of the final product itself, with many economic implications for an organization. Therefore, the robustness and validity of decisions taken at a PI level are critical, and the final objective of the DOE + ML framework is to support such decisions. Consequently, the outcomes of the models must be reliable and it is the opinion of the authors that a quantification of prediction uncertainty and at least an attempt to shed light on the black box of ML models is necessary.

Unfortunately, most ML models do not by nature provide a quantification of prediction uncertainty; this is a well-known problem, which worsens if we consider the small data setting of this study, in which less information is available to support point predictions. In the examined database, only 11.1% of the studies provide a quantification of uncertainty (Figure 16(A)), mainly by leveraging the Bayesian framework in BO applications or by employing Bootstrap or ensemble methods. In one case, a hybrid modeling approach of ML and ANOVA is employed that provides a quantification of uncertainty (Table 5).

The situation gets better if the problem of causal investigation is considered, with 48.2% of articles trying to investigate the rationale which drives the algorithms (Figure 16(B)). It should be stressed that causality is a problematic topic in ML because the models are by definition empirical and solely rely on correlative relationships. However, some tools are at the analyst's disposal to at least try to attenuate such problems, by questioning the rationale which drives the algorithms. Since a complete discussion on the possibility of performing causal claims in a ML setting is out of the scope of the present

1091638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 The methodologies used for causal investigation in terms of percent value (**%**) and number of papers (**N**)

Description	%	*N*
Using traditional DOE tools (also graphical ones)	42.5	17
Variable importance by interpretation of ML algorithm parameters (e.g., weights in ANN)	27.5	11
Using domain knowledge	20	8
Variable selection (both with DOE and ML tools)	12.5	5
Bayesian framework	10	4
Variable importance by sensitivity analysis	10	4

Abbreviations: ANN, artificial neural network; DOE, Design of Experiments; ML, machine learning.

study, we decided to identify those works that use an inquisitive attitude and try to shed light on the black box of ML algorithms, that is, explicitly or implicitly acknowledge the need for a causal investigation. In the literature this is done by assessing the importance of each input factor, by performing variable selection, and by using graphical tools to estimate, even qualitatively, the contribution of the factors on the model or the relationships among different factors (Table 6). Other attempts are made by corroborating the ML predictions with field knowledge or previous literature results or by using a Bayesian framework. The most adopted strategy for causal investigation uses tools of traditional DOE analysis (main effect plots, interaction plots, contour plots, etc.), often also applied to the PR model, to support the results of a ML algorithm. The limitations of such an approach are clear, however, it at least gives more insights than a blind black box application of the algorithm. One last thing to point out on this topic is that, given that DOE is based on randomized controlled experiments, some level of causality is intrinsic to the results since this is not the typical application of ML on purely observational data.

Taking the above discussion into account, we believe that a direction for future research concerns the study and development of proper ML algorithms to use in a DOE + ML framework. The analyst should strive to select or develop ML algorithms that can quantify uncertainty of predictions and can be, at least to a certain extent, interpretable. Some works in the literature already acknowledge the importance of these aspects,[6] and great potential is shown by the AL and BO approaches since both take the uncertainty of predictions into account when suggesting the next experimental configuration to test.[3,27] Furthermore, over the years, some model-agnostic tools have been proposed to make ML more interpretable,[40] but this is a growing field of research. A special mention should be given to GP models, since they are recognized as being fairly interpretable[41,42] in ML research.

Another problem to be addressed, and very specific to this context, is the scarcity of data. The usual application of ML is on big data, but this SLR highlighted the potential for use of ML algorithms on small data as well. The ability to quantify uncertainty, investigate causality and model small datasets are all factors that, together with the accuracy of predictions, should be considered when selecting which algorithm to use in a DOE + ML framework application in PI.

One final issue that deserves attention is the choice of data partitioning method for ML model tuning and assessment. As discussed in Section 2.3, some concerns have been raised about the appropriateness of splitting DOE data for ML model training. Nevertheless, a set of hold-out data is necessary to properly tune the hyperparameters of the ML model and provide a fair estimation of the generalization error of the algorithm.[22] Otherwise, there is a risk that the model overfits the training data, and we suspect that this may be the case for the 20% of papers that did not partition the training data, but rather validated the model using only a handful of configurations by means of additional experiments, often carried out in the vicinity of the identified optimal combination (Figure 14). The VS approach also presents drawbacks, mainly related to the variability embedded in the execution of the random split and the fact that not all the observations are used to train the models.[22] k-fold CV and LOOCV are, in general, preferable choices, with k-fold CV being the most balanced solution both in terms of computational effort and reduction in variance of the error estimation.[22] Nevertheless, we recognize the need for future research on the convenience and impact of partitioning DOE data and on the study of alternative methods that can keep the design structure while at the same time enabling a proper training and assessment of the ML algorithms.

6 | CONCLUSIONS

The present work constitutes the first SLR on the topics of DOE and ML in PI. Several valuable insights can be extracted from this study.

First, it has been shown that this is a growing field of research with mainly case study applications, but few papers propose new methodological advances. Advantages and challenges characterize the application of ML in comparison to traditional statistical tools, but often these tend to be valid for the specific application and not in general.

Furthermore, not enough attention is given to the links between the DOE and ML phases in the DOE + ML framework. In the literature, the choice of DOE strategy and type, and the adoption of ML models show that these two macro-phases are treated as if they were separate, but in reality, the quality of the final results depends on both. The initial design choice should be made considering the ML algorithm to be applied afterwards, and the ML algorithm should be selected considering the peculiarities of the data collected by the DOE. Further research is needed on ways to integrate these two phases, also in terms of data partitioning and model assessment methods, as several advantages of this framework on the PI process are found in terms of time, cost, return on investment, and robustness of final decisions. Particularly high potential is embedded in the sequential DOE strategy and AL framework.

