Is MLOps different in Industry 4.0? General and Specific Challenges

Leonhard Faubel, Klaus Schmid and Holger Eichelberger
University of Hildesheim, Germany

Keywords: MLOps, Machine Learning, Industry 4.0, Challenges.

Abstract: An important part of the Industry 4.0 vision is the use of machine learning (ML) techniques to create novel capabilities and flexibility in industrial production processes. Currently, there is a strong emphasis on MLOps as an enabling collection of practices, techniques, and tools to integrate ML into industrial practice. However, while MLOps is often discussed in the context of pure software systems, Industry 4.0 systems received much less attention. So far, there is no specialized research for Industry 4.0 in this regard. In this position paper, we discuss whether MLOps in Industry 4.0 leads to significantly different challenges compared to typical Internet systems. We identify both context-independent MLOps challenges (general challenges) as well as challenges particular to Industry 4.0 (specific challenges) and conclude that MLOps works very similarly in Industry 4.0 systems to pure software systems. This indicates that existing tools and approaches are also mostly suited for the Industry 4.0 context.

1 INTRODUCTION

Industry 4.0 is aimed at the next industrial evolution in manufacturing, this time based on digital technologies. A core part of it is the use of machine learning (ML) to enable more intelligent, flexible, and efficient industrial processes. Scenarios like lot-size one, predictive maintenance, or supply-chain optimization can significantly transform business models in Industry 4.0 (Borgmeier et al., 2017). Currently, Machine Learning Operations (MLOps) as a collection of methods, techniques, and tools for integrating ML into software development practice is widely discussed as an enabler for large-scale ML applications, however, this happens mostly in the context of pure software systems.

As Industry 4.0 aims at applying ML to industrial production, the need for MLOps in this context is clear. Thus, an important question is whether the specific context of Industry 4.0, i.e., complex, large-scale Cyber-Physical Systems (CPS), changes the challenges to the application of MLOps. The aim of this paper is to discuss challenges of applying MLOps in an Industry 4.0 context. As a result we identify challenges to MLOps that are specific to the Industry 4.0 context or to specific scenarios within Industry 4.0 (specific challenges) and challenges that are roughly comparable to MLOps in other contexts (general challenges). This serves as a basis for finding solutions to address the identified novel challenges.

The next section introduces our understanding of MLOps, which relies on existing models, but with an adaptation to the Industry 4.0 context. Section 3 is the core of the paper and presents the challenges we could identify. We discuss these from a cross-sectional perspective in Section 4, while Section 5 concludes.

2 MLOps

MLOps aim to enhance the automation and quality of intelligent systems (Meedeniya and Thennakoon, 2021). It combines principles from DevOps with machine learning. The flexibility provided by DevOps principles is beneficial to machine learning (ML) as, typically, several iterations need to occur to identify well-working ML models and then adapt them over time as the situation changes in the application.

The first step in MLOps is always manual and is performed in an analysis environment (Fig. 1). This serves to first understand the problem and possibilities of ML. Based on the results of this step, a high-level architecture in which MLOps operate must be created. This takes the ML method, SE architecture, hardware architecture, configuration, and, if applicable, even the architecture of the CPS into account and

Faubel, L., Schmid, K. and Eichelberger, H.
Is MLOps different in Industry 4.0? General and Specific Challenges.
DOI: 10.5220/0011589600003329
In *Proceedings of the 3rd International Conference on Innovative Intelligent Industrial Production and Logistics (IN4PL 2022)*, pages 161-167
ISBN: 978-989-758-612-5; ISSN: 2184-9285

may even evolve them, if necessary. In particular, this defines the deployment of the various MLOps components and under which conditions they are triggered. One input to this architecture definition is whether automated retraining of models should be supported (Step B). This is on the hand dependent on the outcome of the model development and on the other it depends on business decisions. Also the type of algorithms that are identified in the model development significantly influence the high-level system architecture, e.g., neural networks have very different resource implications vs. random forest classification. Based on monitoring information, automated model adaptations can be triggered (automated MLOps) or it can be signaled that such adaptations may need to be done manually, leading to a redefinition of the machine learning approach. Finally, there is always a model application stage (Step C), which will typically, but not always, be performed on edge devices. The details of the architecture will have a strong influence on the hardware resources as well as the ML components used, as we will discuss in Section 3.

