Comparative Analysis of Small Data Acquisition Strategies in Machine Learning Regression Tasks Addressing Potential Uncertainties

Xukuan Xu
Aschaffenburg University of Applied Sciences
Aschaffenburg, Germany
e-mail: xukuan.xu@th-ab.de

Felix Conrad
Dresden University of Technology
Dresden, Germany
e-mail: felix.conrad@tu-dresden.de

Xingyu Xing
Aschaffenburg University of Applied Sciences
Aschaffenburg, Germany
e-mail: xingyuxing0630@gmail.com

Oskar Loeprecht
Dresden University of Technology
Dresden, Germany
e-mail: oskar.loeprecht@yahoo.de

Michael Moeckel
Aschaffenburg University of Applied Sciences
Aschaffenburg, Germany
e-mail: michael.moeckel@th-ab.de

***Abstract*—As the algorithms mature, the bottleneck in applying Machine Learning (ML) to engineering, in particular to process analysis, monitoring and control, is often caused by the limited availability of suitable data and the cost of data acquisition. For many ML projects, datasets have been collected independently of subsequent analysis. In laboratory-based development, data acquisition and coverage of possible process uncertainties pose challenges to the preparation of datasets suitable for ML. This paper benchmarks existing design of experiments (DOE) strategies based on data generated by a simulation model, discussing their aptitude for training accurate ML regression models. 11 representative sampling strategies have been investigated to provide guidance for data collection under data acquisition constraints, including consideration of possible measurement uncertainties. As the optimal DOE depends on available data volume and the uncertainty level, recommendations for DOE selection are given.**

Keywords-Small-data; Process uncertainty; Design Of Experiments; Machine learning; Model-based sampling; Auto-sklearn.

I. INTRODUCTION

ML makes it possible to efficiently mine valuable information from data due to its powerful data analysis capabilities. With the prosperous advancement of algorithm research, model building is no longer a challenge limiting ML applications [1][2]. In fact, according to a survey from Crowdflower in 2016 [3], the efforts of data scientists are mainly (60%) consumed by data organizing and data cleaning. After this, 19% of the time is spent collecting datasets. This shows that data preparation involves considerable effort of ML applications in the current stage. However, this difficulty is often overlooked by the informatics community. In most cases, the datasets are pre-existing. With this standpoint, they simply optimate the algorithm at the software side for data analysis. However, the dataset's quality determines the upper limit of data analysis. Therefore, in some cases, it may be unfeasible to look at a solution only from the ML model side. Only recently, the intersection of experimental design towards data collection and ML has come to the fore. R. Arboretti et al. systematically reviewed the joint application of DOE and ML in areas such as industrial production, which identified the current status of research in terms of DOE selection for ML [4]. In this context, a preliminary study of the relationship between DOE selection and ML was conducted based on simulation models [5]. Roberto Fontana et al. benchmarked the performance of ML models obtained from data collected with different DOE strategies, where the potential of an Active Learning (AL) approach for dataset acquisition was investigated [6]. However, their experiments were limited to a specific amount of data without further guidance of DOE selection for varying data volumes.

It is both a challenge and an advantage to look at data preparation from the perspective of a production engineer. Collecting a single element of the dataset requires that a product is physically produced and the relevant data is measured during the manufacturing process. In practice, an extra number of products is required to account for deficient outcomes. This limits the amount of usable data for ML analysis. The cost considerations often constrain the overall amount of data. However, pre-existing knowledge, experience or even intuition of the process often allows an engineer to focus the data generation on particularly relevant subsets of an overly complex parameter space.

Purpose-built datasets for ML modeling may address two possible directions [7]:

I. Finding the control variables and their optimal values that result to an optimal response
II. Exploring the neighborhood around the optimal values to generate knowledge for monitoring, anomaly detection and control

This article investigates the latter under the constraint of limited resources (e.g., time, budget) for data acquisition and fixed overall statistical process uncertainty. Based on the data obtained from an experimental lithium-ion battery (LIB) production line realized within the KIproBatt project [8], we describe the practical difficulties in preparing datasets for ind-

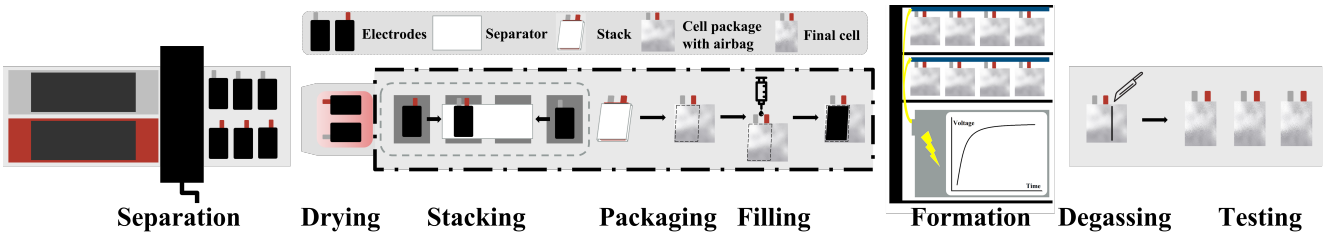

Figure 1. LIB cell assembly process from separation to EOL-tests

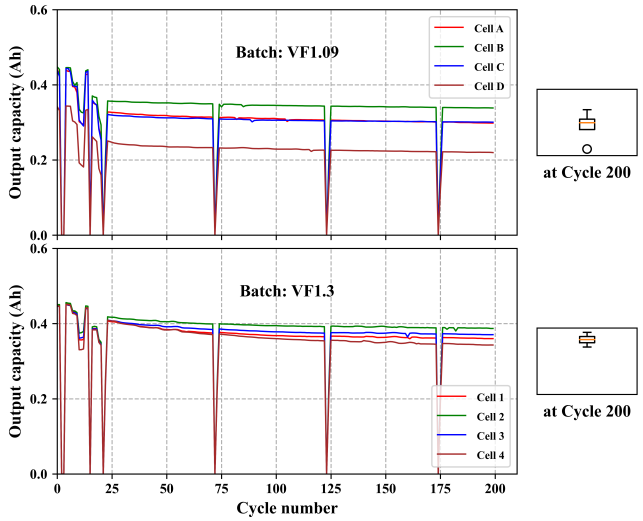

Figure 2. Cell output capacity related to cycle number in a cycling test

ustrial process development in Section II. In Section III, existing DOE approaches are described. The ML package Auto-sklearn [9] and its modelling capabilities are introduced. Finally, the experimental setups for comparing the potential of different DOEs with predefined data resources are presented. The experiments include both cases with and without process uncertainties. The results of the experiments and the discussions are documented in Section IV. Guidelines for DOE selection based on experimental data are given at its end. Considerations on the application of an iterative DOE in case of small data are given with Emukit DOE as an example. Section V summarizes the contributions of this article and possible further expansions of this research are envisaged.