Future research directions can be summarized with the following research questions and related sub-questions:

- What are the most appropriate DOE strategies (sequential vs. nonsequential) and DOE types for a DOE + ML framework?
- What are the most appropriate ML algorithms in a DOE + ML framework? How does one choose them?
  - How does one quantify uncertainty of predictions in a DOE + ML framework?
  - In a DOE + ML framework, how does one shed light on the black box of ML algorithms? How does one investigate causality?
- What are the most appropriate methods for data partitioning and ML model assessment in a DOE + ML framework?
  - What is the true impact of the application of traditional resampling methods such as k-fold CV on DOE data?

It is the authors' intention to build on the results of the present SLR and to focus further research on the identified problems and limitations in order to advance the current state of the art.

ACKNOWLEDGMENT

The authors gratefully thank Fondazione Cariparo for partially supporting this research.

Open Access Funding provided by Universita degli Studi di Padova within the CRUI-CARE Agreement.

CREDIT STATEMENT

Rosa Arboretti: methodology, validation, investigation, resources, writing – review & editing, supervision, project administration. **Riccardo Ceccato**: validation, visualization, data curation, writing – review & editing. **Luca Pegoraro**: Conceptualization, methodology, validation, investigation, formal analysis, data curation, writing – original draft, writing – review & editing. **Luigi Salmaso**: conceptualization, validation, investigation, resources, writing – review & editing, supervision, project administration.

ORCID

Luigi Salmaso https://orcid.org/0000-0001-6501-1585

REFERENCES

1. Mehdiyev N, Emrich A, Stahmer BP, Fettke P, Loos P. *iPRODICT – Intelligent Process Prediction based on Big Data Analytics*.BPM (Industry Track); 2017:13–24.
2. Yuan R, Tian Y, Xue D, Xue D, Zhou Y, Ding X, Sun J, Lookman T. Accelerated search for BaTiO3-based ceramics with large energy storage at low fields using machine learning and experimental design. *Adv Sci*. 2019;6:1901395. https://doi.org/10.1002/advs.201901395
3. Tian Y, Yuan R, Xue D, Zhou Y, Ding X, Sun J, Lookman T. Role of uncertainty estimation in accelerating materials development via active learning. *J Appl Phys.* 2020;128:014103.
4. Montgomery DC. *Design and Analysis of Experiments*. Hoboken, New Jersey, USA: John Wiley & Sons; 2017.
5. Freiesleben J, Keim J, Grutsch M. Machine learning and design of experiments: alternative approaches or complementary methodologies for quality improvement?. *Qual Reliab Eng Int*. 2020;36:1837–1848. https://doi.org/10.1002/qre.2579
6. Arboretti R, Ceccato R, Pegoraro L, Salmaso L, Housmekerides C, Spadoni L, Pierangelo E, Quaggia S, Tveit C, Vianello S. Machine learning and design of experiments with an application to product innovation in the chemical industry. *J Appl Statist*. 2021:1–26. https://doi.org/10.1080/02664763.2021.1907840

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

7. Fisher RA. Design of experiments. *Br Med J*. 1936;1:554–554.
8. Bisgaard S. Industrial use of statistically designed experiments: case study references and some historical anecdotes. *Qual Eng*. 1992;4:547–562. https://doi.org/10.1080/08982119208918936
9. Box GEP, Wilson KB. On the experimental attainment of optimum conditions. *J R Stat Soc Ser B Methodol*. 1951;13:1–45. http://www.jstor.org/stable/2983966
10. Myers RH, Montgomery DC, Vining GG, Borror CM, Kowalski SM. Response surface methodology: a retrospective and literature survey. *J Qual Technol*. 2004;36:53–77. https://doi.org/10.1080/00224065.2004.11980252
11. Taguchi G, Wu Y. *Introduction to Off-line Quality Control*. Nagaya, Japan: Central Japan Quality Control Assoc; 1979.
12. Taguchi G. *Introduction to Quality Engineering: Designing Quality into Products and Processes* [technical report]. 1986.
13. Taguchi G. *System of Experimental Design; Engineering Methods to Optimize Quality and Minimize Costs* [technical report]. 1987.