While Fig. 1 provides a high-level overview of MLOps in Industry 4.0, we need a more detailed MLOps model to define the individual challenges. Various life-cycle models have been proposed for MLOps (van der Goes, 2021). Two MLOps life cycles are predominant in the literature. Symeonidis et al. depict an MLOps life-cycle with three stages: ML, development, and operations (G. Symeonidis et al., 2022). Van der Goes describes a variant with four stages (van der Goes, 2021). Here, the ML stage is subdivided into data management and modeling. Each stage consists of a cycle with tasks that connect cycles to each other. These models do not address the specific activities of MLOps, but make it clear that ML is added to DevOps principles. Moreover, they do not include cross-cutting activities. Therefore, we propose a new life-cycle. Our life-cycle model defines the phases (Fig. 2): Data Engineering, Model Engineering, and Operations, each consisting of multiple activities; they are complemented by supporting activities.

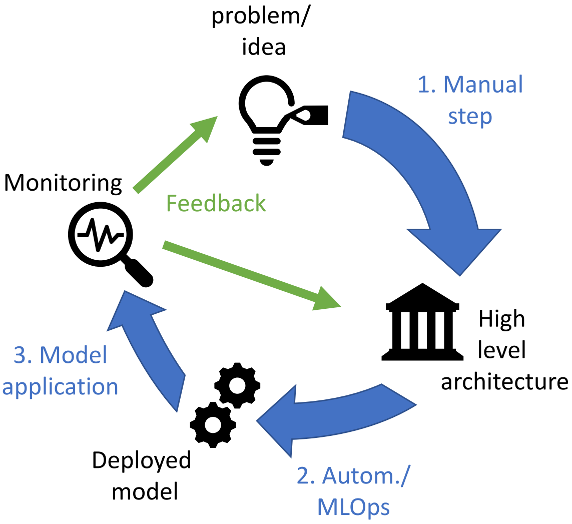

Figure 1: High level structure of the MLOps cycle.

Data Engineering

Data Collection: This provides a basis for machine learning. The collected data can include machine data, product information, customer data, behavioral data, etc. The relevant data needs to be determined based on the use case and typically provides insights into the effectiveness and quality of the production process.

Data Analysis: This step aims at understanding of the data and its quality, e.g., identifying outliers. Typically, statistical methods are used.

Data Preparation: Transformations are applied, e.g, for data cleaning or value imputation. Feature transformation can also be done here (Cardoso Silva et al., 2020).

Model Engineering: In the structure given in Fig. 1 this can either be performed manually (especially in the first iteration) or in an automated way.

Model Building: Model building aims at creating the necessary machine learning models. This includes the identification of the relevant approaches (e.g., neural networks vs. decision trees), determining corresponding model structures, and potentially determining hyperparameters.

Model Training: Candidate ML models are trained and fitted to the data.

Model Evaluation: ML models are evaluated on test data (Sun et al., 2022).

Model Selection: The most appropriate (usually the best best performing) ML model is selected (or multiple models, if there are several problems, which are addressed by ML techniques). Potentially further fine-tuning is performed (Cardoso Silva et al., 2020).

Model Packaging: The final ML models are packaged as one or more application components or as a ”model as a service” (Sato et al., 2019).

Operations

CI/CD-testing: As part of continuous integration and deployment, special tests for features and data, for models, for ML infrastructure, and