II. DESCRIPTION OF SMALL-DATA CONTEXT

A. Small data problem

Small-batch production is often unavoidable in laboratory research, on a pilot production stage prior to upscaling [10], or in customer-specific (individualized) manufacturing [11]. Often, data acquisition is limited by budget or time constraints to datasets with less than one thousand elements. The particular choice of selected data points affects the outcomes of subsequent analysis. For illustration, we consider the project KIproBatt as an example of a typical small-scale data generation: a total of ca. 500 Li-ion battery cells is to be produced with a semi-automatic production line in a laboratory environment. Research questions include the impact of process deviations on the quality of final cells as well as the exploration of complex correlations among process parameters. Note that one cannot define the "small-data problem" by sole reference to a fixed amount of data. Instead, the characteristics and complexity of both the research objectives and the applied ML methods have to be considered.

B. Lack of process knowledge & complexity of the production process

The number of required data depends on the complexity of the process. A large number of features, non-linear relationships and interactions between features increase the complexity of the process and thus the number of data points required. These conditions are often found in industrial production processes [11]. The assembly process of a LIB pouch cell is an example of such a complex process and is depicted in Fig. 1: cell assembly starts with electrode separation. Then, the anodes and cathodes are dried and fed into a glove box with a controlled atmosphere. Next, a stacking machine assembles the electrodes with a separator into cell stacks (Z-fold stacking). After the packaging, sealing and electrolyte filling, the cell is activated by the first charge and discharge (formation). The gas generated in this procedure is removed and the cell is finally sealed.

The complexity of this multi-step process leads to manifold variable interdependencies. Hence, an effective analysis should be based on a ML approach. However, it is challenged by limited data, which may lead to under sampling of the parameter space and a lack of convergence of the ML models. We define this as the fundamental characteristic of small-data context.

C. Process uncertainty

Complex processes are normally investigated for a limited set of process parameters only. While the remaining parameters are, in theory, assumed to remain constant, their unavoidable fluctuations contribute to statistical uncertainty in all measured data. Other sources for uncertainties lie, for instance, in the measurement uncertainties of the used sensors. This uncertainty is manifested in the data as identical input parameters will lead to a statistical spreading in the target responses.

In the KIproBatt project, using the injected electrolyte volume as the only tunable factor with two levels, we produced four cells at each level while ensuring that the rest of the process parameters were unchanged. Each cell was then tested according to the same cycling protocol to evaluate its performance. The cycling protocol also includes non-cycling tests such as pulse tests, c-rate test and quick charge test. Pulse tests are designed to obtain information regarding battery resi-

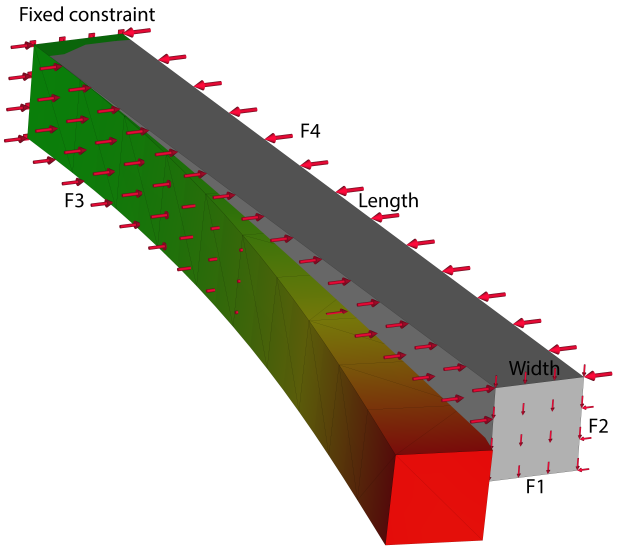

Figure 3. Constraints and forces in the FEM model.

TABLE I. INPUT PARAMETERS IN THE FEM SIMULATION

	Variables	Range	Unit
Input	Ym	50,000 – 600,000	Mpa
	Pr	0.1 – 0.45	1
	L	8000 – 10,000	mm
	W	1000 – 2000	mm
	F1	1000 – 10,000	kN
	F2	1000 – 5000	kN
	F3	1000 – 10,000	kN
	F4	1000 - 5000	kN
Output	Displacement	Ca. 0.7 – 400	mm

Stance, which are labelled as 0 during data processing. The results, using output capacity (OC) as an indicator, are shown in Fig. 2.

It can be seen that the performance of the battery cells within each batch varies. As the box plot illustrates, the process uncertainty is so evident in batch VF1.09 that cell D can be judged as an outlier (box plot).

The reasons for this might be processing errors due to human operations, a lack of process understanding that leaves some potential variables uncontrolled, or measurement errors in the hardware. But in the end, what emerges is the uncertainty of the OC.

When the process uncertainty exceeds the variation imposed on control variables, no direct conclusion can be derived. Normally, uncertainty reduction could be achieved either by optimizing hardware or by repeated measurement and averaging. However, for fixed measurement capacity, the latter implies a reduced ability for parameter space exploration. Therefore, DOE strategies can be developed further to find new compromises between resource allocation for uncertainty reduction and for parameter space sampling.

III. SETTING OF THE EXPERIMENTS

In this section, the potential of various sampling methods to build a regression model under different data volumes and levels of uncertainty are investigated. The analysis is divided into two parts:

1. The first part is to understand the performance of different DOEs through training an optimal regression model as the data volume varies.
2. The second analysis is to investigate the potential of these DOEs under varying levels of uncertainty, where different uncertainties are introduced to the target parameter.

An independent test dataset is obtained using Latin Hypercube Sampling (LHS), which consists of 2,500 data points. The root mean square error (RMSE) of the predicted displacement versus the output from the Finite Element Method (FEM) simulation is used to measure the true error of the ML model. The R^2 Score is also employed to evaluate the model [12]. The best achievable performance with the given training dataset of these models on the test dataset is considered as the potential of the corresponding DOEs.

