5th Conference on Production Systems and Logistics

Machine Learning Driven Design Of Experiments For Predictive Models In Production Systems

Sebastian Maier[1], Patrick Zimmermann[1], Rüdiger Daub[2]
[1]*Fraunhofer Institute for Casting, Composite and Processing Technology IGCV, Augsburg, Germany*
[2]*Institute for Machine Tools and Industrial Management, Technical University of Munich*

Abstract

Machine learning (ML) describes the ability of algorithms to structure and interpret data independently or to learn correlations. The use of ML is steadily increasing in companies of all sizes. However, insufficient market readiness of many ML solutions inhibits their application, especially in production systems. Predictive models apply ML to understand the complex behavior of a system through regression from operational data. This enables determining the relationship between factors and target variables. Accurate predictions of these models for production systems are essential for their application, as even minor variations can significantly affect the process. This accuracy depends on the available data to train the ML model. Production data usually shows a high epistemic uncertainty, leading to inaccurate predictions unfit for real-world applications. This paper presents ML-driven, data-centric Design of Experiments (DoE) to create a process-specific dataset with low epistemic uncertainty. This leads to improved accuracy of the predictive models, ultimately making them feasible for production systems. Our approach focuses on determining epistemic uncertainty in historical data of a production system to find data points of high value to the ML model in the factor space. To identify an efficient set of experiments, we cluster these data points weighted by feature importance. We evaluate the model by running these experiments and using the collected data for further training of a prediction model. Our approach achieves a significantly higher increase in accuracy compared to continuing the training of the prediction model with the same amount of regular operating data.

Keywords

Machine Learning; Design of Experiments; Epistemic Uncertainty; Predictive Models; Production Systems

1. Introduction

At the latest since the publication of ChatGPT by OpenAI, Machine Learning (ML) has received much attention [1]. This type of generative artificial intelligence (AI) is considered a technical revolution with a potential impact on a wide variety of industries and society as a whole [2]. In line with this trend, the use of AI continues to increase in companies of all sizes, although primarily in large companies with more than 1000 employees [3]. Barriers to deploying AI in companies include a lack of expertise in the ML area, insufficient market maturity of AI solutions, and a lack of data [4].

After the IT industry, manufacturing ranks second among the most important application areas for AI in Germany [5]. However, compared to the IT industry, manufacturing requires significantly more effort to generate data on the operational technology (OT) level and to transfer it to the application level. It becomes clear that the algorithms behind ML models might not always be the limiting factor but rather the limited data available to train these models. Instead of improving a model on a given dataset, data-centric AI focuses

DOI: https://doi.org/10.15488/15289
ISSN: 2701-6277

on data quality and reliability [6]. The methodology presented in this paper utilizes this data-centric approach and addresses the creation of robust predictive ML models for complex production processes.

2. Data in production systems and its effect on predictive models

Production systems use actuators to control the process and sensors to generate data. These input data points are called factors and consist of disturbance variables and control variables, as shown in Figure 1.

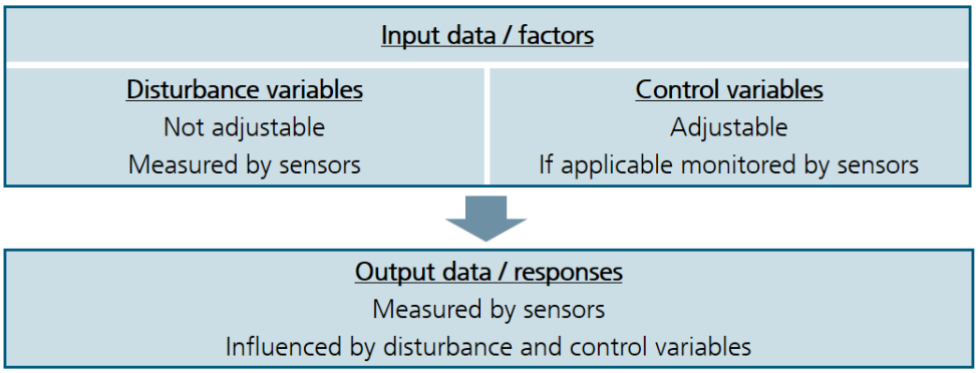

Figure 1: Types of data in a production system

Disturbance variables are factors that the operator of a production system cannot adjust. In contrast, control variables can be set to predefined values. Control variables are actuators that are often equipped with sensors to monitor their current state. All factors influence one or more output data points called responses. In production systems, responses are usually one or more measurable product quality characteristics. Production systems typically rely on the empirical knowledge of experts and plant personnel who adjust the control variables to achieve the desired responses.

