Implementing MLOps practices for effective machine learning model deployment: A meta synthesis

Danylo O. Hanchuk[1], Serhiy O. Semerikov[1,2,3,4,5]

[1]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

[2]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*

[3]*Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine*

[4]*Kryvyi Rih National University, 11 Vitalii Matusevych Str., Kryvyi Rih, 50027, Ukraine*

[5]*Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

Abstract

The successful deployment of machine learning (ML) models in production environments remains a significant challenge despite advancements in ML algorithms and model development. MLOps, a set of practices that combines machine learning and DevOps principles, has emerged as a promising approach to address this challenge. This paper presents a meta-synthesis of systematic reviews to provide a comprehensive understanding of MLOps practices, tools, and challenges for effective ML model deployment. The meta-synthesis was conducted following Chrastina's [1] methodology and included three systematic reviews and one review of MLOps products and providers. The meta-synthesis identified key MLOps principles, such as automation, reproducibility, collaboration, continuous learning, and data governance. It also revealed the main stages of the MLOps workflow, including data collection and processing, model development and training, deployment, monitoring, and retraining. Various frameworks and architectures that facilitate MLOps implementation were discussed, such as open-source platforms, cloud computing platforms, containerization, and container orchestration. The study highlighted the main features offered by MLOps tools, including automation, experiment tracking, versioning, monitoring, and model deployment. The most common methods of deploying ML models in production environments were identified as the use of container technologies, cloud platforms and services, and deployment of models as web services. The meta-synthesis also discussed the importance of adapted software development maturity models for assessing the maturity level of MLOps processes in organizations. Furthermore, the paper emphasized the critical roles and responsibilities involved in ML model operationalization activities, such as data scientists, data engineers, DevOps engineers, and domain experts. The main challenges encountered when deploying ML models in production environments were discussed, including managing the model lifecycle, ensuring scalability and performance, monitoring and maintaining models in real-world conditions. Open issues and challenges in MLOps were also identified, such as the need to develop standards and best practices, ensure interpretability and responsible use of models, and effectively manage data. The findings of this meta-synthesis can guide organizations in adopting MLOps practices to improve the efficiency, reliability, and scalability of their ML model deployments.

Keywords

MLOps, machine learning, model deployment, meta-synthesis, systematic review, DevOps, automation, monitoring, versioning, scalability

1. Introduction

Machine learning (ML) has become an increasingly important technology, finding applications in various domains such as finance, healthcare, manufacturing, retail, and more [2, 3, 4]. The successful deployment of ML models in production environments is crucial for organizations to reap the benefits of this technology. However, despite significant progress in developing ML algorithms and models, their effective deployment in production environments remains a challenging task [5, 6]. This is due to several factors, such as the need for scalability, reproducibility, security, and reliability of models, as well as the complexity of integrating development and operation processes.

To address these challenges, the methodology of MLOps (Machine Learning Operations) has emerged, aiming to apply DevOps principles and practices to the processes of developing and deploying ML models [7, 5]. MLOps covers a wide range of practices, such as automation of ML pipelines, versioning of data and models, monitoring of model performance, experiment management, and more [7, 6]. Studies show that applying MLOps practices can significantly improve the efficiency and reliability of deploying ML models in production environments [5, 6].

However, despite significant interest in MLOps from both researchers and practitioners, there are still gaps and unresolved issues in this field. In particular, there are no generally accepted standards and best practices for implementing MLOps, the issues of integrating MLOps with other approaches (DataOps, ModelOps, AIOps, etc.) are insufficiently explored, and there is a need to develop new tools and platforms to automate MLOps processes [7, 5, 6].

This paper aims to address the relevant problem of defining and analyzing the MLOps practices needed for effective deployment of machine learning models. The basis for this work is the need to systematize and generalize knowledge about MLOps practices, as well as the need to develop recommendations for their implementation in organizations to improve the efficiency and reliability of deploying machine learning models in production environments.