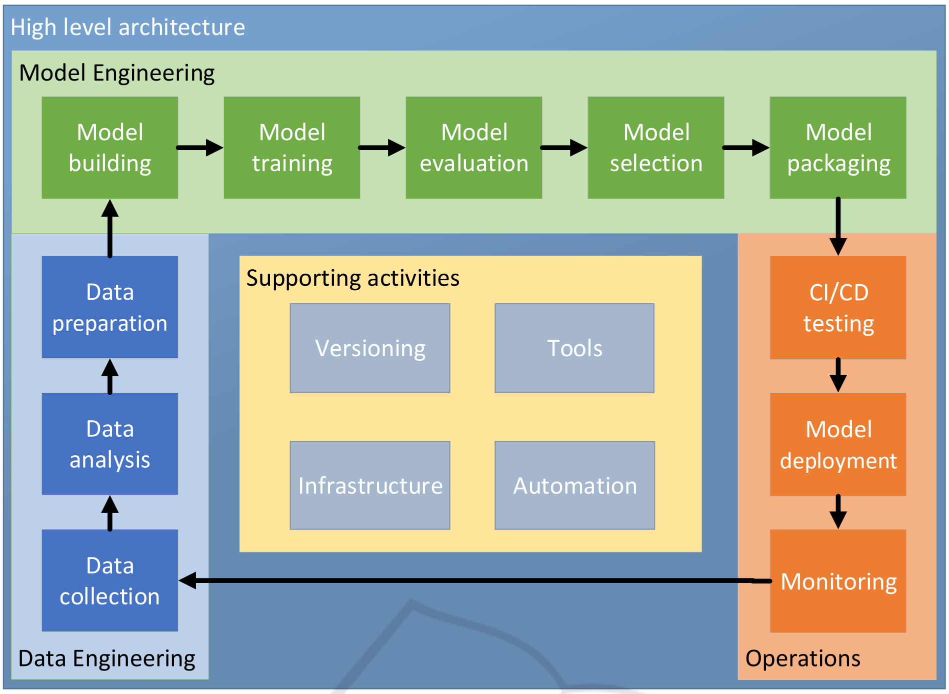

Figure 2: MLOps activities.

monitoring for ML are run to ensure quality of the deployed system.

Model Deployment: The production-ready ML models are deployed as part of the production system (Rahman and Kandogan, 2022).

Monitoring: The performance (quality) of the ML models is continuously monitored in order to determine whether a manual or automated intervention is necessary (Rahman and Kandogan, 2022). Although ML methods are highly appreciated in Industry 4.0 settings, according to our experience, the application in production settings is sometimes a bit conservative, i.e., automatically adjusting an existing or deploying a new version of a working ML model is judged rather sceptically. Thus, at least the option to manually approve such an intervention is often requested.

Supporting Activities

Infrastructure: The necessary infrastructure components, including the relevant hardware, required to develop, deploy, and run complex ML systems must be selected, acquired, installed and maintained. Often, this also involves data integration activities, e.g., in Industry 4.0 settings to obtain additional data from ERP or MES systems.

Versioning: Versioning of code, data, and the ML model as well as configuration information is needed (Oluyisola et al., 2022). Versioning is often desirable to be able to go back to the last working model to avoid or at least minimize (factory) downtimes. Versioning is also viewed as a ML safety mechanism, in particular if manual approvals of changed ML models are demanded.

Automation: Various steps in the overall lifecycle are often automated. This requires the implementation of this steps, either using existing automation capabilities or implementing them in special ways.

Tools: Tools for developing of ML applications are needed. This can include domain-specific tools like domain-oriented modeling or simulation (e.g., for factories or machines in an Industry 4.0 scenario).

3 CHALLENGES

This section presents challenges encountered when using MLOps in an Industry 4.0 environment. We identified them based on 2.5 years of project experience as well as analysis of the involved tasks. The challenges are organized based on the MLOps activities model in Fig. 2. The focus of the discussion is always: which activity does not cause additional difficulties and what additional difficulties exist in MLOps for Industry 4.0 over more traditional MLOps scenarios. Of course, this may vary according to project-specific requirements. The challenges are divided into general and specific challenges. General challenges also exist in situations in other contexts, while specific challenges relate exclusively to case-dependent specific situations in Industry 4.0 scenarios.

3.1 Data-related Tasks

Here, we describe the challenges in the data-related MLOps tasks.

3.1.1 Data Collection

Depending on the use case specific substantial technical challenges exist here. Data acquisition requires suitable sensors and data transmission in the factory environment since data collection starts in the manufacturing machine and ends in the software system. Depending on the application, real-time requirements exist and large amounts of data are transmitted or stored.

