Received: 8 February 2019 | Accepted: 27 August 2019

DOI: 10.1002/qre.2579

DISCUSSION PAPER

WILEY

Machine learning and Design of Experiments: Alternative approaches or complementary methodologies for quality improvement?

Johannes Freiesleben | Jan Keim | Markus Grutsch

University of Applied Sciences St. Gallen, Institute for Quality Management and Business Administration, Rosenbergstrasse 59, St. Gallen, Switzerland

Correspondence
Johannes Freiesleben, Loorenstrasse 37
8053 Zürich.
Email: jfreiesleben@gmx.de

Funding information
FHS St. Gallen University of Applied Sciences

Abstract

Machine Learning (ML), or the ability of self-learning computer algorithms to autonomously structure and interpret data, is a methodological approach to solve complicated optimization problems based on abundant data. ML is recently gaining momentum as algorithmic applications, computing potency, and available data sets increased manifold over the past two decades, providing an information-rich environment in which human reasoning can partially be replaced by computer reasoning. In this paper, we want to assess the implications of ML for Design of Experiments (DoE), a statistical methodology widely used in Quality Management for quantifying effects and interactions of factors with influence on the production quality or the process yield. We specifically want to assess the future role and importance of DoE: Will it remain unaltered by ML, will it be made obsolete, or will it be reinforced? With this, we want to contribute to the discussion of the future use of traditional Quality Management methodologies in production, as our ML assessment can in principle be applied to other statistical methodologies as well. While we are convinced that ML will heavily impact the field of Quality Management and its predominant set of statistical methodologies, we find reason to expect that this impact will be a mutual one. As this is the first paper addressing the joint force potential of the two methodologies ML and DoE, we expect a range of follow-up papers being written on the subject and a spark in specialized applications addressing DoE's ML-enhanced vital functionality for process improvements.

KEYWORDS

Design of Experiments, in-process monitoring, machine learning, optimization, quality improvement, quality maintenance

1 | INTRODUCTION

Machine Learning (ML), or the ability of self-learning computer algorithms to autonomously structure and interpret data, may well be regarded as the next technological revolution given its potential impact on the way we conduct business[1] and organize other societal routines.[2] Even for those who are skeptical about the degree to which human

The authors would like to dedicate this work to the memory of Soren Bisgaard intelligence can be exceeded by artificial intelligence, the growing amount of scientific papers published on the topic might be convincing enough that a vast scope of new possibilities is awaiting us.[3] Already we find a multitude of applications of ML, ranging from spam filtering, credit card fraud detection, weather forecasts, and social media interactions to speech and image recognition. Other applications might follow soon, for instance fully autonomous cars or automated professional services such as market research and legal advice, replacing human input by the automated processing of large amounts of data. The spread of applications into other realms seems just to be a question of time when economic advantages outweigh implementation costs.

The quest for data and its analysis is constitutional to the scientific process and to all applied sciences. In the quality management discipline, researchers have long emphasized the overarching importance of data for understanding and managing production processes.[4,5] Some—including the authors of this paper—would even define the core of quality management as information processing since profound understanding of the production process is the basis for problem detection and in-depth knowledge of the surrounding technological landscape the basis for finding the adequate improvement option. With this resemblance to data processing, it seems likely that the quality field will also be impacted by applications of ML algorithms.

The investigative part of quality management, specifically in preproduction process design and in-process problem identification, is an example of massive data utilization and relies on science-based methodologies.[4] In this paper, we want to examine one specific methodology widely used in the investigative part, which is Design of Experiments (DoE), and assess whether and how ML will impact its future use. DoE is a statistical methodology used for quantifying main effects and interactions of factors with influence on the production quality or the process yield. We specifically want to assess the future role and importance of DoE: Will it remain unaltered by ML, will it be made obsolete, or will it be reinforced? With this, we also want to contribute to the discussion of the future use of traditional quality management methodologies in production, as our assessment can in principle be applied to other statistical methodologies as well.

The paper is organized as follows. Section 2 will provide a short literature overview on the two methodologies and the use of big data in DoE, outlining our field of interest. Section 3 will then compare DoE and ML from a functional point of view, highlighting main similarities and differences. Section 4 will address our main research question about the future importance of DoE. In Section 5, we will conclude our thoughts with an outlook and discussion points.

2 | LITERATURE OVERVIEW

Ronald A. Fisher developed a statistical methodology for planning experiments in both efficient and effective manner in his famous Rothamstead Research in the 1920s, subsequently refining the "Design of Experiments" methodology and publishing the results in a book[6] that sparked the rapid development of the field. Especially the factorial DoE, with its promising economic aspects of reduced experimental setup and running time due to measure point selection and statistical interpretation of the results, received widespread attention in industry and helped replace the ubiquitous and costly "one factor at a time" method of searching for the factors with the greatest impact on the objective function.[7] Main contributors to the refinement of the DoE methodology include statisticians such as George E. P. Box, Søren Bisgaard, William G. Hunter, and Genichi Taguchi.[8] Today, DoE counts as one of the most widely applied statistical methodologies to help industrial engineers in optimizing process performance by way of cleverly designed experimental setups.

