

A Multivocal Review of MLOps Practices, Challenges and Open Issues

BEYZA EKEN*, Faculty of Computer and Information Sciences, Sakarya University, Turkey

SAMODHA PALLEWATTA*, CREST - The Centre for Research on Engineering Software Technology and The University of Adelaide, Australia

NGUYEN KHOI TRAN, CREST - The Centre for Research on Engineering Software Technology and The University of Adelaide, Australia

AYSE TOSUN, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Turkey

MUHAMMAD ALI BABAR, CREST - The Centre for Research on Engineering Software Technology and The University of Adelaide, Australia

With the increasing trend of Machine Learning (ML) enabled software applications, the paradigm of ML Operations (MLOps) has gained tremendous attention of researchers and practitioners. MLOps encompasses the practices and technologies for streamlining the resources and monitoring needs of operationalizing ML models. Software development practitioners need access to the detailed and easily understandable knowledge of MLOps workflows, practices, challenges and solutions to effectively and efficiently support the adoption of MLOps. Whilst the academic and industry literature on the MLOps has been growing rapidly, there have been relatively a few attempts at systematically synthesizing and analyzing the vast amount of existing literature of MLOps for improving ease of access and understanding. We conducted a Multivocal Literature Review (MLR) of 150 relevant academic studies and 48 gray literature to provide a comprehensive body of knowledge on MLOps. Through this MLR, we identified the emerging MLOps practices, adoption challenges and solutions related to various areas, including development and operation of complex pipelines, managing production at scale, managing artifacts, and ensuring quality, security, governance, and ethical aspects. We also report the socio-technical aspect of MLOps relating to diverse roles involved and collaboration practices across them through the MLOps lifecycle. We assert that this MLR provides valuable insights to researchers and practitioners seeking to navigate the rapidly evolving landscape of MLOps. We also identify the open issues that need to be addressed in order to advance the current state-of-the-art of MLOps.

CCS Concepts: • **Software and its engineering** → **Software development methods**; • **General and reference** → **Empirical studies; Surveys and overviews**.

Additional Key Words and Phrases: MLOps, Machine Learning Operations, DevOps, CI/CD

1 INTRODUCTION

Currently, there is a huge trend of incorporating Artificial Intelligence (AI) and Machine Learning (ML) components into software systems to take advantage of their well established inferential and predictive capabilities. These systems, known as ML-enabled systems, provide enhanced business value through data-driven decisions. The increasing demand for AI/ML-enabled solutions and the availability of novel technologies/tools to support the engineering of such systems have increased AI/ML adoption. IBM Global AI Adoption Index for 2022 reports that AI/ML adoption is increasing rapidly with 44% of the organizations working on embedding AI/ML in their products or processes

*Both authors contributed equally to this research.

Authors' addresses: Beyza Eken, beken@sakarya.edu.tr, Faculty of Computer and Information Sciences, Sakarya University, Turkey; Samodha Pallegatta, samodha.pallegatta@adelaide.edu.au, CREST - The Centre for Research on Engineering Software Technology and The University of Adelaide, Australia; Nguyen Khoi Tran, nguyen.tran@adelaide.edu.au, CREST - The Centre for Research on Engineering Software Technology and The University of Adelaide, Australia; Ayse Tosun, tosunay@itu.edu.tr, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Turkey; Muhammad Ali Babar, ali.babar@adelaide.edu.au, CREST - The Centre for Research on Engineering Software Technology and The University of Adelaide, Australia.

in 2022 [2]. However, the productionalization of the AI/ML components, which is the process of transforming experimental ML models into an ML-enabled software product operating in a production environment remains one of the main challenges. The majority of the organizations that build ML pilots fail to deploy their models in production or fail to maintain successful operation of deployed models [27] due to novel operationalizing challenges of ML-enabled software systems compared to traditional software. These challenges include ML-specific technical debt introduced during model development, the need for continuous monitoring, model retraining and delivery due to the data-dependant probabilistic nature of ML systems, collaboration among different actors such as data scientists, software engineers, and software operators, the critical need of ensuring responsible and trustworthy AI, and the complexity of integration and scaling of ML-enabled systems [2, 27]. To tackle these challenges, MLOps has emerged as a paradigm that introduces practices, processes, techniques and technologies for enabling rapid and reliable productionalization of ML systems.

MLOps was first proposed by the research community in 2015 as a solution to speed up the ML lifecycle management and to achieve the high level of scalability expected by business applications [11, 28] by streamlining time and resources required to operationalize ML models. Similar to DevOps, the concept of MLOps has emerged from production challenges, driven by the industry, and encompasses a wide range of practices, techniques, and technologies. For instance, the prominent aspect of MLOps is the definition and management of an automated pipeline, such as for testing and deploying ML-enabled software to production. Such pipeline can be considered the codification of the repetitive yet hidden processes between teams responsible for the life cycle of ML-enabled software. Automating these processes could potentially reduce human errors and increase their velocity. Moreover, the act of defining pipelines could be catalytic to facilitate the communication between teams, and to help identify bottlenecks and pitfalls. Another practice often cited as related to MLOps is the establishment and utilization of shared version-control repositories for datasets, features, and models to provide a consistent source of truth for the automation pipelines and increase the visibility across teams to reduce duplicate work. Another relevant practice under MLOps umbrella is on-demand monitoring services, which could initiate and increase the fidelity of feedback loops. When connected to the pipeline, the improved monitoring could lead to an increase in velocity. Research shows that proper operationalization of ML-enabled systems depends on utilizing such robust MLOps practices, processes, and technologies [27]. According to the survey conducted by Databricks, while the percentage of ML models making it to production is still low, it sees a considerable increase with the rapid advancements in MLOps solutions (e.g., MLFlow, Huggingface, SageMaker, Google Vertex AI). As of January 2023, 34% of the experimental models are registered as candidates for production which is a significant increase compared to 20% in 2022 [PS114].

The apparent success of MLOps in bridging the gap between development and deployment of ML-enabled software systems has sparked a lot of interest among both researchers and practitioners, who have been increasingly reporting different types of MLOps practices and tools for supporting MLOps. As building ML systems is yet not as mature as traditional software systems and due to their inherent complexity, different organizations tend to follow a variety of approaches for developing and operationalizing ML systems based on their unique use cases and organizational structure. MLOps spans across a broad spectrum of activities (starting from data collection to model maintenance), a plethora of practices, tools and technologies supporting and automating each activity and interdisciplinary collaborations (data scientists, ML engineers, software engineers, MLOps engineers). Such inherent complexity and diversity make it harder for practitioners and researchers to navigate current MLOps landscape, which hinders effective adoption and further

improvements of MLOps practices and solutions. Despite the increasing amount of the peer-reviewed and gray literature on MLOps, there is significant confusion surrounding MLOps due to its complexity, diversity, and broad scope [PS154].

Hence, it is important to systematically select, analyze and organize the MLOps literature covering its various aspects such as how MLOps is being perceived, MLOps practices and technologies, and MLOps adoption challenges and the related solutions to overcome the reported adoption challenges. Whilst there have been a few attempts to review the State-Of-The-Art (SOTA) of MLOps such as reported in these papers [15, 20–22, 25], these studies have reviewed only academic literature and do not cover several important aspects of MLOps. Given MLOps is quite applied paradigm about which there is an increasing amount of gray literature, mainly published by industry. Moreover, the peer-reviewed literature on MLOps has also grown significantly since the previous reviews [15, 20–22, 25], there is an important and urgent need of a comprehensive review of the peer-reviewed and gray literature for building an evidence-based body of knowledge on MLOps.

We conducted a Systematic Multivocal Literature Review (MLR) covering peer-reviewed and gray literature on MLOps published up to September 2023. This study aimed at systematically selecting, synthesizing and analyzing the state-of-the-art MLOps perception, best practices, adoption challenges and solutions for building a body of knowledge to be used by researchers and practitioners interested in MLOps.

The main contributions of our study, linked to our Research Questions (RQs), are as follows:

- It conducts a MLR for building an evidence-based body of knowledge of MLOps from peer-reviewed and gray literature.
- It helps improve our understanding of how MLOps is perceived by both researchers and practitioners, consisting of MLOps definitions (RQ1.1), activities (RQ1.2), roles, and responsibilities (RQ1.3).
- It systematically identifies and analyzes the state-of-the-art best practices (i.e., processes, techniques, and technologies) reported to support MLOps (RQ2).
- It determines the challenges faced by practitioners during MLOps adoption and derives the future research directions for improving MLOps adoption (RQ3).

2 RELATED LITERATURE REVIEWS

We consider existing systematic review studies related to MLOps and conduct a qualitative comparison with our work in terms of the type of the considered primary studies, research questions covered in the review study, number of primary studies used in the review, and considered time period for primary study selection. We identified a total number of 8 related studies, consisting of three Multivocal Literature Reviews (MLR) (i.e., these studies collected data from both academic and gray literature), and five Systematic Literature Reviews (SLR) (i.e., they solely collected data from academic literature). Table 1 presents the comparison of these related reviews with our work to highlight our contributions.

Kreuzberger et al. [21] reported a mixed-method research, where they conducted an interview with practitioners in addition to a SLR. Their work targeted one RQ: “what is MLOps”, which they answered with an overview of MLOps activities, roles and responsibilities and used these insights to create a unifying definition for MLOps. While they also discussed several open challenges related to adopting MLOps, it is not the main objective of the study. Lima et al. [22] reported a SLR, where their RQs aimed to identify activities related to ML model deployment, MLOps related roles and responsibilities, operationalization tools for ML models, and challenges in ML model deployment. However, this study lacks comprehensive data synthesis to provide indepth analysis of roles and responsibilities. Additionally, their discussion on MLOps related technologies is limited

Table 1. Comparison of Related Work

Study	Is multivocal	RQ1.1	RQ1.2	RQ1.3	RQ2	RQ3	No. Studies	Primary	Timeframe
[21]	No	○	●	●		○	27 A	Up to May 2021	
[22]	No	○		○	○	○	30 A	Up to Jul, 2021	
[20]	No				○		24 A	2017 - 2021	
[15]	No	●	○		○		69 A	Jan, 2012 - May, 2022	
[25]	No	●	○	○			18 A	2015 - 2022	
[32]	Yes	○	●			○	79 A, 72 G	2010 - 2021	
[18]	Yes		○		○		8 A, 15 G	Jan, 2015 - Mar 31, 2021	
[24]	Yes				○		1 A, 8 G	Gray lit.: Up to Apr, 2020, Ac. lit.: Up to Feb, 2019	
Our work	Yes	●	●	●	●	●	150 A, 60 G	2013 - Sep, 2023	

●: Full coverage of RQ, ○: Partial coverage of RQ
A: Academic literature study, G: gray literature study

to MLOps platforms extracted from academic literature and their discussion on challenges does not elaborate on the challenges related to MLOps adoption. Kolltveit et al. [20] reported an SLR on the operationalization of ML systems, which covers all steps following model training and evaluation. Specifically, they answered four RQs focused on tooling and infrastructure related aspects of MLOps such as; how are ML models operationalized, what are the main challenges in operationalizing ML models, what tools and software infrastructure are used to operationalize ML models, what are the feature gaps in the tooling and infrastructure used to operationalize ML models. However, since the focus is only on tooling and infrastructure for operationalizing ML models, this work does not provide a comprehensive view of MLOps best practices considering the entire ML lifecycle. Faubel et al. [15] reported an SLR on the technical overview of MLOps. Specifically, the study presented MLOps definitions and listed the most common activities, and listed tools for automating activities in MLOps. However, this review does not consider roles and task distributions in MLOps, thus lacking emphasis on socio-technical aspects in MLOps. Mboweni et al. [25] reported an SLR on the state-of-the-art DevOps practices in ML-enabled systems. Specifically, this work aimed to report how MLOps is understood and differs from DevOps. The reported content provides a higher-level overview of MLOps definitions, presents key roles of MLOps and associated tasks, and describes MLOps activities by constructing an MLOps architecture covering ML lifecycle.

Steidl et al. [32] reported a mixed-method investigation (i.e., SLR and a semi-structured review study) focusing specifically on continuous development pipelines of AI models, which is one aspect of MLOps practices. This work focused on analyzing definitions of various terms related to continuous development pipelines for AI such as DevOps for AI, CI/CD for AI, MLOPs, etc., tasks handled by these pipelines, and the challenges related to implementation, adaption and usage of pipelines for continuous development of AI. John et al. [18] delivered a MLR on the adoption of MLOps in practice. More specifically, the authors derived a framework for MLOps that shows the activities that take part in the continuous development of ML models and maps the practices of companies to MLOps maturity model to show how practices evolve through stages. However, they provided a high-level picture rather than elaborating on details regarding these questions. Lwakatare et al. [24] presented an MLR on the RQ of “how to enable modern software development process with continuous delivery for AI-enabled systems”. The authors considered one academic and eight gray literature items. The study discusses tools and techniques used in the continuous delivery of ML systems. Important details are available concerning the development stages of ML-enabled systems.

Our analysis of the secondary studies presented in Table 1 demonstrates, that no existing work contributes to building a holistic and timely depiction of MLOps landscape by answering all three research questions. Moreover, the majority of the related work relies solely on the academic literature [15, 20–22, 25]. Given that MLOps is a rapidly evolving, practice-oriented field, insights from gray literature sources such as industry white papers and company websites offer valuable and timely perspectives on how MLOps is perceived by the community and adopted in real-life settings. Furthermore, the rapid improvements in MLOps call for a systematic review of the most recent literature to provide a timely view of the MLOps landscape. However, coverage by existing reviews extends only until 2022.

Considering these, our MLR aims to provide a thorough understanding of MLOps encompassing its definition, tasks, roles, state-of-the-art best practices and adoption challenges. We combine theoretical analysis from 150 academic studies with more practical knowledge from 48 gray literature items published up to September 2023. Also, we characterize MLOps as covering emerging practices such as socio-technical practices, complex pipeline management, and ethical and governance considerations. Additionally, we discuss future directions in detail which facilitates further advancement in the field.

3 METHODOLOGY

We conducted a Multivocal Literature Review (MLR) to provide a comprehensive overview of MLOps in terms of its definition, activities, involved people and their responsibilities, state-of-the-art best practices for implementing MLOps, and MLOps adoption challenges faced by the practitioners, mapping them to solutions proposed in the literature.

An MLR study allows researchers to collect and present information from the industry when the formal/academic literature is not sufficient [17]. Since MLOps is an emerging topic in the field of AI and software engineering, only considering academic publications is not sufficient to cover all the knowledge related to this topic. Industry practitioners have been implementing and getting experience in MLOps to integrate ML-enabled systems into their software development pipelines, their knowledge is a crucial element in understanding MLOps. Thus, gray literature such as industrial white papers, blog posts, and company websites, are the key sources for expanding the knowledge obtained from academic literature. Therefore, we believe an MLR would be the most suitable methodology for the MLOps topic. We performed our MLR following the guidelines reported by Garousi et al. [17].

Consequently, our MLR consists of three main stages: 1) *planning the review*, 2) *conducting the review*, and 3) *reporting the review*. The planning stage includes identifying the need for the review, specifying the RQs, developing a review protocol containing search strings for the peer-reviewed and gray literature, and defining inclusion/exclusion criteria for primary study selection. In the second stage, we conducted the search for peer-reviewed and gray literature and selected the primary studies. Afterward, data extraction and synthesis were performed following Thematic Analysis guidelines proposed by Braun et al. [8]. In the final stage, we have organized and reported the synthesized data. Figure 1 represents our MLR methodology.

3.1 Research Questions

Our work addresses the RQs listed in Table 2 to provide a comprehensive overview of the current MLOps landscape from academic and gray literature.

3.2 Study Search and Selection Process

We based our search process on the guidelines provided by Garousi et al. [17] for conducting MLR. We aimed at retrieving the maximum number of academic, peer-reviewed, and gray literature

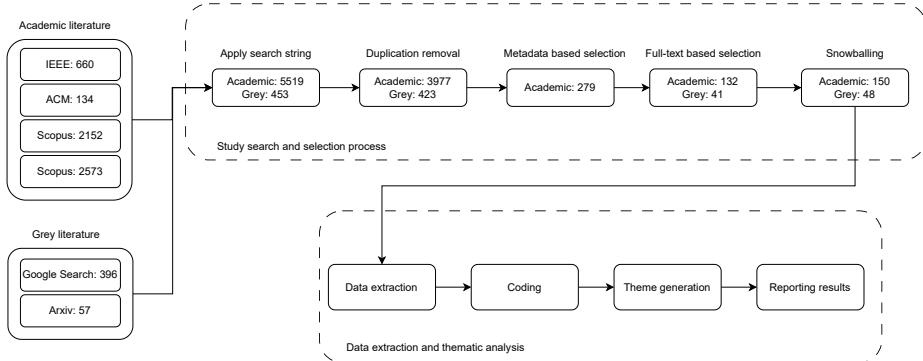


Fig. 1. Research Methodology

Research Questions	Motivation
RQ1 - How is MLOps perceived by researchers and practitioners?	RQ1 focuses on collecting, synthesizing, and reporting an overview of MLOps and its core principles that would reflect the views of both researchers and practitioners under 3 sub-RQs.
<i>RQ1.1 - How is MLOps defined in the literature?</i>	RQ1.1 aims to provide insights on common and varying definitions of MLOps to facilitate better understanding of its foundational aspects to contribute towards creating a standardized usage by the community.
<i>RQ1.2 - What are the primary activities and tasks of MLOps?</i>	RQ1.2 reports the state-of-the-art activities that take part in MLOps to help practitioners adopt MLOps in their contexts more efficiently and allow the readers to gain a clear business value/perspective.
<i>RQ1.3 - What are the practitioner roles and responsibilities in MLOps?</i>	RQ1.3 highlights the responsibilities of people who take part in MLOps settings. Establishing a common understanding of roles and their responsibilities leads to more productive collaboration practices within the complex MLOps team structure.
RQ2 - What are the state-of-the-art practices and techniques proposed for MLOps?	A complete view of best practices, processes and supporting technologies is required for companies to plan for successful implementation of MLOps by minimizing the technical debt and improving the maintainability of the ML systems as they grow in complexity over time. To this end, RQ2 aims to provide state-of-the-art MLOps practices and techniques covering social and technical aspects across the entire life cycle of ML models.
RQ3 - What are the challenges practitioners face when adopting MLOps?	The motivation behind this question is to gain a deeper understanding of the challenges practitioners face when implementing MLOps to see the bottlenecks in MLOps adoption and identify solutions proposed in the current literature. Addressing the state-of-the-art challenges would be beneficial to practitioners to overcome those when they adopt MLOps for productionalizing their ML-enabled software systems. Further, existing challenges often lead to new and highly impactful research directions and innovative solutions.

Table 2. Research Questions addressed in this MLR

related to the research questions. Our search strategy consists of three main steps: automatic search method using inclusive search strings, primary study selection by applying quality assessment and snowballing to identify further studies [35].