FreeCAD was chosen as the platform for building simulation. It supports building models with python code and provides an application programming interface to facilitate the import and export of data. The simulation model includes eight input parameters: Young's modulus (Ym), Poisson's ratio (Pr), length (L) and width (W) of the beam with four force constraints applied to the beam. The displacement magnitude of the beam is defined as the target parameter. Table I and Fig. 3 provide further information about this simulation model.

Twelve algorithms covered by Auto-sklearn are used to build the regression models for the prediction of the target parameter in the parameter space [13]. In order to provide an objective comparison among the DOEs without potential deviations during the training process, the settings of the hyperparameters in Auto-sklearn should be tuned to appropriate values [14]. Thus, it can be ensured that the potential of DOEs are effectively compared without the influence of non-optimal model training.

A. *Tested DOE strategies*

DOE is an established approach to systematically collect information about a system or process. It aims at delivering the most relevant experimental data for addressing a given research objective. The origin of classical DOE can be traced back to the Analysis of Variance (ANOVA) proposed by FISHER in the 1920s [15]. Conventional DOE has a set of proven paradigms: screening design, e.g., full factorial design (FFD) for identifying relevant parameters; response surface design, including central composite design (CCD), Box–Behnken design (BBD), for detailed investigation of optimal parameter configurations [16]. With the development of data science and easier access to data, ML tools have been

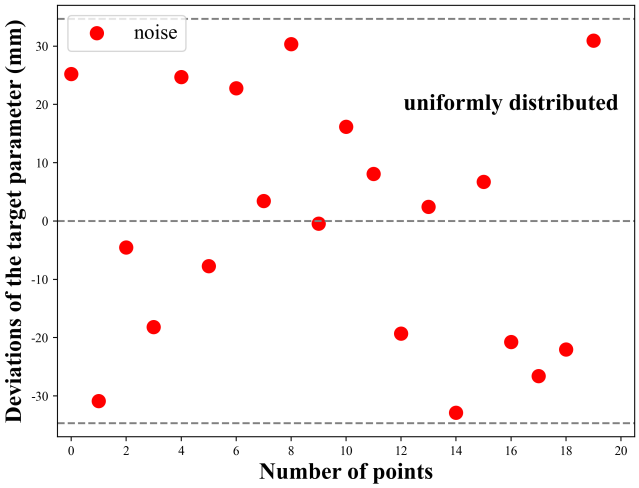

Figure 4. Deviations of displacements from simulated values due to a 10% uncertainty (up to 10% of the maximal change of the displacement in the predefined input dataspace)

successfully applied to many data analysis problems. ML has unparalleled efficiency advantages in analyzing big data (compared to the volume of data in conventional DOEs) with complex interdependencies. However, little attention has been paid to the interplay of data set generation and ML-based data analysis. Represented by LHS, space-filling design is able to partition the data space isometrically into multiple levels [17][18]. This feature makes LHS well-suited to drive ML schemes. A series of studies have conducted the generation of datasets for ML based on conventional DOEs in the past five years [19][20]. In addition, motivated by some ML algorithm developments, iterative data acquisition schemes have been discussed.

Emukit provides such a model-based iterative DOE scheme within a Bayesian optimization framework [21][22]. The Emukit DOE tool starts from a set of given initial data points and iterates the following three steps to generate sample points in a given input space:

- fit a prediction model to the existing data
- find the next point with the highest marginal predictive variance as predicted by the prediction model
- add this new data point to the existing dataset

Such iteration allows for the most efficient allocation of a limited number of data points based on certain metrics, such as marginal predictive variance of the model. This model-based scheme works well with ML data analysis since a prediction model (e.g., Gaussian process model, GP model) is used to predict the target response and calculate the variance during each iteration of data acquisition [22].

Table II contains a summary of the different DOEs which have been tested. Different settings for the CCD, criteria in the LHS and different acquisition functions in Emukit were considered as different DOEs. The range of the training data volume is set from 40 to 320. Since conventional DOEs (FFD, BBD, CCD) are predetermined by the number of input factors, levels and the DOE strategies, it is not possible to change the data volume continuously to build multiple datasets with a specified amount of data. As an example, given 8 variables, the dataset generated according to FFD must consist of 2^8 data points. The adopted solution was to use the D-optimal criterion [23] to filter the required optimal design. For example, the use of the D-optimal criterion enables the construction of any subsets with less than 256 data points, which makes it possible to continuously change the amount of data within a certain range.

TABLE II. DOEs AND THEIR ABBREVIATION CODES

Abbreviation	Sub	Descriptions
FFD		Full-Factorial design
CCD	CCD_c	Central-Composite design, where the star points are at the same distance from the center
CCD	CCD_i	A scaled down CCD_c design with each factor level of the CCD_c design divided by a given constant
CCD	CCD_f	Star points are at the center of each face of the factorial space
BBD		Box-Behnken design
LHS	LHS_c	Latin-Hypercube sampling, which centers the points within the intervals
LHS	LHS_m	Maximize the minimum distance between points, randomly distribute points within the intervals
LHS	LHS_cm	Maximize the minimum distance between points, centered them within the intervals
LHS	LHS_cor	Minimize the maximum correlation coefficient
Emukit	Emukit_us	Iterative sampling strategy, choose the next point according to the marginal predictive variance of a GP model [23]
Emukit	Emukit_ivr	Choose the next point such that the total variance of the model is reduced maximally [25]

B. *ML modeling*

The model training using Auto-sklearn is repeated five times. The best performance among them, i.e., the performance of the best model that can be obtained for this training dataset, will be recognized as the potential of the corresponding DOE used for collecting the training dataset. The experiments were conducted on a Dell workstation (Intel® Xeon® W-2295 Processor: 3.00 GHz * 36, memory: 128GiB). The settings of hyperparameters in Auto-sklearn used for modeling are shown in Table III.