The idea of learning machines was technically introduced by Arthur Samuel in 1959, referring to "[...] the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning" (Samuel[9], p 71). Shortly before this, Alan Turing devised a test to determine if a computer has real intelligence.[10] The so-called "Turing-Test" is still the ultimate test of Artificial Intelligence: To pass it, the computer must be able to fool a human into believing it is also human. Google's AI Assistant Duplex recently passed the Turing test, making reservations in restaurants while perfectly mimicking a human voice.[11] At the beginning a rather theoretical field, Artificial Intelligence and its subfield ML evolved rapidly in the past decades because of the great prospects of data analysis aided by computers. The first wave of ML algorithms centered on methods such as symbolic representation and neural networks that basically are applications of generalized linear modeling widely used in statistics.[12,13] In the 1980s, statistics-based methods came out of favor and expert systems emulating human decision making gained popularity, while research into applicable statistics-based methods continued outside the field, eg, in the form of pattern recognition.[14] In the 1990s, ML witnessed a refocus on statistics and an ensuing growth as an own recognized scientific field. This was mainly due to the improved performance of its statistical approaches, the increasing availability of larger data sets, and the decreasing cost of computing power.[15] Combining computer science and statistics, it is one of today's most rapidly growing fields of technology with commercial use prospects, leading to better decision-making in health care, manufacturing, education, financial modeling, and many other disciplines.[15] The importance of big data and ML also for management processes has been highlighted by various authors.[16]

Given the obvious link between DoE and ML—both are concerned with data analysis and the application of statistical methods—there are surprisingly few papers on the intersection of the two fields. Addressing the potential usefulness of a combination of the two concepts, Staelin[17] applies DoE principles to identify optimal or near optimal initial parameter settings in an example of support vector machines, a common class of ML algorithms. In a similar quest, Packianather et al[18] optimize the design parameters in an example of neural networks, another class of ML algorithms, by Taguchi's offline DoE approach, finding that this vastly outperforms the classical "trial and error" approach in setting parameter levels. Applying the same optimization logic to other examples of neural networks, Sukthomya and Tannock[19], Ortiz-Rodriguez et al,[20] and Belastrassi et al[21] all reach the conclusion that the DoE approach allows for gaining a profound understanding of the effects of parameters on the network performance and hence enables better parameter adjustments. Whereas such optimizations of ML algorithms by DoE are evident, an area where DoE design can benefit from ML is Latin Hypercube design, a special experimental design used in computer simulation.[22,23] Here, ML algorithms help identify the optimal experimental points such that significant results can be obtained from the experiment. Comparisons of the two methodologies as alternative approaches to optimization are for instance made in Korany et al[24] to determine significant factors for optimization of a condensation reaction.

DoE-related fields, such as tolerance design, have recently been shown to benefit from ML approaches,[25] but the DoE methodology as such has not been assessed in this regard. The above-cited papers compare or combine the two concepts in specific areas of interest or for specific problem investigations, but a paper producing a generalizable assessment of how the two methodologies can be applied jointly to increase investigative potency in production, or discussing the implications of ML for the future use of DoE, has not been put forward so far. This paper aims at filling the gap.

3 | ML AND DOE CHARACTERISTICS

For our aim of comparing the potency of either methodology and detect its potential for combination, a starting point is to deduct similarities and differences from a comparison of its basic functionality. Being the more specific methodology, the aim of DoE is to identify the set of process factors being most relevant to the process performance and to determine the optimal factor levels to maximize performance, thereby providing a powerful and cost-effective method for understanding and optimizing manufacturing processes.[26] The key idea thereby is to select factors to be included in the experiment, determine two or more levels on which each factor is varied, and conduct experimental runs with different combinations of factor levels. The main factor effects on the process performance and the potential factor interactions, ie, relationships between two or more factors that lead to a jointly nonadditive effect on performance, can with relative ease be calculated from differences in mean performance values at the examined factor levels.[27] For this, an experiment must be conducted with measurements of factors on at least two factor levels, producing a total of l^k measurements for l levels and k factors in a full factorial design. Saving on measurements and hence experiment expenditures, a fractional factorial design only requires a total of l^{k-p} measurements (with parameter p determining the fraction 0.5^p of measurements as compared to a full factorial), but it may require additional statistical analysis to obtain significant results, with increasing analytical effort for increasing p. Figure 1 shows a typical full factorial design for three factors A, B, and C varied on both high (+) and low (−) level, yielding a total of eight combinations (ie, experimental runs) to be included in the experiment.

To apply DoE effectively, the process engineers need to possess a high level of process knowledge, mainly to subject the right factors to the experiment and to set the experimental factor levels such that meaningful results can be obtained. Subjecting the right factors to the experiment is usually easier in preproduction process design than in in-process problem identification as the problem root causes are often not known a priori; in fact, DoE can be seen as a methodology to identify the factors contributing to the problem. In the process design phase, the main factors contributing to the performance of, eg, a machine are usually given by the machine's technical specifications. If the process engineers had perfect knowledge about the production process, finding the right factors would be no challenge even in in-process problem identification and the above distinction would be irrelevant. In practice, however, perfect process knowledge seldom exists. It can often only be approximated over time by continuous quality improvements and the automated monitoring of identified quality-relevant process factors, thus indicating any abnormal factor variability in time before defects are produced.

Consider, as an illustrative example, the investigation of a quality problem that manifests itself in the cracked coating of a metal frame used in sporting bike production, causing significant inspection and rework effort. Given imperfect