3.2.1 Automatic Search. To identify the primary studies from both academic and gray literature, two different search strings were utilized on the databases, namely IEEEXplore Digital Library, ACM Digital Library, Web of Science, Scopus for academic literature search and Google Search, and Arxiv for gray literature search. Search strings evolve through an exploratory search process. To ensure that key studies are identified, we refined search strings by including relevant keywords (e.g., continuous test, continuous release, and mlsecops). Furthermore, we did not use restricting keywords (e.g., technologies, challenges, and tools) to ensure the relevant studies are captured

through the automatic search. Moreover, the automatic search process was performed by two authors to avoid researcher bias. Our search strings are:

Academic literature	<code>mllops OR "machine learning operations" OR mlsecops OR mlflow OR kubeflow OR cd4ml OR ((devops OR "continuous software engineering" OR cse OR "CI/CD" OR "continuous integration" OR "continuous deployment" OR "continuous delivery" OR "continuous test" OR "continuous verification" OR "continuous security" OR "continuous compliance" OR "continuous evolution" OR "continuous use" OR "continuous trust" OR "continuous monitoring" OR "continuous release" OR "continuous build" OR "continuous systems engineering") AND ("machine learning" OR "artificial intelligence" OR "deep learning" OR ml OR ai OR dl))</code>
Gray literature	<code>mllops OR "machine learning operations" OR (devops AND "machine learning") OR ("continuous software engineering" AND "machine learning")</code>

3.2.2 Primary Study Selection from Academic Literature. As we searched the academic literature across multiple databases, some studies were duplicated. At the beginning of the selection process, we removed those duplicated items from our pool of studies, which reduced the number of academic studies from 5519 to 3977. Afterwards, we applied a two-step study selection process to the academic literature: 1) metadata-based selection and 2) full text-based selection. The metadata-based selection is conducted by considering only the paper title, paper keywords, and abstract of the paper, whereas the full-text selection is carried out by reviewing the full text of studies.

During the metadata-based selection, we followed the inclusion and exclusion criteria listed in Table 3. On the other hand, during the full text-based selection, we assessed the quality of each study considering the research aim, research design, methodology, results, limitations of the study, and the value added to research or practice by the study. A more detailed list of quality assessment criteria can be found in our online appendix [3]. These criteria are adopted from [4] and [17], discussed and agreed by all authors.

At the beginning of both metadata and full-text-based selection steps, a pilot selection is carried out by all authors on a sample of the study pool. This pilot study is crucial to increase the inter-rater agreement and establish a common terminology among authors who later continue with study selection individually on a subset of relevant studies. Pilot study samples were picked randomly to avoid the bias that might occur during database search since the most relevant studies are naturally listed at the top of the search. We picked six studies for metadata-based selection and eight for full-text selection. These studies were reviewed by all the authors and the results were discussed.

After completing the pilot study, the study pool was split into three by considering the bias of study order, and the selection processes were carried out separately by three authors. When a decision cannot be made on a study by a single author, a second and if necessary a third reviewer assessed the same study, and the final decision was made accordingly. If authors were unable to reach a consensus, the study was added to the final list of primary studies in order to avoid the exclusion of any relevant content. We ended up having 279 studies after the metadata-based selection process. After reading the full text and assessing the quality of the studies based on our criteria, there were 132 academic studies left.

As the final step in academic literature selection, we applied forward and backward snowballing based on the guidelines by Wohlin et al. [35]. We received 39 more studies as a result of snowballing on the selected academic studies. Following the same inclusion and exclusion, and quality criteria we used in the prior steps, we included 18 out of the 39 studies found by snowballing. Figure 1 depicts a summary of the selection process with the number of papers selected at each step.

3.2.3 Primary Study Selection from Gray Literature. First, we collected a list of studies by applying our search string on Google Search and Arxiv. We stopped our search on Google when the results reached theoretical saturation [17], i.e., meaning that no new content related to MLOps was found.

Table 3. Inclusion and Exclusion Criteria

Academic literature	Gray literature
Inclusion criteria	
I1: Full text available and peer-reviewed articles	I1: Freely accessible full text articles
I2: Studies that relate to any phase of the ML systems production	
Exclusion criteria	
E1: Studies that are duplicate	
E2: Studies that are written in a language other than English	
E3: Studies that are not free to access the full article	
E4: Studies that only focus on enhancing machine learning models	
E5: Secondary studies (e.g., reviews, surveys, and mapping studies)	E5: Videos, audios, question-answering websites, theses
	E6: Studies do not present significantly new knowledge or insights

Table 4. Quality Assessment Criteria for Gray Literature

Category	Criteria
Authority of the producer	Author or publishing authority is reputable or has expertise in the domain Author or publishing authority has published other relevant articles
Methodology	The source has a clearly stated aim The source is supported by authoritative, contemporary references
Date	The item has a clearly stated date
Position related sources	Key related gray literature or formal sources have been linked/discussed
Novelty	The item enriches or adds something unique to the research The item strengthens or refutes a current position
Outlet type	Gray literature category (i.e., Tier 1, 2 or 3)
Tier 1 - High credibility: Books, magazines, reports, white papers	
Tier 2 - Moderate credibility: Annual reports, news articles, presentations, videos, Q/A sites, wiki articles	
Tier 3 - Low credibility: blogs, emails, tweets	

The gathered results roughly covered 50% of all search results. At the end of the study collection, we had a total number of 453 studies including search engine results and pre-prints. Second, we applied the inclusion and exclusion criteria reported in Table 3 to select primary studies for our work. We removed video contents, contents that require non-free access, and duplicate studies. We ended up having 423 studies. Third, we applied our quality assessment criteria, listed in Table 4 which are adopted from [17], on the remaining studies. We iteratively simplified the quality criteria provided by Garousi et al. [17] to ensure we captured all the relevant content. For example, we excluded many of the methodology related criteria as MLOps is still an emerging field and many are still trying to define it and its associated technologies and concepts. Hence studies tend to skip reporting experimental data or methodology. After the quality assessment, we selected studies that met at least seven out of nine criteria listed in the table.

Similar to the academic study selection process, we established consensus for gray literature selection by having two authors independently review the first 50 of the search engine results, and comparing their results with each other at the beginning of the quality assessment. Later, two authors separately continued to review the remaining studies and ended up with 41 primary studies. As the last step, a snowballing process was performed on the selected gray literature items. 27 more studies were collected from snowballing and 19 of them were selected as primary studies.

3.3 Data Extraction and Thematic Analysis

Thematic analysis is a qualitative data analysis method that is widely used in systematic literature reviews [10, 12]. For synthesizing and analyzing the qualitative data extracted from the reviewed studies, we followed the six-phase process of the thematic analysis as per the guidelines provided by Braun et al. [8].

As the first step of the process, selected articles were divided among three authors and each author familiarised them-self with data by reading the full-text and identifying content related to our research questions using a predefined data-extraction form [4, 17]. To ensure the inter-rater agreement regarding the data extraction process, a pilot data extraction was carried out on six papers by each author and results were discussed. Later, data extraction was carried out separately by splitting the primary studies among all the authors. As the second step, initial codes were generated ensuring that all data extracts from the primary studies are coded and collated together within each code. We used a shared spreadsheet as a Code Book to record codes and their definitions. This allowed authors to coordinate, validate the generated codes, and re-use codes when applicable. We also maintained an Obsidian¹ vault containing a collection of notes, where each note was created per generated code along with related data extracts from the primary studies. These notes collated data extracts to generated codes. In the third step, we analyzed the generated codes and

¹<https://obsidian.md/>

combined them to generate a collection of candidate hierarchical themes (themes, sub-themes, and sub-sub-themes). Three authors carried out this step, where each author generated the themes related to one of the research questions. The generated themes were reviewed by the other two authors to refine and validate them. The fourth step is for defining the final names and definitions of the refined themes followed by the final step of reporting the synthesized data.

3.4 Overview of the Selected Studies

There are a total number of 220 primary studies selected from both the academic and gray literature; 150 of them are selected from the academic literature while 48 of them correspond to the gray literature. Figure 2 reports the distribution of studies over time and according to their literature types. The publication year distribution of our selected studies shows that MLOps-focused studies have been increasing each year since 2019. Note that, the year 2023 column shows the statistics of studies published up to September 2023.

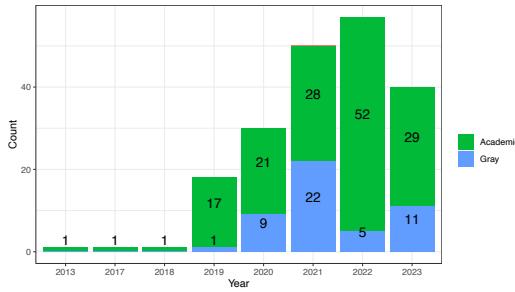


Fig. 2. Publication counts by year and literature type

4 RQ1: HOW DO RESEARCHERS AND PRACTITIONERS PERCEIVE MLOPS?

4.1 RQ1.1 - How is MLOps Defined in the Literature?

MLOps is often defined by the development, deployment, and operationalization of ML systems, as well as the application of cultural practices in ML projects to enable productionalization of ML systems. The most common definitions that occur in the literature are *continuous ML*, *DevOps for ML*, and *automation of ML pipelines*. In the following subsections, we define MLOps in more detail, from the most common definitions to less common ones, e.g., 34 studies define MLOps as “continuous ML”, 32 studies define it as “DevOps for ML”, 18 studies refer to it as cultural and organizational practices for ML. We provide the full statistics in our online appendix [3].

4.1.1 Continuous ML. Continuous ML draws from the well-known software engineering practice CI/CD (continuous integration and delivery) which aims to integrate the software frequently to deliver high-quality products (see Appendix B.1). This concept takes place in building ML systems as well, and it is often defined by the combination of CI/CD with continuous training (CT) [PS80]. MLOps brings continuous practice into many phases of ML life-cycle [PS95] (i.e., model design [PS5], model training [PS1], serving [PS67], validating [PS86], monitoring [PS5], model delivery, [PS28] and integration of pipelines [PS102]). MLOps aims at rapid development and operations through continuous practices, adapting existing solutions quickly to changing conditions (e.g. business needs, concept shift) [PS87], [PS85], [PS102]. Moreover, continuous practices allow feedback loops between different phases that are critical in ML projects due to the experimental nature of ML

engineering. For example, ML models are often updated after the initial phase of monitoring, and feature engineering and model serving tasks are repeated with different model needs [PS10].

4.1.2 DevOps for ML. MLOps originates from the DevOps concept [PS1], [PS29], [PS32], [PS141] and targets development (ml) and operationalization (ops) of ML systems [PS30], [PS88], [PS102]. MLOps adopts practices and culture associated with DevOps [PS32], [PS85], [PS156] aiming to standardize streamline development, deployment, operation, and management of ML systems [PS108], [PS115]. MLOps unifies development and operational cycles carried out by different roles, to ensure automated, frequent, and continuous delivery of high-performing ML systems [PS67].

4.1.3 Automation of ML pipelines. Another highlighted aspect of MLOps in the literature is the automation of pipelines. Automation leads to more rapid development and deployment [PS28], and easier management of the infrastructure that ML models are served [PS72]. MLOps advocates automation [PS102] at all steps of ML system development and operations including activities from data collection, model building and deployment, and monitoring [PS1], [PS17], [PS30]. Automation eases the deployment by providing a more simple release and environment configuration process, thus, researchers can focus more on model optimization [PS44].

4.1.4 Cultural and organizational practices to deliver ML systems. Developing and operationalizing an ML-enabled system necessitate collaboration between many different roles, e.g., business experts, ML engineers, and software engineers. As the number of ML applications and their complexity grows in the community, the need for establishing a standard for new roles and responsibilities, and organizational collaboration practices becomes more prevalent. MLOps is reported as a set of tools and organizational practices that lead to more sustainable development and operationalization of ML-enabled systems [PS70], through enabling collaboration across diverse teams, e.g., data science team and IT professionals [PS45].

4.1.5 Intersection of Data Science and ML. As emphasized by several studies [PS81], [PS87], [PS86], [PS96], MLOps shares a similar philosophy with DevOps, but has unique aspects related to data and model. Hence, definitions of MLOps point out that MLOps is the intersection of data engineering, data science, machine learning, DevOps, and software engineering [PS157]. Further, MLOps is closely linked with agility in data science [PS30], [PS32], [PS36], [PS83]. Specifically, producing and releasing ML models in an iterative manner is the application of agile principles [16] in ML projects.

4.1.6 ML system productionalization. One prominent definition of MLOps is the process of putting ML solutions in production [PS7]. Productionalization of ML refers to serving the ML solution to customers in a live environment, which involves tasks and challenges different than development. MLOps targets minimizing the leap between experimental development and production environment [PS163], and transferring the developed ML model to production at scale by ensuring reproducibility, reliability, and efficiency [PS71], [PS23]. Thus, MLOps involves practices to unify the development, deployment, and operationalization of ML systems through processes, tools, automation, and repetition to create a streamlined workflow from development through operationalization [PS22], [PS41], [PS89], [PS96].

4.1.7 ML life-cycle management. Another commonly emphasized definition of MLOps is the management of life-cycle, which refers to the end-to-end processes of development, deployment, and monitoring [PS37], [PS44], [PS54], [PS61]. These life-cycle phases work in flows and interact with each other through triggers, loops, and common artifacts [PS61]. Further, serving ML solutions requires integration with middleware, software systems and services [PS2], ensuring fairness of predictions [PS152], scaling ML solution accordingly to infrastructure (e.g., decentralized development

[PS64], MLOps on the edge [PS144]). In addition to that, the nature of ML models includes many complexities, some of which are noise in data, model tuning, and large amounts of artifacts. MLOps is the management, standardization and optimization of all these complexities considering reproducibility, stability, optimization, and scalability of pipelines [PS91], [PS94], [PS95], and managing various artifacts generated over the life-cycle of a project (i.e., multiple versions of experiments, datasets, models, and metadata related to models) [PS144].

4.1.8 Unify of ML development, deployment, and operations. Several studies define MLOps as unifying ML-enabled systems' development, deployment, and operations. This definition refers to supporting practices for developing [PS89], deploying [PS102], monitoring [PS25], and scaling [PS86] ML models quickly and efficiently, and managing their infrastructure configuration complexity [PS90] to ensure ML-enabled systems operate optimally and securely [PS90] within the predefined business thresholds [PS86].

4.2 RQ1.2 - What are the Primary Activities and Tasks of MLOps?

Figure 3 shows a comprehensive MLOps pipeline that includes MLOps activities. We elaborate more on these activities in the following subsections.

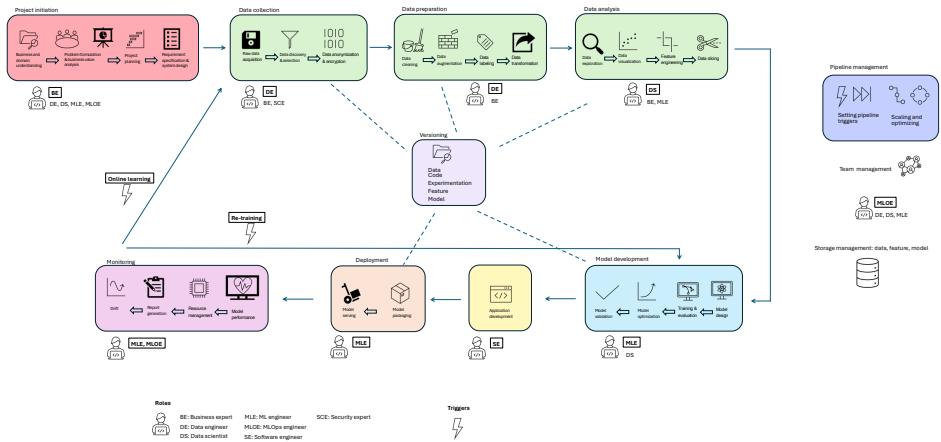


Fig. 3. A pipeline for ML development and operations

4.2.1 ML project initiation. ML project initiation mainly includes activities such as business analysis, domain understanding, planning the project, specifying system requirements, and designing the ML system. Many stakeholders take part in these activities, i.e., business analysts, business owners, domain experts, data scientists, and ML and MLOps engineers.

Specification of business and system requirements is crucial for the planning and execution of an ML project, and for the successful deployment and operation of the ML system as well [PS105], [PS111]. One of the initial steps in identifying business requirements is the formulation of the business problem to establish a clear idea of the problem to be addressed with the ML system [PS111], [PS41]. Formulation of the business problem helps the team evaluate the business value to be gained by solving the problem [PS111], analyzing whether the problem could be addressed by building an ML system, and the feasibility of engineering of an ML system [PS116]. Identification of business requirements includes further steps such as obtaining security objectives of the company,

e.g. confidentiality and privacy requirements of the data used through ML system [PS105], and regulatory rules [PS99].

Planning an ML system also includes tasks related to software project management knowledge areas [7]; i.e., determining the scope of the project [PS90], specifying ML use cases [PS70], designing the model architecture, pipeline, and infrastructure of the system [PS90], [PS70], identification of performance metrics that will be used to assess the ML model [PS111], determining how statistics will be collected during the monitoring phase and how the feedback of business owners will be received [PS70].

4.2.2 Data collection. Data collection is the first step in the ML pipeline, referring to the *acquisition of raw data* that will be processed in later steps to train a ML model. Often, raw data is gathered from relevant data sources, e.g., sensors [PS24], GitHub code repositories [PS27], or cameras [PS34].

Some studies have more exploratory data collection steps. For example, data collection is combined with *data discovery* and *data selection*, i.e., collecting data for a drug discovery task requires performing experiments to produce data and selection of data by domain experts to be used in later phases [PS37]. On the other hand, there are cases that follow a prior data management plan prepared by data engineers to collect data to be fed into the ML pipeline [PS117]. For security-critical systems, data collection process is applied with anonymization and encryption of the collected information to ensure data privacy [PS31].

4.2.3 Data preparation. This step aims to prepare data for the ML model building phase, by processing the data including cleaning the raw data, data augmentation, labeling data points, data transformation, and feature engineering according to the requirements of the ML model.

Data cleaning refers to detecting and fixing errors in raw data, e.g., duplication removal [PS24], removing invalid and irrelevant data points [PS27], [PS111], imputing missing values in data [PS37], [PS117], [PS111]. This step is also referred to as *data filtering* in the literature [PS27].

Data augmentation is an activity that helps to increase the data sample size to overcome problems that might occur due to limited data availability. Applying oversampling and using simulation are two techniques mentioned [PS55].

Data labeling is the task of marking data points according to the observations in the problem domain. Data engineers, data scientists, and domain experts take part in this task [PS117]. While data labeling is a manual task in many cases, there are tools that support automated labeling [PS111].

Data transforming means adjusting the format of the data in accordance with the needs of the ML system to be built, e.g., removing some characters to make the data content available for use in other file systems [PS27]. This step also includes converting data into another representation format, i.e., applying distributional transformation techniques on numerical data [PS31], and applying data discretization for converting the data into smaller units [PS31].

4.2.4 Data analysis. Data analysis aims to explore the data statistics, e.g., the mean and standard deviation in data, quantiles of data [PS30], [PS62] to deeply comprehend the data before the model development step. It helps to understand and assess which features [PS112] and data slices [PS30] are best for the ML model requirements. Data analysis also offers *data visualization* using dashboards to the team to make their evaluations of data easier [PS25], [PS42], [PS114], [PS109]. Moreover, data analysis helps data validation, which aims to assess the health of data, i.e., data might involve anomalies or unexpected properties that might affect the model quality [PS62].

Feature engineering focuses on processing features in data to create successfully performing ML models, i.e., improving the ML model's prediction performance and outcome accuracy [PS64], [PS134], [PS112]. In the literature, feature engineering is mentioned with respect to the following

sub-steps: feature transformation, feature selection, feature extraction, feature construction, and feature storage [PS162], [PS114], [PS115], [PS102]. Our analysis shows that feature engineering is mostly seen as the responsibility of data engineers and scientists. Also, as with many data-related MLOps activities, feature engineering uses the expertise of other roles such as security [PS105] and product experts [PS31].