TABLE III. HYPERPARAMETERS IN AUTO-SKLEARN

Hyperparameters in Auto-sklearn	Value
time_left_for_this_task	300s
per_run_time_limit	30s
initial_configurations_metalearning	25
memory_limit	20480 MB
resampling_strategy	"cross validation"
resampling_strategy_arguments	"folds: 5"
n_jobs	18

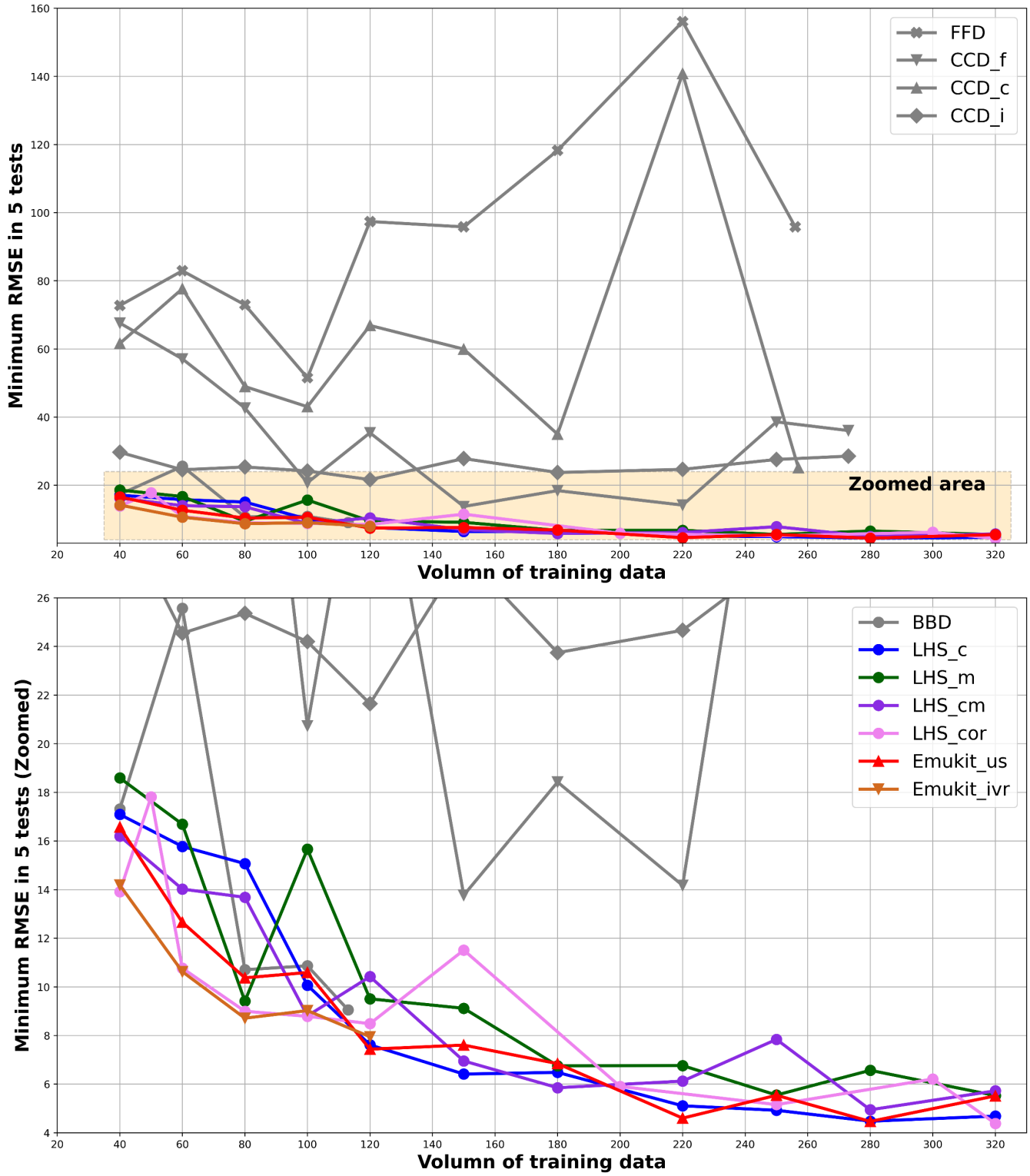

Figure 5. The potential of tested DOE strategies

C. *Settings of the Uncertainty*

Uniform distributed noise was added to the target parameter to mimic the process uncertainty described in Section II C. The reason for choosing uniform distribution over Gaussian distribution lies in the fact that Gaussian noise will produce a large number of low-level noise points around zero. Such noise points cannot represent the set level of uncertainty.

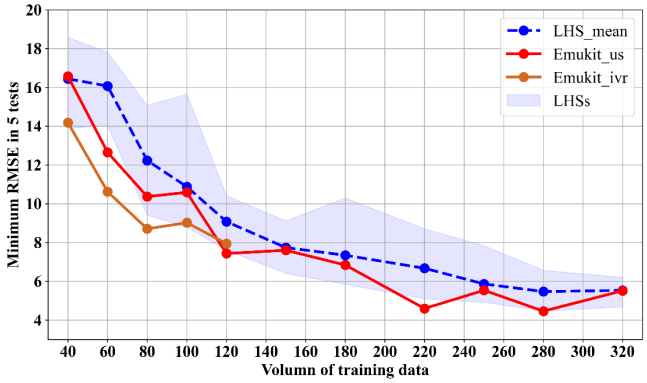

Figure 6. The potential of Emukit strategies compared to the average potential of LHSs

The tested range of process uncertainty was set to be 0-20% of the variation range of the target parameter. To generate the process uncertainty, the random function in the Python Numpy library was employed. Fig. 4 shows the distribution of 20 generated uncertainty points at the 10% level of uncertainty.

Owing to the instability in the noise points, ten sets of noise points were generated independently and added to the target parameter to create the ten different training datasets for each DOE strategy. For each training dataset, the modeling process was performed only once.

It should be noted that for conventional DOEs or LHS strategies, the uncertainty addition scheme adopted is to generate all data points without uncertainty all at once. The uncertainties are then added directly to the output displacement as the final step in the data generation. For Emukit, however, this scheme does not correspond to the actual experiment procedure. The following data generation scheme is iterated to generate uncertainty-containing training data for Emukit:

- fit a prediction model to the existing data
- find the next point with the highest marginal predictive variance as predicted by the prediction model
- add this new data point to the existing dataset

The test dataset without uncertainty was used to evaluate the trained models. The average performance of the ten trained models is recognized as the potential of the corresponding DOE used for the training dataset.

IV. Results and Discussions

A. Without uncertainty

For a relatively complex parameter space consisting of eight input factors, most of the conventional DOE methods cannot build a promising training dataset. As can be seen from the first half in Fig. 5, the performance of conventional DOEs (FFD, CCD_f, CCD_c, CCD_i) are not comparable to that of LHS or Emukit under the same amount of data. BBD is the best strategy among conventional DOEs, which performs almost similarly. However, as mentioned above, one of the major drawbacks of conventional DOEs is their inability to generate a specified amount of data as required. With the aid of D-optimal design, the BBD strategy is also only capable of planning data points within its given range. Such a drawback greatly limits the use of conventional DOE in the ML domain.