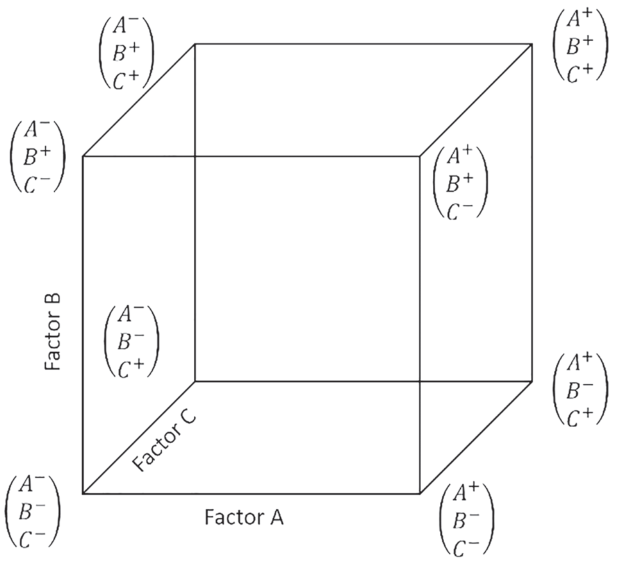

$\begin{pmatrix}A^-\\B^+\\C^+\end{pmatrix}$
$\begin{pmatrix}A^+\\B^+\\C^+\end{pmatrix}$
$\begin{pmatrix}A^-\\B^+\\C^-\end{pmatrix}$
$\begin{pmatrix}A^+\\B^+\\C^-\end{pmatrix}$
Factor B
$\begin{pmatrix}A^-\\B^-\\C^+\end{pmatrix}$
$\begin{pmatrix}A^+\\B^-\\C^+\end{pmatrix}$
Factor C
$\begin{pmatrix}A^-\\B^-\\C^-\end{pmatrix}$
Factor A
$\begin{pmatrix}A^+\\B^-\\C^-\end{pmatrix}$

FIGURE 1 Design of Experiments (DoE) space for a 2^3 full factorial experiment with eight experimental runs, superscript (+) indicating a high and (−) a low factor level

process knowledge, a variety of causes need to be considered as potentially contributing to the problem: a rough metal frame surface, the chemical lacquer composition, the appliance of the lacquer to the frame, the oven coating process, the bike assembly process, and so on. A process engineer would now use DoE to examine each potential cause and to determine which of the factors in each potential cause might contribute to the problem. Let us say the engineer examines the electric convection oven and registers abnormally low temperatures in certain spots, leading to the hypothesis that this is what caused the lacquer to be unevenly coated and made prone to cracking. He then identifies the factors potentially contributing to this uneven temperature distribution: the conduction current, the batch load, the insulation layer, and the recirculation air swirl. He then designs an experiment in which he varies these four factors on two factor levels, with the conduction current levels set at 1.8 and 2.2 A, the batch load at densely packed and lightly packed levels, the insulation layer using type A and B insulation, and the recirculation air swirl at medium and high levels. A full factorial experiment with 16 experimental runs is concluded with the result that two factors—the batch load and the recirculation air swirl—have a significant effect on the coating quality for the given oven design, and that the interaction between these two factors is high, meaning they always must be tuned to each other. Although experiments can encompass a multitude of factors and therefore reach high degrees of complexity, DoE is a decidedly human-centered methodology: The process engineer applies his process knowledge or intuition to select the factors to be examined, sets the factor levels for the experiments, and conducts the experiment.

Opposed to that, ML is decidedly nonhuman. Its aim is the automated detection of patterns in data, forming a model of the data based on both input and output data. This is in sharp contrast to classical programming, where the computer is given input data and a specifically programmed algorithm to perform a certain transformation with the data and return the desired output. ML detects the model by which input is transformed into output and hence produces the logic that is sought for. For this task, it uses an ML algorithm that aims at detecting patterns in data. As a broad classification, two main algorithmic approaches can be distinguished: supervised and unsupervised learning. Supervised learning refers to the fact that the algorithm is trained on a set of training examples before it develops its own model of the data; in other words, the sequence of input data x_1,x_2,...,x_n is accompanied by desired training outputs y_1,y_2,...,y_m, and the objective of the algorithm is to learn to produce the correct output given a new input x_{n+1}.[28] With this initial training input, it has an element of human bias as engineers or IT specialists select the training data, but as ML algorithms produce better and faster results when trained, this slight bias is generally accepted.[29] Training output data are categorized and labeled, such as photos with facial feature categorizations and additional information such as tagged names. As an example, having learned from a small variety of a person's photos and a vast variety of other photos what makes a human face unique, the face recognition algorithm of the social media platform Facebook can recognize a person's face in any incoming new photo, even when untagged.

Unsupervised learning means the algorithm performs pattern detection without prior training on labeled output, making sense of vast amounts of unlabeled and uncategorized data x_1,x_2,...,x_n. Here, human bias is prevented, and there is no distinction between training data, test data, and augmented data. Although the absence of any predefined structure may be interpreted as a disadvantage of the unsupervised algorithm, its functionality lies in building representations of the input data that can be utilized in form of decision support and predictions and in general in finding patterns beyond what a human would classify as white noise.[28] Unsupervised learning is for instance used in biometric pattern recognition, with the objective of identifying a human individual by analyzing physical (facial features, hand geometry, fingerprints, iris, and so on) and behavioral (signature, voice, keystroke, and so on) characteristics of the individual.[30] A good example for the applicability of unsupervised learning is speech recognition. Recorded speech is characterized by a plethora of features (such as phonetic variability, acoustic environment, technical transmission, physical differences in the voice production organs, manners of speaking, language, and dialect) and hence the so-called "curse of dimensionality": The number of required training samples for reliable estimations grows exponentially with the number of features.[31] One of the key results of applying unsupervised learning algorithms is precisely a dimensionality reduction, as they often use Principal Component Analysis to structure the raw data.[30]

Unsupervised learning can be thought of as learning a probabilistic model of the input data: It estimates a model that represents the probability distribution for a new input x_t given previous inputs $x_1,\ldots,x_{t-1}$, ie, it models $P(x_t | x_1,\ldots, x_{t-1})$.[28] Its algorithms therefore find widespread application, eg, in the fields of weather or stock price forecast. Furthermore, unsupervised learning offers great chances to finding solutions to problems, where we as humans cannot foresee all important categorizations or, what is equally important, are simply not able to express them; many tasks are performed subconsciously, without us being aware of the exact reasoning behind our performance.[32]