Another mentioned activity in the literature that falls under data discovery through data analysis is *data slicing* [PS137]. Data slicing is the process of finding the subset of data that performs according to the use case. It overlaps with the model building and evaluation steps since it explores the interest in data through evaluating the ML model's prediction performance on different data slices.

4.2.5 Model development. This activity is the traditional step in ML life-cycle that aims to construct the ML model(s). Model building includes model design, training, evaluation, selection, validation, and model re-training activities. The model building partially overlaps with data analysis since the building of an ML model often involves exploring multiple training approaches, i.e., ML algorithm and hyperparameter configurations [PS114].

Model design refers to designing the appropriate architecture or selecting the appropriate tools for ML model construction to ensure successful model training and prediction performance. Examples are designing the architecture of neural network models [PS39], selecting the right algorithm for the problem domain [PS24], and designing the best performing ensemble [PS31].

Model training is building an ML model that will often be used for prediction. Traditionally in the ML life-cycle, the model training step takes the prepared dataset and by utilizing a learning algorithm it creates a prediction model learned from the prepared data [PS111], [PS103], [PS120], [PS125].

Model evaluation is assessing the predictive performance of the ML model. This step is traditionally done by testing the ML model with a held-out test set, which is a separate set of data points that are not used to train the model [PS29, PS139, PS141, PS114, PS102]. The model's performance on the test set is evaluated using appropriate metrics for the use case [PS120].

Model optimization refers to improving a model to make it more successful in terms of performance. One of the most common approaches used in the literature is hyperparameter tuning [PS24], [PS102].

Model selection step deals with deciding the most appropriate model for the problem domain among various models built for the same ML system [PS1], [PS7], [PS152].

Model validation step is the evaluation of the model according to the business goals [PS70], [PS110], quality standards [PS62], [PS114], and regulatory requirements [PS110], [PS114]. Deployment of a model is decided based on the model validation results. Model validation often requires the supervision of a manager [PS34], and the collaboration of data scientists, business stakeholders, and ML engineers [PS114]. The common method for validating the quality of an ML model is comparing its predictive performance against a baseline model or a fixed threshold value [PS62], [PS157], [PS102].

4.2.6 Deployment. This is the process of deploying trained ML models into production environments, where the model predictions are accessible to users. Depending on the use case, deployment is done by integrating an ML model into a software application or serving the model as a service.

As reported in [PS117], "the deployment stage is a pivotal moment in the MLOps process", mainly due to the differences between the phases prior to the deployment and after the deployment. The majority of the tasks prior to the deployment phase correspond to the traditional ML tasks, i.e., data collecting, data preparing, and model building, that are under the responsibility of data and ML-related roles (i.e., data engineers, data scientists, and ML engineers). However, deployment

related tasks differentiate from the routine ML activities [PS143], mostly include model packaging [PS138] and engineering software infrastructures, e.g., deployment in cloud [PS102]. Deployment is the phase in which engineers from different backgrounds, i.e., ML engineers and software engineers, are involved in the process. In prior phases, engineers and data scientists mostly work experiments in a research environment. During the deployment phase, it is software engineers and MLOps engineers working on the smooth integration of the trained ML model into a running production environment [PS117], [PS110], [PS108].

The deployment phase addresses various aspects of ML operations, i.e., model packaging to share the trained models with the community [PS24], packaging and creating an image of the model that also includes its dependent libraries [PS70], [PS141], [PS144], [PS139], [PS138].

Deployment methodologies and complexity of the deployment vary in the literature depending on the target ML system [PS24], [PS127]. A common way of making the predictions of an ML model available to users is providing an application that enables users to query the ML model and serve its predictions [PS29]. Many studies prefer the *model as a service* approach which allows serving the model through REST API end point [PS29], [PS37], [PS141].

4.2.7 Monitoring. Monitoring is one of the key activities of MLOps that is applied by many studies and supported by many tools [PS26], [PS32], [PS152], [PS114], [PS109], [PS148]. It helps to improve the system's health through receiving feedback which often leads to repetition of the necessary ML tasks.

Monitoring activity aims to ensure the service quality and operational stability of the ML system in production by continuously observing the system resources, dependencies, and performance to capture the issues in a timely manner [PS145], [PS117]. The task results in a notification of the situation of the system [PS109] and often leads to an invocation of earlier tasks in the pipeline [PS102], [PS114]. Besides the detection of erroneous cases, monitoring is also used to gain perspective on how pipeline tasks are running and how accurate the model's performance is [PS91]. Monitoring also helps to increase users' and business stakeholders' trust by enabling system observability [PS26], [PS146], [PS100].

The most commonly monitored subjects in ML systems are concept drift, quality of data and model, performance of the model, usage of computational resources such as disk usage, and non-functional requirements of the system such as security [PS24], [PS32], [PS114], [PS111], [PS148].

Monitoring is under the responsibility of MLOps engineers [PS137], [PS111]. However, in many cases, other roles are involved in monitoring, i.e., product performance experts, site reliability engineers, and solution engineers consume the output of the monitoring to examine the issues [PS31].

Systems and organizations that are at a more mature stage e.g., highly regulated industries like financial services [PS99], are also concerned with the explainability of ML models as part of monitoring operations [PS99], [PS25]. Explainability helps a better understanding of the model's predictions, i.e., visualization of the models' decisions [PS17], and providing a rule-based root-cause analysis [PS23].

4.2.8 Pipeline management. An MLOps pipeline is a flow of traditional ML life-cycle tasks and ML operations. These pipelines can easily become complex since there is a variety of activities that connect with each other through the MLOps life-cycle. Management of pipelines aims to execute MLOps tasks efficiently and reliably by addressing pipeline design, orchestration, and optimization.

Pipeline automation aims to reduce human intervention to a minimum degree by enabling the execution of MLOps tasks automatically and easing the handling of repeating complex ML tasks and operations [PS42]. Many studies work on the automation of MLOps tasks through a pipeline designed according to their use case. For example, one study focuses on the automated pipeline of

data downloading, data processing, model training, and testing steps [PS3], and another focuses on the automated pipeline of the model configuration, serving, and monitoring [PS44].

Automation requires *pipeline triggers* to enable the automated repetition of the pipeline tasks. At the most basic level, triggers are set to execute a task in the pipeline when the prior one is completed [PS42]. But pipelines can also be triggered based on multiple factors depending on the use case [PS102]. Tasks in the pipeline can be triggered periodically according to a scheduled period [PS25], [PS157], [PS152], deployment can be triggered when a newly trained model is available [PS148], the training task can be triggered when new data is available or the model performance is reduced [PS141], [PS69], training can be triggered when the distribution of data significantly changes (concept drift) [PS102].

Scaling pipelines has importance in today's infrastructures. More specifically, ML-enabled systems work on edge devices, and many ML systems and tools are served to different customers. So, the same pipeline structure may not fit all. Scaling and adjusting the complexity of pipelines is important. Keeping data pipelines scalable to manage the evolving customer requirements is proposed [PS42], Airflow supports the scalability of pipelines to handle complex and dependent structures [PS153].

4.2.9 Artifact management. Versioning the code, data, models, experiments, metadata, and documents is an important practice in ML projects as it is in traditional software engineering projects. The aims of this activity involve keeping a record of history and versions which allow easier collaboration among team members [PS33], [PS36], debugging the errors and backing up the system [PS157], [PS102], artifact reuse [PS32], creating an audit trail for quality and risk assessments [PS24], [PS134], providing transparency [PS41] in the project.

Specifically, studies focus on versioning data science experiments [PS122], [PS108], [PS100], codes [PS24], [PS36], ML models [PS95], [PS65], training metadata [PS134], parameter weights used for model training [PS157], tracking software versions running with the ML model [PS41], [PS42], logging ML pipeline executions [PS157], [PS102], logging monitoring events [PS33], and logging container data [PS90]. Further, data [PS24] and model [PS190] provenance are stored which are the records that keep the origin and lineage of data and model in ML enabled systems [PS66].

4.2.10 Quality assurance. The quality of ML models and systems through their lifecycle is important. Quality assurance covers data quality and model quality, as well as covering the overall system's quality. Data quality addresses accessibility, completeness, consistency in data format and structure, accuracy (e.g., realistic values), data heterogeneity, fairness of collected data [PS24], [PS152]. Model quality mostly addresses the assessment of the predictive performance of the models through the evaluation of performance metrics and test set [PS21], [PS107]. Model quality assessment often results in model optimization. Model quality also refers to bias and fairness [PS100], [PS152].

4.2.11 Collaboration. As MLOps become popular in recent years, organizations needed new collaboration practices for smoother communication across people from different backgrounds (i.e., ML engineers and software engineers) to manage the MLOps projects smoothly [PS151]. One of the important aspects regarding the collaboration theme is identifying collaboration points for MLOps teams to organize the communication between roles from different backgrounds [PS151]. These collaboration points ease communication across different people.

4.3 RQ1.3 - What are the Practitioner Roles and Responsibilities in MLOps

4.3.1 Data Engineer. Mainly, data engineers are responsible for fundamental data activities and enabling data operations in a high-quality, secure, and scalable way [PS1], [PS24], [PS42], [PS85].

In a typical MLOps project, data engineers' main set of tasks are identifying suitable data sources, checking the legality of data, collecting data from various sources, cleaning them, and preparing the labeled data [PS5], [PS51], [PS93]. Here, data engineering work is characterized as "extract-transfer-load" process [PS112], [PS106]. It also involves transferring data into convenient formats to make it consumable by data scientists and ML engineers [PS112], [PS106]. The essential part is gathering and preparing the data for further steps which are data analytics, ML engineering, and operations. These steps are often executed in collaboration with domain and business experts, data scientists, and ML engineers.

Some organizations have specialized roles within their teams whose responsibilities overlap with a typical data engineer's responsibilities. For instance, *data curator* provides data and implements custom data acquisition adapters [PS19] while *data expert* applies specific noise reduction techniques to prepare data for further steps [PS5].

The responsibilities of data engineers are not limited to basic data-collection and preparation activities, they are also involved into data operations, i.e., setting up data storage infrastructures and building automated data pipelines [PS1]. Ensuring that the ML system's infrastructure is right for data storage and other data operations is managed by data engineers [PS69]. Leading by the size of data and structure of the system, data engineers build data storages and pipelines, e.g., storing data curated from distributed sources in a centralized data lake [PS69], enabling big data streaming platforms [PS125]. Designing and constructing automated data pipelines is critical for making data accessible to the MLOps team for ML operations at scale [PS72], [PS85], [PS106], [PS112], [PS109].

Data engineers are also responsible for security controls of data, data privacy, data quality, and governance [PS86], [PS106]. While data security, quality, and governance responsibilities are shared among many roles, i.e., data engineers, MLOps engineers, domain and business experts, and security experts, there are two specific roles defined for mainly these purposes; *data steward* who is responsible for data quality management [PS24], [PS117], and *data governance officer* who ensures data governance and privacy [PS114].

4.3.2 Data Scientist. Data scientists are mainly responsible for exploring and analyzing the data made available by data engineers and then training, tuning, and evaluating a model to derive insights from the data [PS122], [PS124], [PS187], [PS161], [PS104].

In a typical MLOps project, the initial responsibility of a data scientist is to understand the business problem and its domain, and then evaluate the feasibility of solving the problem through building an ML-enabled system [PS24], [PS114]. At this step, data scientists work with domain and business experts to discuss objectives and key results (OKRs) [PS86]. During this step, experiments conducted by data scientists play a significant role in evaluating the suitability of the collected data to address the business problem, as well as evaluating the effect of feature engineering and model design for more feasible ML solutions.

Feature engineering is often conducted by data scientists to analyze the best combination of features in data to train more feasible ML models. Later, data scientists conduct model training and evaluation tasks to generate various models to select the most appropriate, e.g., the best-performing, one among them. During this stage, dataset splitting into training, test, and validation sets, choosing the suitable ML algorithm, training the model, choosing suitable model evaluation metrics (e.g., accuracy), evaluating the model accordingly, and selecting the suitable model are the responsibilities of data scientists.

These responsibilities of data scientists necessitate collaboration with other roles, namely business experts and ML engineers. For example, during model validation, data scientists need to assess whether the trained model fits technical, business, or regulatory requirements [PS114].

Experiments conducted by data scientists typically iterated through multiple iterations until a defined goal (e.g., accuracy threshold for the model) is reached [PS24]. Writing experiment codes and ensuring the trackability of multiple versions of these codes, datasets, and models are the responsibility of data scientists.

Moreover, as data engineers are responsible for the initial part of an ML pipeline, data scientists are responsible for designing and building scalable and high-quality data analytics pipelines for experimentation and model training. A data scientist's last responsibility in the ML pipeline is early monitoring [PS114], [PS93]. When a model is deployed, monitoring its early results is often accompanied by data scientists as well as ML and MLOps engineers, since the early results may not be as expected which often leads to returning to the feature engineering, model training, and validation steps.

4.3.3 ML Engineer. ML engineer and data scientist roles may overlap, or sometimes these roles are mentioned interchangeably [PS135], [PS141]. However, many studies distinguish definitions of both roles [PS114], [PS154]. According to those, an ML engineer's task is primarily model optimization, deployment, and monitoring.

ML engineers' tasks are experimenting with datasets [PS118, PS106], experimenting with models [PS118], and optimizing the model and its parameters [23], [5], [PS118], [PS107], [PS108]. These are also the responsibilities of a typical data scientist in an MLOps project.

In MLOps projects, ML engineers' responsibilities are not limited to these: They contribute to the monitoring and production activities [PS114], [PS107]. ML engineers collaborate with data scientists and business stakeholders to ensure the model selected for deployment confirms the necessary level of performance and business requirements. ML engineers take responsibility in the deployment and monitoring stages of the ML pipeline. ML engineers take responsibility for executing a suitable deployment for the business use case, e.g., online serving, and monitoring the deployed model's performance in the production environment.

4.3.4 MLOps Engineer. MLOps engineers are responsible for building automated pipelines, maintaining the operational stability of the entire ML system, and collaboration across different roles [PS106], [PS117], [PS151].

An automated pipeline for data collection, data analysis, model development, and deployment phases is designed and practiced by data engineers, data scientists, and ML engineers respectively. However, the main responsibility for building and optimizing these automated pipelines belongs to MLOps engineers. Particularly, MLOps engineers integrate different services to pipelines (e.g., cloud, container), maintain the management of dependent packages, versioning and tracking mechanism of the project, and overall platform governance [PS106].

Monitoring the data, model, and application is often carried out by MLOps engineers along with ML engineers [PS117]. MLOps engineer steps in when human intervention is needed during monitoring, i.e., model needs re-training in the case of data shift, dependencies between data and model need to be resolved [PS105], [PS137]. All these tasks require managing the communication across different roles in an MLOps project [PS105].

4.3.5 Software and application developers. take crucial responsibilities in many MLOps teams, e.g., [PS144], [PS104]. They develop the final applications and services that help ML production. In many cases, they collaborate with data scientists and ML engineers in analyzing and processing data, and building models [PS115], [PS151]. Besides these, software engineers are also responsible for the deployment, packing, releasing, configuring, and monitoring of ML systems [PS161], [PS144], [PS154]. Further, the security of the ML system is maintained by software engineers specialized in security [PS105].

4.3.6 Operational engineers. focus on mainly monitoring activities [PS104], while they also take roles in deployment operations, i.e., packaging, releasing, and configuring of ML systems [PS144].

4.3.7 Technical and solution architects. are responsible for choosing the right technologies and design paradigms for building ML systems, integrating ML systems into cloud-based applications, and ensuring the building of secure ML systems [PS105], [PS155].

4.3.8 Quality, governance, risk, and compliance roles. are not typically engaged in day-to-day MLOps operations [PS114], as their primary tasks revolve around quality assurance, risk management, and governance. Studies define various roles for quality and governance across multiple MLOps aspects such as data, model serving, and model performance, etc. Particularly, *data governance officer*, *data steward*, and *data quality board* are the roles mentioned in the literature for being responsible for data governance, data quality management, and data privacy [PS117], [PS114]. *Compliance teams* are committed to mitigate unfairness and bias in ML systems across model training, deployment, and monitoring phases [PS108], [PS152]. Recently, *risk professionals* become involved in MLOps projects to manage data and model-related risks [PS155]. Specifically, *model risk manager* is responsible for ML model inventory management, model explainability, and pipeline management [PS111], [PS106], while other studies only mention risk professionals for managing any data and technology related risks [PS155].

4.3.9 Domain experts. These are the people who have expertise and knowledge about the problem domain, e.g., clinical neurologists for an ML-enabled psychological monitoring system [PS174], urban planners for a smart city system [PS30]. Domain experts play a crucial role in domain discovery [PS41], engineering of domain specific requirements [PS24], and fulfilling business goals [PS41], [PS117] due to their deep understanding of the problem domain and environment. They are “irreplaceable actor within ML projects” [PS24]. Moreover, they also take part in collecting, cleaning, and labeling data, interpreting data distributions, developing and validating hypotheses about data, model evaluation and interpretation of the results [PS32], [PS151], [PS111]. According to the literature, domain experts either collect and process data and transfer it to the MLOps team, or they are directly accessible to the MLOps team for collaboration, especially with data engineers, data scientists and ML engineers [PS144], [PS174], [PS173].

4.3.10 Business analyst and experts. Analyzing the problem that needs to be solved, defining clear objectives, and assessing whether an ML-enabled system could answer the problem are the primary tasks of these people [PS86], [PS175], [PS154], [PS93], [PS97]. The MLOps team collaborates with business experts to gain a better understanding of the problem and convert it to a successful ML solution that aligns with the business goals. Depending on the domain and project organization, business experts’ and domain experts’ responsibilities may refer to the same tasks; both help people to understand the problem (business) domain.

4.3.11 Collaborative activities among roles. In a typical MLOps pipeline (Figure 3), various roles collaborate with each other for various responsibilities. Data engineers, data scientists, and ML engineers are mainly responsible for carrying out data collection and preparation, data analysis, and model development and deployment stages, respectively. However, these roles often collaborate with business and domain experts. Particularly, data engineers work together with data scientists and ML engineers to decide on the appropriate data format necessary for further steps. Data scientists visualize the data analysis results and receive feedback from ML engineers on the potential model performances. Data scientists also provide insights to ML engineers regarding model design and optimization. Data engineers, scientists, and ML engineers often collaborate with business and domain experts to interpret the data and model. ML engineers and sometimes data scientists

accompany the MLOps engineers at the initial monitoring to ensure the stability of the model and step in when the model or features need optimization. On the other hand, MLOps engineers interact with every role to ensure the optimization of the automated pipeline, the operational stability of the ML system, and the smooth collaboration across various roles.

5 RQ2: WHAT ARE THE STATE-OF-THE-ART PRACTICES AND TECHNIQUES PROPOSED FOR IMPLEMENTING MLOPS?

Here, we report the state-of-the-art work practices, processes, techniques, and supporting technologies that aid organizations in implementing MLOps. These MLOps practices cover a broad spectrum of social and technical aspects across the life cycle of ML models. We have grouped and reported them under seven categories below.

5.1 Adapting Team Structure and Work Practices for MLOps

The productionalization of ML models is a multifaceted challenge that requires the expertise of stakeholders with diverse skill sets, from domain knowledge to data science, software engineering, system operation, and cyber security. Following best practices are introduced to enable the successful implementation of MLOps by fostering effective collaboration among such diverse stakeholders.