Also, the LHS and Emukit strategies outperform the conventional DOEs except for BBD at any amount of data. For the LHS family, with the exception of a few data points (LHS_m at 100, LHS_cor at 150), the LHSs perform essentially similarly with the same amount of data. It cannot be concluded that one certain LHS is necessarily better than other LHS strategies. As a kind of space-filling design, LHS is able to evenly distribute the limited data resource in a given data space to explore as much data space as possible. It is certainly a DOE suitable for ML data analysis.

Both Emukit strategies (Emukit_us & Emukit_ivr) are safe choices compared to the LHSs. In other words, Emukit strategies never perform the worst at any data volume, not to mention that the Emukit_ivr has the top performance with small data volumes (40 - 100).

Fig. 6 demonstrates this conclusion more clearly. The dashed line in Fig. 6 shows the average performance of the four LHS strategies. Both Emukit strategies outperform the average performance of LHSs over their data volume interval. This difference is particularly noticeable when the amount of data is relatively small (<120). Whereas, when the amount of data is sufficient (>250), the performance of LHSs can converge to Emukit_us. It can be concluded that one of Emukit's advantages is its ability to efficiently allocate data resources when data volumes are insufficient.

Both LHS and Emukit can generate DOEs with the requirement of training data volume based on the number of input factors. As an iterative scheme, Emukit is more flexible than space filling DOE: it can continuously generate additional data points besides existing data. In contrast, LHS requires that the amount of data volume be specified at the beginning, which isn't compatible with additional data generation.

But Emukit is not always the optimal choice. It needs an initial amount of data for subsequent iterations. If the model trained with the initial dataset does not drive Emukit correctly, then the results out of the iterations can be disastrous. This is further discussed in paragraph C in this section.

B. With uncertainty

According to the uncertainty generation scheme in Section III. A, 10 different sets of noisy data were generated for each dataset at each data volume. Most conventional DOEs showed inferior performance compared to LHSs. Thus, only CCD_i was selected from conventional DOEs for comparison in this phase. LHS_c from the LHS family was selected as a representative strategy. Since the CCD is a pre-set conventional DOE, the test range of CCD_i in the uncertainty test was set to 40-250. For LHS_c, the upper limit on the amount of data is extended to 700 for observing the improvement in model performance despite the existence of