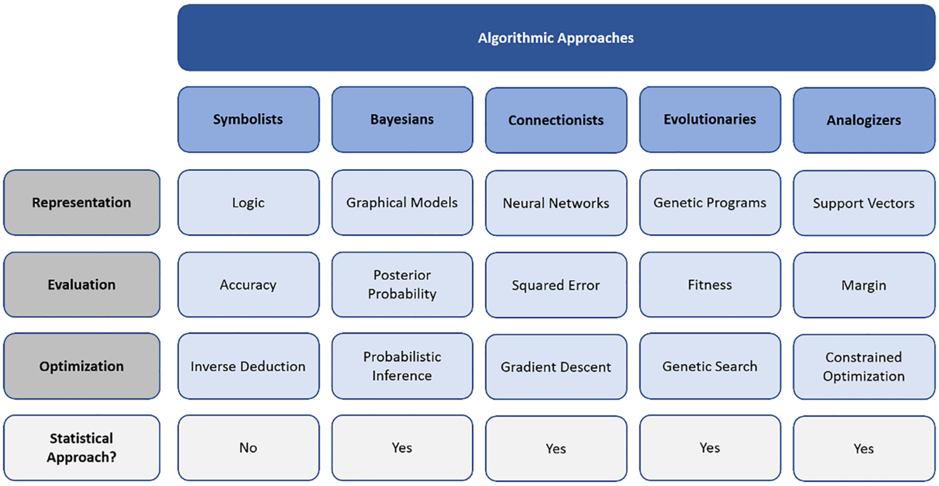
The diverse algorithmic landscape of both supervised and unsupervised learning encompasses different fundamental techniques, most of which are rooted in statistics and are well-studied and understood. Hence, applying ML is no magic but requires understanding the potency of each technique and selecting the adequate one for the existing problem. Figure 2 provides a short overview of the main clusters of techniques, highlighting their representation (formal language in which the algorithm presents its models), the evaluation (the scoring function saying how good the model is), the optimization (the algorithm searching for the highest scoring model), and its base in statistics.

Looking at the applicability of ML, its algorithms are of high utility whenever we deal with

a. unknown patterns between input and output data, for instance white noise;
b. inability to model the pattern mathematically by classical programming, for instance highly complex or nonlinear problems or activities with large parts of inherent knowledge such as driving or recognizing faces;
c. large amounts of unlabeled or uncategorized data, for instance relating to highly complex phenomena such as weather prediction or turning medical archives into medical knowledge.[29]

4 | HOW BENEFITS CAN BE CREATED BY COMBINING ML AND DOE

As stated above, the obvious link between ML and DoE is the common concern with data analysis and—with the exception of the symbolist cluster of ML algorithms based on inverse deduction—the application of statistics. A superficial assessment might conclude that this is all common ground there is; after all, the two methodologies have different



	Algorithmic Approaches				
	Symbolists	Bayesians	Connectionists	Evolutionaries	Analogizers
Representation	Logic	Graphical Models	Neural Networks	Genetic Programs	Support Vectors
Evaluation	Accuracy	Posterior Probability	Squared Error	Fitness	Margin
Optimization	Inverse Deduction	Probabilistic Inference	Gradient Descent	Genetic Search	Constrained Optimization
Statistical Approach?	No	Yes	Yes	Yes	Yes

FIGURE 2 Algorithmic clusters and their classification (based on Domingos[32]) [Colour figure can be viewed at wileyonlinelibrary.com]

aims, DoE being focused on the identification of quality-relevant process factors and ML being focused on the identification of pattern in large amounts of sometimes unstructured data. In this regard, DoE is the more focused methodology and ML the broader. Furthermore, DoE is human-centered while dealing with "small" data, whereas ML is machine-centered while dealing with "big" data. As has been pointed out in the literature section, DoE techniques have indeed been used to optimize initial parameter settings in ML algorithms, but very little has DoE benefitted from the application of ML.

Another assessment might then discover some high-level similarities between the methodologies. For instance, on an abstract level, DoE aims at dimensionality reduction as it tries to detect the factors most relevant to the objective function in the multitude of potential factors, similar to how unsupervised ML algorithms reduce dimensionality in detecting the data features most relevant to the model out of all possible data features. Concerning the knowledge base of the methodologies, DoE is more effective the more the process engineers have acquired prior in-depth knowledge about the process, similar to what in a supervised ML approach would be characterized as training input. These similarities motivate to look deeper into the joint potential of the two methodologies, especially the potential for DoE to be made more efficient by the application ML.

For this aim, we depicted in Figure 3 the schematized core processes for both methodologies, highlighting the human-based and the software-based parts. Essentially, we need to assess which part of the DoE core process can potentially be supported by ML or if the human-based parts of DoE might even be completely replaced by ML applications.

The assessment thus needs to address ML applicability in the five core parts of the DoE process:

1. the knowledge-based support part leading to better factor selection,
2. the knowledge-based support part leading to better level adjustment,
3. the sequential part 1 of factor selection,
4. the sequential part 2 of factor variation (the experiment) and
5. the sequential part 3 of effect and interaction calculation.

As it is often already software-based and concerns predefined calculation, the fifth core part of effect and interaction calculation can be dropped from our assessment. To discuss all other core parts, it is important to stress that we presume a quality control system based on in-process monitoring as the basis for data generation. In-process monitoring is sensor-based and keeps track of process parameters with influence on the production quality during production itself. Usually, this is done by checking factor data for variability, detecting data abnormality by exceedance of predefined specification limits.[5] Without an extensive monitoring system, a company cannot claim to be "in control" of its process as it lacks data about the current process state; other quality control technologies, such as inspection, cannot provide this information in the same accuracy or timely manner. In most, if not all industrial production processes, sensor networks allowing for in-process monitoring can be integrated into the process flow as industrial production logic is based