14. Robinson TJ, Borror CM, Myers RH. Robust parameter design: a review. *Qual Reliab Eng Int*. 2004;20:81–101.
15. Box G, Bisgaard S, Fung C. An explanation and critique of Taguchi's contributions to quality engineering. *Qual Reliab Eng Int*. 1988;4:123–131.
16. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*. 1979;21:239–245. http://www.jstor.org/stable/1268522
17. Joseph VR. Space-filling designs for computer experiments: a review. *Qual Eng*. 2016;28:28–35. https://doi.org/10.1080/08982112.2015.1100447
18. Box GEP, Liu PYT. Statistics as a catalyst to learning by scientific method part I – an example. *J Qual Technol*. 1999;31:1–15. https://doi.org/10.1080/00224065.1999.11979889
19. Box GE. Statistics as a catalyst to learning by scientific method part II – a discussion. *J Qual Technol*. 1999;31:16–29.
20. Vining G. Technical advice: Design of experiments, response surface methodology, and sequential experimentation. *Qual Eng*. 2011;23:217–220. https://doi.org/10.1080/15226514.2011.555280
21. Jensen WA. Open problems and issues in optimal design. *Qual Eng*. 2018;30:583–593.
22. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. vol. 112. New York, USA: Springer; 2013.
23. Polydoros AS, Nalpantidis L. Survey of model-based reinforcement learning: applications on robotics. *J Intell Rob Syst*. 2017;86:153–173.
24. Arboretti RG, Ceccato R, Pegoraro L, Salmaso L. A study of design of experiments and machine learning methods to improve fault detection algorithms. In: *DoE and Statistics – Methods and Applications*. SimStat, Salzburg: Springer; 2019. accepted.
25. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, USA: Springer Science & Business Media; 2009.
26. Olsson F. *A Literature Survey of Active Machine Learning in the Context of Natural Language Processing* [technical report]. Swedish Institute of Computer Science; 2009.
27. Greenhill S, Rana S, Gupta S, Vellanki P, Venkatesh S. Bayesian optimization for adaptive experimental design: a review. *IEEE Access*. 2020;8:13937–13948.
28. Lemkus T, Ramsey P, Gotwalt C, Weese M. Self-validated ensemble models for design of experiments. 2021. h ttp://arxiv.org/abs/2103.09303.
29. Xu L, Gotwalt C, Hong Y, King CB, Meeker WQ. Applications of the fractional-random-weight bootstrap. *Am Stat*. 2020;74:345–358. https://doi.org/10.1080/00031305.2020.1731599
30. Searle SR. An overview of variance component estimation. *Metrika*. 1995;42:215–230.
31. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
32. Xiao Y, Watson M. Guidance on conducting a systematic literature review. *J Plan Educ Res*. 2019;39:93–112. https://doi.org/10.1177/0739456X17723971
33. Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic mapping studies in software engineering. In: *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. 2008:1–10.
34. Berelson B. *Content Analysis in Communication Research*. New York, USA: Free Press; 1952.
35. Seuring S, Gold S. Conducting content-analysis based literature reviews in supply chain management. *Supply Chain Manag: Int J*. 2012;17:544–555.
36. Ilzarbe L, Álvarez MJ, Viles E, Tanco M. Practical applications of design of experiments in the field of engineering: a bibliographical review. *Qual Reliab Eng Int*. 2008;24:417–428.
37. Politis SN, Colombo P, Colombo G, Rekkas DM. Design of experiments (doe) in pharmaceutical development. *Drug Dev Ind Pharm*. 2017;43:889–901.
38. Salmaso L, Pegoraro L, Giancristofaro RA, Ceccato R, Bianchi A, Restello S, Scarabottolo D. Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study. *Commun Stat Simul Comput*. 2019:1–13. https://doi.org/10.1080/03610918.2019.1656740
39. Dao SD, Abhary K, Marian R. A bibliometric analysis of genetic algorithms throughout the history. *Comput Ind Eng*. 2017;110:395–403. https://doi.org/10.1016/j.cie.2017.06.009
40. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: *Proceedings of the ECML PKDD 2020 Workshops*; 2020:417–431.
41. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts, USA: The MIT Press; 2006.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