Establishing balanced project teams: The AWS Well-Architected Machine Learning framework [PS111] and Azure Machine Learning best practices [PS116] recommend establishing *project teams*, which comprise of domain expert, data engineers, data scientists, ML engineers, MLOps engineers, IT auditor, and cloud security engineers to work with ML-based software features and systems, rather than building functional silos for each capability, e.g. ML engineering team, security team. A project team should be configured with balanced capabilities such that it can carry out all phases of the life cycle of an ML model.

Establishing specialized ML engineering teams as ML projects grow: As organizations grow in ML utilization, the number of ML projects maintained by a project team could rapidly outgrow the ability of the team to scale. At this state, a helpful practice is establishing a specialized ML engineering team that focuses on the deployment, operations, and maintenance of ML models. These capabilities are common across ML projects and uncoupled to domain specific ML problems. By centralizing these capabilities into a specialized team, an organization can develop, manage, and utilize these capabilities more efficiently. Specialized ML engineering team along with standardized processes and reusable patterns and artifacts lay a foundation for establishing “AI Factory”, scaling ML productionalization to ML *industrialization* [PS116].

Adopting a Git-based workflow: ML project teams can leverage Git-based workflows to coordinate and carry out the day-to-day activities of ML projects. Git allows team members to work on isolated branches without disrupting the stable version of the project as well as enabling the automation of CI/CD activities. Gunny et al., [PS132] proposes a branch structure consisting of two stable branches (i.e., dev and main) and several short-lived feature branches to integrate model training, testing, and deployment activities into pull requests under two main stages. Pull requests from feature branches to dev branch would trigger model training and testing within a staging environment, thus ensuring the model’s performance. Pull requests from dev to main branch would trigger the packaging of the model as software containers and deployment of the packaged model into a production environment.

Integrate practices for cross-silo concerns across ML life cycle: Cross-silo concerns include AI fairness, AI explainability, and security. Instead of addressing these concerns at any particular phase, they should be a cross-silo effort, which each part of the life cycle implementing different

aspects. Amazon AI Fairness and Explainability Whitepaper [PS152] recommends integrating bias measurement before-, during-, and after-training, as well as in production. The detected bias then can be addressed using pre-processing, training tuning, and post-processing.

Security is another cross-silo concern of ML life cycle. Zhang and Jaskolka [PS105] introduce the concept of Secure Machine Learning Operations (SecMLOps) based on the People, Processes, Technology, Governance, and Compliance (PPTGC) framework. SecMLOps embeds security controls across the whole life cycle. In the requirement phase, security objectives and requirements of ML models and data are defined, and security models are established. These models drive the development of security policies regarding data, model, physical security, personnel security, administrative security, and network security across the ML life cycle and all involved parties. These models also drive the development and integration of security monitoring and incident response components that monitor ML models in production.

5.2 Improve Transparency and Provenance Across ML Life Cycle

The goal of improving the transparency across the ML life cycle is twofold. Firstly, improving transparency helps project teams avoid or reduce duplicated work, such as building and running data pre-processing and feature engineering pipelines. A closely related advantage is avoiding unexpected dependencies between projects, such as via a shared feature set. Secondly, maintaining the provenance of data, models, and features facilitates project teams to identify the root causes of production issues and roll back ML systems to the previous working state when production issues happen. The improvement of transparency and provenance can happen along three axes: models and experiments, datasets, and features.

Applying version control to experiment codes and artifacts: Existing guidelines [PS116], [PS111] advocated using version control systems such as Git to manage all inputs that go into experiments, such as software codes for model training and testing, datasets, environment configurations, and other experiment parameters. Reproducibility is a key benefit of version control. However, the benefit of this practice extends beyond reproducibility. It provides project teams with the ability to roll back the system to the last known functional version of an ML system, thus facilitating recoverability and improving the overall fault tolerance of these systems.

Maintaining history of experiment runs: The probabilistic nature of many ML models makes it possible for a set of version controlled source codes and artifacts to produce ML models with varying performance characteristics across experiment runs. Therefore, it is beneficial to maintain the historical data of experiment runs of the model [PS116], in addition to the version controlled source codes and artifacts. Experiment versioning plays an important role in ensuring the repeatability of experiments and improving their conclusiveness [PS132]. The historical data of experiment runs could provide additional transparency into the decision to upgrade models. This data would also be invaluable for planning the resource provisioning and optimization for the model training process. Experiment management tools such as MLflow [PS164] and IBM AI Factsheets [PS165] enable experiment tracking along with model and data versioning [PS154]. By integrating tools like AutoML with Experiment management tools, practitioners can automate the generation of the best suited models [PS154].

Tracking and Maintaining Lineage and Provenance of ML Models: Lineage and dependencies among models and datasets are other aspects that project teams need to capture to develop a comprehensive picture of ML model life cycle. Closely related to lineage is the model's provenance, containing its machine-readable metadata and historical events within its life cycle, such as how and by whom it was trained and tested. The IBM MLOps handbook [PS165] highlights two main

needs for capturing metadata and recording historical events regarding the life cycle of ML models: improving the productivity of ML model productionization by making ML model lifecycle related metadata visible to all stakeholders, and facilitating the regulatory compliance checking and explainability of the models. Academic literature proposes multiple provenance extraction tools such as Vamsa [PS130], Geyser [PS130] and MLflow2PROV [PS164]. Vamsa is a knowledge-base driven *static* provenance extraction tool. Geyser extends Vamsa for dynamic provenance extraction, provenance storage and querying. MLflow2PROV was designed for extracting provenance from version control system, experiment management systems and store in a central database for querying and visualization. IBM AI Factsheets [PS165] capture model and process metadata such as model details, training information, metrics, etc., and store in a central meta-store.

Defining concrete model management strategy: A model management strategy defines how models are identified, named, and labeled. An organization might also include requirements on the type and format of metadata, provenance, and lineage information to be maintained by project teams. A critical element of a model management strategy is the process and criteria for replacing models. For example, the Azure Machine Learning best practices [PS116] recommended labeling a baseline model as “champion” and new models as “challengers.” Strategies can be defined to leverage experiment results captured in the historical data of experiment runs to decide when and how to replace the “champion” model.

Establishing centralized data repositories: Centralized repositories provide project teams with visibility into the available data and its utilization to avoid redundant effort and potential dependency problems [PS155]. Centralization of data also helps organizations to enforce a consistent access control policy and allows them to implement best practices regarding data privacy and bias that are relevant across projects. Centralizing data also facilitates the implementation of techniques such as real-time monitoring and alerting of data drifts [PS106] and automated detection of data dependencies. For instance, Boue et al. [PS146] propose a dependency mapping framework to analyze and notify MLOps engineers of changes in data sources and their effects. Another advantage of centralized data repositories is resource utilization efficiency because they allow organizations to employ specialized and scalable data technologies, such as transactional databases supporting structured or semi-structured data [PS134], data lakes [PS110], and data warehouses [PS154] where version controlled datasets are stored to maintain links between data, code, and models.

Establishing feature stores: Features can become a point of friction between the data science and operation sides of an ML model, especially if the feature construction code is not well documented and lacks scalability. In these cases, the feature engineering code might need to be reconstructed. However, this process might introduce subtle changes to the features that make model’s performance deviate from the experiment results. At the same time, reusing features and feature engineering pipelines might lead to undocumented coupling between projects and unexpected consequences of changes. Establishing centralized feature stores lays a foundation to address these challenges, because they provide visibility into existing features and feature engineering pipelines to avoid redundant effort. Moreover, feature stores facilitate feature governance by making features visible and reusable among different stakeholders in a secure manner [PS112], [PS113], [PS154]. Some existing feature store platforms and technologies include Amazon SageMaker Feature Store² and Databricks Feature Store³.

²<https://aws.amazon.com/sagemaker/feature-store/>

³<https://www.databricks.com/resources/demos/tutorials/data-science-and-ai/feature-store-and-online-inference>

5.3 Continuous Monitoring Across ML life cycle

Establishing continuous model performance monitoring: The concept of “model performance” has multiple facets, such as prediction accuracy, robustness, fairness, explainability, and safety. Once the required performance metrics have been established during project initiation, they can be implemented as automated monitoring jobs. Various tools exist for monitoring production ML models. For instance, AWS Sagemaker Model Monitor [PS107], [PS166] enables detecting data drift, concept drift, bias drift, and feature attribution drift of deployed ML models. Sagemaker Clarify [PS152] can detect bias to ensure fairness of models deployed on Amazon Sagemaker. Similarly IBM OpenScale [PS165] monitors models for model accuracy, fairness and data drift. Databricks Lakehouse Monitoring [PS114] maintains inference tables with request-response data to detect data and model drift. Jayalath and Ramaswamy [PS145] highlight the importance of user feedback for the success of deployed ML models and propose a monitoring framework to that utilizes user feedback to investigate data deficiencies. Visual analytic tools such as Slice Teller [PS137] would be invaluable in aiding human experts in such cases when human inspection of model’s *local* performance on some particular segments of data is necessary.

Establishing continuous data monitoring and validation: As organizations have more data over time, they often face issues of data drift, bias, data quality and security (e.g., poisoning). Continuous monitoring and validation of datasets are practices to address these challenges. Data validation tools such as Tensorflow DataValidation⁴ and Amazon DeeQu⁵ are used to write declarative data quality constraints to catch data quality issues as they occur. Further, Tu et al. [PS131] extend this concept to auto-program data quality constraints by leveraging statistical information from past executions.

Leveraging events to detect and response to model and data issues: The continuous monitoring of models and datasets can produce the signals to facilitate further analysis and response. Event-driven programming approach is used to take advantage of these signals [PS116]. In particular, monitoring platforms are considered as event emitters. These events can be published on a centralized event bus that is available to all teams. Authorized team members can subscribe to events from the bus to be notified for further analysis and response.

5.4 Systemizing the Provisioning and Configuration of Infrastructure and Environment

The performance of ML models, particularly deep learning variants, depends not only on the input data and model parameters but also on the software environment surrounding them. Therefore, it is critical to maintain the consistency of the software dependencies and configurations across training, testing, staging, and production environments. Practices such as infrastructure-as-code (IaC), automated infrastructure provisioning, and Inference-as-a-Service provide the necessary systematization to address the environment consistency challenge.

Leveraging IaC to ensure consistency between environments: The IaC practice ensures that software environments are constructed and configured consistently according to predefined specifications. These specifications can be captured as software artifacts, stored in source code repositories, and managed by version control systems. Treating environment specifications as code artifacts enables project teams to collaborate on the development, maintenance, and improvement of these configurations. Existing IaC tools such as Ansible and Terraform can be readily adapted to ML projects. Existing literature has also proposed ML-specific tools, such as the Pinto utility [PS132] that automatically creates software environment based on the specification of an ML project.

⁴https://www.tensorflow.org/tfx/tutorials/data_validation/tfdv_basic

⁵<https://aws.amazon.com/blogs/big-data/test-data-quality-at-scale-with-deequ/>

Leveraging Inference-as-a-Service: Inference-as-a-Service is an approach to centralize the execution of ML models to a service provider. Instead of setting up local environment to run a model, entities and software systems needing model's inference would request the inference results from the service provider via a predefined service API instead. By removing the need to run models locally, Inference-as-a-Service can reduce the number of production environments to be setup, thus mitigating the risk of organizations using inconsistent environments. Another advantage of this approach is the maximization of resource utilization: the centralization of AI inference can replace multiple under-utilized infrastructure across entities that run the model locally with a fully-saturated infrastructure at the centralized service provider [PS132]. NVIDIA Triton Inference Server is an open-source software that could be leveraged to implement AI inference services.

Automate provisioning of training infrastructure: An advantage of the systematization of environment creation is the possibility to provide data scientists with the ability to provision training infrastructures such as virtual machines by themselves. For instance, AWS [PS111] provides managed training infrastructures for optimum automated resource provisioning during iterative training process through Amazon Sagemaker. Sagemaker monitors the resource utilization of GPUs, CPUs and network bandwidth to make auto scaling of training infrastructure, thus abstracting the infrastructure management activities from the users and reducing training costs through use of on-demand computing resources. Raffin et al., [PS134] propose the use of Infrastructure As Code (IaC) and declarative configurations (i.e., Yaml manifests of Kubernetes defining the desired state of operation) to automate provisioning of underlying infrastructure in a reliable, reproducible and scalable manner.

A related problem is the monitoring of system metrics (GPU, CPU and I/O utilization) to provide insights into overall performance of the training infrastructure. Rauschmayr et al., [PS167] introduces a profiling tool for Deep Learning model training clusters, which can be integrated to Amazon Sagemaker as an add-on. Profile monitors the system utilization metrics related to training jobs to identify performance bottlenecks (GPU, CPU and I/O bottlenecks) and provide insights for optimizing resource usage during model training phase.

5.5 Systemizing Model Deployment

The process of translating data science codes into production codes and maintaining them brings about a variety of challenges. Some patterns for model deployment and maintenance have emerged from the literature and best practices.

Using software container to package and deploy ML code:

A common practice is packaging ML code with all of its dependencies within software containers, and then applying optimizations such as parallelization and load balancing upon the containerized workloads [PS179], [PS172]. Containerized workloads can also be operated at scale by software container orchestrate platforms such as Kubernetes [PS156]. By encapsulating ML models as docker containers practitioners can reap the benefits of container management platforms such as AWS Fargate which provides serverless options for ML model deployment [PS111]. Kubernetes enables automated deployments, monitoring, and scaling of ML models to ensure efficient operation of deployed ML models, thus overcoming limitations of manual model deployments [PS128]. Syrigos et al. [PS128] propose an extension for Kubernetes Scheduler where container scheduling decisions are made in a latency-sensitive manner in addition to CPU and RAM of the computing resources to deploy ML models.

Abstracting data acquisition: Deployed ML systems do not exist in vacuum. Instead, they need to pull data from various sources and perform pre-processing tasks such as merging datasets and

cleaning. Therefore, containerizing ML codes by itself is inadequate for addressing productionalization challenges. An approach to abstracting the data acquisition activities as services can help attain scalability of data processing with increasing volumes of data. [PS179] proposes a pattern to abstract data sources and introduce a software agent that pulls and feeds data to containerized ML code. This design introduces the necessary decoupling to parallelize the containerized ML code.

Applying patterns for upgrading models: To prevent potential failures after updates to ML models, various deployment strategies for a safe upgrade of ML models in production are proposed. AWS [PS111] highlights the importance of conducting a trade-off analysis to compare multi-version deployment strategies like A/B testing, canary deployments, and blue/green deployment to select the best approach. Such strategies facilitate rollback or roll-forward decisions to be made after evaluation of the performance of the updated model without affecting the availability of the deployed service. To this end, Guissouma et al. [PS168] propose a multi-version execution platform that facilitates the execution of shadow versions alongside the current active version, measures safety metrics of the updates through monitoring, and safe activation of the updated version if it outperforms the current active version.

5.6 Setting up Automated Pipelines for Cross-silo Activities

Developing and utilizing automated pipelines: MLOps pipelines can be implemented with various tools. At a basic level, software scripts (e.g., Bash or Python) for submitting training or testing jobs can form a rudimentary version of automated pipelines. Moving up a level, organizations can adapt the existing tools and platform for CI/CD to MLOps workflows. GitHub Actions, CircleCI, ArgoCD, and Jenkins were among the tools suggested by the existing literature and industry best practices for adaptation to MLOps. For instance, Stieler and Bauer [PS136] proposed a Git workflow for an “active learning life cycle”, which contains four interconnected iterations of data management, model training, software development, and system operations. In the proposed workflow, a CI/CD tool hosted on a resource-rich machine with GPUs executes ML pipelines for training, testing, and integrating ML models based on the code committed by developers to a source control management system. Alternative, generic workflow engines such as Apache Airflow⁶ can be adopted to MLOps pipelines. For instance, the FLScalize [PS172], an MLOps platform for federated learning, utilizes Airflow to orchestrate the interaction sequences among federated learning servers and clients as well as other monitoring workflows. Finally, there have been more ML-specific workflow orchestration tools (e.g., SageMaker Pipelines, IBM Watson Pipelines, Google Vertex AI Pipelines, Netflix Metaflow, ZenML) for defining and running MLOps pipelines.

Optimising and automating MLOps pipeline creation and management: By specifying pipelines explicitly, it is possible to automate, optimize, and manage the execution of MLOps pipelines. The automation helps overcome the uncertainties that ML workflows face such as changing requirements, changing system context due to other workflows competing for resources, workflow orchestrator’s failure, or hardware crashes [PS170]. For instance, Scanflow-K8s [PS170] was proposed as a framework for autonomic management and supervision of ML workflows in Kubernetes clusters. Moreover, automating the creation of MLOps workflows contributed to reducing the setup overhead, particularly during the initial states of ML development. Carqueja et al. [PS129] proposed a lightweight Quick Machine Learning framework, QML, to address this problem. Automated creation also facilitates optimized deployment of data analytic pipelines against the heterogeneous infrastructure available for training or operating the models. Díaz-de-Arcaya et al. [PS169] proposed Orfeon, an AIOps framework for optimizing the deployment of analytical

⁶<https://airflow.apache.org/use-cases/mlops/>

pipelines, taking as input the static pipeline definition by data scientists and the dynamic metrics from the target infrastructure.

Simplifying the integration of software tools and components into MLOps pipelines. Software-centric integration approach [PS171] is proposed in Looper where predefined API end points are provided for interconnecting steps in MLOps pipelines. Modularisation through microservice-based design is recommended in [PS111], [PS120],[PS127] to enhance the integration of pipeline components.

5.7 Abstracting MLOps Complexity

Integrated development platform for ML Systems. Integrated development platforms provide a collaborative environment for whole-life-cycle-management for ML models and abstract the complexity from end users (e.g., data scientists, researchers, and other ML practitioners) [PS171]. They provide interactive control planes accessible by ML practitioners to manage end-to-end ML life cycle [PS154]. Managed cloud-based platforms such as Amazon SageMaker [PS108] and IBM Watson [PS115] provide support for data preparation, model training and experiment tracking, automated deployment on managed infrastructure and monitoring throughout the life cycle. Sagemaker [PS108] is developed with a modular and customizable architecture to support the integration of open-source tools to meet practitioner requirements. MLflow, Kubeflow, and Metaflow are open-source end-to-end platforms to simplify the ML lifecycle [PS109]. Databricks Lakehouse Platform [PS110] provides a data-native platform which together with MLflow provides a single platform for managing data, conducting data analysis, model training, and model serving. Peng et al. [PS196] proposes a platform for deploying models within Kubernetes and Istio to enable cloud-native environments by providing Source to Image (S2I) and model deployment support.

Reference Architecture of an MLOps system. Functional decomposition and reference architecture of MLOps systems guide the composition or development of MLOps pipelines and platforms depending on domain-specific requirements. To this end, academic literature presents domain specific architectures for various domains such as manufacturing domain [PS134], Industry 4.0 [PS127], transnational machine learning in health domain [PS133], and Gravitational Wave Physics [PS132].

6 RQ3: WHAT ARE THE CHALLENGES PRACTITIONERS FACE WHEN ADOPTING MLOPS?

In RQ2 we discuss work practices, techniques and technologies that organizations need to implement to enable MLOps. However, there are challenges that hinder teams from efficiently adopting MLOps into their organization's ML-based software system productionalization process. Thus, we identify such challenges and highlight reported solutions from the literature, and summarize them in Table 5.

6.1 Socio-technical Challenges

MLOps is a continuously evolving, technology-intensive domain which requires a considerable cultural shift in organization for its successful adoption (as discussed in section 5.1). Our thematic analysis of literature resulted in three major challenges arising due to interactions between social and technical aspects when adopting MLOps.