If suitable technical conditions exist, i.e., machines with the appropriate equipment, in which the sensors and the respective networking are adequately dimensioned, the application of MLOps is more straightforward, as one only has to think about potential problems in storing the relevant amounts of data.

If technical prerequisites are not met, data collection for MLOps becomes very difficult as conversion or adaptation of existing hardware and software may involve a significant effort, if possible at all. For example, adding additional sensors to a machine may require breaking guarantee seals or safety certifications, which would render the machine unusable or imply additional effort or costs. In the worst-case scenario, whole machines, infrastructures, or manufacturing lines must be redeployed or exchanged.

3.1.2 Data Analysis

Typically, this is performed with a sample data set outside the industry 4.0 environment. Thus, there is no unusual complexity in this task.

3.1.3 Data Preparation

In data preparation, features are provided online for inference and offline for experiments and training. This is often implemented in the form of feature stores to realize reproducibility and versioning using feature and metadata registries (Kreuzberger et al., 2022). Features usually describe characteristics of the production process. The following specific challenges arise in this regard. When data preparation is done as a pre-processing step in model inference in production real-time requirements typically apply. In this case, the frequency and size of samples required for further steps as well as the capabilities of existing hardware influences how challenging this activity is. With high real-time requirements, high-performance hardware and software measures may be required in the production environment. However, applicability and availability of such hardware may be limited by technical factory floor regulations, e.g., electrical or mechanical norms, as well as by budget restrictions.

3.2 Model-related Tasks

Here, we describe the identified challenges in the model-related MLOps activities.

Generally, if the model building, model training, model evaluation and model selection activities are part of an automated pipeline, there is a need for additional software, hardware, and infrastructure considerations (see Section 3.4.1 below).

3.2.1 Model Building, Training, Evaluation, and Selection

If these activities take place outside the Industry 4.0 environment, which is usually the case, there is no impact on regular operation. However, there is a general challenge in this respect. If the steps are performed in an automated way within a factory environment, e.g., as automated retraining, then additional hardware resources and software infrastructures like GPUs or ML implementations for embedded devices are needed within the factory.

3.2.2 Model Packaging

In model packaging, the particular constraints of the available infrastructure in the factory like hardware resources (e.g., GPUs/TPUs/ FPGAs, special operation systems, and implementations, or available network bandwidth and storage capacitites) must be taken into account.

3.3 Operations-related Tasks

The following challenges address operations-related tasks that relate to activities required for deployment and during ongoing operations in the MLOps environment.

3.3.1 CI/CD Testing

As large parts of a CI/CD testing environment will be carried out in a classical IT-environment, no challenges stem from this part. However, there is a general challenge: the heterogeneity of hardware and operating systems typical of Industry 4.0 can make these tasks more complex than usual. Also, intelligent methods require additional CI/CD tests. These include tests for features and data, model development, ML infrastructure, and monitoring tests for ML serving, e.g., performance. A particular challenge is that ML elements are typically analyzed on a statistical basis, while traditional testing requires correctness of each individual test case. Moreover, in automated adaptation scenarios even these tasks may happen within the factory environment, making this a rather complex task.

3.3.2 Model Deployment

MLOps activities can be performed on the factory floor, in the cloud, or in corporate IT environments. Specifically, in the case of the factory floor, edge resources are needed, leading to the typical problems of sufficient and appropriate resources to ensure necessary technical performance requirements.

3.3.3 Monitoring

In Industry 4.0, monitoring physical processes is an integral part. Ideally, the existing monitoring solution is suitable for MLOps or can be extended easily. If this is not the case, e.g., if physical separation of groups or networks is required, a specific challenge arises. Additional hardware requirements or development efforts become necessary.

3.4 Support Activities

Support Activities are cross-cutting activities related to infrastructure, tools, versioning and automation. Here, we address challenges refering to support activities.

3.4.1 Infrastructure

A major general challenge related to the infrastructure is the heterogeneity, which is even larger than for MLOps in information systems. Some parts need to run in an embedded context, some in corporate IT-environments. In the case that model engineering activities are also automated to some extend, there may even exist the case that for the same steps multiple different infrastructures are needed (IT vs. embedded).