In this work, we specifically address a subset of ML, namely predictive models, which learn the behavior of the plant using regression. These models use historical data collected during the operation of the production system to determine the relationship between factors and responses. Thus, the trained predictive model can quantify the effect of factor changes on responses independently of the process experts' empirical knowledge.

Since the predictions take place simulatively without affecting the real-world production system, finding the factors for the desired responses based on the predictions of a trained model is an optimization problem. Given the current disturbance values of the real-world production system, a set of control variables is sought, reproducing the desired responses according to the prediction model as accurately as possible. These simulatively determined control variables can be set in the actual production system afterward. The goal of this optimization is to avoid scrap or to increase efficiency. In production, this method is called predictive quality. However, applying these predictive models in production is associated with some challenges.

Most of the data generated in production systems is available as time series data. Programmable Logic Controllers (PLCs) cyclically retrieve individual measured values and provide them with a time stamp. The combination of all process factors forms the input space, also called design space or factor space. Figure 2 shows a simplified 2D factor space. In this simple example, the factors x1 and x2 create the factor space. Within the green area there are data points, represented as black dots. These data points are recordings during the regular operation of the production system. The distribution of data points within the green area is not homogeneous, so there are areas with an accumulation of data points as well as areas with few or no data points.

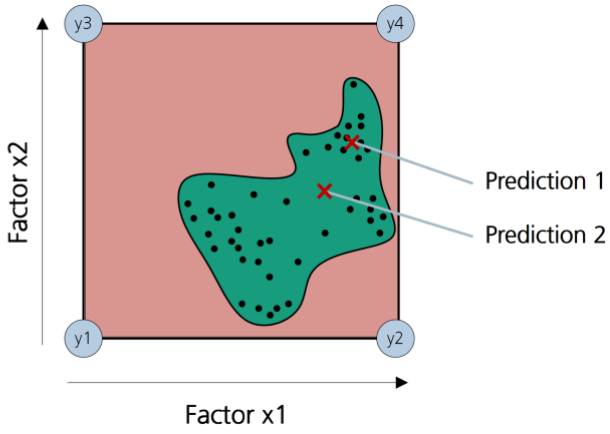

Figure 2: 2D factor space with data points illustrated as black dots

This distribution of data for training predictive models significantly impacts their robustness [7]. Areas with many data points lead to accurate predictions due to the prevalence of low uncertainty, like prediction 1 in Figure 2. Prediction 2 in Figure 2, on the other hand, lies in an area with few data points nearby, leading to higher uncertainty and less accurate predictions. This type of uncertainty arises from various forms of insufficient knowledge or incomplete data and is referred to as “epistemic uncertainty” [8]. Therefore, providing additional data reduces epistemic uncertainty. Predictions within the green range can also have high epistemic uncertainty and thus produce inaccurate predictions. All predictions within the green region are also referred to as "in-distribution" predictions [9]. However, this does not mean they are independent and identically distributed [10]. Accurate predictions within this range are essential for applying predictive models since even slight fluctuations can significantly affect the process. The prediction of desired target variables, therefore, only allows minor errors. High epistemic uncertainty leading to high inaccuracy makes predictions unusable for production systems.

Predictions in the red region of the factor space have a very high epistemic uncertainty and produce incorrect predictions because of factor compositions that the training data does not cover. All predictions within the red region are referred to as "out-of-distribution" (OOD) predictions [9,10]. Valid predictions in the OOD region can still be highly relevant in applications where process experts want to identify new operating conditions. In this case, the objective is to find factor compositions in the red region that are more efficient than previous operating regions "in-distribution" or more consistently achieve the desired product quality.