6.1.1 Cultural Changes. The introduction of MLOps practices disrupts the processes followed during the development and management of traditional software systems. Thus, the successful adoption of MLOps practices depends on the openness and willingness of the teams to change and

Table 5. MLOps adoption challenges and proposed solutions

MLOps Adoption Challenges		Primary Aspects	Proposed Solutions
Socio-technical Challenges	Cultural Changes: Disruptions caused by MLOps related culture and work practices in traditional software management process.	Openness and willingness of teams to change [PS103] Continuous collaboration among team members from multiple disciplines [PS104], [PS108] Successful implementation of cultural changes [PS103]	Available MLOps processes, practices and tools to promote collaboration [PS104], [PS108] [PS103], and identify/resolve collaboration conflicts [PS104] Follow a human-centric approach to implement cultural changes [PS154]
	MLOps Fragmentation: Fragmentation of practices, technologies and tools due to rapid growth in MLOps.	Inherent complexity of MLOps [PS154], [PS132] Accidental complexity due to fast evolution of MLOps [PS154]	MLOps Standardization [PS154] Creating domain-specific MLOps reference architectures [PS134], [PS133], [PS127]
	Responsible MLOps: Challenges related to achieving regulatory compliance, AI fairness, transparency, accountability and security throughout the MLOps life cycle.	Lack of expertise and standardized MLOps practices related to compliance, fairness, explainability and security [PS105], [PS121] Limitations in end-to-end MLOps tools for facilitating responsible AI [PS152] [PS114]	Frameworks integrating responsible AI practices into MLOps [PS108], [PS117]
MLOps Pipeline Related Challenges	Complexity: Complexity of ML pipelines hinders their use during initial stages of the ML development	High setup and maintenance overhead. [PS129] High level of expertise required to operate pipelines. [PS129] High infrastructure costs [PS129]	Frameworks for creating lightweight ML pipelines. [PS129]
	Optimized deployment: Optimal placement of pipeline modules within distributed environments.	Operation across distributed and heterogeneous resources [PS127], [PS128] Manual provisioning of pipelines in distributed and dynamic environments is error prone and sub-optimal [PS127, PS128] Training challenges due to ML models growth [PS126]	Strategic modularization [PS127] Dynamic provisioning of containerized pipeline modules [PS127] Dynamic resource provisioning for parallel or distributed model training [PS126]
	Managing large scale pipelines: Maintaining thousands of production pipelines without failures.	Failure-prone, recurring and inter-dependent production pipelines. [PS131] [PS176] Manual, time-consuming monitoring and resolution of failures. [PS131] [PS177]	Automation of pipeline operation monitoring and failure root cause analysis. [PS131] [PS177]
MLOps Platform Related Challenges	Continual learning: Operation of efficient continual learning (online learning) pipelines for high velocity and high volume streaming data	Synchronization between learning, serving and maintenance activities [PS124], [PS125] Overwhelming model updating due to incoming high-velocity data streams [PS125]	Horizontal and vertical scaling of online-learning tasks based on predictions demand, data velocity, and volume [PS125]
	Automation Decision Dilemma: Deciding which activities of the pipelines to automate and which needs human involvement	Automation of certain ML life cycle activities (model development, model validation) reduces incorporation of human judgment which can negatively effect the use of responsible AI [PS110], [PS114]	Design and implementation of pipelines in which the tedious tasks are automated and human insights are incorporated wherever it is needed throughout the lifecycle [PS110], [PS114] [PS117]
	Platform Customization: Decision on whether to use a managed MLOps platform or create a customized platform.	Complexity and cost of managing own platforms add a heavy burden [PS108] Cloud-based managed platforms lacks fine-grain control and customization support [PS108]	Building managed platforms with integration and customization support [PS108] Adopt a modular architecture [PS108] Increase support for open-source ML tools and their integration [PS108]
	Scalability: Ability of the MLOps platforms to maintain smooth execution of MLOps services as the number of service requests increases.	Lack of scalability as the number of service requests fluctuates in monolithic architecture [PS120]	Using Microservice Architecture for platform design [PS120]
	Artifact Management: Management of data, models and associated metainformation to ensure ML experiment tracking/governance and reproducibility	Challenging versioning, traceability and reproducibility due to increased number of artifacts [PS85], [PS164] Reproducibility based on conjugated management of the related artifacts. [PS122]	Tools/ platforms supporting version controlled and conjugated management of artifacts [PS122]
	Distributed computing environments: Challenges with design, deployment and maintenance of MLOps pipelines in decentralized edge environments	Edge-AI has given rise to paradigms like TinyML, Federated ML at edge, which have MLOps related requirements different from the cloud-based platforms. [PS123]	Build platforms that support MLOps support for usecases such as TinyML and Federated ML at the edge. [PS123]
	Migration to MLOps platforms: Challenges faced by practitioners when integrating existing projects to MLOps platforms	Tedious process of retrieving information from existing project that are required as input for the platform [PS178] Selecting a suitable MLOps platform for the project [PS178] Incompatibility between MLOps and DevOps tools [PS160] [PS19]	Document the ML components and artifacts and use it for migration process [PS178] Selecting a platform based on the architecture of the existing project [PS178] Use tools that offer hybrid CI/CD environment [PS178]

embrace the cultural changes introduced by MLOps [PS103]. Works such as [PS103], [PS104] and [PS108] discuss the inter-disciplinary collaboration requirements of the MLOps practices as one of the major cultural changes. Successful adoption of MLOps requires continuous collaboration among interdisciplinary teams that previously used to work in silos, which may result in conflicts. As a solution to this, it is important to have MLOps best practices in place to enable smooth cultural transitions. One way to achieve this is the introduction of a clear and well-structured communication plan, and having a clear definition of the business objectives/goals and metrics

as possible solutions to help teams understand novel MLOps practices introduced to organization culture [PS103]. Furthermore, methods need to be implemented for swiftly identifying and resolving conflicts arising related to responsibility boundaries and different goals/ expectations of multi-disciplinary stakeholders [PS104], [PS151]. [PS104] proposes a ⁱ* Strategic Actor Rationale to model the actor relationships to identify and resolve conflicts during collaboration. However, addressing the challenges associated with cultural changes is unique to each organization. For example, in the financial services domain, due to the complex regulatory environment, teams have to operate in silos [PS108]. Thus, financial service institutions benefit from tools and platforms (e.g., ValidMind⁷) that enable continuous collaboration between data scientists and compliance teams to achieve efficient risk management throughout the productionalization lifecycle. Hence, necessary cultural changes must be implemented (through MLOps best practices and tools) after careful analysis of the existing organizational structure and strategically implement the changes to maximize their positive impact [PS103]. To this end, Nvidia [PS154] recommends a human-centric approach to guide the selection and adoption of MLOps tools, techniques, and practices.

6.1.2 MLOps Fragmentation. Rapid evolution and growth in MLOps practices, solutions and tools create industry-wide fragmentation of MLOps due to the increase in **complexity** and **diversity** of MLOps. Being a broad domain consisting of practices, processes, solutions and tools to productionalize ML systems, MLOps is inherently complex [PS154]. Moreover, the popularity of MLOps and the use of ML in various domains that lack expertise in ML (e.g., physical sciences [PS132]) result in the introduction of diverse approaches and tools particular to different domains, use cases and organizations [PS154]. This results in accidental complexities, and it is challenging to separate accidental complexity from essential complexity [PS154]. Standardizing MLOps process, tooling, etc. and generation of domain-specific reference architectures to capture domain-specific ML productionalization requirements contribute to solving these issues. For instance, [PS134] extends generic MLOps to meet the requirements of the Manufacturing Domain where ML-enabled software is deployed within distributed Edge-Cloud environments. However, as production ML systems are novel and the rapid evolution of MLOps, standardization needs a huge effort and should be done by bringing practitioners, researchers and industry MLOps solution providers together.

6.1.3 Responsible MLOps. Achieving Responsible MLOps is a challenging multidisciplinary effort that brings people, technology and processes together throughout the MLOps life cycle [PS105]. This includes various aspects such as compliance with regulations and standards, AI fairness assurance through bias detection/mitigation, transparency and accountability through explainable AI and security. However, the MLOps process lacks explicit consideration for responsible AI practices and provides only limited guidance concerning these aspects, which makes their implementation challenging for practitioners [PS105] [PS121]. Being a fast-evolving field, there's a shortage of experts who can navigate these challenging aspects successfully to implement responsible MLOps [PS121]. Moreover, varying regulatory requirements across different domains demand different levels of collaboration with compliance teams and quality assurance teams throughout the MLOps lifecycle. For instance, highly regulated domains such as financial services require compliance by design as an MLOps practice. Developing comprehensive MLOps frameworks by integrating responsible AI practices can help practitioners overcome this. As an initial step, academic literature presents conceptual frameworks that integrate security (SecMLOps framework presented in [PS105]) and explainable AI practices (Explainable AI framework presented in [PS117]) into the MLOps process.

⁷<https://validmind.com/>

To facilitate Responsible MLOp, practitioners require tools that support end-to-end assurance of these aspects, which includes provenance support tools during developments and experiment activities [PS114], tools for bias mitigation during different stages of the ML life cycle and explain models and their predictions [PS152], and tools for automated ML compliance checking [PS105]. Moreover, detailed documentation regarding responsible MLOps support of the tools and MLOps platforms can enable users to make informed decisions [PS105] on which tools to use. As an example, [PS106] provides the architectural components of a secure MLOps platform using AWS services, which can be used by practitioners to decide if AWS services meet their responsible AI requirements.

6.2 MLOps Pipeline Related Challenges

MLOps best practices recommend using automated pipelines for managing the life cycle of the ML-enabled software systems (as discussed in section 5.6). This section details challenges faced by practitioners during the implementation of MLOps pipelines.

6.2.1 Complexity. While there are many powerful tools such as Kubeflow, FedML, Polyaxon for creating MLOps pipelines, high infrastructure cost, high expertise required for pipelines operation, high setup and maintenance overhead of the pipelines slows down their adoptions [PS129]. This is prevalent especially during the initial stages of ML projects or during ML projects driven by small companies. In order to avoid these complexities, practitioners tend to choose manual workflows over automated pipelines. IBM's 2022 Global AI adoption Index report [2] also highlights that the success rate of MLOps adoption of small companies is considerably lower compared to large corporations due to existing complexity. Design and development of MLOps pipeline creation frameworks with low setup, maintenance overhead and low infrastructure cost can overcome these challenges. DagsHub⁸ and Quick Machine Learning framework (QML) [PS129] are lightweight frameworks designed to act as an abstraction layer between ML practitioners and their chosen pipelines and enable creation of lightweight pipelines to meet ML workflows.

6.2.2 Optimized deployment. IoT domains require MLOps pipelines to operate within distributed computing environments which include resource constrained and geographically distributed edge devices and centralized cloud data centers [PS127]. This creates the challenge of optimally placing modularized pipeline components across these resources, especially to meet the increasing resource requirements during model training and model serving phases [PS128]. This challenge is further amplified by the availability of continuous dynamic data streams. Thus, manual provisioning of pipeline components is error prone and provides sub-optimal results. Strategic modularization of pipeline components to suit the business domain characteristics and designing dynamic resources provisioning frameworks and algorithms can help overcome these issues [PS127], [PS128]. Furthermore, training pipelines need to adjust dynamically (i.e., resource optimized model training, model splitting, etc. [PS126]) to enable parallel or distributed model training to ensure efficient operation of the MLOps pipelines.

6.2.3 Managing large scale pipelines. Large-scale enterprises that adopt MLOps practices have to operate thousands of recurring and interdependent production pipelines. Such pipelines are prone to silent failures caused by data quality issues in the data pipelines such as schema-drift, increasing nulls, change of units, change of value standards, and change of data volume [PS131], [PS176]. Such issues generate cascading failures in pipelines. Thus, practitioners face the challenge of maintaining large-scale pipelines without failures. Moreover, due to the sheer scale of pipelines and their complexities, manual monitoring of data pipelines and manual root cause analysis for

⁸<https://dagshub.com/docs/>

model mispredictions incur extra human cost [PS131] [PS177]. To overcome these challenges, practitioners should automate pipeline operations monitoring as much as possible. This includes automation of data quality issue detection [PS131] and automated root causes analysis for detected errors [PS177].

6.2.4 Continual learning. With the popularity of IoT, streaming big data is available for training powerful ML models to derive insights from data [PS125] [PS124]. However, high volume and velocity of streaming big data, make batch training infeasible due to cost of storage and cascading lag created by storing and retraining stages on recurring pipelines. While continual learning pipelines have emerged as a solution to this, they need to be scalable enough to withstand high-velocity data streams and make swift model updates accordingly. To enable executing model updates without pausing the inference pipelines, practitioners need to create and use elastically scalable online learning pipeline architectures with dynamic approaches to scale pipelines components vertically and horizontally (e.g., StreamMLOps [PS125]).

6.2.5 Automation decision dilemma. Although the practitioners tend to automate wherever possible when implementing MLOps, ML lifecycle consists of activities where human feedback is crucial and full automation can create negative impacts, especially considering the responsible AI practices [PS110]. For example, model validation processes require communication and approval from regulatory agencies, fairness and bias identification and analysis, root cause analysis from monitoring alerts, etc. Moreover, companies may have to prioritize other tasks over allocating time and effort on automating some of the tasks. Thus, caution must be taken during automation such that the tedious tasks are automated and human insights are incorporated wherever it is needed throughout the lifecycle [PS114], [PS110], [PS117].

6.3 MLOps Platform Related Challenges

MLOps platforms are designed for end-to-end ML lifecycle management, as a means of abstracting the complexity of implementing MLOps as discussed in 5.7. This section highlights challenges and limitations related to creating and using such platforms covering platform provider and practitioner perspectives.

6.3.1 Platform customization. For companies looking to build and deploy ML models, the main goal is to generate business value from data. Thus, for most of the practitioners, the underlying MLOps platform is only instrumental and their aim is to minimize the effort on implementing and managing the platform. Cloud-based managed MLOps platforms like Sagemaker are advantageous to meet this requirement. However, using fully managed MLOps platforms means the practitioners would lose fine-grained control over the technology stack and the services supported by the platform may not meet specialized requirements of many businesses [PS108]. This drives practitioners to implement and manage their own platforms which add additional complexity and cost. Practitioner expectation in this scenario is easy customization of managed platforms by integrating open-source, proprietary or homegrown tools. To meet this requirement managed MLOps platform providers need to design and implement their platforms following a modular architecture, integration and customization support to enable practitioners to integrate open-source tools. As an example, AWS highlights the design of Amazon SageMaker which provides practitioners the flexibility to integrate their own workflows with open-source APIs supporting TensorFlow, PyTorch, Kubeflow, etc [PS108].

6.3.2 Scalability. Services provided by MLOps platforms need to maintain performance requirements as the amount of service requests fluctuates and MLOps platforms should facilitate practitioners to create scalable workflows. In such scenarios, the lack of scalability of the underlying platform design becomes a significant challenge. As an example, the use of monolithic architecture

would considerably impact the scalability of a system. Thus, for MLOps platforms Microservices Architecture (MSA) demonstrated high scalability with independently scalable but interconnected services [PS120]. It is important to highlight, scalability of MSA based MLOps platforms can be achieved by using dynamic pipeline deployment approaches (discussed in section 5.6) to ensure the efficiency of MLOps pipelines to enable large-scale learning and continuous deployment of ML models.

6.3.3 Artifact Management. Compared to traditional software, types and amount of artifacts related to ML-enable software system is considerably higher. This includes software codes, ML models and related hyper parameters, data sets, and other metadata such as provenance data used for improving the trust of the ML-enabled system [PS85]. This is further complicated by the highly explorative and iterative nature of the ML model building process where the experiments include different datasets, data preparation techniques, algorithms, hyper parameters, etc. Thus, versioning, traceability, and reproducibility become challenging [PS164]. Lack of support for conjugated management of these artifacts is a challenge faced by practitioners when using MLOps platforms [PS122]. To overcome this, MLOps platforms need to be built with version control, storage, and governance (access control, authentication, and authorization) support for data, model and associated meta-information [PS122].

6.3.4 Distributed computing environments. Edge computing is introduced as a means of improving data privacy by processing data closer to the sources. This has created paradigms such as TinyML (i.e., ML models running on very resource constrained heterogeneous edge devices) and Federated Learning (i.e., a method of collaboratively training a shared model without transmitting data to a central location). These aspects affect certain ML lifecycle activities compared to ML model deployment in cloud. For example, in distributed edge computing scenarios, the model selection depends on the edge devices it is deployed on due to the heterogeneous nature of the devices, and model retraining needs to be carried out in decentralized manner using distributed streaming data. Thus, existing Cloud-based managed platforms do not meet the requirements for TinyML and Federated ML scenarios at the edge. Thus, platforms need to be designed and developed (TinyMLOps platform [PS123]) considering such special requirements arising in distributed computing environments [PS124], [PS125].

6.3.5 Migration to MLOps Platforms. As MLOps is a new and upcoming practice, many companies are migrating already existing projects to MLOps platforms to enable efficient production of the ML components of their software systems. Moreover, in ML-enabled software systems ML components are a small part of larger traditional software systems where CI/CD is achieved through DevOps practices and tools. So these companies utilize DevOps tools used for traditional software development for initial management of ML projects. Thus, when migrating MLOps platform requires additional information (artifacts and their links, workflow for creating pipelines, underlying resource requirements) needs to be extracted, which is a time consuming task [PS178]. It is important for practitioners to conduct the migration process strategically by carefully documenting ML components and artifacts available in the current project, understanding the current architecture of the project to determine the best suited platform and utilizing tools supporting CI/CD support for software systems containing both traditional software components and ML components [PS178].

7 FUTURE RESEARCH DIRECTIONS

Based on the MLOps practices, adoption challenges and proposed solutions, we identify the future research directions to facilitate MLOps adoption. To provide a concise view of our findings, Fig. 4 presents a mapping of MLOps best practices, adoption challenges and their related solutions from RQ2 and RQ3, along with future research directions derived based on the existing gaps.