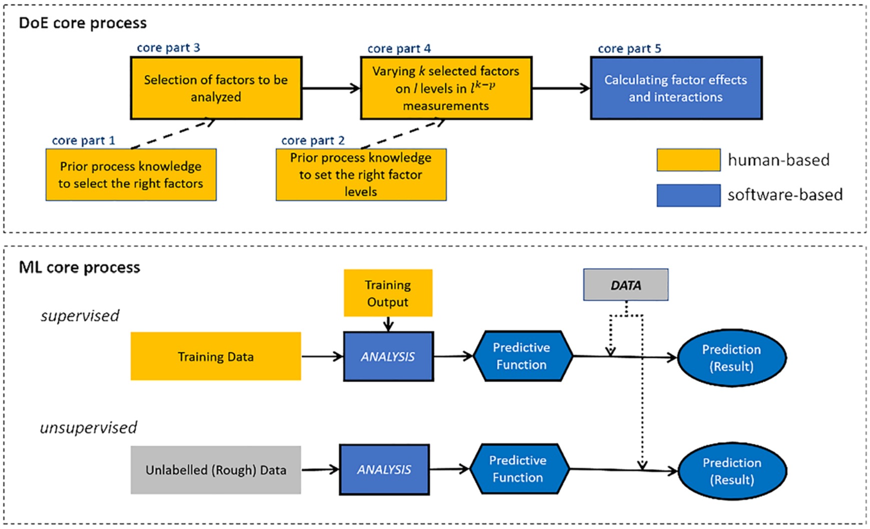


FIGURE 3 Schematic core processes of Design of Experiments (DoE) and machine learning (ML) with human-based and software-based parts [Colour figure can be viewed at wileyonlinelibrary.com]

on repeat processes and standardized automated procedures. This provides a data-rich environment concerning the state of the process and hence the basis for the application of ML.

How could ML be applicable to the core parts of DoE? Key to answering this question is to recognize that both methodologies need varying factor data that allow them to yield analytical results—DoE by setting the factor levels for the experiment and ML by observing the fluctuating data that are already available. In other words, we can think of two different kinds of data variability: the normal, natural variability of a factor as observed in the sensor data, and the human-induced "variability" of different factor levels set for the experiment—in fact, DoE can be viewed as an extreme temporary manifestation of variability. The first kind of variability is easily accessed and analyzed by an ML system through the monitoring system, whereas the second kind requires either human intervention or automated level settings. Figure 4 depicts these two kinds of variability schematically. As can be seen in the left part of the figure, where the observed variability of factors A and B are depicted, this natural variability might produce all possible combinations of high/low levels of B with high/low levels of A at certain times—for instance, the highlighted combinations of $\left(B_n^+, A_n^-\right)$ and $\left(B_n^-, A_n^+\right)$, with subscript n signifying the natural generation of the data. On the right side of the figure, this is contrasted by the set factor levels in an experiment, again highlighting the combination of $\left(B_d^+, A_d^-\right)$ and $\left(B_d^-, A_d^+\right)$, with subscript d signifying the data generation by design. As can be readily seen, the only difference is the order of magnitude, giving rise to the question which order of magnitude is needed to produce significant findings.

Applying this question to the core parts of DoE helps us assessing the applicability of ML in DoE. Core part 1, the prior knowledge about the production process leading to better factor selection, is the amount of accumulated process knowledge, which helps process engineers identify the right factors to submit to the experiment. However, this accumulated process information might be erroneous or insufficient, reflecting limited process understanding or human errors leading to wrong conclusions or correlation assumptions. Hence, more valid data about the process might aid the engineers in their decision making. Targeted to this, ML can increase both the quantity and the quality of available process information. After gaining access to monitoring data, it can automatically search for patterns a human or a programmed software might not have recognized, thereby reducing the "white noise" or unknown part of the production function. By analyzing past data and simultaneously observing factor measurements across the whole production process, the ML algorithm might identify important factors and correlations between factors previously unknown. As this is based on the observed data variability, it presupposes an extensive monitoring system keeping track of the majority of process factors, such that the first requirement would be to subject as many factors as possible to an ML-controlled sensor network. With such a system in place, the process engineers could complement their intuition and process knowledge with ML-derived suggestions regarding factor selection.

In an ideal scenario, the natural factor variability would already be analytically so significant that the ML algorithm could derive all factor effects and interactions from the sensor data, thereby potentially replacing the entire DoE system in its functionality as quality problem analyzer. In the realistic case, ML will find patterns in the data that aid in the selection of factors to be further analyzed, yielding indications for significant effects based on extrapolations from microeffects already detectable in the natural process data. As can be seen in Figure 4, the factor variability data might chronologically coincide to obtain the necessary combinations of high and low factor level measurements, thereby mimicking a DoE design without any actual design and experimental effort. The longer factor variabilities are observed, the higher the probability that all necessary high-low combinations are included in the observed data.

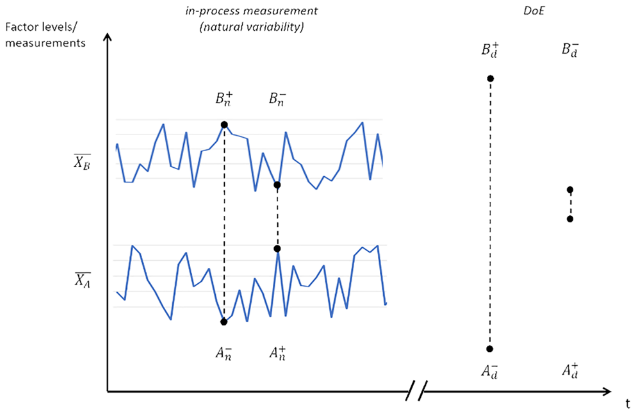


FIGURE 4 Schematic depiction of machine learning (ML) measurements based on natural factor variability (subscript n) and Design of Experiments (DoE) factor levels (subscript d) for two factors A and B [Colour figure can be viewed at wileyonlinelibrary.com]

The same logic also applies to core part 2, the knowledge-based support part leading to better level adjustment. Based on the monitored factor data, the ML algorithm might be able to indicate the suggested dimension of the experimental level settings for further factor analysis, based on extrapolations of the already noted microeffect at the observed variability levels. Setting the right factor levels in the experiment is a key task in DoE and a difficult one as such.[8] Again, in-depth process knowledge or intuition would be replaced by suggestions derived from the continuous flow of process data. Level selection is understandably more important for factors that are continuous, as discrete factors present fewer such opportunities.