The rest of the paper is organized as follows. Section 2 describes the methodology used for the meta-synthesis. Section 3 presents the results of the meta-synthesis, including the definition of MLOps, stages of the MLOps workflow, frameworks and architectures facilitating MLOps implementation, MLOps tools, features offered by MLOps tools, methods of deploying ML models, maturity models for assessing the level of automation, roles and responsibilities in ML model operationalization activities, challenges encountered when deploying ML models, open issues and challenges in MLOps, and opportunities and future trends in MLOps. Section 4 concludes the paper and outlines directions for future research.

2. Methodology

The meta-synthesis was performed according to Chrastina [1] in the following steps:

1. *Defining the research subject*: MLOps practices for effective model deployment.
2. *Identifying relevant sources*: systematic literature reviews [7, 5, 6] and a review of MLOps products and providers [8].
3. *Close examination* to determine the common time period, commonalities and differences in the aims, research questions, sources, inclusion, exclusion and quality criteria, MLOps definitions and stages.
4. *Determining the relationship between the works* by identifying and grouping key themes.
5. *Reciprocal translation of the results of different works* by defining common terminology, explaining contradictions in the results from different works, and generalizing the results from different works.
6. *Synthesis of results*.
7. *Expressing the synthesis*.

The meta-synthesis included three systematic reviews [7, 5, 6] and one review of MLOps products and providers [8]. The systematic reviews were selected based on a search in the Scopus database using the following query: TITLE ((systematic OR review OR survey) AND mlops). The review of MLOps products and providers [8] was included to provide additional insights into the practical aspects of MLOps implementation.

The meta-synthesis focused on the following key themes: MLOps definition, stages of the MLOps workflow, frameworks and architectures facilitating MLOps implementation, MLOps tools, features offered by MLOps tools, methods of deploying ML models, maturity models for assessing the level of automation, roles and responsibilities in ML model operationalization activities, challenges encountered when deploying ML models, open issues and challenges in MLOps, and opportunities and future trends in MLOps.

The results of the meta-synthesis were expressed in the form of a narrative synthesis, summarizing the key findings and insights from the included reviews. The synthesis aimed to provide a comprehensive understanding of MLOps practices, tools, and challenges for effective ML model deployment, as well as to identify open issues and future trends in this field.

3. Results

3.1. MLOps definition

Despite some differences in emphasis and formulations, all the reviews examined define **MLOps** as an **approach for managing, automating, and operationalizing the processes of developing, deploying, and maintaining machine learning models based on practices from software engineering and DevOps**. MLOps is a key component for the successful implementation of machine learning solutions in an industrial environment.

3.2. MLOps workflow stages

Despite different levels of detail and grouping, the reviews demonstrate overall consistency regarding the main stages of the MLOps workflow. These stages cover the entire lifecycle of machine learning models, from data collection and processing to deployment, monitoring, and retraining of models. Differences in the presentation of stages reflect different approaches to structuring and describing the MLOps workflow.

The main stages of the MLOps workflow identified in the reviews include:

- Data collection and processing: collecting, cleaning, transforming, and enriching data for training models.
- Model development and training: selecting algorithms, developing model architecture, training and validating models.
- Model evaluation and testing: evaluating model performance on test data, conducting reliability, security, and compliance tests.
- Model deployment: packaging models with required dependencies, deploying to target environments.
- Model monitoring and maintenance: tracking model performance, detecting and resolving issues, updating models as needed.
- Model lifecycle management: coordinating all stages of model development, deployment, and maintenance, ensuring compliance with regulatory requirements.

3.3. Frameworks and architectures facilitating MLOps implementation

There are many frameworks and architectural approaches that facilitate the implementation of MLOps, from open platforms and libraries to commercial solutions and cloud services. Key factors are support for automation, scalability, portability, and integration with existing systems and tools. The choice of appropriate frameworks and architectures depends on the specific requirements and constraints of the organization, as well as the maturity level of its MLOps processes.

The reviews highlight the following frameworks and architectures as key components of the MLOps ecosystem:

- Open-source platforms and frameworks, such as MLflow, Kubeflow, and TensorFlow Extended (TFX).
- Cloud computing platforms and services, such as AWS, Google Cloud, and Azure, for deploying and scaling MLOps solutions.
- Containerization (e.g., using Docker) and container orchestration (e.g., using Kubernetes) architectures for ensuring portability and scalability of MLOps solutions.