If frameworks like TensorFlow are used for machine learning in the manual environment, they are typically not available as implementations on the factory floor. Rather this requires special frameworks available for edge computing that may not bring the same features, e.g., TensorFlow-light. The higher the level of automation and the requirements for the properties associated with MLOps become, the more difficult it is to deploy them in the environment of IoT and edge devices.

Another general problem is the lack of widely and homogeneously adopted standards; while there are standards, many different standards are around and applied in inhomogeneous fashion, partially also due to (expensive) legacy machines or retrofitting of factory equipment.

This also leads to a lack of standardization of tools, making tool selection in the context of Industry 4.0 a major challenge.

The extent to which parts of the MLOps cycle are automated varies significantly among cases. Of course, several steps, like deployment or productive operation are usually automated. However, so far our observation is that while many companies envision a high degree of automation, and even are interested in full automation of model adaptation, so far none, we are aware of, implemented this degree of automation in production.

We envision for the future that some degree of automated adaptation will become standard practice also in Industry 4.0. Nevertheless, difficult questions about what may be changed independently in the model and productive operation will remain. In particular, changes by self-adaptation may impact hardware requirements and reliability.

MLOps emphasizes the management of interdependencies among data, models, and code. Thus, versioning this information is important. If these information should be available at edge nodes, e.g., as feature stores, appropriate versioning infrastructure must also be available there.

3.4.2 Tools

Here, we present a general challenge. Typically, a more complex tool environment is required for MLOps in Industry 4.0. This is due to the fact that some steps will be done manually in an IT-environment, but also corresponding tools in the embedded environment are needed. Also tools that address the specifics of the industry environment (e.g., cross-compilation) are required. Special tools like simulation environments may also be needed in order to study the impact of the ML solutions. In particular, if they influence the factory behavior.

4 DISCUSSION

Typically, existing MLOps life-cycle models do not describe that some activities can be either performed manually or in an automated way at different points in time. With our revised model of MLOps, we tried to capture this.

In our view, an important part of MLOps in Industry 4.0, which is typically not present in other MLOps models, is the definition of a high-level architecture. This is particularly relevant as a connection to software engineering. Hence, we introduced this here.

The application of MLOps principles in an Industry 4.0 context is not very special. Challenges exist mainly for the reason that it is a heterogeneous environment and many individual solutions are involved. A full-scale architecture and implementation needs to cover the whole environment of the cyber-physical system or at least interface with relevant existing solutions in this context. This is particularly challenging due to severe resource constraints on the embedded devices and the complexity that is introduced due to the many different hardware platforms and operating environment as well as due increasingly distributed computing. If self-adaptation is introduced and corresponding model engineering happens on edge devices, the complexity of the environment becomes even more severe.

The majority of MLOps activities are of similar complexity in an Industry 4.0 case as in information systems. The main reason for this is that they are typically performed outside the factory infrastructure, especially, if they are not automated. This applies to ”Data analysis” and the majority of model-related tasks.

Challenges arise in other tasks under certain conditions. It is highly case-dependent whether MLOps is difficult to implement in Industry 4.0. Also the starting-level of the technical infrastructure as well as the aimed at degree of automation influence very significantly the overall complexity of implementing MLOps in Industry 4.0. For example, in data-related tasks, technical challenges arise when already existing devices are not ready for data collection for the specific use case of an ML model or for its execution. Operations are problematic when special implementations are required or no sufficient resources are available.

Today, some guidance, best practices, frameworks, and platforms already exist to facilitate MLOps in IoT environments (and thus Industry 4.0): Ruf et al. discuss a selection of benefits using MLOps in industrial scenarios (Ruf et al., 2022). A digital twin architecture with MLOps techniques is proposed by Fujii et al. (Fujii et al., 2022). An MLOps framework for automated ML at the edge is described by Raj et al. (Raj, E et al., 2021). None of these takes a broad view of problems in MLOps in Industry 4.0 as we do here. Hence, not only our analysis of challenges, but also the specific MLOps model can be regarded as a contribution of our work.