Core part 3, the factor selection, is largely a function of core part 1, the prior process knowledge. However, ML can contribute to factor selection in an even stronger way: by making selection as such unnecessary. DoE is constrained by its core part 5, the results calculation, as l^{k-p} measurements can soon overwhelm the methodology in terms of number of experimental runs—but this transgression from small data to big data for increasingly large l and k actually aides the application of ML, which performs better the larger the database. Hence, in an ideal scenario, there would be no factor selection but an inclusion of *all* factors and *all* potentially interesting factor levels into the analysis. This would presuppose the normal process data are already analytically significant enough—if not, the need for factor selection persists, and real experiments need to be conducted supplementary to what the ML algorithm could identify.

Core part 4 then concerns the experimental measurements. Although we deal with physical experimental runs, ML might still be able to improve this core part. A key problem of DoE is the exponential growth in number of experimental runs with the number of factors included and hence gives rise to fractional factorial designs with l^{k-p} measurements reducing the experimental effort by multiple $(1 - 0.5^p)$. As this means leaving out many experiments, which might turn out to be important, the experimental points must be well chosen not to miss out on important factors and still allow for interpretable results; sometimes this necessitates conducting follow-up experiments to resolve ambiguities arising from blurring of factor effects and interactions.[33] Hence, finding the optimal experimental points can be defined as a key challenge for fractional factorial designs and is, like the selection of the factor levels in the experiment, dependent on human skill and ingenuity. Yet again, this human input might be replaceable by data mining: Based on what it could learn from the factor effects and interactions reflected in natural variability, an ML algorithm might be able to suggest refined settings of a fractional factorial design to keep the number of experimental runs at a minimum, for instance by suggesting the likely significant factors while automatically considering the interaction effects. ML could also help in finding the optimal experimental points by simply using its computational potency to vary theoretical design setups and assessing their adequacy to the problem in question.

From this brief assessment of ML applicability in DoE, some key insights emerge. First, ML functions well if it has access to sensor data from all or most of the process parameters of the respective investigation area, as only available data can be analyzed and hence considered. This is the fundamental prerequisite of effective ML applicability. Second, by analyzing the continuous flow of sensor data at natural variability levels, the ML algorithm might already be able to detect patterns, such as factor effects and interactions, during production, without the need for experimental setups and potential production stops. We want to call this the first-level benefit of ML. Third, if these microdata are not significant enough to yield direct results, it might still be useful in assessing whether a factor can be excluded from further analysis or be selected for inclusion into an experiment; we want to call this the second-level benefit of ML. Fourth, optimizing the fractional factorial design setup for experiments with the remaining factors would be the third-level benefit of ML.

A defining difference between the two methodologies is that DoE is discrete while ML is continuous: It observes an endless flow of data. We could imagine an ideal future scenario, where all factors could be varied automatically by the ML algorithm itself to detect factor effects and interactions, exceeding the normal factor variability and including some experimental elements from the DoE logic. This would mean factor levels could be varied continuously and smoothly, the number of regarded levels would not matter, and several factors could be modified at the same time to yield insights for the fine-tuning the process. The whole course of production would be equivalent to an enormous amount of test runs leading to a continuous self-optimization without disruption of the process flow. This would require an adequate production steering software and automatization tool to let the ML algorithm directly influence production settings. Although this autonomously self-improving factory is still a vision, we might not be so far from the first applications: Scientists have already constructed robots using ML algorithms to plan, conduct, and interpret physical experiments by themselves.[34]

It is conceivable that for our objective of making DoE more effective, or even fully automate it in future, ML seems highly appropriate. In general terms, both supervised and unsupervised approaches are applicable and could yield complementary insights. For the supervised algorithms, training sets could consist of structured prior process knowledge, eg, sensor data connected to relevant factors with known effects and interactions. Unsupervised algorithms might be applied to the process data that are unlabeled and uncategorized, shedding light into the “white noise” of the production function and helping us identify previously unidentified factors, their effects, and interactions. They might even indicate the existence of hitherto unmonitored factors when observed effects cannot be explained by data from the monitoring system. Unsupervised algorithms might thereby fill in the blanks in the complex web of factors and processes and help us assess how they fit into the overall logic of the production system. They can furthermore reduce dimensionality and thereby make DoE more focused.

As for the specific ML techniques, as highlighted in Figure 2, all algorithmic approaches seem to offer some potential for application in our area of interest. It needs to be pointed out that applications of ML algorithms are still largely customized solutions to specific problems, and standard software based on ML algorithms for common classes of problems has yet to be established. However, some general insights might guide the way to a customized solution. Inverse deduction algorithms can for instance be applied in all rule-based systems, and a production unit falls under this description. The rules of physics, which form the basis of any engineering application, lay the groundwork for the logical deduction of correlations. Inverse deduction algorithms may help explain interaction of factors in any production unit, such as in the example we used from the coating oven. Past data can be analyzed by probabilistic inference algorithms to detect interaction effects and refine the found patterns by continuous updating on new data. Similarly, gradient decent algorithms can form a process model, which is subsequently refined and potentially modified by incoming new data, including factor effects. Genetical algorithms are mostly applicable to help find optimal design setups for fractional factorials, as highlighted in core part 4 of the DoE process. They might also be the adequate approach if all factors are known, and the globally optimal individual factor adjustments are sought. Lastly, constrained optimization algorithms might be used in identifying clusters of factors, helping for instance in the selection of factors for further experimental analysis by assigning new factors to existing clusters.

