Industrial Edge MLOps: Overview and Challenges

Fatima Rania,\*, Nicolas Cholletb, Lucas Vogta, Leon Urbasa

*aFaculty of Electrical and Computer Engineering, Chair of Process Control Systems & Process Systems Engineering Group, Technische Universität Dresden, 01069 Dresden, Germany  
bLyRIDS (Laboratory of Research in Interdisciplinary Digital engineering & Sciences) ECE Paris engineering school, 75015, Paris, France*

\* [fatima.rani@tu-dresden.de](mailto:fatima.rani@tu-dresden.de)

Abstract

Machine Learning Operations (MLOps) is not a buzzword anymore. In the last few years, there has been a lot of booms in different MLOps tools and frameworks. Basically, it’s a paradigm that focuses on the automation and operationalization of AI development, including model packaging, monitoring, and deployment. While there have been advancements in MLOps tools and frameworks, there are still challenges and research gaps when it comes to deploying MLOps pipelines on edge devices. This paper provides an overview of the Edge MLOps area. Our aim is also to define the basic architecture and components needed for the industrial MLOps deployment. In this context, we also highlight some tools and frameworks. Moreover, we address some limitations and research gaps in the Industrial edge MLOps area.

**Keywords**: MLOps, Edge Computing, IIoT, Industry 4.0, Process Industry.

* 1. Introduction

Presently, Machine Learning (ML) and Artificial Intelligence (AI) are inevitable, and there is a bloom in Industrial Internet of Things (IIoT) devices, resulting in substantial growth in real-time series industrial plant data. Besides this, ML models are booming in the process industry, and to compensate for the hidden technical debts in ML systems, MLOps is the way to go. In the last few years, MLOps has been an emerging paradigm that aims to operationalize the machine learning models into production (Rani et al., 2023). Kreuzberger et al. (2023) emphasize the need to automate and operationalize ML products and provide an aggregated overview of the principles, components, roles, architecture, and workflows of MLOps. However, there are still some challenges presents in deploying the ML models on the cloud (Paleyes et al., 2020). Hence, in the ever-evolving data-driven landscape of industrial plants, the convergence of edge computing (Barakat et al., 2023) and MLOps, or edge MLOps, is essential. As a result, Edge Computing enables us to make decisions at the data sources rather than send the data to the cloud and centralized database, which allows us to reduce the issues of latency, bandwidth, economics, and privacy (especially to meet regulations of countries like GDPR in Europe), and reliability and energy consumption as transmitting large data file is usually a power-hungry process for IoT devices running on battery. In addition, consumers need low latency for real-time experiences, and businesses require local processing to operate securely and reliably while complying with government privacy regulations (Ahmed et al., 2017). Thus, in this contribution, we ask the research question:

**RQ:** *What is the industrial edge MLOps?* *What are the best practices that can be implemented in Industrial Edge MLOps?* *"What are the key challenges in it?*

To answer this question, we conducted a literature survey to (a) identify the core principles of Edge MLOps and (b) its challenges, which contribute to an understanding of Industrial Edge MLOps concepts. In this article, we provide a comprehensive and aggregated overview of the architecture components that facilitate the MLOps on edge devices. We also highlight the open challenges and research gaps for deploying the MLOps pipelines on edge devices. Finally, this work points out some use cases to deploy the MLOps on Edge and its tools and framework.

* 1. Methodology

To derive the main insights for Industrial Edge MLOps, we conduct (a) a structured literature review and (b) tool review (details of these are out of the scope of this article due to page limitation). On the basis of this methodology, we elaborate on our finding of core components required for Edge MLOps architecture, its challenges, use cases, tools, and framework in the next sections.

* 1. Edge MLOps Architecture

The architecture of the typical Industrial Edge MLOps involves the following components:

* + 1. Data Collection and Preprocessing
       1. Edge Devices

This stage involves collecting data from various devices located at the edge of the industrial network sources, such as sensors, machines, actuators, different databases, and control systems, where the data is either generated or already stored (Raj et al., 2021).

* + - 1. Data preprocessing

Then, the different lightweight preprocessing techniques (like cleaning, transformation etc.) are applied to make data suitable for ML model training at the edge device (Raj et al., 2021).

* + 1. Model Training and Development
       1. Machine Learning Models

In this stage, ML models are developed and trained using preprocessed data. This may involve selecting appropriate algorithms, optimizing hyperparameters, and evaluating model performance. Basically, these are predictive models developed by utilizing ML algorithms to analyze and take actions based on the data collected from industrial processes. In many cases, a hybrid approach is employed, for the deployment of ML models (Rani et al., 2023; Paleyes et al., 2020).

* + - 1. Edge Compute

This is the computational infrastructure located close to the edge devices, where machine learning models could be deployed and executed. It could include edge servers or gateway devices (Raj et al., 2021).