42. Liu H, Ong Y-S, Shen X, Cai J. When Gaussian process meets big data: a review of scalable GPs. *IEEE Trans Neural Netw Learn Syst*. 2020;31:4405–4423.
43. Abed AM, Elattar S, Gaafar TS, Alrowais FM. The neural network revamping the process's reliability in deep lean via internet of things. *Processes*. 2020;8:729. https://doi.org/10.3390/PR8060729
44. Abuhimd H. Chemical vapor deposition parameters dependent length control of hexagonal graphene. *Proc Inst Mech Eng Part E J Process Mech Eng*. 2017;231:1187–1196.
45. Aksu B, Paradkar A, Matas MD, Özer Ö, Güneri T, York P. A quality by design approach using artificial intelligence techniques to control the critical quality attributes of ramipril tablets manufactured by wet granulation. *Pharm Dev Technol*. 2013;18:236–245. https://doi.org/10.3109/10837450.2012.705294
46. Aksu B, Paradkar A, Matas MD, Özer Ö, Güneri T, York P. Quality by design approach: application of artificial intelligence techniques of tablets manufactured by direct compression. *AAPS PharmSciTech*. 2012;13:1138–1146. https://doi.org/10.1208/s12249-012-9836-x
47. Alaeddini A, Craft E, Meka R, Martinez S. Sequential Laplacian regularized v-optimal design of experiments for response surface modeling of expensive tests: an application in wind tunnel testing. *IISE Trans*. 2019;51:559–576. https://doi.org/10.1080/24725854.2018.1508928
48. Alam FM, McNaught KR, Ringrose TJ. An artificial neural network based metamodel for analysing a stochastic combat simulation. *Int J Enterp Inf Syst (IJEIS)*. 2006;2:38–57.
49. Altan M. Reducing shrinkage in injection moldings via the Taguchi, ANOVA and neural network methods. *Mater Des*. 2010;31:599–604. https://doi.org/10.1016/j.matdes.2009.06.049
50. Al-Zoubi N, Kachrimanis K, Younis K, Malamataris S. Optimization of extended-release hydrophilic matrix tablets by support vector regression. *Drug Dev Ind Pharm*. 2011;37:80–87. https://doi.org/10.3109/03639045.2010.492396
51. Amar Y, Schweidtmann AM, Deutsch P, Cao L, Lapkin A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chem Sci*. 2019;10:6697–6706 https://doi.org/10.1039/c9sc01844a.
52. Aroonsingkarat K, Hansupalak N. Prediction of styrene conversion of polystyrene/natural rubber graft copolymerization using reaction conditions: central composite design versus artificial neural networks. *J Appl Polym Sci*. 2013;128:2283–2290. https://doi.org/10.1002/app.38148
53. Azcarate SM, Pinto L, Goicoechea HC. Applications of mixture experiments for response surface methodology implementation in analytical methods development. *J Chemom*. 2020;34:e3246. https://doi.org/10.1002/cem.3246
54. Babutzka J, Bortz M, Dinges A, Foltin G, Hajnal D, Schultze H, Weiss H. Machine learning supporting experimental design for product development in the lab. *Chem-Ing-Tech*. 2019;91:277–284. https://doi.org/10.1002/cite.201800089
55. Balasundar I, Raghu T, Kashyap BP. Taguchi based optimisation of artificial neural network to establish a direct microstructure: mechanical property correlation in a near-α titanium alloy. *Trans Indian Inst Met*. 2016;69:1929–1941. https://doi.org/10.1007/s12666-016-0852-5
56. Barmpalexis P, Karagianni A, Karasavvaides G, Kachrimanis K. Comparison of multi-linear regression, particle swarm optimization artificial neural networks and genetic programming in the development of mini-tablets. *Int J Pharm*. 2018;551:166–176. https://doi.org/10.1016/j.ijpharm.2018.09.026
57. Bayer B, Striedner G, Duerkop M. Hybrid modeling and intensified doe: an approach to accelerate upstream process characterization. *Biotechnol J*. 2020;15:2000121. https://doi.org/10.1002/biot.202000121
58. Besseris GJ. Profiling multiple static and transient puff-pastry characteristics with a robust-and-intelligent processor. *J Food Eng*. 2015;164:40–54. https://doi.org/10.1016/j.jfoodeng.2015.04.011
59. Blanco JML, González-r PL, García CMA, Cózar-Bernal MJ, Suárez MC, Ortiz DC, Álvarez AMR, Rodríguez MLG. Artificial neural networks as alternative tool for minimizing error predictions in manufacturing ultradeformable nanoliposome formulations. *Drug Dev Ind Pharm*. 2018;44:135–143. https://doi.org/10.1080/03639045.2017.1386201
60. Bourquin J, Schmidli H, Hoogevest PV, Leuenberger H. Application of artificial neural networks (ANN) in the development of solid dosage forms. *Pharm Dev Technol*. 1997;2:111–121. https://doi.org/10.3109/10837459709022616
61. Bourquin J, Schmidli H, Hoogevest PV, Leuenberger H. Comparison of artificial neural networks (ANN) with classical modelling techniques using different experimental designs and data from a galenical study on a solid dosage form. *Eur J Pharm Sci*. 1998;6:287–300. https://doi.org/10.1016/S0928-0987(97)10025-2
62. Bourquin J, Schmidli H, Hoogevest PV, Leuenberger H. Advantages of artificial neural networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form. *Eur J Pharm Sci*. 1998;7:5–16. https://doi.org/10.1016/S0928-0987(97)10028-8
63. Bourquin J, Schmidli H, Hoogevest PV, Leuenberger H. Pitfalls of artificial neural networks (ANN) modelling technique for data sets containing outlier measurements using a study on mixture properties of a direct compressed dosage form. *Eur J Pharm Sci*. 1998;7:17–28. https://doi.org/10.1016/S0928-0987(97)10027-6
64. Broniatowski DA, Tucker C. Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity. *Syst Eng*. 2017;20:483–496. https://doi.org/10.1002/sys.21414
65. Chen J, Wong DSH, Jang SS, Yang SL. Product and process development using artificial neural-network model and information analysis. *AIChE J*. 1998;44:876–887. https://doi.org/10.1002/aic.690440413
66. Deng L, Feng B, Zhang Y. An optimization method for multi-objective and multi-factor designing of a ceramic slurry: combining orthogonal experimental design with artificial neural networks. *Ceram Int*. 2018;44:15918–15923. https://doi.org/10.1016/j.ceramint.2018.06.010
67. Dixit SR, Das SR, Dhupal D. Parametric optimization of Nd:YAG laser microgrooving on aluminum oxide using integrated RSM-ANN-GA approach. *J Ind Eng Int*. 2019;15:333–349. https://doi.org/10.1007/s40092-018-0295-1