5 | CONCLUSION

Our analysis of the joint applicability of ML algorithms and DoE highlighted differences and similarities between the two methodologies but most importantly found mayor common ground: DoE can aid in making ML algorithms more effective by finding the optimal algorithmic parameter settings, and ML can greatly support the aim of DoE in detecting factor effects and interactions. We assessed the five core parts of DoE for applicability of ML and found that it can aid in all four human-based parts, facilitating factor selection by identifying important factors and interactions based on the observed variability of monitoring data, setting the right factor levels in the experiment, and finding the optimal fractional factorial design setup for larger quantities of selected factors. This essentially means DoE can be made more effective and efficient by the application of ML. We sketched out some algorithmic approaches to achieve this.

We have also pointed out the potential of ML to replace DoE. First, if the factor variabilities observed in the monitoring data are already significant enough to obtain factor effects and interactions, DoE would be made obsolete. Second, ML could theoretically include *all* factors and *all* potential levels into its analysis, which would run counter to the idea of DoE and rather be a comprehensive “all factors at all times” concept. ML is therefore a much finer analytical tool than DoE, and the continuous automated analysis of all factors will positively impact the quality level and the efficiency of resource use in production. Combined with automated production steering software and connected to the production machinery, ML could transform the production into a self-optimizing entity independent of human input, automatically checking on factor effects and interactions, determining and implementing the optimal factor adjustments whenever deemed justified. If this scenario becomes reality, the ideas of DoE and human knowledge, skill, and intuition would be replaced by data mining. Once the ML algorithm has completely understood the production process and all its factors, continuing with any of DoE's core ideas like selection of factors or setting of levels would not be necessary anymore. DoE would vanish from the set of useful methodologies for process engineers, as would other statistical methodologies.

How likely is this scenario? For all professionally run production units with established monitoring system, applying ML is no less viable than applying DoE. With increasing factor knowledge and experience with ML, a gradual replacement of human-based statistical analysis by ML seems likely, as it offers diverse benefits such as sharply reduced cost and effort. However, the pace of this process is constrained by a set of limitations. The first limitation concerns the data availability. To find the relevant patterns, ML requires not only input data from factor variability but also output data, signifying an extension of the monitoring system to output. However, not all output characteristics might be automatically checkable or require a large additional investment in monitoring. The second limitation concerns algorithmic challenges present in ML, such as overfitting (erroneous interpretations of data) and biasing (yielding trivial insights). A third limitation might concern bad data quality, insufficient data, or excess requirement of computational resources.

All these limitations are, however, only temporary. Sensor technology advances rapidly, increasing the options for cost-efficient input and output monitoring—sometimes, it is ML itself that provides the best solution, as cases of automated output inspection show.[35] Overfitting and biasing are known problems and can be counteracted by careful fine-tuning of the algorithm itself. Additionally, a large stream of research dedicated to these problems is likely to alleviate or solve them in future. Bad data quality is largely an internal issue of the user and affects the explanatory power of DoE in a similar manner. Insufficient data can be addressed by comprehensively extending the monitoring system, including factors process engineers normally discard as irrelevant. The potency of ML specifically lies in being able to analyze such big data. As this lastly requires computational power, the technological progress and the allocation of adequate financial resources to the issue might address it sufficiently.

For the time being, DoE and other statistical methodologies will continue to be applied by engineers, and some companies will start using ML to reinforce them. A skillful application of ML will accelerate knowledge generation and usage and enable the gradual automation of process adjustments. Technological progress and solving of ML's limitations might make its use more widespread, and the first companies to have implemented a full-scale ML system for production steering and control might enjoy a significant competitive advantage, thus exerting competitive pressure on those who have not.

ML will thereby greatly speed up the process of quality management, from quality problem recognition and improvement to quality control and maintenance. It will help overcome productivity barriers that exist because of human errors, incomplete knowledge, and limited analytic capabilities. It will replace DoE and other statistical methodologies based on human input. But this is no reason for lamentations. In the end, both ML and DoE aim for the same objective: a world where poor quality is synonymous with the past.

FUNDING INFORMATION

This work was supported by the FHS St. Gallen University of Applied Sciences.

ORCID

Johannes Freiesleben https://orcid.org/0000-0003-3969-4129

REFERENCES

1. Athey, S., 2018. Why business leaders shouldn't have blind faith in AI. *Stanford University Business Insights*. Posted on www.gsb.stanford.edu/insights on May 23rd, 2018 (last accessed on Aug 8th, 2018).
2. Faraj S, Pachidib S, Sayegha K. Working and organizing in the age of the learning algorithm. *Information and Organization*. 2018;28(1):62-70. https://doi.org/10.1016/j.infoandorg.2018.02.005
3. Dean, J., Patterson, D. & Young, C., 2018. A new golden age in computer architecture: empowering the machine-learning revolution. *IEEE Micro*, 38(2), 21–29. https://doi.org/10.1109/MM.2018.112130030
4. Bisgaard, S., 1988. The quality detective: a case study. *CQPI Report* no. 32, University of Wisconsin-Madison.
5. Box GEP, Luceño A. *Statistical Control by Monitoring and Feedback Control*. New York: Wiley; 1997.
6. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd; 1935.
7. Rose JM, Bliemer MC. Constructing efficient stated choice experimental designs. *Transport Reviews*. 2009;29(5):587-617. https://doi.org/10.1080/01441640902827623
8. Montgomery DC. *Design and Analysis of Experiments*. 8th ed. New Jersey: John Wiley & Sons; 2017.
9. Samuel AL. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*. 1959;3(3):210-229. https://doi.org/10.1147/rd.33.0210
10. Turing AM. Computing machinery and intelligence. *Mind*. 1950;59(236):433-460. https://doi.org/10.1093/mind/LIX.236.433
11. Vincent, J., 2018. Google's AI sounds like a human on the phone—should we be worried? Published on www.theverge.com on May 9th, 2018 (last accessed on Aug 8th, 2018).
12. Cheng, B. & Titterington, D. M.,1994. Neural networks: a review from a statistical perspective. statistical science, 9(1), 2–30. http://www.jstor.org/stable/2246275
13. Sarle WS. *Neural Networks and Statistical Models. Proceedings of the Nineteenth Annual SAS Users Group International Conference*. North Carolina: Cary; 1994.
14. Langley P. The changing science of machine learning. *Machine Learning*. 2011;82(3):275-279. https://doi.org/10.1007/s10994-011-5242-y
15. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255-260. https://doi.org/10.1126/science.aaa8415
16. Sheng J, Amankwah-Amoah J, Wang X. A multidisciplinary perspective of big data in management research. *International Journal of Production Economics*. 2017;191:97-112. https://doi.org/10.1016/j.ijpe.2017.06.006
17. Staelin, C., 2003. *Parameter selection for support vector machines*. Hewlett-Packard Company.