* + 1. Model Deployment and Orchestration

This is the crucial stage in which trained ML models are deployed to edge devices, which may be microcontrollers, embedded systems, or specialized AI accelerators. Furthermore, tools for deploying ML models onto edge devices and orchestrating their execution (Kreuzberger et al., 2023). This step may involve containerization technologies like Docker and orchestration tools like Kubernetes.

* + 1. Model Monitoring and Management

After the deployment of the ML model, there is a continuous need either centralized or decentralized retraining for the deployed models to be continuously monitored for performance, accuracy, and potential issues. This may involve real-time data analysis, anomaly detection, and model retraining. A key issue of Edge ML is its sensitivity to model drift. Model Drift is the inevitable degradation of a model’s performance due to the ever-changing nature of data (Rajapakse at el., 2023). This means that the model, once deployed, needs to be updated over time with new and local data. Also, systems are used to monitor the performance of deployed models, collect logs, and provide insights into the behavior of the models in real-time.

* + 1. Model Versioning and Governance

ML models are versioned and managed to ensure reproducibility, traceability, and compliance with regulatory requirements by tools like DVC, MLflow, TensorBoard etc.

* + 1. Security and Compliance

Ensuring the security by means of data encryption, access control, containerization etc., of the deployed models and compliance with industry regulations and standards.

A diagram of a diagram

Description automatically generated

Figure 1- The main components of the Industrial Edge MLOps Architecture(Adopted from Hymel et al., 2022).

* 1. Edge MLOps Challenges

There are several challenges that need to be addressed while implementing the industrial edge MLOps.

* + 1. Resource Constraints

The biggest challenge is that edge devices often have limited computing, memory, and storage resources, which can constrain the complexity and performance of ML models.

Most of the time, due to these limited computational resources, deploying and running complex machine learning models becomes complicated, and hardware must be carefully selected to fit the application requirements. Another issue with resource-constrained hardware is their inability to train the Edge ML model, a process that usually requires significant computation power and that is required to avoid the model drift issue (Rajapakse et al., 2023). Another important resource constraint is energy consumption. IoT devices usually run on battery, and therefore, the implementation of the MLOps procedure should keep power usage to a minimum.

* + 1. Data Availability and Quality

Industrial edge environments may have limited or noisy data, which can impact ML model accuracy and reliability (Raj et al., 2021). Generic application models should always be refined through additional training with locally collected data from a device to cover for device drift.

* + 1. Connectivity and Latency

Edge devices may operate in remote or offline scenarios, requiring reliable communication and low-latency data processing. Achieving low-latency processing is crucial for real-time decision-making, which may be challenging in resource-constrained edge environments (Shafique et al., 2021). Another issue is the usage of wireless communication for sensors running on batteries, as it is usually an energy-hungry procedure. Low Power Wide Area Networks are often used by such sensors to allow communication at minimal power cost but only offer low data rates that can be insufficient for MLOps procedures (Hymel et al., 2022).

* + 1. Data Security and Privacy

Industrial edge systems handle sensitive data and control critical processes, requiring robust security and privacy measures. Ensuring the security of sensitive industrial data when processing it at the edge is a critical concern. Moreover, governments tend to enforce needed but constraining regulations to avoid the over usage of private data.

* + 1. Continuous Integration and Continuous Delivery (CI/CD)

Integrating and deploying ML models to edge devices in a continuous and automated manner can be challenging. Especially, when updating edge devices over the air after their deployment.

* + 1. Model Versioning and Updates

Managing the versions of deployed models and updating them seamlessly without disrupting operations.

* + 1. Interoperability & Scalability

Integrating different edge devices, protocols, and ML models from various vendors can be challenging. ML can be performed at the edge through numerous software, hardware, or hybrid methods. Ensuring that the MLOps infrastructure can scale to accommodate the growing number of edge devices and the increasing complexity of ML models.

A diagram of a model

Description automatically generated

Figure 2- The cloud and edge network challenges of the Industrial Edge MLOps Architecture (Adopted from Raj at el., 2021).

* 1. IIoT Use Cases

MLOps can be applied to various domains by streamlining and automating the lifecycle of machine learning models, ensuring their efficient development, deployment, and maintenance across diverse applications such as the one presented in Fig. 3:

* Industry and Factory operations: Edge ML can be used in automation, predictive maintenance, and quality control processes (Kreuzberger et al., 2023).
* Healthcare: Wearable edge devices can be used in patient diagnosis, treatment personalization, and healthcare data analysis (Khattak at al., 2023).
* ADAS: Car manufacturers can apply MLOps for Advanced Driver-Assistance Systems (ADAS) to manage and maintain Edge AI applications across the vast fleet of vehicles (Gupta et al., 2021).

A group of icons of various types of technology

Description automatically generated with medium confidence

Figure 3– Edge MLOps domains of application in a nutshell.