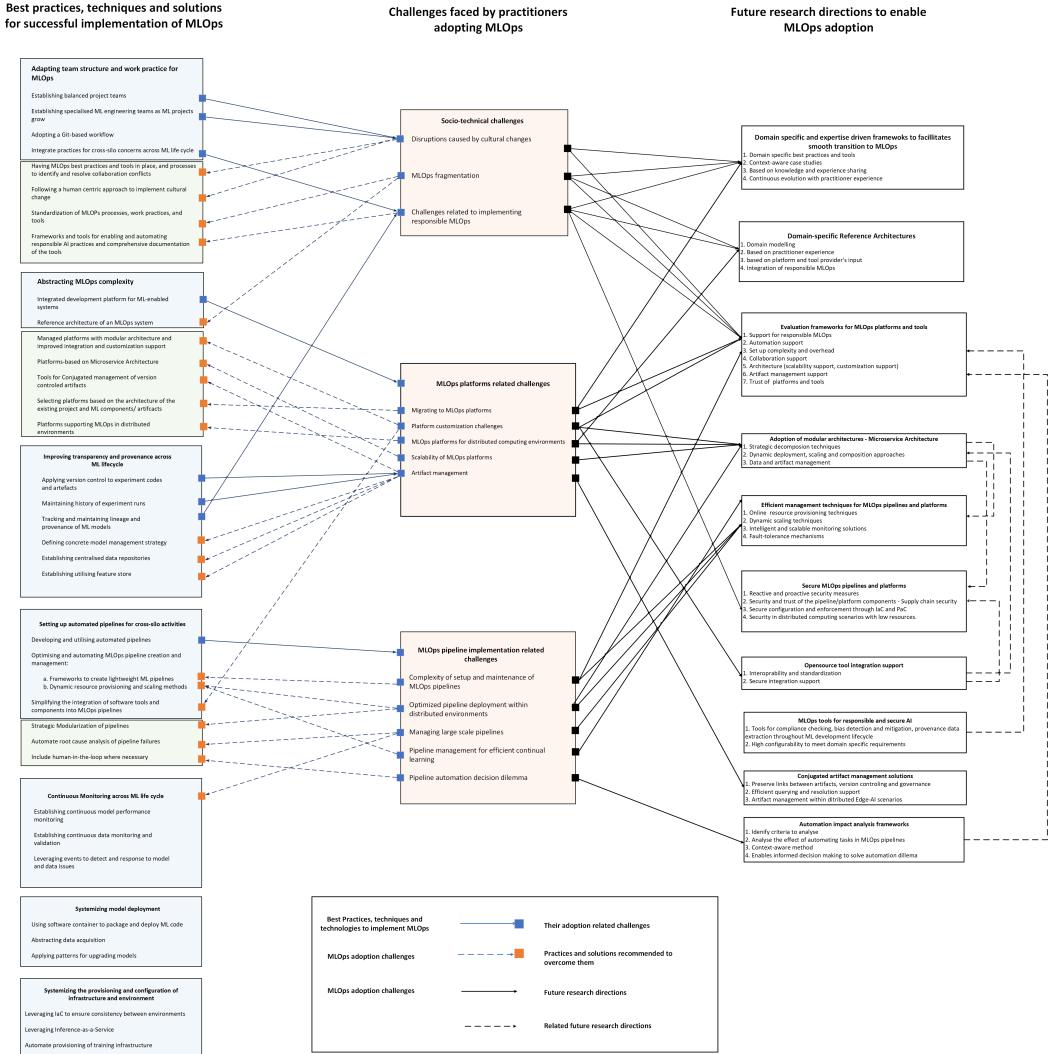


Fig. 4. Mapping of MLOps best practices, adoption challenges, solutions and Future research directions to improve MLOps adoption

Domain specific and expertise driven frameworks to facilitate cultural changes. Challenges such as cultural shifts, integration of responsible and secure AI practices, MLOps fragmentation, and transitioning from traditional DevOps technologies highlight the importance of sharing practitioner experience and knowledge to drive standardized MLOps adoption. To this end, future research can contribute to developing frameworks consisting of case studies to facilitate cultural transitions in a context-aware manner, and address best practices, solutions and tools employed to meet domain specific requirements. These frameworks need to be rooted in knowledge and experience sharing and designed to evolve as MLOps adoption increases. Studies such as [PS121] and [PS178] attempt to organize practitioner experience, but do not provide appropriate frameworks that can sufficiently capture the said aspects.

Domain specific MLOps reference architectures. MLOps architectures, tools, and technologies developed for different domains, e.g., manufacturing domain, health domain capture unique requirements for ML system productionization. While academic literature such as [PS134], [PS127], [PS132], and [PS133] makes contributions towards this, the industry-driven and practitioner-oriented nature of MLOps demands reference architectures that combine practitioner input with platform and tool provider expertise. Future work on reference architectures should integrate important aspects like security, ethics, transparency, compliance and governance, etc. to enable responsible and secure AI throughout the ML life cycle.

Evaluation frameworks for MLOps platforms and tools. The current literature lacks comprehensive frameworks that enable evaluation of MLOps platforms and tools, hindering standardization and effective navigation of practitioners in this space. Academic works such as [PS105], [PS129] and [PS122] evaluate existing MLOps platforms based on criteria like secure ML productionalization support, pipeline setup complexity, and artifact management support. Industry white papers and case studies such as [PS110], [PS106], [PS152], and [PS108] highlight the platform capabilities in automation, secure pipeline creation, responsible AI support, collaboration and architecture. However, this knowledge is scattered and does not provide a structured and comprehensive method for practitioners to evaluate which platforms and tools suit their needs the most.

Automation impact analysis frameworks. Although the proposed best practice for automation dilemma is to automate tedious tasks and incorporate human feedback where possible [PS110], existing literature does not provide a structured approach to do this. Given that the automation decisions depend on contextual factors such as industry domain, business objectives, regulatory requirements, organization culture, availability of tools, etc., impact analysis framework needs to be created to analyze the effect of automation in a context-aware manner. Such analysis framework would help practitioners assess benefits and drawbacks using quantifiable metrics representing different criteria such as efficiency, scalability, cost, cascading risks, etc., thus enabling them to make informed decisions to avoid under or over automation.

Adoption of modular architectures such as Microservice Architecture. Adoption of Microservice Architecture for MLOps pipelines and platforms facilitates their efficient and scalable deployment across distributed edge and cloud environments to support novel paradigms such as TinyML and Federated ML. Academic research on this demonstrates the potential of adopting microservice architecture [PS120], [PS179]. However, the success of this approach depends on using strategic decomposition techniques for determining microservice boundaries without introducing unnecessary complexities. Moreover, management of such modular and distributed architectures requires development and integration of dynamic deployment, scaling, orchestration, and load balancing techniques along with data and artifact management to ensure their secure availability across distributed microservices.

Efficient management techniques for MLOps pipelines and platforms. Efficient pipeline component management techniques play a vital role in navigating the complexities of modular architectures, large scale pipelines, and continual learning and delivery of ML models in Big Data environments. Future research should focus on online and dynamic resource provisioning, dynamic scaling, and fault tolerance enabled by distributed and scalable monitoring of pipelines. Moreover, these techniques need to consider the resource heterogeneity, and limited and distributed resource availability at the edge.

Secure MLOps pipelines and platforms. Practitioners can combine managed platforms with open-source, vendor supplied, or in-house tools to create pipelines. Security and trust of these components ultimately affect the security of the ML models produced. Security breaches can

compromise data privacy, model integrity, overall system availability, etc. Proactive and reactive mechanisms need to be developed to monitor these pipelines and platforms and mitigate security breaches. As MLOps pipelines are created by integrating tools from various vendors, software supply chain vulnerabilities also become a considerable security concern. Intelligent approaches are needed to ensure secure configuration using technologies such as Infrastructure as Code (IaC) and Policy as Code (PaC). Moreover, with the rise of Edge-AI applications, pipelines and platforms need to be secured within distributed computing environments with limited resources.

Open source tool integration support. To meet different business requirements, customization of MLOps platforms is required where practitioners attempt to integrate open-source tools with managed MLOps platforms [PS108]. In this case, interoperability is achieved through standardization of interfaces, model and data formats, protocols, etc. Moreover, integration of open-source tools can cause security risks, requiring analysis before integration and continuous audits to detect and mitigate any risks after integration.

MLOps tools for responsible AI. As highlighted in both academic and gray literature such as [PS152], [PS105], [PS164], responsible AI tools (i.e., bias detection and mitigation, provenance extraction, compliance checking) need improvement. Future work on the responsible AI tools requires input from domain experts from different domains having various levels of security, fairness, regulatory, governance requirements that span across the ML life cycle.

Conjugated artifact management solutions. With the increased types and number of artifacts generated during MLOps productionization, conjugated artifact management solutions are required to track the artifacts while maintaining relationships among them. Future solutions should also focus on governance along with efficient and scalable storage of artifacts across distributed environments.

8 CONCLUSION

The complexity of ML solutions due to the experimental nature of data and ML, the dynamic nature of production environments, the availability of continuous streams of new data and decentralized computing infrastructure (i.e., edge-cloud integrated environments) makes productionalization of ML models extremely challenging for the practitioners. Hence, MLOps introduces practices and technologies to create continuous integration, continuous delivery, continuous training and continuous monitoring-related autonomous workflows for the smooth transition of ML models from development to deployment, with increased transparency, model quality, and responsible and secure AI support. This has made MLOps a popular topic among practitioners and academia. However, with the innate complexity of MLOps further enhanced by rapid contributions made to MLOps from industry and academia, it is challenging to navigate the current landscape of MLOps.

We have systematically selected and analyzed 150 relevant academic studies and 48 gray literature to provide a comprehensive overview of the current MLOps landscape by addressing three research questions: 1) How do researchers and practitioners perceive MLOps, 2) What are the enabling solutions and best practices proposed for MLOps?, 3) What are challenges practitioners face when adopting MLOps?. We derived future research directions based on the comprehensive analysis of MLOps adoption challenges and existing solutions to drive future research in MLOps. Practitioners and researchers can utilize rigorous analysis and insights of MLOps provided in this work to guide the adoption of MLOps into their ML-enabled application productionalization process with increased understanding of cultural, technical and socio-technical aspects of MLOps and drive future advances to improve MLOps adoption.

REFERENCES

- [1] 2021. MLOps with Azure Machine Learning. (2021). [White Paper].

- [2] 2022. IBM Global AI Adoption Index 2022. <https://www.ibm.com/watson/resources/ai-adoption>
- [3] 2024. Online appendix of A Multivocal Literature Review on MLOps: Overview, Solutions and Challenges. [https://github.com/beken/MLOpsMLR.git/](https://github.com/beken/MLOpsMLR.git)
- [4] Stuart Charters Barbara Kitchenham. 2007. Guidelines for performing systematic literature reviews in software engineering.
- [5] Nils Baumann, Evgeny Kusmenko, Jonas Ritz, Bernhard Rumpe, and Moritz Benedikt Weber. 2022. Dynamic data management for continuous retraining. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*. 359–366.
- [6] Kent Beck. 2000. *Extreme programming explained: embrace change*. addison-wesley professional.
- [7] Pierre Bourque and RJNICS Fairley. 2004. Swebok. *Nd: IEEE Computer society* (2004).
- [8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77.
- [9] Patricia Ortegon Cano, Ayrton Mondragon Mejia, Silvana De Gyves Avila, Gloria Eva Zagal Dominguez, Ismael Solis Moreno, and Arianne Navarro Lepe. 2020. A Taxonomy on Continuous Integration and Deployment Tools and Frameworks. In *International Conference on Software Process Improvement*. Springer, 323–336.
- [10] Roland Croft, Yongzheng Xie, and Muhammad Ali Babar. 2022. Data preparation for software vulnerability prediction: A systematic literature review. *IEEE Transactions on Software Engineering* 49, 3 (2022), 1044–1063.
- [11] IBM Data and AI Team. 2023. MLOps and the evolution of data science. IBM, [blog].
- [12] Nesara Dissanayake, Asangi Jayatilaka, Mansooreh Zahedi, and M Ali Babar. 2022. Software security patch management- A systematic literature review of challenges, approaches, tools and practices. *Information and Software Technology* 144 (2022), 106771.
- [13] Paul M Duvall, Steve Matyas, and Andrew Glover. 2007. *Continuous integration: improving software quality and reducing risk*. Pearson Education.
- [14] Julian Ereth. 2018. DataOps-Towards a Definition. *LWDA* 2191 (2018), 104–112.
- [15] Leonhard Faubel and Klaus Schmid. 2023. Review protocol: a systematic literature review of MLOps. (2023).
- [16] Martin Fowler, Jim Highsmith, et al. 2001. The agile manifesto. *Software development* 9, 8 (2001), 28–35.
- [17] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and software technology* 106 (2019), 101–121.
- [18] Meenu Mary John, Helena Holmström Olsson, and Jan Bosch. 2021. Towards mlops: A framework and maturity model. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 1–8.
- [19] Ioannis Karamitsos, Saeed Albarhami, and Charalampos Apostolopoulos. 2020. Applying DevOps practices of continuous automation for machine learning. *Information* 11, 7 (2020), 363.
- [20] Ask Berstad Kolttveit and Jingyue Li. 2022. Operationalizing machine learning models: a systematic literature review. In *Proceedings of the 1st Workshop on Software Engineering for Responsible AI*. 1–8.
- [21] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. 2023. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access* (2023).
- [22] Anderson Lima, Luciano Monteiro, and Ana Paula Furtado. 2022. MLOps: Practices, Maturity Models, Roles, Tools, and Challenges-A Systematic Literature Review. *ICEIS* (1) (2022), 308–320.
- [23] Paul-Philipp Luley, Jan M Deriu, Peng Yan, Gerrit A Schatte, and Thilo Stadelmann. 2023. From concept to implementation: The data-centric development process for AI in industry. In *2023 10th IEEE Swiss Conference on Data Science (SDS)*. IEEE, 73–76.
- [24] Lucy Ellen Lwakatare, Ivica Crnkovic, Ellinor Rånge, and Jan Bosch. 2020. From a data science driven process to a continuous delivery process for machine learning systems. In *Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21*. Springer, 185–201.
- [25] Tsakani Mboweni, Themba Masombuka, and Cyrille Dongmo. 2022. A systematic review of machine learning devops. In *2022 international conference on electrical, computer and energy technologies (ICECET)*. IEEE, 1–6.
- [26] Elie Neghabi, Zerui Wang, Jun Huang, and Yan Liu. 2023. Linking Team-level and Organization-level Governance in Machine Learning Operations through Explainable AI and Responsible AI Connector. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 1223–1230.
- [27] Khalid Salama, Jarek Kazmierczak, and Donna Schut. 2021. Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning. (2021). [White Paper].
- [28] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).
- [29] Mojtaba Shahin, Muhammad Ali Babar, and Liming Zhu. 2017. Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE access* 5 (2017), 3909–3943.

- [30] Helge Spieker and Arnaud Gotlieb. 2019. Towards testing of deep learning systems with training set reduction. *arXiv preprint arXiv:1901.04169* (2019).
- [31] Daniel Ståhl and Jan Bosch. 2014. Modeling continuous integration practice differences in industry software development. *Journal of Systems and Software* 87 (2014), 48–59.
- [32] Monika Steidl, Michael Felderer, and Rudolf Ramler. 2023. The pipeline for the continuous development of artificial intelligence models—Current state of research and practice. *Journal of Systems and Software* 199 (2023), 111615.
- [33] Manasi Vartak. 2022. What is MLOps? DataOps? And Why do They Matter? <https://devops.com/what-is-mlops-dataops-and-why-do-they-matter/> [Online; accessed Mar 2024].
- [34] Christoph Windheuser, Danilo Sato, Alex Sadovsky, David Cooperberg, Greg Willis, Josh Poduska, Jim Falgout, and Dylan Tong. 2020. MLOps: Continuous Delivery for Machine Learning on AWS. (2020). [White Paper].
- [35] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 1–10.

Appendix A PRIMARY STUDIES

- [PS1] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. Who needs mllops: What data scientists seek to accomplish and how can mllops help? In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 109–112. IEEE, 2021.
- [PS2] Damian A Tamburri. Sustainable mllops: Trends and challenges. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 17–23. IEEE, 2020.
- [PS3] Yue Zhou, Yue Yu, and Bo Ding. Towards mllops: A case study of ml pipeline platform. In *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pages 494–500. IEEE, 2020.
- [PS4] Aquilas Tchanjou Njomou, Alexandra Johanne Bifona Africa, Bram Adams, and Marios Fokaefs. Msr4ml: Reconstructing artifact traceability in machine learning repositories. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 536–540. IEEE, 2021.
- [PS5] Ziqi Guo, Tingwen Bao, Wenlong Wu, Chao Jin, and Jay Lee. Iai devops: A systematic framework for prognostic model lifecycle management. In *2019 Prognostics and System Health Management Conference (PHM-Qingdao)*, pages 1–6. IEEE, 2019.
- [PS6] Rupesh Raj Karn, Prabhakar Kudva, and Ibrahim Abe M Elfadel. Dynamic autoselection and autotuning of machine learning models for cloud network analytics. *IEEE Transactions on Parallel and Distributed Systems*, 30(5):1052–1064, 2018.
- [PS7] Lucas Cardoso Silva, Fernando Rezende Zagatti, Bruno Silva Sette, Lucas Nildaimon dos Santos Silva, Daniel Lucrédio, Diego Furtado Silva, and Helena de Medeiros Caseli. Benchmarking machine learning solutions in production. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 626–633. IEEE, 2020.
- [PS8] Tamás Karácsony, Anna Mira Loesch-Biffar, Christian Vollmar, Soheyl Noachtar, and João Paulo Silva Cunha. Deepelpi: Towards an epileptologist-friendly ai enabled seizure classification cloud system based on deep learning analysis of 3d videos. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE, 2021.
- [PS9] Chandrasekar Vuppulapati, Anitha Ilapakurti, Karthik Chillara, Sharat Kedari, and Vanaja Mamidi. Automating tiny ml intelligent sensors devops using microsoft azure. In *2020 ieee international conference on big data (big data)*, pages 2375–2384. IEEE, 2020.
- [PS10] Tuomas Granlund, Aleksi Kopponen, Vlad Stirbu, Lalli Myllyaho, and Tommi Mikkonen. Mlops challenges in multi-organization setup: Experiences from two real-world cases. In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 82–88. IEEE, 2021.
- [PS11] Lucy Ellen Lwakatare, Ivica Crnkovic, and Jan Bosch. Devops for ai-challenges in development of ai-enabled applications. In *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6. IEEE, 2020.
- [PS12] Benjamin Benni, Mireille Blay-Fornarino, Sébastien Mosser, Frédéric Précisio, and Günther Jungbluth. When devops meets meta-learning: A portfolio to rule them all. In *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 605–612. IEEE, 2019.
- [PS13] Ilias Gerostathopoulos, Stefan Kugele, Christoph Segler, Tomas Bures, and Alois Knoll. Automated trainability evaluation for smart software functions. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 998–1001. IEEE, 2019.
- [PS14] Álvaro López García, Jesus Marco De Lucas, Marica Antonacci, Wolfgang Zu Castell, Mario David, Marcus Hardt, Lara Lloret Iglesias, Germán Molto, Marcin Plociennik, Viet Tran, et al. A cloud-based framework for machine learning workloads and applications. *IEEE access*, 8:18681–18692, 2020.
- [PS15] Amine Barrak, Ellis E Eghan, and Bram Adams. On the co-evolution of ml pipelines and source code-empirical study of dvc projects. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 422–433. IEEE, 2021.
- [PS16] Fuyuki Ishikawa and Yutaka Matsuno. Evidence-driven requirements engineering for uncertainty of machine learning-based systems. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 346–351. IEEE, 2020.
- [PS17] Markus Borg, Ronald Jabangwe, Simon Åberg, Arvid Ekblom, Ludwig Hedlund, and August Lidfeldt. Test automation with grad-cam heatmaps-a future pipe segment in mllops for vision ai? In *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 175–181. IEEE, 2021.
- [PS18] Eero Kauhanen, Jukka K Nurminen, Tommi Mikkonen, and Matvei Pashkovskiy. Regression test selection tool for python in continuous integration process. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 618–621. IEEE, 2021.
- [PS19] Bojan Karlaš, Matteo Interlandi, Cedric Renggli, Wentao Wu, Ce Zhang, Deepak Mukunthu Iyappan Babu, Jordan Edwards, Chris Lauren, Andy Xu, and Markus Weimer. Building continuous integration services for machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2407–2415, 2020.