- Pipeline orchestration platforms, such as Apache Airflow, Jenkins, and Polyaxon.
- Specific frameworks and platforms, such as Kafka-ML and MLModelCI, for managing the lifecycle of ML models.

3.4. MLOps tools for building ML pipelines and model operationalization

There is a wide range of MLOps tools for building machine learning pipelines and operationalizing models, from open platforms such as MLflow to commercial solutions from cloud providers and specialized companies. The choice of specific tools depends on the needs and scale of the organization, as well as compatibility with the existing technology stack.

The reviews highlight the following popular MLOps tools:

- MLflow, a popular open-source platform for managing the lifecycle of ML models, experiments, and deployment.
- Cloud platforms from major providers, such as AWS SageMaker, Google Cloud AI Platform, and Azure Machine Learning, for operationalizing models.
- Containerization tools, such as Docker, and orchestration tools, such as Kubernetes, for deploying models.
- Other tools, such as Kubeflow, Polyaxon, Comet.ml, Kafka-ML, and MLModelCI, for managing pipelines and deploying models.
- Popular commercial MLOps platforms, such as Iguazio, Domino Data Lab, Comet, Valohai, and others.

3.5. Key features offered by MLOps tools

MLOps tools provide a wide range of features to support the machine learning model lifecycle, with a focus on automation, experiment tracking, versioning, monitoring, and model deployment. Some tools offer more specialized features, such as hyperparameter optimization or data management. The choice of a tool with an appropriate set of features depends on the specific needs and goals of the organization in the field of MLOps.

The reviews identify the following key features offered by MLOps tools:

- Experiment tracking and model versioning.
- Automation and orchestration of MLOps workflows, such as training and deployment pipelines.
- Monitoring of model performance and degradation in production environments.
- Data management features, such as data extraction, transformation, and monitoring.
- Support for various machine learning libraries and frameworks.
- Automated hyperparameter optimization and model testing.
- Integration with existing systems and support for collaborative work of teams.

3.6. Methods of deploying ML Models in production environments

The most common methods of deploying machine learning models in production environments are the use of container technologies, cloud platforms and services, and deployment of models as web services. The choice of a specific approach depends on the requirements for latency, scalability, and availability of models, as well as the existing infrastructure and ecosystem of tools in the organization.

The reviews describe the following key methods and aspects of deploying ML models:

- Deploying models using container technologies, such as Docker, which ensures model mobility and isolation.
- Deploying models in cloud environments using platforms and services from major providers, such as AWS, Google Cloud, and Azure.

- Deploying models as web services using REST API or other protocols to provide access to predictions in real-time.
- Using orchestration platforms (Apache Airflow, Jenkins, Kubeflow, MLflow, Polyaxon, Seldon Core, Valohai) for automated scaling and container management.
- Built-in deployment capabilities of some MLOps tools, such as MLflow, Kubeflow, and Kafka-ML.
- Deploying models not only in the cloud but also on edge devices using specialized frameworks, such as TensorFlow Lite and Core ML.
- The main stages and features of the ML model deployment process using CI/CD pipelines

3.7. Maturity models for assessing the level of automation in deploying ML models

All systematic reviews [7, 5, 6] indicate that the level of automation of MLOps processes is one of the key factors in assessing an organization's maturity in this area. Although the reviews do not provide a comprehensive description of MLOps maturity models, they emphasize the importance of assessing the level of automation in the processes of developing, testing, and deploying models as a key factor in an organization's maturity in this area. Adapting existing software development maturity models to the specifics of MLOps can be an effective approach to assessing and improving machine learning processes in an organization.