18. Packianather MS, Drake PR, Rowlands H. Optimizing the parameters of multilayered feedforward neural networks through Taguchi design of experiments. *Quality and Reliability Engineering International*. 2000;16(6):461-473. https://doi.org/10.1002/1099-1638(200011/12)16:6<461::AID-QRE341>3.0.CO;2-G
19. Sukthomya W, Tannock J. The optimisation of neural network parameters using Taguchi's design of experiments approach: an application in manufacturing process modelling. *Neural Comput Applic*. 2005;14(4):337-344. https://doi.org/10.1007/s00521-005-0470-3
20. Ortiz-Rodríguez, J. M., Martínez-Blanco, M. R., & Vega-Carrillo, H. R., 2006. Robust design of artificial neural networks applying the Taguchi methodology and DoE. Electronics, Robotics and Automotive Mechanics Conference (CERMA'06), 131-136. https://doi.org/10.1109/CERMA.2006.83
21. Balestrassi PP, Popova E, Paiva AD, Lima JM. Design of experiments on neural network's training for nonlinear time series forecasting. *Neurocomputing*. 2009;72(4–6):1160-1178. https://doi.org/10.1016/j.neucom.2008.02.002
22. Bates, S., Sienz, J., & Toropov, V., 2004. Formulation of the optimal Latin hypercube design of experiments using a permutation genetic algorithm. *In* 45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference. Palm Springs, California. https://doi.org/10.2514/6.2004-2011
23. Viana FA, Venter G, Balabanov V. An algorithm for fast optimal Latin hypercube design of experiments. *Int J Numer Methods Eng*. 2010;82(2):135-156. https://doi.org/10.1002/nme.2750
24. 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(9):6385-6394. https://doi.org/10.1039/C4RA12750A
25. Otsuka A, Nagata F. Quality design method using process capability index based on Monte-Carlo method and real-coded genetic algorithm. *International Journal of Production Economics*. 2018;204:358-364.
26. Antony J. Training for design of experiments using a catapult. *Quality and Reliability Engineering International*. 2002;18(1):29-35. https://doi.org/10.1002/qre.444
27. Box GEP, Hunter JS. The 2^{k-p} fractional factorial designs. *Dent Tech*. 1961;3(3):311-351. https://doi.org/10.1080/00401706.1961.10489951
28. Ghahramani, Z., 2004. Unsupervised Learning. In: Bousquet, O., von Luxburg, U. & Rätsch, G. (eds.) Advanced Lectures on Machine Learning. ML 2003. *Lecture Notes in Computer Science, 3176, 72–112*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-28650-9_5
29. Shalev-Shwartz S, Ben-David S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press; 2014.
30. Ortiz N, Hernández RD, Jimenez R, Mauledeoux M, Avilés O. Survey of biometric pattern recognition via machine learning techniques. *Contemporary Engineering Sciences*. 2018;11(34):1677-1694. https://doi.org/10.12988/ces.2018.84166
31. Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*. 2010; 52(2010):12-40. https://doi.org/10.1016/j.specom.2009.08.009
32. Domingos P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York City: Basic Books; 2015.
33. Box GEP, Bisgaard S. Statistical tools for improving designs. *Mechanical Engineering*. 1988;110(1):32-40.
34. Scammell, R., 2018. AI robot aids scientists in malaria discovery. Published on www.drugdevelopment-technology.com on Jan 19th, 2018 (last accessed on Aug 8th, 2018).
35. Duan G, Wang H, Liu Z, Chen YW. A machine learning-based framework for automatic visual inspection of microdrill bits in PCB production. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*. 2012;42(6):1679-1689. https://doi.org/10.1109/TSMCC.2012.2216260

AUTHOR BIOGRAPHIES

Johannes Freiesleben is a professor of business administration with a focus on Machine Learning applications, especially in production. He is currently a Senior Consultant and works based in Zurich, Switzerland. Dr Freiesleben has extensively researched the economic implications of better production quality and has published 20+ papers in international journals regarding the topic. He sees great value in looking at the quality question from the perspective of novel technologies and other disciplines.

Jan Keim is a graduate student at Trinity College, Dublin, and engages vividly in projects using Machine Learning as enabling technology.

Markus Grutsch is professor of business administration with a focus on quality and process management and currently works as Senior Consultant based in St. Gallen, Switzerland.

How to cite this article: Freiesleben J, Keim J, Grutsch M. Machine learning and Design of Experiments: Alternative approaches or complementary methodologies for quality improvement? *Qual Reliab Engng Int*. 2020; 1–12. https://doi.org/10.1002/qre.2579