- [PS20] Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, et al. Developments in mlflow: A system to accelerate the machine learning lifecycle. In *Proceedings of the fourth international workshop on data management for end-to-end machine learning*, pages 1–4, 2020.
- [PS21] Cedric Renggli, Frances Ann Hubis, Bojan Karlaš, Kevin Schawinski, Wentao Wu, and Ce Zhang. Ease. ml/ci and ease. ml/meter in action: Towards data management for statistical generalization. *Proceedings of the VLDB Endowment*, 12(12):1962–1965, 2019.
- [PS22] Huaizheng Zhang, Yuanming Li, Yizheng Huang, Yonggang Wen, Jianxiong Yin, and Kyle Guan. Mlmodelci: An automatic cloud platform for efficient mlaas. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4453–4456, 2020.
- [PS23] Maurits van der Goes. Scaling enterprise recommender systems for decentralization. In *Fifteenth ACM Conference on Recommender Systems*, pages 592–594, 2021.
- [PS24] Philipp Ruf, Manav Madan, Christoph Reich, and Djaffar Ould-Abdeslam. Demystifying mlops and presenting a recipe for the selection of open-source tools. *Applied Sciences*, 11(19):8861, 2021.
- [PS25] Emmanuel Raj, David Buffoni, Magnus Westerlund, and Kimmo Ahola. Edge mlops: An automation framework for aiot applications. In *2021 IEEE International Conference on Cloud Engineering (IC2E)*, pages 191–200. IEEE, 2021.
- [PS26] Silverio Martínez-Fernández, Xavier Franch, Andreas Jedlitschka, Marc Oriol, and Adam Trendowicz. Developing and operating artificial intelligence models in trustworthy autonomous systems. In *International Conference on Research Challenges in Information Science*, pages 221–229. Springer, 2021.
- [PS27] Romeo Kienzler and Ivan Nesić. Claimed, a visual and scalable component library for trusted ai. *arXiv preprint arXiv:2103.03281*, 2021.
- [PS28] Alexandra Posoldova. Machine learning pipelines: From research to production. *IEEE Potentials*, 39(6):38–42, 2020.
- [PS29] Yan Liu, Zhijing Ling, Boyu Huo, Boqian Wang, Tianen Chen, and Esma Mouine. Building a platform for machine learning operations from open source frameworks. *IFAC-PapersOnLine*, 53(5):704–709, 2020.
- [PS30] Mirella Sangiovanni, Gerard Schouten, and Willem-Jan van den Heuvel. An iot beehive network for monitoring urban biodiversity: vision, method, and architecture. In *Symposium and Summer School on Service-Oriented Computing*, pages 33–42. Springer, 2020.
- [PS31] Amitabha Banerjee, Chien-Chia Chen, Chien-Chun Hung, Xiaobo Huang, Yifan Wang, and Razvan Chevesean. Challenges and experiences with {MLOps} for performance diagnostics in {Hybrid-Cloud} enterprise software deployments. In *2020 USENIX Conference on Operational Machine Learning (OpML 20)*, 2020.
- [PS32] Willem-Jan van den Heuvel and Damian A Tamburri. Model-driven ml-ops for intelligent enterprise applications: vision, approaches and challenges. In *International Symposium on Business Modeling and Software Design*, pages 169–181. Springer, 2020.
- [PS33] Waldemar Hummer, Vinod Muthusamy, Thomas Rausch, Parijat Dube, Kaoutar El Maghraoui, Anupama Murthi, and Punleuk Oum. Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*, pages 113–120. IEEE, 2019.
- [PS34] Junsung Lim, Hoejoon Lee, Youngmin Won, and Hunje Yeon. {MLOp} lifecycle scheme for vision-based inspection process in manufacturing. In *2019 USENIX Conference on Operational Machine Learning (OpML 19)*, pages 9–11, 2019.
- [PS35] Behrouz Derakhshan, Alireza Rezaei Mahdiraji, Tilmann Rabl, and Volker Markl. Continuous deployment of machine learning pipelines. In *Eldb*, pages 397–408, 2019.
- [PS36] Stuart Jackson, Maha Yaqub, and Cheng-Xi Li. The agile deployment of machine learning models in healthcare. *Frontiers in big Data*, page 7, 2019.
- [PS37] Ola Spjuth, Jens Frid, and Andreas Hellander. The machine learning life cycle and the cloud: implications for drug discovery. *Expert opinion on drug discovery*, 16(9):1071–1079, 2021.
- [PS38] GeumSeong Yoon, Jungsu Han, Seunghyun Lee, and JongWon Kim. Devops portal design for smartx ai cluster employing cloud-native machine learning workflows. In *Advances in Internet, Data and Web Technologies: The 8th International Conference on Emerging Internet, Data and Web Technologies (EIDWT-2020)*, pages 532–539. Springer, 2020.
- [PS39] Krzysztof Czarnecki. Software engineering for automated vehicles: Addressing the needs of cars that run on software and data. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 6–8. IEEE, 2019.
- [PS40] Jacopo Tagliabue. You do not need a bigger boat: Recommendations at reasonable scale in a (mostly) serverless and open stack. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 598–600, 2021.
- [PS41] Philipp Kohl, Oliver Schmidts, Lars Klöser, Henri Werth, Bodo Kraft, and Albert Zündorf. Stamp 4 nlp—an agile framework for rapid quality-driven nlp applications development. In *International Conference on the Quality of Information and Communications Technology*, pages 156–166. Springer, 2021.
- [PS42] Aiswarya Raj Munappy, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, and Anas Dakkak. From ad-hoc data analytics to dataops. In *Proceedings of the International Conference on Software and System Processes*, pages

165–174, 2020.

- [PS43] Haoyu Cai, Chao Wang, and Xuehai Zhou. Deployment and verification of machine learning tool-chain based on kubernetes distributed clusters: This paper is submitted for possible publication in the special issue on high performance distributed computing. *CCF Transactions on High Performance Computing*, 3(2):157–170, 2021.
- [PS44] Chaoyu Wu, E Haihong, and Meina Song. An automatic artificial intelligence training platform based on kubernetes. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology*, pages 58–62, 2020.
- [PS45] Mlops: Methods and tools of devops for machine learning, Jul 2020. [Online; accessed 1. Apr. 2022].
- [PS46] Youngjune Lee, Oh Joon Kwon, Haeju Lee, Joonyoung Kim, Kangwook Lee, and Kee-Eung Kim. Augment & valuate: A data enhancement pipeline for data-centric ai. *arXiv preprint arXiv:2112.03837*, 2021.
- [PS47] Tuomas Granlund, Vlad Stirbu, and Tommi Mikkonen. Towards regulatory-compliant mlops: Oravizio’s journey from a machine learning experiment to a deployed certified medical product. *SN computer Science*, 2(5):342, 2021.
- [PS48] Sindhu Ghanta, Sriram Subramanian, Lior Khermosh, Swaminathan Sundararaman, Harshil Shah, Yakov Goldberg, Drew Roselli, and Nisha Talagala. MI health monitor: taking the pulse of machine learning algorithms in production. In *Applications of Machine Learning*. International Society for Optics and Photonics, 2019.
- [PS49] Grigori Fursin, Herve Guillou, and Nicolas Essayan. Coderef: an open platform for portable mlops, reusable automation actions and reproducible benchmarking. *The Workshop on MLOps Systems at MLSys’20*, 2020.
- [PS50] Denis Baylor, Kevin Haas, Konstantinos Katsiapis, Sammy Leong, Rose Liu, Clemens Menwald, Hui Miao, Neoklis Polyzotis, Mitchell Trott, and Martin Zinkevich. Continuous training for production ml in the tensorflow extended (tfx) platform. In *2019 USENIX Conference on Operational Machine Learning (OpML ’19)*, pages 51–53, 2019.
- [PS51] Rama Akkiraju, Vibha Sinha, Anbang Xu, Jalal Mahmud, Pritam Gundecha, Zhe Liu, Xiaotong Liu, and John Schumacher. Characterizing machine learning processes: A maturity framework. In *International Conference on Business Process Management*, pages 17–31. Springer, 2020.
- [PS52] R Ciucu, FC Adochiei, Ioana-Raluca Adochiei, F Argatu, GC Serițan, B Enache, S Grigorescu, and Violeta Vasilica Argatu. Innovative devops for artificial intelligence. *The Scientific Bulletin of Electrical Engineering Faculty*, 19(1):58–63, 2019.
- [PS53] Emmanuel Raj, David Buffoni, Magnus Westerlund, and Kimmo Ahola. Edge mlops: An automation framework for iot applications. In *IEEE International Conference on Cloud Engineering (IC2E)*, pages 191–200. IEEE, 2021.
- [PS54] Leonel Aguilar, David Dao, Shaoduo Gan, Nezihe Merve Gurel, Nora Hollenstein, Jiawei Jiang, Bojan Karlas, Thomas Lemmin, Tian Li, Yang Li, et al. Ease. ml: A lifecycle management system for mldev and mlops. *Proc. of Innovative Data Systems Research*, 2021.
- [PS55] Rohith Sothilingam, Eric Yu, and Arik Senderovich. Towards higher maturity for machine learning: A conceptual modelling approach. *The iJournal: Student Journal of the Faculty of Information*, 5(1):80–97, 2019.
- [PS56] Emmanuel Raj, Magnus Westerlund, and Leonardo Espinosa-Leal. Reliable fleet analytics for edge iot solutions. *arXiv preprint arXiv:2101.04414*, 2021.
- [PS57] Willem-Jan van den Heuvel and Damian A Tamburri. Model-driven ml-ops for intelligent enterprise applications: vision, approaches and challenges. In *Business Modeling and Software Design: 10th International Symposium, BMSD 2020, Berlin, Germany, July 6–8, 2020, Proceedings 10*, pages 169–181. Springer, 2020.
- [PS58] Cedric Renggli, Bojan Karlaš, Bolin Ding, Feng Liu, Kevin Schwawinski, Wentao Wu, and Ce Zhang. Continuous integration of machine learning models with ease. ml/ci: Towards a rigorous yet practical treatment. *Proceedings of Machine Learning and Systems*, 1:322–333, 2019.
- [PS59] Cedric Renggli, Luka Rimanic, Nezihe Merve Gürel, Bojan Karlaš, Wentao Wu, and Ce Zhang. A data quality-driven view of mlops. *IEEE Data Engineering Bulletin*, 2021.
- [PS60] Fuyuki Ishikawa and Yutaka Matsuno. Continuous argument engineering: Tackling uncertainty in machine learning based systems. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, pages 14–21. Springer, 2018.
- [PS61] Christian Weber, Pascal Hirmer, Peter Reimann, and Holger Schwarz. A new process model for the comprehensive management of machine learning models. In *ICEIS (1)*, pages 415–422, 2019.
- [PS62] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, et al. Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1387–1395, 2017.
- [PS63] Tim Kraska, Ameet Talwalkar, John C Duchi, Rean Griffith, Michael J Franklin, and Michael I Jordan. Mlbase: A distributed machine-learning system. In *Cidr*, volume 1, pages 2–1, 2013.
- [PS64] Micah J Smith, Jürgen Cito, Kelvin Lu, and Kalyan Veeramachaneni. Enabling collaborative data science development with the ballet framework. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021.
- [PS65] Can Kaymakci, Simon Wenninger, and Alexander Sauer. A holistic framework for ai systems in industrial applications. In *International Conference on Wirtschaftsinformatik*, pages 78–93. Springer, 2021.

- [PS66] Pulkit Agrawal, Rajat Arya, Aanchal Bindal, Sandeep Bhatia, Anupriya Gagneja, Joseph Godlewski, Yucheng Low, Timothy Muss, Mudit Manu Paliwal, Sethu Raman, et al. Data platform for machine learning. In *Proceedings of the 2019 international conference on management of data*, pages 1803–1816, 2019.
- [PS67] Sridhar Alla and Suman Kalyan Adari. *Beginning MLOps with MLFlow*. Springer, 2021.
- [PS68] *MLOps for IT Teams: How to Transform the Machine Learning Lifecycle*. DataRobot, 2021.
- [PS69] *Practical MLOps - How to get ready for production models*. Valohai, 2020.
- [PS70] Niko Vuokko, Jukka Remes, Alexander Finn, Harry Souris, Hossein Yousefi, Joanna Purosto, and Pauliina Alanen. Reliable and scalable ai with mlops, 2021. SILO AI, [ebook].
- [PS71] Zechu Li, Xiao-Yang Liu, Jiahao Zheng, Zhaoran Wang, Anwar Walid, and Jian Guo. Finrl-podracer: High performance and scalable deep reinforcement learning for quantitative finance. *arXiv preprint arXiv:2111.05188*, 2021.
- [PS72] Vlad Stirbu, Tuomas Granlund, Jere Helén, and Tommi Mikkonen. Extending soup to ml models when designing certified medical systems. In *2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH)*, pages 32–35. IEEE, 2021.
- [PS73] Markus Borg. Agility in software 2.0-notebook interfaces and mllops with buttresses and rebars. In *International Conference on Lean and Agile Software Development*, pages 3–16. Springer, 2022.
- [PS74] Frances Ann Hubis, Wentao Wu, and Ce Zhang. Quantitative overfitting management for human-in-the-loop ml application development with ease. ml/meter. *arXiv preprint arXiv:1906.00299*, 2019.
- [PS75] Yizheng Huang, Huaizheng Zhang, Yonggang Wen, Peng Sun, and Nguyen Bin Duong TA. Modelci-e: Enabling continual learning in deep learning serving systems. *arXiv preprint arXiv:2106.03122*, 2021.
- [PS76] Dennis Muiruri, Lucy Ellen Lwakatare, Jukka K Nurminen, and Tommi Mikkonen. Practices and infrastructures for machine learning systems: An interview study in finnish organizations. *Computer*, 55(6):18–29, 2022.
- [PS77] Nikhil Muralidhar, Sathappah Muthiah, Patrick Butler, Manish Jain, Yu Yu, Katy Burne, Weipeng Li, David Jones, Prakash Arunachalam, Hays' Skip' McCormick, et al. Using antipatterns to avoid mllops mistakes. *arXiv preprint arXiv:2107.00079*, 2021.
- [PS78] Georgios Symeonidis, Evangelos Nerantzis, Apostolos Kazakis, and George A Papakostas. Mlops-definitions, tools and challenges. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0453–0460. IEEE, 2022.
- [PS79] Razvan Ciobanu, Alexandru Purdila, Laurentiu Piciu, and Andrei Damian. Solis—the mllops journey from data acquisition to actionable insights. *arXiv preprint arXiv:2112.11925*, 2021.
- [PS80] Xiao-Yang Liu, Zechu Li, Zhuoran Yang, Jiahao Zheng, Zhaoran Wang, Anwar Walid, Jian Guo, and Michael I Jordan. Elegantrl-podracer: Scalable and elastic library for cloud-native deep reinforcement learning. *arXiv preprint arXiv:2112.05923*, 2021.
- [PS81] Peizheng Li, Jonathan Thomas, Xiaoyang Wang, Ahmed Khalil, Abdelrahim Ahmad, Rui Inacio, Shipra Kapoor, Arjun Parekh, Angela Doufexi, Arman Shojaeifard, et al. Rlops: Development life-cycle of reinforcement learning aided open ran. *arXiv preprint arXiv:2111.06978*, 2021.
- [PS82] Chuhong Min, Akhil Mathur, Utku Günay Acer, Alessandro Montanari, and Fahim Kawsar. Sensix++: Bringing mllops and multi-tenant model serving to sensory edge devices. *ACM Transactions on Embedded Computing Systems*, 22(6):1–27, 2023.
- [PS83] Grigori Fursin. The collective knowledge project: Making ml models more portable and reproducible with open apis, reusable best practices and mllops. *arXiv preprint arXiv:2006.07161*, 2020.
- [PS84] Fan Yun, Toshiki Shibahara, Yuichi Ohsita, Daiki Chiba, Mitsuaki Akiyama, and Masayuki Murata. Understanding machine learning model updates based on changes in feature attributions. 2020.
- [PS85] Christoph Windheuser, Danilo Sato, Alex Sadovsky, David Cooperberg, Greg Willis, Josh Poduska, Jim Falgout, and Dylan Tong. Mlops: Continuous delivery for machine learning on aws. 2020. [White Paper].
- [PS86] Sumeet Agrawal and Anant Mittal. Mlops: 5 steps to operationalize machine learning models. 2020. [White Paper].
- [PS87] Mlops with azure machine learning. 2021. [White Paper].
- [PS88] Rohit Panikkar, Tamim Saleh, Maxime Szybowski, and Rob Whiteman. Operationalizing machine learning in processes. *McKinsey&Company*, September, 2021.
- [PS89] Khalid Salama, Jarek Kazmierczak, and Donna Schut. Practitioners guide to mllops: A framework for continuous delivery and automation of machine learning. 2021. [White Paper].
- [PS90] Building versus buying an ml management platform, 2020. Algorithmia Inc, [white paper].
- [PS91] The complete guide to machine learning operations, 2020. run.ai, [white paper].
- [PS92] Stephen Watts. What Is MLOps? Machine Learning Operations Explained, Jul 2020. [Online; accessed 1. Apr. 2022].
- [PS93] Harshit Tyagi. What is MLOps? Machine Learning Operations Explained, Mar 2021. [Online; accessed 1. Apr. 2022].
- [PS94] Alessandro Gaggia. Mlops essentials: four pillars for machine learning operations on aws, Oct 2021. [Online; accessed 1. Apr. 2022].