Reducing epistemic uncertainty with new data in the factor space is difficult to achieve in an application. Instead of only having two factors, as shown in Figure 2, real-world applications often deal with tens or even hundreds of factors. This high dimensionality and the often complex interrelationship of the factors limit the applicability in production systems. This work focuses on improving “in-distribution” predictions through new data points. We present a method that evaluates feature importance and model uncertainty in the factor space to create an effective set of experiments. The effect of gathering data during these experiments shows a more significant decrease in model error compared to training the model further with regular training data.

3. State of the art

This chapter presents the state of the art regarding our research. First, we discuss the role of uncertainties, feature importance, and their corresponding methods for explainable AI. Subsequently, we point out current research in the field of ML-based experimental design.

3.1 Understanding ML models

Explainable AI aims to provide interpretable insights into the decision-making process of AI models to understand them better. Explainability is particularly important in production systems since incorrect predictions can cause property damage or even cause injuries. There are many tools and methods to evaluate the correctness and relevance of a prediction. In this work, however, we focus on two approaches: uncertainty and feature importance. Feature importance indicates the influence of individual factors on the responses. In the context of feature importance, global model-agnostic methods describe the average behavior of models [11]. The most prominent global model-agnostic method is permutation feature importance. By randomly shuffling a factor's value and measuring the model error, permutation importance provides the importance of each factor for the responses [12]. Local model-agnostic methods, on the other hand, explain individual model predictions [11]. SHapley Additive exPlanations (SHAP) is a method that uses Shapley values, known from coalitional game theory, to find the contribution of each factor to the prediction [13].

According to [14], the shell model of uncertainties divides the uncertainty of AI or ML applications into three areas: The innermost layer, "Model Fit," indicates whether the used model can represent the complexity of the data. The subsequent layer, "Data Quality," indicates whether the data contains the information and correlations related to the target variable. The outermost layer, "Scope Compliance," includes uncertainty due to differences between the model and application context. This work concerns the "Data Quality" layer, which describes epistemic uncertainty. Bayesian methods are essential in determining the amount of uncertainty [15,16]. However, since Bayesian approaches are often more computationally expensive and less performant than conventional ML models, [17] demonstrated that dropout in neural networks can be used to approximate uncertainty, combining the advantages of conventional ML with Bayesian methods.

3.2 ML-driven Design of Experiments

Factorial or screening experimental plans quickly reach their limits when there are many factors and complex processes. ML-based Design of Experiments (DoE) attempts to solve this problem and helps to determine process interrelationships through targeted experiments. Determining the most efficient experiments is called "Optimal Experimental Design" (OED). Approaches in this area use, for example, iterative exploration of areas in the factor space with high information content by predicting the variance [18] or introducing estimators for the information gain of an experiment [19]. ML-based DoE already finds application in the field of materials science [20], chemical reactions [21], or in the selection of fast-charge protocols [22]. However, these approaches are limited to discrete factor spaces and do not account for continuous factor spaces as they prevail in production systems.

4. Methodology

This section introduces our ML-driven DoE for predictive models in production systems. Figure 3 shows the workflow of the method.

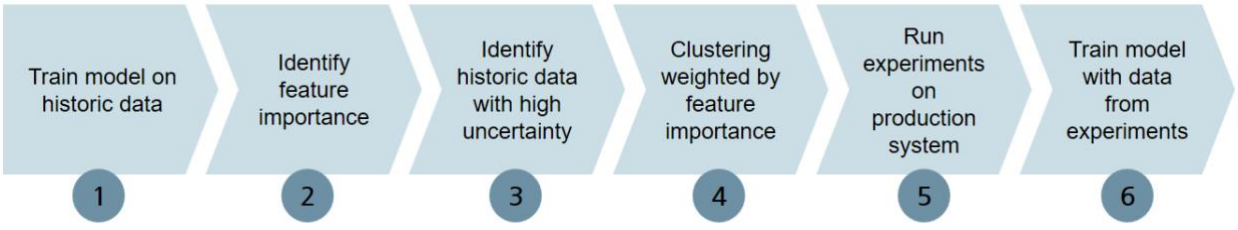

Figure 3: Workflow for the ML-driven approach to DoE for predictive models in production systems