The reviews mention several maturity models for assessing the level of improvement in the machine learning solutions development process:

- The maturity model proposed by Amershi et al. [9], based on the Capability Maturity Model (CMM) and the Six Sigma methodology, which checks whether an activity: (1) has defined goals, (2) is consistently implemented, (3) is documented, (4) is automated, (5) is measured and tracked, and (6) is continuously improved.
- According to Dhanorkar et al. [10], organizations can be classified into three levels of maturity in developing machine learning solutions: (1) data-oriented, (2) model-oriented, (3) pipeline-oriented.
- Lwakatare et al. [11] describe five stages of development practices improvement: (1) a manual process driven by data science, (2) a standardized process of experimental-operational symmetry, (3) an automated ML workflow process, (4) integrated software development and ML workflow processes, and (5) an automated and fully integrated CD and ML workflow process.
- Akkiraju et al. [12] proposed an adaptation of the CMM model with the definition of five maturity levels for each assessed capability: (1) initial, (2) repeatable, (3) defined, (4) managed, and (5) optimizing.

3.8. Roles and responsibilities identified in ML model operationalization activities

Despite some differences in the detail of roles, the reviews examined recognize the need to involve professionals from various fields - software development, data engineering, machine learning, subject matter experts, and management - for the successful operationalization of machine learning models. Close collaboration and communication between these roles is critical for the implementation of MLOps practices in organizations.

The reviews identify the following key roles and responsibilities in ML model operationalization activities:

- Data scientists/ML scientists, who are responsible for developing, training, and experimenting with ML models.
- Data engineers/data providers, who are responsible for extracting, processing, transforming, and ensuring the quality of data for training models.
- DevOps engineers, ML/MLOps engineers, and software engineers, who are responsible for operationalizing models, automating deployment processes, creating pipelines, and managing environments.

- Managers, leadership, and business stakeholders, who define requirements for models regarding their deployment, make decisions, and support the MLOps strategy.
- Domain experts, who provide domain knowledge and participate in data labeling in specific domains.

Some reviews also mention additional roles, such as computational scientists/engineers, ML scientists/engineers, provenance specialists, application developers, and deployment leads, who have specific responsibilities in the MLOps process.

3.9. Challenges encountered when deploying ML models in production environments

The reviews demonstrate that deploying machine learning models in production environments is associated with a number of challenges, such as managing the model lifecycle, ensuring scalability and performance, monitoring and maintaining models in real-world conditions. Addressing these challenges requires a comprehensive approach that includes automating MLOps processes, selecting appropriate infrastructure, ensuring data security and privacy, and effective communication with business stakeholders.

The main challenges identified in the reviews include:

- The complexity of managing the machine learning model lifecycle, including versioning, tracking, and reproducibility of models and data, as well as the problem of ensuring the scalability and performance of models in real-world conditions with large volumes of data and requests.
- Challenges related to monitoring and maintaining models in a production environment, including detecting data drift and model performance degradation.
- Challenges related to integrating software development with the machine learning pipeline.
- Challenges related to ensuring data security and privacy when deploying machine learning models.
- The quality, availability, preparation, labeling, and integration of data from various sources is a significant challenge that requires a lot of time and resources, and the problem of interpreting and explaining the results of model operation to end-users and business stakeholders is highlighted.
- The gap between software engineering and machine learning skills - data scientists often do not understand the requirements of certain production environments, and software developers do not have sufficient machine learning skills.
- The effective distribution, parallelization, and orchestration of data and ML tasks.
- The diversity of computing infrastructure.

3.10. Open issues, challenges and particularities in MLOps

The reviews identify a number of open issues and challenges in MLOps, such as the need to develop standards and best practices, ensure interpretability and responsible use of models, and effectively manage data. The particularities of MLOps, such as the difference from traditional software development, the importance of the human factor, and the need to integrate knowledge from various fields, require consideration when implementing MLOps practices in organizations.

The main open issues and challenges in MLOps identified in the reviews include:

- The need to develop methods and tools to ensure interpretability, reproducibility, and responsible use of machine learning models in the context of MLOps.
- The importance of developing approaches to data management in MLOps, including ensuring data quality, privacy, and security.
- The need to develop and implement MLOps standards and best practices to ensure consistency and compatibility between different tools and platforms.