5 CONCLUSION

MLOps is an important set of practices and activities, which are key to the implementation of modern ML-based software solutions. In this paper, we discussed challenges from the perspective of MLOps in Industry 4.0 and discuss how they differ from MLOps challenges in other contexts.

Overall, we conclude that most Industry 4.0 MLOps challenges exist in a similar manner in the more traditional software engineering context. Some additional challenges exist, at least in some application scenarios. In particular, we could identify significant (specific) challenges for four activities. Most of the challenges are not unique to Industry 4.0, a positive indicator for using existing technologies and practices in this context as well. We plan to study these and the corresponding ways to address them in more detail in the future. For this purpose, we will conduct industry studies around MLOps and apply what we have learned to platforms and frameworks that implement MLOps for Industry 4.0.

ACKNOWLEDGMENTS

This work was partially supported by HAISEM-Lab (grant 01—S19063B) and EXPLAIN (01—S22030E) funded by the Federal Ministry of Education and IIP-Ecosphere funded by the German Ministry of Economics and Climate Action (BMWK) under grant number 01MK20006C.

REFERENCES

Borgmeier, A., Grohmann, A., and Gross, S. F. (2017). *Smart Services und Internet der Dinge: Geschäftsmodelle, Umsetzung und Best Practices: Industrie 4.0, Internet of Things (IoT), Machine-to-Machine, Big Data, Augmented Reality Technologie*. Carl Hanser.

Cardoso Silva, L., Rezende Zagatti, F., Silva Sette, B., Nildaimon dos Santos Silva, L., Lucrédio, D., Furtado Silva, D., and de Medeiros Caseli, H. (2020).

Benchmarking machine learning solutions in production. In *International Conference on Machine Learning and Applications*, pages 626–633.

Fujii, T. Y., Hayashi, V. T., Arakaki, R., Ruggiero, W. V., Bulla, R., Hayashi, F. H., and Khalil, K. A. (2022). A Digital Twin Architecture Model Applied with MLOps Techniques to Improve Short-Term Energy Consumption Prediction. *Machines*, pages 455–462.

G. Symeonidis, E. Nerantzis, A. Kazakis, and G. A. Papakostas (2022). MLOps - Definitions, Tools and Challenges. In *Annual Computing and Communication Workshop and Conference*, pages 453–460.

Kreuzberger, D., Kühl, N., and Hirschl, S. (2022). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *arXiv:2205.02302 [cs]*.

Meedeniya, D. and Thennakoon, H. (2021). Impact factors and best practices to improve effort estimation strategies and practices in DevOps. In *International Conference on Information Communication and Management*, pages 11–17.

Oluyisola, O. E., Bhalla, S., Sgarbossa, F., and Strandhagen, J. O. (2022). Designing and developing smart production planning and control systems in the industry 4.0 era: A methodology and case study. *Journal of Intelligent Manufacturing*, pages 311–332.

Rahman, S. and Kandogan, E. (2022). Characterizing practices, limitations, and opportunities related to text information extraction workflows: A human-in-the-loop perspective. In *CHI Conference on Human Factors in Computing Systems*.

Raj, E, Buffoni, D, Westerlund, M, and Ahola, K (2021). Edge MLOps: An Automation Framework for AIoT Applications. In *International Conference on Cloud Engineering*, pages 191–200.

Ruf, P., Reich, C., and Ould-Abdeslam, D. (2022). Aspects of Module Placement in Machine Learning Operations for Cyber Physical Systems. In *Mediterranean Conference on Embedded Computing*, pages 1–6.

Sato, D., Wider, A., and Windheuser, C. (2019). Continuous delivery for machine learning. visited 2022-06-18.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W. H., Ma, X., Rao, Y., Bednar, J. A., Tan, A., Wang, J., Purushotham, S., Gill, T. E., Chastang, J., Howard, D., Holt, B., Gangodagamage, C., Zhao, P., Rivas, P., Chester, Z., Orduz, J., and John, A. (2022). A review of earth artificial intelligence. *Computational Geoscience*.

van der Goes, M. (2021). Scaling Enterprise Recommender Systems for Decentralization. In *Conference on Recommender Systems*, pages 592–594.