For this method to be applicable, the production system must run and provide the data generated during the process. All data points are snapshots of the production system at a certain time and consist of *n* factors. The first step is to train an ML model using historical data from the operation of the production system. The type of selected regression model must either provide the uncertainties in its predictions (i.e. Bayesian methods) or allow to calculate the uncertainty of its prediction with additional methods (i.e. neural networks with the approximation of [17]). These requirements are important at a later stage in the workflow. In some cases, this regression model might already produce sufficient predictions. However, the predictions are too inaccurate for many use cases in production systems, which is where our method comes into place. In step two, we calculate the feature importance of all control variables. Since we are interested in the feature importance over all data points, we apply a global model-agnostic method like permutation importance.

In the third step, we determine the uncertainty of the model over historical data points. Ideally, these data points are a test set which the model has not been trained on. Then we pick *m* data points with the highest uncertainty for further investigation in the next step. The amount of data points to choose varies depending on the number of experiments and fluctuation in the data. If we plan a high amount of experiments and the data fluctuates considerably, an increased number of data points is necessary for step 4.

Since it is expensive to perform many experiments, we do not want to run separate experiments for *m* data points. Instead, we want to design an efficient set of experiments that generate as much useful information for the ML model as possible. This is why we cluster the data points with high uncertainty in step 4. By clustering similar data points, we aim to compress the information of *m* data points into fewer experiments. Algorithms like k-means minimize the variance of data points within a cluster. In this case, we want to minimize the variance of all control variables since we cannot influence disturbance variables during experiments. Additionally, we weight all control variables by their feature importance. This ensures efficiency since factors with a high impact on model estimates also have a higher impact during clustering. To get a single experiment from each cluster, we calculate the average for the factors in each cluster. The set of all averaged clusters acts as an experiment plan for the production system. This plan provides a set of experiments that target areas in the factor space where the model is uncertain while focusing on the factors that influence the responses the most.

In step 5, we set the control variables according to the first experiments from the experiment plan and record all data during production. After completing all experiments, we use the data from the experiments to further train the model in step 6.

5. Evaluation

In this chapter, we evaluate our method based on its effect on the error of predictive models in production systems. We apply the method at a geothermal power plant (GPP) that produces power by extracting heat from hot thermal water. We use 22 sensors in the process. The process operators have written the operational data cyclically into a database every 5 minutes for 268 days. These sensor values are the historical data and consist of different physical properties like water flow rate, pressure, temperatures, etc. Out of these 22 values, we use three adjustable factors as the control variables for our method. In addition, one of the values that the process operators recorded is the net energy gain of the GPP, which acts as the response of the predictive model. After training a basic feed-forward neural network on this data, we evaluate the feature importance of all control variables using permutation importance. Table 1 shows the feature importance of all control variables.

Table 1: Feature Importance of the control variables

Control variable	Feature Importance
Number of active air condensers	33.6068
Blade angle of air condenser ventilators	32.9897
Receiver condensate level	29.4302

In the next step, we predict the net energy gain based on unseen data. We use the dropout method of [17] to estimate the epistemic uncertainty of these predictions. Figure 4 shows the uncertainty of the predictions over a period of 30 days.