- The importance of the human factor in MLOps, including the need to ensure effective communication and collaboration between different roles and teams and to train qualified personnel with cross-functional skills in programming, data processing, and operational activities.

The reviews also highlight the following particularities of MLOps:

- MLOps must take into account the specifics of the machine learning model development process, which differs from traditional software development.
- MLOps should be considered in the context of the overall digital strategy of the organization and emphasize the need to align MLOps practices with business goals and needs.

3.11. Opportunities, future trends and fields of application of MLOps

The reviews outline a number of opportunities and trends in the development of MLOps, such as the creation of standardized platforms, application in the context of distributed learning, and integration with other approaches to data and model lifecycle management. Current and future areas of MLOps application include a wide range of industries, from finance and healthcare to IoT and natural language processing, indicating the significant potential impact of this approach.

The main opportunities and future trends in MLOps identified in the reviews include:

- The potential for the development of standardized MLOps platforms and tools that will simplify and accelerate the deployment of machine learning models in production.
- The prospects for integrating MLOps with other approaches, such as DataOps, ModelOps, and DevSecOps, to ensure comprehensive management of the machine learning model lifecycle.
- Significant opportunities for further academic research and development due to the fact that MLOps is still at an early stage.
- The expectation of an increase in demand for MLOps tools and platforms with the spread of artificial intelligence solutions.
- The emergence of new roles and competencies related to MLOps as the industry develops.
- Opportunities for applying MLOps practices in the context of distributed and federated model learning, which will allow efficient use of decentralized data.
- Involving business units and educating management on MLOps principles.
- The use of hardware platforms such as FPGA and IoT to improve performance and privacy.

The reviews identify the following current and future areas of MLOps application:

- Industries such as finance, healthcare, retail, marketing, and manufacturing, where machine learning models are already actively used to solve real business problems.
- The potential for applying MLOps in the field of IoT and edge computing, where machine learning models can be deployed on resource-constrained devices.
- 5G and 6G technologies, educational and scientific activities.
- Transport and logistics.

4. Conclusion

This paper performed a meta-synthesis of systematic reviews and a review of MLOps products and providers to summarize knowledge about implementing MLOps practices for effective deployment of machine learning models. The main findings obtained as a result of the meta-synthesis are as follows:

1. MLOps is an approach for managing, automating, and operationalizing the processes of developing, deploying, and maintaining machine learning models based on practices from software engineering and DevOps. MLOps is based on a set of principles, processes, and practices that ensure the effective development, deployment, and maintenance of machine learning models.

2. The main stages of the MLOps lifecycle include data collection and processing, model development and training, deployment, monitoring, and retraining.

3. Various frameworks and architectures are used to implement MLOps, such as open source platforms (MLflow, Kubeflow, TensorFlow Extended), cloud computing platforms (AWS, Google Cloud, Azure), containerization (Docker), and container orchestration (Kubernetes).

4. MLOps tools provide a wide range of features to support the machine learning model lifecycle, with a focus on automation, experiment tracking, versioning, monitoring, and model deployment.

5. The most common methods of deploying machine learning models in production environments are the use of container technologies, cloud platforms and services, and deployment of models as web services.

6. Adapted software development maturity models, such as CMM, can be used to assess the maturity level of MLOps processes in organizations.

7. Successful implementation of MLOps requires the involvement of professionals from various fields - software development, data engineering, machine learning, subject matter experts, and management.

8. The main challenges in deploying machine learning models in production environments are managing the model lifecycle, ensuring scalability and performance, monitoring and maintaining models in real-world conditions.

9. The open issues and challenges in MLOps are the need to develop standards and best practices, ensure interpretability and responsible use of models, effectively manage data, and integrate knowledge from various fields.

10. The main opportunities and trends in the development of MLOps are the creation of standardized platforms, application in the context of distributed learning, and integration with other approaches to data and model lifecycle management. Current and future areas of MLOps application include a wide range of industries, from finance and healthcare to IoT and natural language processing.

The meta-synthesis showed that MLOps is a promising approach for the effective deployment of machine learning models in production environments, which requires further research and development to address existing challenges and realize potential opportunities.