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

68. Domagalski NR, Mack BC, Tabora JE. Analysis of design of experiments with dynamic responses. *Org Process Res Dev*. 2015;19:1667–1682. https://doi.org/10.1021/acs.oprd.5b00143
69. Gong X-D, Liao D-M, Chen T, Zhou J-X, Yin Y-J. Optimization of steel casting feeding system based on bp neural network and genetic algorithm. *China Foundry*. 2016;13:182–190.
70. Green DV, Pickett S, Luscombe C, Senger S, Marcus D, Meslamani J, Brett D, Powell A, Masson J. Bradshaw: a system for automated molecular design. *J Comput Aided Mol Des*. 2020;34:747–765. https://doi.org/10.1007/s10822-019-00234-8
71. Gubernatis JE, Lookman T. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Phys Rev Mater*. 2018;2:120301. https://doi.org/10.1103/PhysRevMaterials.2.120301
72. Guo Y, Xu J, Zhang Y, Xu H, Yuan Z, Li D. Medium optimization for ethanol production with clostridium autoethanogenum with carbon monoxide as sole carbon source. *Bioresour Technol*. 2010;101:8784–8789. https://doi.org/10.1016/j.biortech.2010.06.072
73. Ho WH, Tsai JT, Hsu GM, Chou JH. Process parameters optimization: a design study for tio2 thin film of vacuum sputtering process. *IEEE Trans Autom Sci Eng*. 2010;7:143–146. https://doi.org/10.1109/TASE.2009.2023673
74. Hsieh KL. Parameter optimization of a multi-response process for lead frame manufacturing by employing artificial neural networks. *Int J Adv Manuf Technol*. 2006;28:584–591. https://doi.org/10.1007/s00170-004-2383-1
75. Hsieh KL. Multi-response problem with adaptive weight consideration by using weight aggregator and ANNs. *Concurr Eng Rese Appl*. 2009;17:203–211. https://doi.org/10.1177/1063293X09344129
76. Hsu SF, Weng MH, Yang RY, Fang CH, Chou JH. An experimental design for processing parameter optimization for cathode arc plasma deposition of ZnO films. *IEEE Trans Autom Sci Eng*. 2016;13:1588–1593. https://doi.org/10.1109/TASE.2016.2572223
77. Hu Q, Liu Y, Zhang T, Geng S, Wang F. Modeling the corrosion behavior of Ni-Cr-Mo-V high strength steel in the simulated deep sea environments using design of experiment and artificial neural network. *J Mater Sci Technol*. 2019;35:168–175. https://doi.org/10.1016/j.jmst.2018.06.017
78. Hussain AS, Yu X, Johnson RD. Application of neural computing in pharmaceutical product development. *Pharm Res: Off J Am Assoc Pharm Sci*. 1991;8:1248–1252. https://doi.org/10.1023/A:1015843527138
79. Jiang ZM, Wang LJ, Gao Z, Zhuang B, Yin Q, Liu EH. Green and efficient extraction of different types of bioactive alkaloids using deep eutectic solvents. *Microchem J*. 2019;145:345–353. https://doi.org/10.1016/j.microc.2018.10.057
80. Kang J, Cao D. Optimization of the silica-gel adsorption technique for the extraction of phytosterol glycosides from soybean lecithin powder using response surface methodology and artificial neural network models. *J Food Sci*. 2020;85:1971–1982. https://doi.org/10.1111/1750-3841.15183
81. Keypour H, Noroozi M, Rashidi A, Beigzadeh R. A modeling study by response surface methodology (RSM) and artificial neural network (ANN) on nitrobenzene hydrogenation optimization using rh nanocatalyst. *Synth React Inorg Met-Org Chem*. 2015;45:1580–1590. https://doi.org/10.1080/15533174.2013.865222
82. Kibria G, Doloi B, Bhattacharyya B. Modelling and optimization of nd:yag laser micro-turning process during machining of aluminum oxide (Al2O3) ceramics using response surface methodology and artificial neural network. *Manuf Rev*. 2014;1:12. https://doi.org/10.1051/mfreview/2014011
83. Kohar CP, Zhumagulov A, Brahme A, Worswick MJ, Mishra RK, Inal K. Development of high crush efficient, extrudable aluminium front rails for vehicle lightweighting. *Int J Impact Eng*. 2016;95:17–34. https://doi.org/10.1016/j.ijimpeng.2016.04.004
84. Koletti AE, Tsarouchi E, Kapourani A, Kontogiannopoulos KN, Assimopoulou AN, Barmpalexis P. Gelatin nanoparticles for nsaid systemic administration: quality by design and artificial neural networks implementation. *Int J Pharm*. 2020;578:119118. https://doi.org/10.1016/j.ijpharm.2020.119118
85. Korany MA, Ragab MA, Youssef RM, Afify MA. Experimental design and machine learning strategies for parameters screening and optimization of Hantzsch condensation reaction for the assay of sodium alendronate in oral solution. *RSC Adv*. 2015;5:6385–6394. https://doi.org/10.1039/c4ra12750a
86. Kroll P, Hofer A, Ulonska S, Kager J, Herwig C. Model-based methods in the biopharmaceutical process lifecycle. *Pharm Res*. 2017;34:2596–2613. https://doi.org/10.1007/s11095-017-2308-y
87. Kundu P, Paul V, Kumar V, Mishra IM. Formulation development, modeling and optimization of emulsification process using evolving RSM coupled hybrid ANN-GA framework. *Chem Eng Res Des*. 2015;104:773–790. https://doi.org/10.1016/j.cherd.2015.10.025
88. Kuo CFJ, Su TL, Li YC. Construction and analysis in combining the Taguchi method and the back propagation neural network in the peek injection molding proces. *Polym Plast Technol Eng*. 2007;46:841–848. https://doi.org/10.1080/03602550701278103
89. Leuenberger H, Lanz M. Pharmaceutical powder technology – from art to science: the challenge of the FDA's process analytical technology initiative. *Adv Powder Technol*. 2005;16:3–25. https://doi.org/10.1163/1568552053166683
90. Li YW, Peng JH, Liang GA, Li W, Zhang SM. Optimization of processing parameters for microwave drying of selenium-rich slag using incremental improved back-propagation neural network and response surface methodology. *J Central South Univ Technol (Engl Ed)*. 2011;18:1441–1447. https://doi.org/10.1007/s11771-011-0859-3
91. Lin MC, Lin YH, Lin CC, Chen MS, Hung YC. An integrated neuro-genetic approach incorporating the Taguchi method for product design. *Adv Eng Inf*. 2015;29:47–58. https://doi.org/10.1016/j.aei.2014.09.002
92. Lin MC, Qiu GP, Zhou XH, Chen CN. Using Taguchi and neural network approaches in the optimum design of product development process. *Int J Comput Integr Manuf*. 2020;33:343–359. https://doi.org/10.1080/0951192X.2019.1639218
93. Ma L, Lv E, Du L, Lu J, Ding J. Statistical modeling/optimization and process intensification of microwave-assisted acidified oil esterification. *Energy Convers Manag*. 2016;122:411–418. https://doi.org/10.1016/j.enconman.2016.06.001