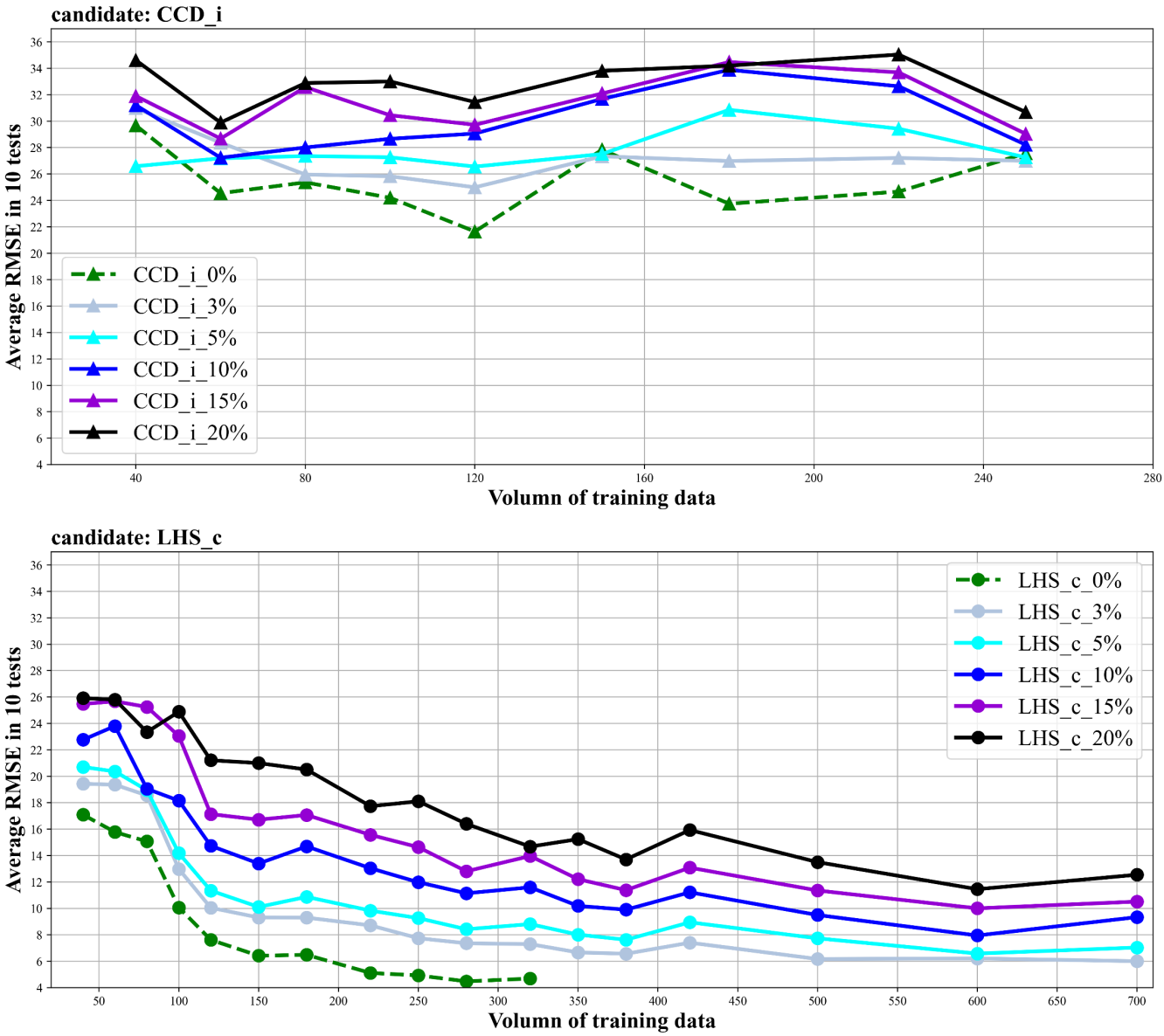

Figure 7. Potential of LHS_c and CCD_i strategies with varying uncertainties

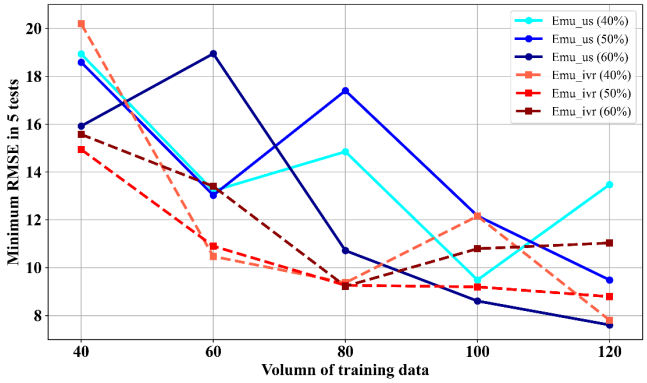

Figure 8. Potential of Emukit strategies with different settings

uncertainties. Emukit's performance under the impact of uncertainty is placed in paragraph D in this section. The experimental results are shown in both Fig. 7 and Fig. 9.

It is clear that for both of the measured DOEs, increasing uncertainty leads to deterioration of model performances. The experiment results of Emukit demonstrate the same trend. Therefore, this conclusion is generalizable to all three types (conventional, space-filling, model-based iterative) of DOE strategies.

It can be observed from the second half of the Fig. 9 that the adverse effect due to uncertainty is gradually compensated for as the amount of training data rises. In the case of LHS_c, for example, the performance of the model obtained using 600 noisy data with a 10% level of uncertainty is approximately the same as the performance of the model trained with 100 training data without any uncertainty. This suggests that "big

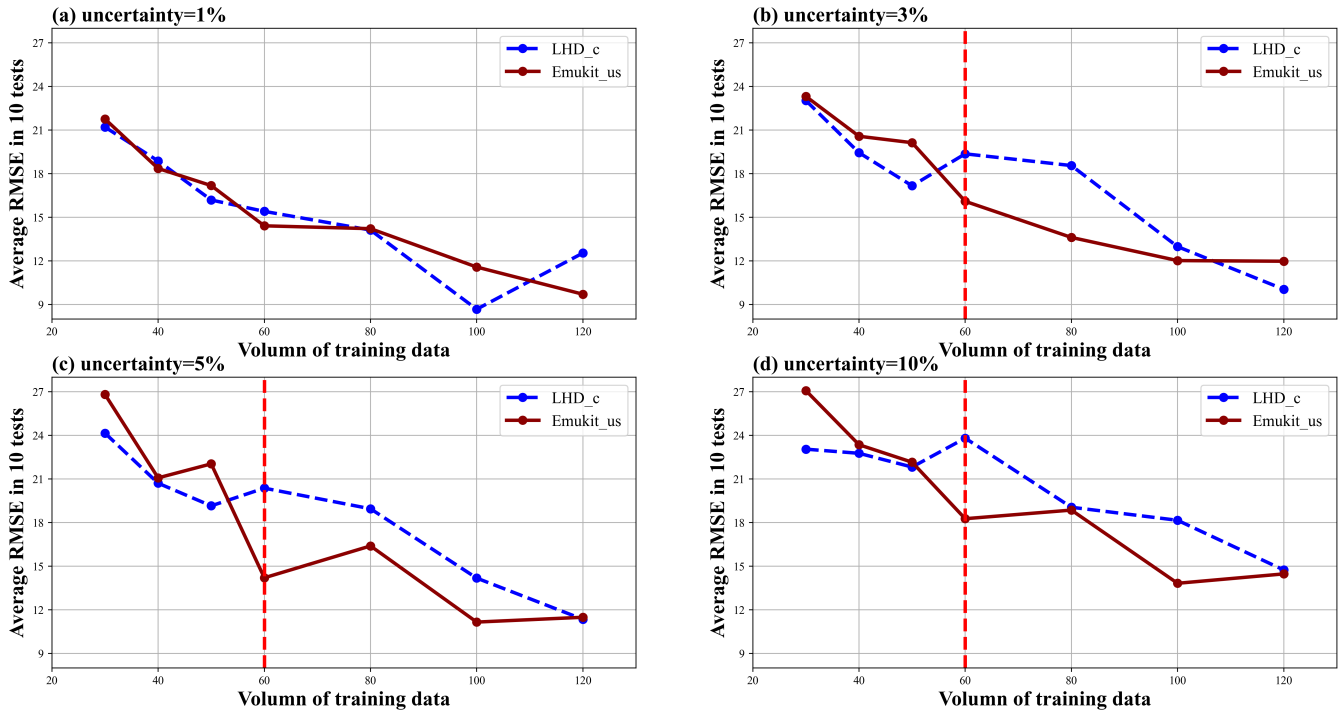

Figure 9. The potential of Emukit_us compared to the Potential of LHS_c in small-data context with varying uncertainties. In the case of small uncertainty (a) no significant differences can be observed. For larger uncertainty (b-d) Emukit outperforms LHS as soon as a critical amount of data (red line) becomes available.

data" can compensate the effect of uncertainty to some extent. However, no such trend appears in the results of CCD_i. Thus, it can be concluded that a space-filling design becomes preferable to a pre-determined conventional DOE under the influence of uncertainty. By employing more data, a predictive model, which is closer to the ground truth can be obtained. Even so, the effect of uncertainty remains at larger data volumes (600-700).

C. *Guide for iterative DOE: Emukit as an example*

As an iterative model-based DOE strategy, Emukit is governed by three hyperparameters:

- the integrated GP-model
- the acquisition functions
- initial data volume as a percentage of total data volume

It is not difficult to imagine that if the GP model has limited predictive power, then its predictions about data points are unreliable. Therefore, the first step in using Emukit is to optimize the GP model. The tuning of the GP model can be found in many references [24-27]. In this regard, the effects of the other two hyperparameters have been explored through experiments. A comparative experiment is conducted within the data volume from 40 to 120. The results are shown in Fig. 8.

As can be seen in Fig. 8, the potential of Emukit_ivr at small data volume (<80) outperforms that of Emukit_us. The advantage of the ivr acquisition function does not exist afterwards, where there is no longer a clear superior choice. Details about these acquisition functions in Emukit are available in [23] [25]. Considering the whole tested range of the training data, 50% initial data share is a safe choice for both acquisition functions. However, in case of extremely small data resources (<40), allocating more resources to the initial dataset seems to be a safe choice. It is also worth noting that the ivr acquisition is time consuming, which consumes at least twenty times as much time as the us acquisition. Considering the efficiency factor, us acquisition is a valid choice when the data resource is large enough.