The results obtained have both theoretical and practical significance. The theoretical significance lies in the generalization and systematization of knowledge about MLOps practices necessary for the effective deployment of machine learning models. The practical significance of the results obtained lies in the possibility of their use by organizations to implement or improve MLOps processes in order to increase the efficiency and reliability of deploying machine learning models in production environments.

Further research may be aimed at developing detailed recommendations for the implementation of individual MLOps practices in organizations, creating new tools and platforms for automating and managing the lifecycle of machine learning models, as well as studying the effectiveness of applying MLOps practices in various industries and areas of machine learning model application.

References

[1] J. Chrastina, Meta-synthesis of qualitative studies: background, methodology and applications, in: NORDSCI Conference proceedings, volume 1 of *NORDSCI Conference*, Saima Consult Ltd, 2018. doi:`10.32008/nordsci2018/b1/v1/13`.

[2] P. V. Zahorodko, S. O. Semerikov, V. N. Soloviev, A. M. Striuk, M. I. Striuk, H. M. Shalatska, Comparisons of performance between quantum-enhanced and classical machine learning algorithms on the IBM Quantum Experience, Journal of Physics: Conference Series 1840 (2021) 012021. doi:`10.1088/1742-6596/1840/1/012021`.

[3] A. E. Kiv, V. N. Soloviev, S. O. Semerikov, H. B. Danylchuk, L. O. Kibalnyk, A. V. Matviychuk, A. M. Striuk, Machine learning for prediction of emergent economy dynamics, CEUR Workshop Proceedings 3048 (2021) I–XXXI.

[4] S. O. Semerikov, M. V. Foki, D. S. Shepiliev, M. M. Mintii, I. S. Mintii, O. H. Kuzminska, Methodology for teaching development of web-based augmented reality with integrated machine learning models, CEUR Workshop Proceedings 3820 (2024) 118–145.

[5] A. Lima, L. Monteiro, A. P. Furtado, MLOps: Practices, Maturity Models, Roles, Tools, and Challenges – A Systematic Literature Review, in: Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 1: ICEIS, INSTICC, SciTePress, 2022, pp. 308–320. URL: https://doi.org/10.5220/0010997300003179. doi:10.5220/0010997300003179.

[6] J. Diaz-de Arcaya, A. I. Torre-Bastida, G. Zárate, R. Miñón, A. Almeida, A Joint Study of the Challenges, Opportunities, and Roadmap of MLOps and AIOps: A Systematic Survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3625289. doi:10.1145/3625289.

[7] G. Recupito, F. Pecorelli, G. Catolino, S. Moreschini, D. D. Nucci, F. Palomba, D. A. Tamburri, A Multivocal Literature Review of MLOps Tools and Features, in: 2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2022, pp. 84–91. doi:10.1109/SEAA56994.2022.00021.

[8] R. Cohen, Digital Strategy, Machine Learning, and Industry Survey of MLOps, in: Digital Strategies and Organizational Transformation, 2023, pp. 137–150. URL: https://tinyurl.com/33z6zpd3. doi:10.1142/9789811271984_0008.

[9] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, T. Zimmermann, Software Engineering for Machine Learning: A Case Study, in: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2019, pp. 291–300. doi:10.1109/ICSE-SEIP.2019.00042.

[10] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, Y. Li, Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle, in: Proceedings of the 2021 ACM Designing Interactive Systems Conference, DIS '21, ACM, New York, 2021, p. 1591–1602. doi:10.1145/3461778.3462131.

[11] L. E. Lwakatare, I. Crnkovic, J. Bosch, DevOps for AI – Challenges in Development of AI-enabled Applications, in: 2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2020, pp. 1–6. doi:10.23919/SoftCOM50211.2020.9238323.

[12] R. Akkiraju, V. Sinha, A. Xu, J. Mahmud, P. Gundecha, Z. Liu, X. Liu, J. Schumacher, Characterizing Machine Learning Processes: A Maturity Framework, in: Business Process Management, volume 12168 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2020, pp. 17–31. doi:10.1007/978-3-030-58666-9_2.