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

94. Manda A, Walker RB, Khamanga SM. An artificial neural network approach to predict the effects of formulation and process variables on prednisone release from a multipartite system. *Pharmaceutics*. 2019;11:109. https://doi.org/10.3390/pharmaceutics11030109
95. Miletić, T, Ibrić S, Durić Z. Combined application of experimental design and artificial neural networks in modeling and characterization of spray drying drug: cyclodextrin complexes. *Dry Technol*. 2014;32:167–179. https://doi.org/10.1080/07373937.2013.811593
96. Moghaddam MG, Ahmad FB, Basri M, Rahman MB. Application of the artificial neural network and response surface methodology models for predicting the enzymatic synthesis of betulinic acid ester: a comparative study. *Asian J Chem*. 2013;25:301–305. https://doi.org/10.14233/ajchem.2013.13010
97. Moghri M, Madic M, Omidi M, Farahnakian M. Surface roughness optimization of polyamide-6/nanoclay nanocomposites using artificial neural network: genetic algorithm approach. *Sci World J*. 2014;2014:485205. https://doi.org/10.1155/2014/485205
98. Momeni M, Guillot M. Development of friction stir welding technique at right angle (RAFSW) applied on butt joint of AA6061-T6 aluminum alloy. *Int J Adv Manuf Technol*. 2018;99:3077–3089. https://doi.org/10.1007/s00170-018-2672-8
99. Momeni M, Guillot M. Effect of tool design and process parameters on lap joints made by right angle friction stir welding (RAFSW). *J Manuf Mater Process*. 2019;3:66. https://doi.org/10.3390/jmmp3030066
100. Olawoye B, Gbadamosi SO, Otemuyiwa IO, Akanbi CT. Improving the resistant starch in succinate anhydride-modified cardaba banana starch: a chemometrics approach. *J Food Process Preserv*. 2020;44:e14686. https://doi.org/10.1111/jfpp.14686
101. Omidbakhsh N, Elkamel A, Duever TA, Reilly PM. Combining design of experiments techniques, connectionist models, and optimization for the efficient design of new product formulations. *Chem Prod Process Model*. 2010;5. https://doi.org/10.2202/1934-2659.1441
102. Ong-on I, Embley B, Chisti Y, Hansupalak N. Prediction of pore properties of hierarchical porous silica templated on natural rubber. *Microporous Mesoporous Mater*. 2016;233:1–9. https://doi.org/10.1016/j.micromeso.2016.06.035
103. Ozcelik B, Erzurumlu T. Comparison of the warpage optimization in the plastic injection molding using ANOVA, neural network model and genetic algorithm. *J Mater Process Technol*. 2006;171:437–445. https://doi.org/10.1016/j.jmatprotec.2005.04.120
104. Patel AD, Agrawal A, Dave RH. Development of polyvinylpyrrolidone-based spray-dried solid dispersions using response surface model and ensemble artificial neural network. *J Pharm Sci*. 2013;102:1847–1858. https://doi.org/10.1002/jps.23526
105. Plumb AP, Rowe RC, York P, Doherty C. The effect of experimental design on the modeling of a tablet coating formulation using artificial neural networks. *Eur J Pharm Sci*. 2002;16:281–288. https://doi.org/10.1016/S0928-0987(02)00112-4
106. Qin X, Wang B, Wang G, Li H, Jiang Y, Zhang X. Delamination analysis of the helical milling of carbon fiber-reinforced plastics by using the artificial neural network model. *J Mech Sci Technol*. 2014;28:713–719. https://doi.org/10.1007/s12206-013-1135-2
107. Rouco H, Diaz-Rodriguez P, Rama-Molinos S, Remuñán-López C, Landin M. Delimiting the knowledge space and the design space of nanostructured lipid carriers through artificial intelligence tools. *Int J Pharm*. 2018;553:522–530. https://doi.org/10.1016/j.ijpharm.2018.10.058
108. Sano S, Kadowaki T, Tsuda K, Kimura S. Application of Bayesian optimization for pharmaceutical product development. *J Pharm Innov*. 2020;15:333–343. https://doi.org/10.1007/s12247-019-09382-8
109. Senthilkumar N, Tamizharasan T. Flank wear and surface roughness prediction in hard turning via artificial neural network and multiple regressions. *Aust JMech Eng*. 2015;13:31–45. https://doi.org/10.7158/M13-045.2015.13.1
110. Shi X, Schillings P, Boyd D. Applying artificial neural networks and virtual experimental design to quality improvement of two industrial processes. *Int J Prod Res*. 2004;42:101–118. https://doi.org/10.1080/00207540310001602937
111. Shieh MD, Yeh YE. Developing a design support system for the exterior form of running shoes using partial least squares and neural networks. *Comput Ind Eng*. 2013;65:704–718. https://doi.org/10.1016/j.cie.2013.05.008
112. Shiva M, Atashi H, Hassanpourfard M. Studying the abrasion behavior of rubbery materials with combined design of experiment-artificial neural network. *Chin J Polym Sci (Engl Ed)*. 2012;30:520–529. https://doi.org/10.1007/s10118-012-1154-y
113. Silitonga AS, Shamsuddin AH, Mahlia TM, Milano J, Kusumo F, Siswantoro J, Dharma S, Sebayang AH, Masjuki HH, Ong HC. Biodiesel synthesis from *Ceiba pentandra* oil by microwave irradiation-assisted transesterification: ELM modeling and optimization. *Renew Energy*. 2020;146:1278–1291. https://doi.org/10.1016/j.renene.2019.07.065
114. Sovány T, Tislér Z, Kristó K, Kelemen A, Regdon G. Estimation of design space for an extrusion–spheronization process using response surface methodology and artificial neural network modelling. *Eur J Pharm Biopharm*. 2016;106:79–87. https://doi.org/10.1016/j.ejpb.2016.05.009
115. Suhaimi SN, Shamsuddin SM, Ahmad WA, Hasan S, Venil CK. Comparison of particle swarm optimization and response surface methodology in fermentation media optimization of flexirubin production. *Jurnal Teknologi*. 2019;81:53–60. https://doi.org/10.11113/jt.v81.10766
116. Susac F, Stan F. Experimental investigation, modeling and optimization of circularity, cylindricity and surface roughness in drilling of PMMA using ANN and ANOVA. *Materiale Plastice*. 2020;57:57–68. https://doi.org/10.37358/MP.20.1.5312
117. Takagaki K, Arai H, Takayama K. Creation of a tablet database containing several active ingredients and prediction of their pharmaceutical characteristics based on ensemble artificial neural networks. *J Pharm Sci.* 2010;99:4201–4214. https://doi.org/10.1002/jps.22135
118. Tang L, Yuan S, Tang Y, Qiu Z. Optimization of impulse water turbine based on GA-BP neural network arithmetic. *J Mech Sci Technol*. 2019;33:241–253. https://doi.org/10.1007/s12206-018-1224-3
119. Tsai JT, Lin CT, Chang CC, Chou JH. Optimized positional compensation parameters for exposure machine for flexible printed circuit board. *IEEE Trans Ind Inf*. 2015;11:1366–1377. https://doi.org/10.1109/TII.2015.2489578
120. Wang X, Sun B, Liu B, Fu Y, Zheng P. A novel method for multifactorial biochemical experiments design based on combinational design theory. *PLoS ONE*. 2017;12:e0186853. https://doi.org/10.1371/journal.pone.0186853