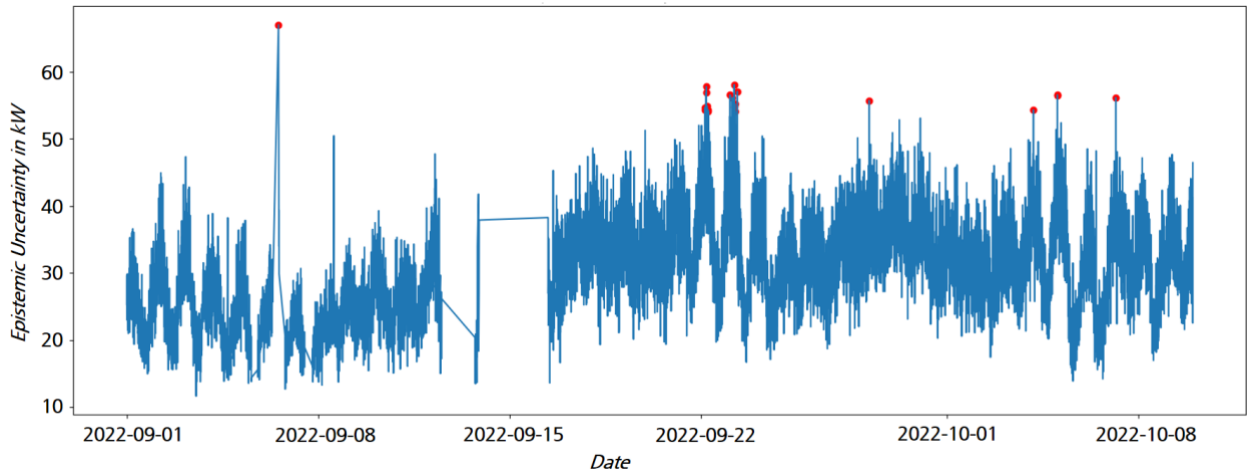

Figure 4: Epistemic uncertainty of the model on test data

Since maintenance work was done on the GPP during this time, we left out some uncertainty data to avoid faulty estimates. The red dots mark 20 predictions with the highest epistemic uncertainty. To reduce the total number of experiments, we build five clusters from these 20 data points weighted by their feature importance and calculate the average of each cluster. The five clusters act as five separate experiments. Therefore, we compress the information of twenty data points into only five experiments. We adjust each control variable according to an experiment and run the GPP with these settings for 24 hours. We record all 21 factors and the corresponding response during the experiments for further training of the feed-forward neural network.

We use the root-mean-square error (RMSE) between the predictions and the actual net power output to evaluate the performance of the model. Table 2 shows the RMSE of three models, which all share the same hyperparameters. We train model 1 on just 268 days of historical data. Model 2 utilizes an additional five days of regular operational data from the GPP. Model 3 uses the 268 days of historical data and the five days of experiments we performed based on our method. This means we trained models 2 and 3 on the same amount of data, whereas model 1 uses five fewer days of data.

Table 2: Evaluation of the models based on their RMSE

Model	RMSE
Model 1: 268 days of data of regular operation data	12.006
Model 2: 268 regular + 5 days of regular operation data	11.935
Model 3: 268 regular + 5 days of experiment data (our method)	10.294

As expected, model 1 with an RMSE of 12.006 performs worse than models 2 and 3 since we provide less training data. While model 2 achieves an RMSE drop of ~0.6% to 11.935, the RMSE of model 3 drops to 10.294, a decrease of ~14.3%. The five days of experimental data provided by our method are more beneficial to the model than five days of regular operating data. This shows that a data-driven approach to predictive models can help lower the error of predictions and make these models more feasible for production systems.

6. Conclusion and future work

It is vital to decide which ML model to use and to find its best hyperparameters for any type of application. However, instead of focusing on models and algorithms, research in data-driven AI focuses on data quality and reliability. In this work, we introduce a method for ML-driven DoE for predictive models in production systems to generate high-quality training data yielding lower prediction errors. We achieve this by considering the epistemic uncertainty of model predictions and clustering them weighted by feature importance. With this, we create data with high value to the model while keeping experimental efforts as low as possible. We have shown that our method yields significantly higher-quality data than regular operation.

This work focused on "in-distribution" predictions, meaning we generated data within the range of regular operation of the production system. However, in some cases, the optimal point of operation does not lie within this area of the factor space. Exploring OOD predictions might produce higher efficiency, better product quality, or higher overall production. Therefore, in future work, we want to focus on OOD predictions to find new operating conditions in production systems.

Acknowledgments

This research has been conducted within the scope of the project "REIF - Resource-efficient, Economic and Intelligent Foodchain," funded by the German Federal Ministry for Economic Affairs and Climate Action. We also want to thank the "Geothermische Kraftwerksgesellschaft Traunreut mbH" for providing the data of their production process and conducting the experiments necessary for this research.

References

[1] J. Schulman et al., 2022. ChatGPT: Optimizing Language Models for Dialogue.

[2] Faraj, S., Pachidi, S., Sayegh, K., 2018. Working and organizing in the age of the learning algorithm. Information and Organization 28 (1), 62–70.

[3] 2021. Mannheimer Innovationspanel: Studie n=1521.