Discussion on the usage of Emukit with small data resources continues with the interference of uncertainties. With this purpose, we conducted experiments with a training data volume of 30-120 and the tested uncertainty level was set to 3%-10%. Uncertainty was added according to the settings described in Section III C. LHS_c and Emukit_us (50% initial data share) were selected as candidates for the experiment. The results are recorded in Fig. 8. The discussion of the results is presented in paragraph D.

D. *Tutorial on DOE selection in small-data context with uncertainty*

In this section, we provide a preliminary generalization towards DOE selection based on our experimental results. Again, it is important to state that our conclusion towards DOE selection is restricted to ML regression models. The goal of the DOE is to explore the predefined parameter space for a prediction model. The selection of DOEs is considered on the

basis of "you only get one chance" principle. Therefore, in addition to comparing the best accuracy of the trained models that each DOE can deliver, the reliability of this DOE in the worst-case scenario also has a decisive influence. More specifically, a DOE strategy that consistently brings the models to an R^2 of around 0.8 regardless of the uncertainties is preferable to a DOE that only in the best-case scenario enables a model to reach 0.9 while, in other cases leaves the models only managing an R^2 score of 0.7.

An empirical conclusion from the ML community regarding the estimation of the required amount of training data is "two subjects per variable" (2SPV) rule of thumb [28] [29]. This rule is certainly influenced by the complexity of the model. The object of the unknown relationship lies in a multidimensional parameter space. A complex relationship between the target parameter and the input parameters demands a larger training dataset. Following this empirical law, an estimation (Est) of the amount of data required to mimic the exemplary FEM using multivariate linear regression with quadratic terms can be determined.

$$Est = 2 * (8 + 8 + 28 + 1) = 90 \quad (1)$$

The first term in brackets in (1) is the number of primary linear coefficients, the second and the third terms are the number of quadratic coefficients. At last, there is one constant coefficient. Therefor, for this FEM model, less than 90 data can be roughly recognized as a small-data context according to the 2SPV rule.

Both LHS family and Emukit strategies are appropriate candidates when the data resources far exceed (>2Est) small-data context. At this point, the main factor affecting the DOE selection is the time efficiency, which has been interpreted in Section IV A. The presence of uncertainty (<20%) leads to deterioration in model performance. To obtain well-performing models it requires more data to compensate for the uncertainty (see Fig.7).

The Emukit is the best choice in terms of best achievable prediction accuracy when the available data resource is 1-2 times the size of the small-data context (Est - 2Est). This choice is safe when the uncertainty level stays below 10%. The application of Emukit demands discretion when the uncertainty level goes higher. In such cases, LHSs are safe candidates.

The impact of uncertainty cannot be ignored in small-data context (<Est), where the available data resource is less than the estimation according to the empirical law. As shown in Fig. 8, for each uncertainty level, the amount of data for which Emukit exceeds LHS for the first time is marked with a red dotted line. It can be found that Emukit outperforms LHS only when the amount of data at its disposal exceeds 60. i.e., Emukit requires a minimal amount of initial training data in order to allocate data points correctly. If the uncertainty remains at a very low level (below 1%), Emukit could still be a good choice compared to LHS. As shown in the first plot of Fig. 8, the potential of LHS and Emukit are comparable within the data amount from 30 – 80. The above discussion on DOE selection is summarized in Table IV, where I denotes iterate sampling (represented by Emukit) and S denotes space-filling design (represented by LHS).

TABLE IV. APPROPRIATE DOE FOR DATA ACQUISITION

Uncertainty / Data volume	< 1 %	3% - 10%	> 10%
< 0.5Est	I ≈ S	S > I *	S
0.5Est - Est	I > S	S ≈ I *	S
Est - 2Est	I > S	I > S	S
> 2Est	I > S	I > S	S

Note that for cases marked with an asterisk in Table IV, iterative sampling is still reliable if the initial training dataset is able to yield a decent model until the effects of uncertainty become significant, or the available initial data is insufficient to enable the core model to deliver an effective predictive model. It is recommended to examine the performance of the model trained with the initial dataset. According to the experiments with Emukit, Gaussian Process models trained with limited initial data perform best if a positive R^2 score ($R^2 > 0$) can be reached.

V. CONCLUSION

This article discussed characteristic aspects of the "small data problem" with process uncertainties. The performance of some existing DOE strategies was tested with data collected from a self-built FEM simulation. The accuracy of different ML regression models trained with data collected according to a specific DOE at given data volume are systematically compared. The effect of uncertainties on different DOEs was also quantified experimentally.

On the basis of the experimental results, a preliminary discussion on how to select an appropriate DOE for data acquisition under the constraints of fixed data volume and a given level of measurement uncertainty is presented. Our study shows that space-filling design and iterative sampling strategy outperform conventional pre-determined DOE schemes for exploring tasks. The iterative sampling strategy is even superior to space-filling design in an ideal scenario with almost no uncertainty (<1%). However, when the effects of process uncertainty cannot be ignored (>3%), model-based iterative sampling strategy requires a certain amount of initial data to obtain a functional kernel model. In such circumstances, space-filling strategy is a safe alternative, particularly when data resources are constrained. Furthermore, we give recommendations on how to correctly drive a model-based iterative sampling strategy.

In subsequent work, we will extend this research procedure to multiple models of varying complexity with a view to generalizing our conclusions about the DOE selection. Other sorts of process uncertainties will be taken into account.

CODE AVAILABILITY

The data generation scripts and the model training scripts mentioned in the paper and the associated data are compiled on Github: https://github.com/xinchengxxc/Small-Dataset-Acquisition-for-Machine-Learning-Analysis.

ACKNOWLEDGMENT

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project KIproBatt (grant number 03XP0309C) and the interdisciplinary PhD school (iDOK) at the University of Applied Sciences Aschaffenburg.

REFERENCES

[1] X. Xu, F. Conrad, A. Gronbach, and M. Möckel, "Small Dataset Acquisition for Machine Learning Analysis of Industrial Processes with Possible Uncertainties" The Ninth International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2023) IARIA, Apr. 2023, pp. 35-38, ISSN: 2519-8386, ISBN: 978-1-68558-041-4.