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

121. Wang H, Ye F, Chen L, Li E. Sheet metal forming optimization by using surrogate modeling techniques. *Chin J Mech Eng (Engl Ed)*. 2017;30:22–36. https://doi.org/10.3901/CJME.2016.1020.123
122. Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, Dai L, Lookman T, Su Y. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater*. 2019;170:109–117. https://doi.org/10.1016/j.actamat.2019.03.010
123. Witek-Krowiak A, Chojnacka K, Podstawczyk D, Dawiec A, Pokomeda K. Application of response surface methodology and artificial neural network methods in modelling and optimization of biosorption process. *Bioresour Technol*. 2014;160:150–160. https://doi.org/10.1016/j.biortech.2014.01.021
124. Xiong F, Wang D, Chen S, Gao Q, Tian S. Multi-objective lightweight and crashworthiness optimization for the side structure of an automobile body. *Struct Multidiscip Optim*. 2018;58:1823–1843. https://doi.org/10.1007/s00158-018-1986-3
125. Zhang Y, Xu J, Yuan Z, Xu H, Yu Q. Artificial neural network-genetic algorithm based optimization for the immobilization of cellulase on the smart polymer Eudragit L-100. *Bioresour Technol*. 2010;101:3153–3158. https://doi.org/10.1016/j.biortech.2009.12.080
126. Zhang J, Zhu H, Yang C, Li Y, Wei H. Multi-objective shape optimization of helico-axial multiphase pump impeller based on NSGA-II and ANN. *Energy Convers Manag*. 2011;52:538–546. https://doi.org/10.1016/j.enconman.2010.07.029

AUTHOR BIOGRAPHIES

Rosa Arboretti is an Associate Professor of Statistics at the Department of Civil, Environmental and Architectural Engineering at the University of Padova. Her research interests include Big Data Analytics and nonparametric statistics.

Riccardo Ceccato holds a Master's Degree in Statistical Sciences from the University of Padova. He is currently a PhD student at the University of Padova, Department of Management and Engineering. His research field is that of ML and nonparametric statistics to enhance PI.

Luca Pegoraro holds an MSc in Management Engineering from the University of Padova where he is currently a PhD student. His current research field is that of DOE and ML with applications to the chemical industry and product development and innovation.

Luigi Salmaso is Full Professor of Statistics and Deputy Chair of the Department of Management and Engineering at the University of Padova. His research interests include Big Data Analytics, DOE, and nonparametric statistics.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Arboretti R, Ceccato R, Pegoraro L, Salmaso L. Design of Experiments and machine learning for product innovation: A systematic literature review. *Qual Reliab Eng Int*. 2022;38:1131–1156. https://doi.org/10.1002/qre.3025

APPENDIX A

See Table A1, which shows the database of papers used for the SLR in alphabetical order.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - CochraneItalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A1 A summary of the database of papers organized by main categories of the content analysis (CA). All abbreviations are explained in the text

Reference	Implications of DOE + ML	DOE strategy	Type of DOE	DOE choice criterion	Data partitioning method	ML method	Advantages of ML	Challenges of ML	Uncertainty assessment	Causality investigation
43	yes	nonseq.	RSM	no	VS	ANN	no	no	no	no
44	no	nonseq.	RSM	no	no part.	ANN	no	no	no	yes
45	no	nonseq.	RSM	no	VS	ANN,GA	no	no	no	yes
46	no	nonseq.	NA	no	VS	ANN,GA	no	no	no	yes
47	yes	seq.	AL	yes	k-fold CV	NA	no	no	NA	NA
48	yes	nonseq.	space-filling	yes	VS	ANN	no	no	yes	no
49	yes	nonseq.	RPD	no	VS	ANN	no	no	no	no
50	no	nonseq.	RSM	no	no part.	SVR	no	no	no	no
51	yes	seq.	AL	yes	LOOCV	GP	no	yes	yes	yes
52	no	nonseq.	RSM	no	VS	ANN	yes	no	no	no
53	no	nonseq.	RSM	no	NA	ANN	no	no	NA	NA
54	yes	seq.	RSM, AL	yes	k-fold CV	GP,PR	no	no	yes	yes
55	yes	nonseq.	RPD	no	no part.	ANN	yes	no	NA	NA
56	no	nonseq.	RSM	no	VS	ANN,GA,PR,other	yes	no	no	yes
57	yes	seq.	iDOE	yes	LOOCV	ANN	no	no	no	yes
58	no	seq.	RPD	yes	no part.	ANN	no	no	yes	no
59	yes	nonseq.	RPD	no	VS	ANN	yes	no	no	yes
60	no	nonseq.	NA	no	no part.	ANN	yes	no	no	no
61	no	nonseq.	RSM	no	VS	ANN	yes	no	no	no
62	no	nonseq.	RSM	no	VS	ANN	yes	no	no	yes
63	no	nonseq.	RSM	no	VS	ANN	no	yes	no	yes
64	yes	NA	NA	NA	NA	NA	no	no	no	yes
65	yes	nonseq.	RSM	no	LOOCV	ANN	yes	no	no	no
66	no	nonseq.	RPD	no	VS	ANN	yes	no	no	no
67	no	nonseq.	RSM	no	k-fold CV	ANN,GA	yes	no	no	no
68	yes	seq.	RSM	no	k-fold CV	GA	yes	no	no	yes
5	yes	NA	NA	NA	NA	NA	yes	yes	NA	no
69	no	nonseq.	RPD	no	VS	ANN,GA	no	no	no	no
70	no	nonseq.	other	no	NA	other	yes	yes	no	no
71	yes	seq.	AL	yes	NA	GP	no	no	yes	yes
72	no	nonseq.	RSM	no	NA	ANN,GA	no	no	no	no