[4] 2022. Künstliche Intelligenz – Wo steht die deutsche Wirtschaft?: Studie n=606.

[5] 2020. State of AI in the Enterprise: Studie n=168.

[6] Zha, D., Bhat, Z.P., Lai, K.-H., Yang, F., Hu, X., 2023. Data-centric AI: Perspectives and Challenges, in: Banerjee, A., Zhou, Z.-H., Papalexakis, E.E., Riondato, M. (Eds.), Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, SIAM, Philadelphia, Pennsylvania, pp. 945–948.

[7] Wiesbaden, S.F., 2013. Kompakt-Lexikon Wirtschaftsmathematik und Statistik: 750 Begriffe nachschlagen, verstehen, anwenden, 1st ed. ed. Springer Gabler, Wiesbaden, 157 pp.

[8] Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 110 (3), 457–506.

[9] Arjovsky, M., 2021. Out of Distribution Generalization in Machine Learning. https://arxiv.org/pdf/2103.02667.

[10] Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P., 2021. Towards Out-Of-Distribution Generalization: A Survey. https://arxiv.org/pdf/2108.13624.

[11] Molnar, C., 2022. Interpretable machine learning: A guide for making black box models explainable, Second edition ed. Christoph Molnar, Munich, Germany, 318 pp.

[12] Breiman, L., 2001. Random Forests. Mach Learn 45 (1), 5–32.

[13] Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions.

[14] Kläs, M., Vollmer, A.M., 2018. Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty, in: Computer Safety, Reliability, and Security. Springer International Publishing, Cham, pp. 431–438.

[15] Bickel, P.J., Blackwell, D., 1967. A Note on Bayes Estimates. Ann. Math. Statist. 38 (6), 1907–1911.

[16] Weiss, R.E., 2010. Bayesian methods for data analysis. American journal of ophthalmology 149 (2), 187-188.e1.

[17] Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: Proceedings of The 33rd International Conference on Machine Learning. PMLR, New York, New York, USA, pp. 1050–1059.

[18] Cohn, D.A., 1996. Neural Network Exploration Using Optimal Experiment Design. Neural networks : the official journal of the International Neural Network Society 9 (6), 1071–1083.

[19] Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y.W., Rainforth, T., Goodman, N., 2019. Variational Bayesian optimal experimental design.

[20] Dieb, T.M., Tsuda, K., 2017. Machine Learning-Based Experimental Design in Materials Science, in: Tomlinson, M., Jung Tjhai, C., Ambroze, M.A., Ahmed, M., Jibril, M. (Eds.), Error-Correction Coding and Decoding: Bounds, Codes, Decoders, Analysis and Applications. Springer, Cham, pp. 65–74.

[21] Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R.S., Cooper, A.I., 2020. A mobile robotic chemist. Nature 583 (7815), 237–241.

[22] Attia, P.M., Grover, A., Jin, N., Severson, K.A., Markov, T.M., Liao, Y.-H., Chen, M.H., Cheong, B., Perkins, N., Yang, Z., Herring, P.K., Aykol, M., Harris, S.J., Braatz, R.D., Ermon, S., Chueh, W.C., 2020. Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature 578 (7795), 397–402.

Biography

Sebastian Maier is a research associate at the Fraunhofer Institute for Casting, Composite, and Processing Technology IGCV since 2020 and works in the Digital Engineering and AI group. He studied mechanical engineering at Ulm University of Applied Sciences and Applied Research on Computer Science at Augsburg University of Applied Sciences.

Patrick Zimmermann is a research associate at the Fraunhofer Institute for Casting, Composite, and Processing Technology IGCV since 2018 and works in the area of plant connectivity, data models for operational technology, and process optimization. He studied mechanical engineering at the Technical University of Munich.

Rüdiger Daub is the holder of the Chair of Production Engineering and Energy Storage Systems at the Technical University of Munich and director of the Fraunhofer IGCV. After studying Electrical Engineering and Information Technology at the Karlsruhe Institute of Technology (KIT), he earned his doctorate at the Technical University of Munich (TUM). From 2012 to 2021, he held several positions at BMW Group Munich, most recently as Head of Technology Development and Prototype Construction of Lithium-Ion Cells.