[2] F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer, and S. Ihlenfeldt, "Benchmarking AutoML for regression tasks on small tabular data in materials design", Sci Rep, vol. 12, no. 1, Art. no. 1, pp. 19350, Nov. 2022, doi: 10.1038/s41598-022-23327-1.

[3] Figure Eight. *CrowdFlower: Data science report*. [Online]. Available from: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf [retrieved: 02, 2023].

[4] R. Arboretti, R. Ceccato, L. Pegoraro, and L. Salmaso, "Design of Experiments and machine learning for product innovation: A systematic literature review", Quality and Reliability Engineering International, vol. 38, no. 2, pp. 1131–1156, Nov. 2022, doi: 10.1002/qre.3025.

[5] R. Arboretti, R. Ceccato, L. Pegoraro, and L. Salmaso, "Design choice and machine learning model performances", Qual Reliab Eng, vol. 38, pp. 3357-3378, Jan. 2022, doi: 10.1002/qre.3123.

[6] R. Fontana, A. Molena, L. Pegoraro, and L. Salmaso, "Design of experiments and machine learning with application to industrial experiments", Stat Papers, vol. 64, pp. 1251-1274, Mar. 2023, doi: 10.1007/s00362-023-01437-w.

[7] A. Dean, D. Voss and D. Draguljić, Design and Analysis of Expermients, 2nd Edition. New York, NY: Springer, 2017.

[8] KIproBatt. *Exploring smart battery cell production based on a generic system architecture and an AI-enhanced process monitoring*. [Online]. Available from: https://doi.org/10.13140/RG.2.2.11573.76006, 2023.11.13.

[9] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: hands-free AutoML via meta-learning", J. Mach. Learn. Res., vol. 23, no. 1, p. 261:11936-261:11996, Jan. 2022.

[10] J. Fleischer, G. Lanza and K. Peter, "Quantified Interdependencies between Lean Methods and Production Figures in the Small Series Production," Manufacturing Systems and Technologies for the New Frontier, pp. 89–92, 2008, doi: 10.1007/978-1-84800-267-8_17.

[11] M. Westermeier, *Qualitätsorientierte Analyse komplexer Prozessketten am Beispiel der Herstellung von Batteriezellen.* [online]. Available from: https://www.mec.ed.tum.de/fileadmin/w00cbp/iwb/Institut/Dissertationen/322_Westermeier_Markus.pdf [retrieved: 02, 2023].

[12] D. Chicco, M. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation", PeerJ Computer Science, 7:e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[13] M. Feurer et al., "Efficient and Robust Automated Machine Learning", Advances in Neural Information Processing Systems, vol. 2, pp. 2962-2970, Dec. 2015, doi: 10.1007/978-3-030-05318-5_6.

[14] F. Fabris and A.A. Freitas, "Analysing the Overfit of the Auto-sklearn Automated Machine Learning Tool", In: International Conference on Machine Learning, Optimization, and Data Science, 2019. [Online]. Available from: https://api.semanticscholar.org/CorpusID:210131212.

[15] R.A. Fisher, The Arrangement of Field Experiments in Breakthroughs in Statistics. New York, NY: Springer, 1992.

[16] G. Box and D. W. Behnken, "Some New Three Level Designs for the Study of Quantitative Variables", Technometrics, vol. 2, no. 4, pp. 455-475, Nov. 1960, 2:4, doi: 10.1080/00401706.1960.10489912.

[17] F. Viana, "A Tutorial on Latin Hypercube Design of Experiments", Qual. Reliab. Engng. vol. 32, pp. 1975-1985, Nov. 2015, doi: 10.1002/qre.1924.

[18] J.-S. Park, "Optimal Latin-hypercube designs for computer experiments", Journal of Statistical Planning and Inference, vol. 39, no. 1, pp. 95-111, Apr. 1994, doi: 10.1016/0378-3758(94)90115-5.

[19] L. Salmaso et al., "Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study", Communications in Statistics - Simulation and Computation, vol. 51, no. 2, pp. 570-582, Feb. 2022, doi: 10.1080/03610918.2019.1656740.

[20] Z. Liu et al., "Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing", Joule, vol. 6, no. 4, pp. 834-849, Apr. 2022, doi: 10.1016/j.joule.2022.03.003.

[21] M. Zhang, A. Parnell, D. Brabazon, and A. Benavoli, "Bayesian Optimisation for Sequential Experimental Design with Applications in Additive Manufacturing". arXiv, Nov. 23, 2021. doi: 10.48550/arXiv.2107.12809.

[22] A. Paleyes et al., "Emulation of physical processes with Emukit". arXiv, Oct. 25, 2021. doi: 10.48550/arXiv.2110.13293.

[23] P.F. de Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart, and R. Phan-Than-Luu, "D-optimal designs", Chemometrics and Intelligent Laboratory Systems, vol. 30, no. 2, pp. 199-210, Oct. 1994, doi: 10.1016/0169-7439(94)00076-X.

[24] C. E. Rasmussen, Gaussian processes in machine learning in Advanced Lectures on Machine Learning (ML 2003), pp. 63-71. Berlin, Heidelberg: Springer, 2004. US

[25] G. Kopsiaftis, E. Protopapadakis, A. Voulodimos, N. Doulamis, and A. Mantoglou, "Gaussian Process Regression Tuned by Bayesian Optimization for Seawater Intrusion Prediction", Computational Intelligence and Neuroscience, vol. 2019, p. e2859429, Jan. 2019, doi: 10.1155/2019/2859429.

[26] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and Analysis of Computer Experiments", Statistical Science, vol. 4, no. 4, pp. 409-423, Nov. 1989, doi: 10.1214/ss/1177012413.

[27] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, "Active Learning for Gaussian Process Considering Uncertainties With Application to Shape Control of Composite Fuselage," in IEEE Transactions on Automation Science and Engineering, vol. 18, no. 1, pp. 36-46, Jan. 2021, doi: 10.1109/TASE.2020.2990401.

[28] R. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. Ngo, "Predicting sample size required for classification performance", BMC Med Inform Decis Mak, vol. 12, no. 8, Feb. 2012, doi: 10.1186/1472-6947-12-8.

[29] P. Austin and E. Steyerberg, "The number of subjects per variable required in linear regression analyses", J Clin Epidemiol, vol. 68, no. 6, pp. 627-636, Jan. 2015, doi: 10.1016/j.jclinepi.2014.12.014.