(Continues)

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A1 (Continued)

Reference	Implications of DOE + ML	DOE strategy	Type of DOE	DOE choice criterion	Data partitioning method	ML method	Advantages of ML	Challenges of ML	Uncertainty assessment	Causality investigation
73	no	nonseq.	RPD	no	no part.	ANN,GA	no	no	no	no
74	no	nonseq.	RPD	no	VS	ANN	yes	no	no	no
75	no	nonseq.	undesigned	no	VS	ANN	no	no	no	no
76	no	nonseq.	RPD	no	no part.	ANN,GA	yes	no	no	no
77	no	nonseq.	RSM	no	VS	ANN	yes	no	no	no
78	no	nonseq.	RSM	no	no part.	ANN	no	yes	no	no
79	yes	nonseq.	RSM	no	VS	ANN,GA	yes	no	no	no
80	no	nonseq.	RSM	no	VS	ANN,GA	yes	yes	no	yes
81	no	nonseq.	RSM	no	VS	ANN	no	no	no	no
82	no	nonseq.	RSM	no	no part.	ANN	yes	no	no	no
83	yes	nonseq.	NA	no	NA	ANN,other	no	no	no	yes
84	no	nonseq.	RSM	no	VS	ANN	yes	no	no	yes
85	no	nonseq.	RSM	no	k-fold CV	ANN,SVR	yes	no	no	yes
86	yes	NA	NA	NA	NA	ANN,SVR	yes	yes	no	yes
87	no	nonseq.	RSM	no	VS	ANN,GA	yes	no	no	yes
88	no	nonseq.	RPD	no	VS	ANN	no	no	no	no
89	yes	nonseq.	NA	no	NA	NA	no	no	no	no
90	no	nonseq.	RSM	no	VS	ANN	no	no	no	yes
91	yes	nonseq.	RPD	no	NA	ANN,GA	no	no	no	no
92	yes	nonseq.	RPD	no	NA	ANN,GA	yes	no	no	no
93	no	nonseq.	RSM	no	VS	ANN	yes	no	no	yes
94	yes	nonseq.	RSM	no	k-fold CV	ANN	no	no	no	yes
95	yes	nonseq.	RSM	no	VS	ANN	yes	no	no	yes
96	no	nonseq.	RSM	no	no part.	ANN	yes	no	no	no
97	no	nonseq.	RSM	no	VS	ANN,GA	no	no	no	yes
98	no	nonseq.	RPD	no	no part.	ANN	no	no	no	yes
99	no	nonseq.	RPD	no	VS	ANN	no	no	no	yes
100	no	nonseq.	RSM	no	VS	ANN,GA	yes	no	no	yes
101	yes	nonseq.	RSM	no	VS	ANN	yes	no	no	no
102	no	nonseq.	RSM	no	VS	ANN	yes	no	no	no
103	no	nonseq.	RPD	no	no part.	ANN,GA	no	no	no	no

(Continues)

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A1 (Continued)

Reference	Implications of DOE + ML	DOE strategy	Type of DOE	DOE choice criterion	Data partitioning method	ML method	Advantages of ML	Challenges of ML	Uncertainty assessment	Causality investigation
104	no	nonseq.	RPD	no	VS	ANN	yes	no	yes	yes
105	no	nonseq.	RSM	yes	VS	ANN	no	yes	no	no
106	no	nonseq.	RSM	no	VS	ANN	no	no	no	yes
107	no	nonseq.	RSM	no	VS	ANN,GA	no	no	no	yes
108	yes	seq.	RSM, AL	yes	LOOCV	ANN,GP	no	no	yes	yes
109	no	nonseq.	RPD	no	VS	ANN	yes	no	no	no
110	yes	nonseq.	RSM	no	VS	ANN	yes	no	no	yes
111	no	nonseq.	RPD	no	LOOCV	ANN,other	no	no	no	yes
112	yes	nonseq.	RSM	no	VS	ANN	no	no	no	yes
113	no	nonseq.	RSM	no	k-fold CV	ANN	no	no	no	yes
114	yes	nonseq.	RSM	yes	k-fold CV	ANN	yes	yes	no	no
115	no	nonseq.	RSM	no	VS	other	yes	no	no	no
116	no	nonseq.	RPD	no	VS	ANN	no	no	no	no
117	no	nonseq.	RSM	no	k-fold CV	ANN	no	no	no	yes
118	no	nonseq.	RPD	no	NA	ANN,GA	no	no	no	yes
119	yes	nonseq.	RSM	no	VS	ANN,GA	no	no	no	no
120	no	nonseq.	other	yes	NA	NA	no	no	no	no
121	yes	seq.	NA	yes	NA	ANN,SVR, GA,other	no	no	no	no
122	yes	seq.	AL	yes	k-fold CV	ANN,SVR, PR,other	no	no	yes	yes
123	no	nonseq.	NA	no	NA	NA	no	no	NA	yes
124	no	nonseq.	space-filling	no	no part.	ANN	no	no	no	no
2	yes	seq.	AL	yes	k-fold CV	SVR	no	no	yes	yes
125	no	nonseq.	RSM	no	NA	ANN,GA	yes	no	no	no
126	no	nonseq.	other	no	no part.	ANN,GA	no	no	no	no

Abbreviations: AL, active learning; ANN, artificial neural network; DOE, Design of Experiments; GA, genetic algorithm; GP, Gaussian process; k-fold CV, k-fold cross-validation; LOOCV, leave-one-out cross-validation; ML, machine learning; PR, parametric regression; RPD, robust parameter designs; RSM, response surface methodology; SVR, support vector regression; VS, validation set.

10991638, 2022, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/qre.3025 by Riccardo Ceccato - Cochraneltalia , Wiley Online Library on [17/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License