* Agriculture: MLOps can help manage sensors in energy-efficient Low Power Wide Area Networks (LPWAN) and, in doing so, keep the sensors fleet running on battery for multiple seasons (Chollet et al., 2022).
* Smart city: Sensors in a smart city can perform tasks like real-time traffic management, environmental monitoring, and public safety surveillance, enhancing urban living and operational efficiency (Goethals et al., 2021).
* Digital twin: Overall, MLOps can be used in all domains of applications to implement the digital twin paradigm (Fujii et al., 2021).
  1. Tools and Frameworks

Numerous tools exist to answer the different needs of each step of the MLOps process. In regard to the complete MLOps framework (Rani et al., 2023), good practices and a few options have already been proposed, and they are gathered by the authors (Kreuzberger et al., 2023). Research conducted by our team has intensively used Edge Impulse (Chollet et al., 2022). It is a SaaS proposed by Google that enables efficient data collection, model training, and validation at the edge, facilitating real-time analytics and decision-making in various applications, and it integrates seamlessly with MLOps workflows for continuous improvement and deployment of these models. For all of these reasons, we can recommend this tool for managing edge AI device operations.

* 1. Conclusion

Despite these challenges, Industrial Edge MLOps are becoming increasingly important as ML becomes more pervasive in industrial applications. By addressing these challenges, organizations can reap the benefits of ML in edge environments, such as improved decision-making, real-time optimization, and predictive maintenance. In a nutshell, Industrial Edge MLOps represents the intersection of edge computing and machine learning operations, addressing the unique challenges and requirements of deploying and managing machine learning models in industrial environments.

References

A. Paleyes, R.G. Urma, N.D. Lawrence, 2022, Challenges in deploying machine learning: a survey of case studies, ACM Computing Surveys (CSUR).

D. Kreuzberger, N. Kühl, S. Hirschl, 2023, Machine learning operations (mlops): Overview, definition, and architecture, IEEE Access.

E. Ahmed, A. Ahmed, I. Yaqoob, J. Shuja, 2017, Bringing computation closer toward the user network: Is edge computing the solution?, IEEE Communications Magazine, 55(11), 138-144.

E. Raj, D. Buffoni, M. Westerlund, K. Ahola, 2021, Edge MLOps: An automation framework for aiot applications, IEEE International Conference on Cloud Engineering (IC2E).

F. Rani,V. Khaydarov, D. Bode, I. Hasan, L. Urbas, 2023, MLOps Practice: Overcoming the Energy Efficiency Gap, Empirical Support Through ecoKI Platform in the Case of German SMEs, PAC World Global Conference.

F.K. Khattak, V. Subasri, A. Krishnan, E. Dolatabadi, D. Pandya, L. Seyyed-Kalantari, F. Rudzicz, 2023, MLHOps: Machine Learning for Healthcare Operations, arXiv preprint, arXiv:2305.02474.

G. Symeonidis, E. Nerantzis, A. Kazakis, G.A. Papakostas, 2022, Mlops-definitions, tools and challenges, In 2022 IEEE 12th Annual CCW, pp. 0453-0460.

M. Barakat, R.A. Saeed, S. Edam, 2023, A Comparative Study on Cloud and Edge Computing: A Survey on Current Research Activities and Applications, IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA).

M. Shafique, A. Marchisio, RV.W. Putra, M.A. Hanif, 2021, Towards energy-efficient and secure edge AI: A cross-layer framework, In 2021 IEEE/ACM ICCAD, pp. 1-9.

N. Chollet, B. Naila, R. Amar, 2022, TinyML Smart Sensor for Energy Saving in Internet of Things Precision Agriculture platform, In 2022 Thirteenth ICUFN, pp. 256-259.

S. Gupta, B. Amaba, M. McMahon, K.Gupta, 2021, May, The evolution of artificial intelligence in the automotive industry, In 2021 Annual RAMS, pp. 1-7.

S. Hymel, C. Banbury, D. Situnayake, A. Elium, C. Ward, M. Kelcey, M. Baaijens, M. Majchrzycki, J. Plunkett, D. Tischler, A. Grande, L. Moreau, D. Maslov, A. Beavis, J. Jongboom, V.J. Reddi, 2022, Edge Impulse: An MLOps Platform for Tiny Machine Learning, arXiv preprint, arXiv:2212.03332.

T. Goethals, B. Volckaert, F.D. Turck, 2021, Enabling and Leveraging AI in the Intelligent Edge: A Review of Current Trends and Future Directions, IEEE Open Journal of the Communications Society, 2, 2311-2341.

T.Y. Fujii, V.T. Hayashi, R Arakaki, W.V. Ruggiero, 2021, A digital twin architecture model applied with MLOps techniques to improve short-term energy consumption prediction. Machines, 10(1), 23.

V. Rajapakse, I. Karunanayake, N. Ahmed, 2023, Intelligence at the extreme edge: A survey on reformable tinyml, ACM Computing Surveys, 55(13s), 1-30.

**Acknowledgments:** This work was Funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under the grant number 03EN2047C.

**Declaration of AI-assisted technologies:** While preparing this manuscript, the authors employed the AI-tools for enhancing the manuscript’s language and readability. The authors carefully reviewed and revised the contents and took the responsibility for the research work and idea for this article content. publication