- [PS95] Erin Wolpert. An overview of machine learning operations - building mature machine learning production systems through mlops, Nov 2021. [Online; accessed 1. Apr. 2022].
- [PS96] Ville Tuulos Outerbounds. Mlops vs. devops: Why data makes it different, Oct 2021. [Online; accessed 1. Apr. 2022].
- [PS97] Navdeep Singh Gill. MLOps Platform - Productionizing Machine Learning Models, June 2023. [Online; accessed Nov 2023].
- [PS98] Mlops: Methods and tools of devops for machine learning, Jul 2020. [Online; accessed 1. Apr. 2022].
- [PS99] Seldon Technologies Ltd. Mlops software: Open source vs enterprise. <https://www.seldon.io/mlops-software-comparison>, 2022. Accessed: 2023.
- [PS100] Danilo Sato, Arif Wider, and Christoph Windheuser. Continuous delivery for machine learning. <https://martinfowler.com/articles/cd4ml.html>, 2019. Accessed: 2023.
- [PS101] Dominick Rocco. The 4 pillars of mlops: How to deploy ml models to production, May 2021. [Online; accessed 1. Apr. 2022].
- [PS102] Mlops: Continuous delivery and automation pipelines in machine learning, Jan 2020. [Online; accessed 1. Apr. 2022].
- [PS103] Beatriz MA Matsui and Denise H Goya. Mlops: A guide to its adoption in the context of responsible ai. In *Proceedings of the 1st Workshop on Software Engineering for Responsible AI*, pages 45–49, 2022.
- [PS104] Rohith Sothilingam, Vik Pant, and Eric Yu. Using i* to analyze collaboration challenges in mlops project teams. 2022.
- [PS105] Xinrui Zhang and Jason Jaskolka. Conceptualizing the secure machine learning operations (secmlops) paradigm. In *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, pages 127–138. IEEE, 2022.
- [PS106] David Ping, Stefan Natu, Qingwei Li, Saeed Aghabozori, Vadim Dabrowski, Simon Zamarin, Ibrahim Gabr, and Amir Imani. Build a secure enterprise machine learning platform on aws, 2021. Amazon Web Services. [AWS Technical Guide].
- [PS107] Bruno Klein. Aws prescriptive guidance - planning for successful mlops, 2021. Amazon Web Services.
- [PS108] Natalie Heard Vin Sharma, Rob Ferguson. Mlops: Emerging trends in data, code, and infrastructure., 2022. Amazon Web Services.
- [PS109] Canonical Ubuntu. A guide to mlops, 2023.
- [PS110] Databricks. How to automate your machine learning pipeline, 2022. [ebook].
- [PS111] Amazon Web Services. Machine learning lens - aws well-architected framework, 2023.
- [PS112] Sergio Zavota, Ajish Palakadan, Salman Taherian, and Mandeep Sehmi. Machine learning lens - aws well-architected framework, 2023. Amazon Web Services.
- [PS113] What is mlops?, 2023. Databricks, [website].
- [PS114] Joseph Bradley, Rafi Kurlansik, Matt Thomson, and Niall Turritt. The big book of mlops, 2023. Databricks, [ebook].
- [PS115] IBM Data and AI Team. Mlops and the evolution of data science, 2023. IBM, [blog].
- [PS116] Machine learning operations (mlops), 2023.
- [PS117] Elie Neghabi, Zerui Wang, Jun Huang, and Yan Liu. Linking team-level and organization-level governance in machine learning operations through explainable ai and responsible ai connector. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1223–1230. IEEE, 2023.
- [PS118] Janvi Prasad, Arushi Jain, and Ushus Elizabeth Zachariah. Comparative evaluation of machine learning development lifecycle tools. In *2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC)*, pages 1–6. IEEE, 2022.
- [PS119] Shailesh Kumar Shivakumar. Web performance monitoring and infrastructure planning. In *Modern Web Performance Optimization*, pages 175–212. Springer, 2020.
- [PS120] Seol Roh, Ki-Moon Jeong, Hye-Young Cho, and Eui-Nam Huh. An efficient microservices architecture for mlops. In *2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 652–654. IEEE, 2023.
- [PS121] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, David Douglas, and Conrad Sanderson. Software engineering for responsible ai: An empirical study and operationalised patterns. In *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice*, pages 241–242, 2022.
- [PS122] Gorka Zárate, Raúl Miñón, Josu Díaz-de Arcaya, and Ana I Torre-Bastida. K2e: Building mlops environments for governing data and models catalogues while tracking versions. In *2022 IEEE 19th International Conference on Software Architecture Companion (ICSA-C)*, pages 206–209. IEEE, 2022.
- [PS123] Sam Leroux, Pieter Simoens, Meelis Lootus, Kartik Thakore, and Akshay Sharma. Tinymlops: Operational challenges for widespread edge ai adoption. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1003–1010. IEEE, 2022.
- [PS124] Mariam Barry, Albert Bifet, and Jean-Luc Billy. Streamai: Dealing with challenges of continual learning systems for serving ai in production. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering*

- in Practice (ICSE-SEIP)*, pages 134–137. IEEE, 2023.
- [PS125] Mariam Barry, Jacob Montiel, Albert Bifet, Sameer Wadkar, Nikolay Manchev, Max Halford, Raja Chiky, Saad EL Jaouhari, Katherine B Shakman, Joudi Al Fehaily, et al. Streammlops: Operationalizing online learning for big data streaming & real-time applications. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3508–3521. IEEE, 2023.
- [PS126] Ayush Shridhar and Deepak Nadig. Heuristic-based resource allocation for cloud-native machine learning workloads. In *2022 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pages 415–418. IEEE, 2022.
- [PS127] Philipp Ruf, Christoph Reich, and Djaffar Ould-Abdeslam. Aspects of module placement in machine learning operations for cyber physical systems. In *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–6. IEEE, 2022.
- [PS128] Ilias Syrigos, Nikolaos Angelopoulos, and Thanasis Korakis. Optimization of execution for machine learning applications in the computing continuum. In *2022 IEEE Conference on Standards for Communications and Networking (CSCN)*, pages 118–123. IEEE, 2022.
- [PS129] Alexandre Carqueja, Bruno Cabral, João Paulo Fernandes, and Nuno Lourenço. On the democratization of machine learning pipelines. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 455–462. IEEE, 2022.
- [PS130] Fotis Psallidas, Megan Eileen Leszczynski, Mohammad Hossein Namaki, Avrilia Floratou, Ashvin Agrawal, Konstantinos Karanasos, Subru Krishnan, Pavle Subotic, Markus Weimer, Yinghui Wu, et al. Demonstration of geyser: Provenance extraction and applications over data science scripts. In *Companion of the 2023 International Conference on Management of Data*, pages 123–126, 2023.
- [PS131] Dezhao Tu, Yeye He, Weiwei Cui, Song Ge, Haidong Zhang, Shi Han, Dongmei Zhang, and Surajit Chaudhuri. Auto-validate by-history: Auto-program data quality constraints to validate recurring data pipelines. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4991–5003, 2023.
- [PS132] Alec Gunn, Dylan Rankin, Philip Harris, Erik Katsavounidis, Ethan Marx, Muhammed Saleem, Michael Coughlin, and William Benoit. A software ecosystem for deploying deep learning in gravitational wave physics. In *Proceedings of the 12th Workshop on AI and Scientific Computing at Scale using Flexible Computing Infrastructures*, pages 9–17, 2022.
- [PS133] Steve Harris, Tim Bonnici, Thomas Keen, Watjana Lilaonitkul, Mark J White, and Nel Swanepoel. Clinical deployment environments: Five pillars of translational machine learning for health. *Frontiers in Digital Health*, 4:939292, 2022.
- [PS134] Tim Raffin, Tobias Reichenstein, Jonas Werner, Alexander Kühl, and Jörg Franke. A reference architecture for the operationalization of machine learning models in manufacturing. *Procedia CIRP*, 115:130–135, 2022.
- [PS135] Rustem Dautov, Erik Johannes Husom, and Fotis Gonidis. Towards mlops in mobile development with a plug-in architecture for data analytics. In *2022 6th International Conference on Computer, Software and Modeling (ICCSM)*, pages 22–27. IEEE, 2022.
- [PS136] Fabian Stieler and Bernhard Bauer. 18th international conference on evaluation of novel approaches to software engineering (enase 2023). 04 2023.
- [PS137] Xiaoyu Zhang, Jorge Piazentin Ono, Huan Song, Liang Gou, Kwan-Liu Ma, and Liu Ren. Sliceteller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):842–852, 2022.
- [PS138] Ruibo Chen and Wenjun Wu. Parallelizing automatic model management system for aiops on microservice platforms. In *European Conference on Parallel Processing*, pages 376–387. Springer, 2021.
- [PS139] John Wickström, Magnus Westerlund, and Emmanuel Raj. Decentralizing machine learning operations using web3 for iot platforms. In *2022 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 238–245. IEEE, 2022.
- [PS140] Nikos Psaromanolakis, Vasileios Theodorou, Dimitris Laskaratos, Ioannis Kalogeropoulos, Maria-Eleftheria Vlontzou, Eleni Zarogianni, and Georgios Samaras. Mlops meets edge computing: an edge platform with embedded intelligence towards 6g systems. In *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 496–501. IEEE, 2023.
- [PS141] Alvaro Luis, Miguel A Patricio, Antonio Berlanga, and José M Molina. Seamless transition from machine learning on the cloud to industrial edge devices with thinger. io. *IEEE Internet of Things Journal*, 2023.
- [PS142] Guillermo Chinarro Álvarez, César Alejandro Achig Ramírez, Javier Andión, Juan C Dueñas, et al. Toward an integrated and supported machine learning process. In *2023 7th International Young Engineers Forum (YEF-ECE)*, pages 37–42. IEEE, 2023.
- [PS143] Sachchidanand Singh, Naveen Singh, and Vinay Singh. Comparative analysis of open standards for machine learning model deployments. In *ICT Systems and Sustainability: Proceedings of ICT4SD 2021, Volume 1*, pages 499–507. Springer, 2022.

- [PS144] Mattia Antonini, Miguel Pincheira, Massimo Vecchio, and Fabio Antonelli. Tiny-mlops: A framework for orchestrating ml applications at the far edge of iot systems. In *2022 IEEE international conference on evolving and adaptive intelligent systems (EAIS)*, pages 1–8. IEEE, 2022.
- [PS145] Hemadri Jayalath and Lakshminish Ramaswamy. Enhancing performance of operationalized machine learning models by analyzing user feedback. In *Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing*, pages 197–203, 2022.
- [PS146] Laurent Boué, Pratap Kunireddy, and Pavle Subotić. Automatically resolving data source dependency hell in large scale data science projects. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 1–6. IEEE, 2023.
- [PS147] DR Niranjan et al. Jenkins pipelines: A novel approach to machine learning operations (mlops). In *2022 International Conference on Edge Computing and Applications (ICECAA)*, pages 1292–1297. IEEE, 2022.
- [PS148] Mlops: Machine learning operations with red hat openshift. 2023. [White Paper].
- [PS149] Data, mlops, and iot for the next-generation insurance industry. 2021. [Report].
- [PS150] Top 5 ways developers and data scientists can collaborate. 2022. [Blog].
- [PS151] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th international conference on software engineering*, pages 413–425, 2022.
- [PS152] Amazon ai fairness and explainability whitepaper. 2023. [White Paper].
- [PS153] Operationalizing machine learning. [Online; accessed 1. Apr. 2022].
- [PS154] William Benton. Demystifying Enterprise MLOps, Mar 2023. [Online; accessed Nov 2023].
- [PS155] Souma Kanti Paul, Sadia Riaz, and Suchismita Das. A conceptual architecture for ai in supply chain risk management. In *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*, pages 1–5. IEEE, 2022.
- [PS156] Ruibo Chen, Yanjun Pu, Bowen Shi, and Wenjun Wu. An automatic model management system and its implementation for aiops on microservice platforms. *The Journal of Supercomputing*, 79(10):11410–11426, 2023.
- [PS157] Juan Pineda-Jaramillo and Francesco Viti. Mlops in freight rail operations. *Engineering Applications of Artificial Intelligence*, 123:106222, 2023.
- [PS158] Sebastian Schelter, Stefan Grafberger, Shubha Guha, Bojan Karlas, and Ce Zhang. Proactively screening machine learning pipelines with arguseyes. In *Companion of the 2023 International Conference on Management of Data*, pages 91–94, 2023.
- [PS159] Tobias Müller, Milena Zahn, and Florian Matthes. On the adoption of federated machine learning: Roles, activities and process life cycle. In *ICEIS (1)*, pages 525–531, 2023.
- [PS160] Rakshith Subramanya, Seppo Sierla, and Valeriy Vyatkin. From devops to mlops: Overview and application to electricity market forecasting. *Applied Sciences*, 12(19):9851, 2022.
- [PS161] Vlad Stirbu, Tuomas Granlund, and Tommi Mikkonen. Continuous design control for machine learning in certified medical systems. *Software Quality Journal*, 31(2):307–333, 2023.
- [PS162] Harsha Moraliyage, Dilantha Haputhanthri, Chamod Samarajeewa, Nishan Mills, Daswin De Silva, Milos Manic, and Andrew Jennings. Automated machine learning in critical energy infrastructure for net zero carbon emissions. In *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*, pages 1–7. IEEE, 2023.
- [PS163] Raúl Miñón, Josu Díaz-de Arcaya, Ana I Torre-Bastida, Gorka Zarate, and Aitor Moreno-Fernandez-de Leceta. Mlpacker: A unified software tool for packaging and deploying atomic and distributed analytic pipelines. In *2022 7th International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–6. IEEE, 2022.
- [PS164] Marius Schlegel and Kai-Uwe Sattler. Mlflow2prov: extracting provenance from machine learning experiments. In *Proceedings of the Seventh Workshop on Data Management for End-to-End Machine Learning*, pages 1–4, 2023.
- [PS165] Dominik Kreuzberger, Julianne Forgo Kreuzberger, Mihai Criveti, Przemek Czuba, and Yvette Machowski. What is mlops? [Online; accessed 1. Nov. 2023].
- [PS166] David Nigenda, Zohar Karnin, Muhammad Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini, and Krishnaram Kenthapadi. Amazon sagemaker model monitor: A system for real-time insights into deployed machine learning models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3671–3681, 2022.
- [PS167] Nathalie Rauschmayr, Sami Kama, Muhyun Kim, Miyoung Choi, and Krishnaram Kenthapadi. Profiling deep learning workloads at scale using amazon sagemaker. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3801–3809, 2022.
- [PS168] Houssem Guissouma, Moritz Zink, and Eric Sax. Continuous safety assessment of updated supervised learning models in shadow mode. In *2023 IEEE 20th International Conference on Software Architecture Companion (ICSA-C)*, pages 301–308. IEEE, 2023.
- [PS169] Josu Díaz-de Arcaya, Ana I Torre-Bastida, Raúl Miñón, and Aitor Almeida. Orfeon: An aiops framework for the goal-driven operationalization of distributed analytical pipelines. *Future Generation Computer Systems*, 140:18–35,

2023.

- [PS170] Peini Liu, Gusseppe Bravo-Rocca, Jordi Guitart, Ajay Dholakia, David Ellison, and Miroslav Hodak. Scanflow-k8s: Agent-based framework for autonomic management and supervision of ml workflows in kubernetes clusters. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 376–385. IEEE, 2022.
- [PS171] Igor L Markov, Hanson Wang, Nitya S Kasturi, Shaun Singh, Mia R Garrard, Yin Huang, Sze Wai Celeste Yuen, Sarah Tran, Zehui Wang, Igor Glotov, et al. Looper: An end-to-end ml platform for product decisions. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3513–3523, 2022.
- [PS172] Semo Yang, Jihwan Moon, Jinsoo Kim, Kwangkee Lee, and Kangyoong Lee. Flscalize: Federated learning lifecycle management platform. *IEEE Access*, 2023.
- [PS173] Hanqing Cao, Joshua Finer, Murad Megjhani, Daniel C Nametz, Virginia Lorenzi, Lena Mamykina, Richard Meyers, Sarah C Rossetti, and Soojin Park. Machine learning model deployment using real-time physiological monitoring: Use case of detecting delayed cerebral ischemia. In *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pages 42–45. IEEE, 2022.
- [PS174] Paul-Philipp Luley, Jan M Deriu, Peng Yan, Gerrit A Schatte, and Thilo Stadelmann. From concept to implementation: The data-centric development process for ai in industry. In *2023 10th IEEE Swiss Conference on Data Science (SDS)*, pages 73–76. IEEE, 2023.
- [PS175] Qi Cheng and Guodong Long. Federated learning operations (flops): Challenges, lifecycle and approaches. In *2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 12–17. IEEE, 2022.
- [PS176] Shir Chorev, Philip Tannor, Dan Ben Israel, Noam Bressler, Itay Gabbay, Nir Hutnik, Jonatan Liberman, Matan Perlmutter, Yurii Romanyshyn, and Lior Rokach. Deepchecks: A library for testing and validating machine learning models and data. *Journal of Machine Learning Research*, 23(285):1–6, 2022.
- [PS177] Keita Sakuma, Ryuta Matsuno, and Yoshio Kameda. A method of identifying causes of prediction errors to accelerate mlops. In *2023 IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest)*, pages 9–16. IEEE, 2023.
- [PS178] Aquilas Tchanjou Njomou, Marios Fokaefs, Dimitry Fumtim Silatchom Kamga, and Bram Adams. On the challenges of migrating to machine learning life cycle management platforms. In *Proceedings of the 32nd Annual International Conference on Computer Science and Software Engineering*, pages 42–51, 2022.
- [PS179] Bo Wen, Yan Koifman, Hongfei Tian, Boris Lublinsky, Raquel Norel, Carla Agurto, Dean Wampler, and Narciso Albarracin. Accelerating automation of digital health applications via cloud native approach. In *Proceedings of the Eighth International Workshop on Container Technologies and Container Clouds*, pages 1–6, 2022.
- [PS180] Gregor Cerar and Jernej Hribar. Machine learning operations model store: Optimizing model selection for ai as a service. In *2023 International Balkan Conference on Communications and Networking (BalkanCom)*, pages 1–5. IEEE, 2023.
- [PS181] Rustem Dautov, Erik Johannes Husom, Fotis Gonidis, Spyridon Papatzelos, and Nikolaos Malamas. Bridging the gap between java and python in mobile software development to enable mlops. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 363–368. IEEE, 2022.
- [PS182] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, and Zhenchang Xing. Towards a roadmap on software engineering for responsible ai. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, pages 101–112, 2022.
- [PS183] Rizal Broer Bahaweres, Aldi Zulfikar, Irman Hermadi, Arif Imam Suroso, and Yandra Arkeman. Docker and kubernetes pipeline for devops software defect prediction with mlops approach. In *2022 2nd International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, pages 248–253. IEEE, 2022.
- [PS184] Bradley Eck, Duygu Kabakci-Zorlu, Yan Chen, France Savard, and Xiaowei Bao. A monitoring framework for deployed machine learning models with supply chain examples. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2231–2238. IEEE, 2022.
- [PS185] Rohan Gawhade, Lokesh Ramdev Bohara, Jesvin Mathew, and Poonam Bari. Computerized data-preprocessing to improve data quality. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 1–6. IEEE, 2022.
- [PS186] Claudia Patricia Ayala Martínez, Besim Bilalli, Cristina Gómez Seoane, and Silverio Juan Martínez Fernández. Dogo4ml: Development, operation and data governance for ml-based software systems. In *Joint Proceedings of RCIS 2022 Workshops and Research Projects Track: co-located with the 16th International Conference on Research Challenges in Information Science (RCIS 2022): Barcelona, Spain, May 17-20, 2022*. CEUR-WS. org, 2022.
- [PS187] Ashish Singh Parihar, Umesh Gupta, Utkarsh Srivastava, Vishal Yadav, and Vaibhav Kumar Trivedi. Automated machine learning deployment using open-source ci/cd tool. In *Proceedings of Data Analytics and Management: ICDAM 2022*, pages 209–222. Springer, 2023.
- [PS188] Raúl Miñón, Josu Diaz-de Arcaya, Ana I Torre-Bastida, and Philipp Hartlieb. Pangea: an mlops tool for automatically generating infrastructure and deploying analytic pipelines in edge, fog and cloud layers. *Sensors*, 22(12):4425, 2022.

- [PS189] Dhia Elhaq Rzig, Foyzul Hassan, Chetan Bansal, and Nachiappan Nagappan. Characterizing the usage of ci tools in ml projects. In *Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 69–79, 2022.
- [PS190] Songzhu Mei, Cong Liu, Qinglin Wang, and Huayou Su. Model provenance management in mlops pipeline. In *Proceedings of the 2022 8th International Conference on Computing and Data Engineering*, pages 45–50, 2022.
- [PS191] Ayan Chatterjee, Bestoun S Ahmed, Erik Hallin, and Anton Engman. Quality assurance in mlops setting: An industrial perspective. *arXiv preprint arXiv:2211.12706*, 2022.
- [PS192] Joran Leest, Ilias Gerostathopoulos, and Claudia Raibulet. Evolvability of machine learning-based systems: An architectural design decision framework. In *2023 IEEE 20th International Conference on Software Architecture Companion (ICSA-C)*, pages 106–110. IEEE, 2023.
- [PS193] Antonio Guerriero, Roberto Pietrantuono, and Stefano Russo. Iterative assessment and improvement of dnn operational accuracy. In *2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 43–48. IEEE, 2023.
- [PS194] Georgios Samaras, Vasileios Theodorou, Dimitris Laskaratos, Nikolaos Psaromanolakis, Marinela Mertiri, and Alexandros Valantasis. Qmp: A cloud-native mlops automation platform for zero-touch service assurance in 5g systems. In *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, pages 86–89. IEEE, 2022.
- [PS195] Nils Baumann, Evgeny Kusmenko, Jonas Ritz, Bernhard Rumpe, and Moritz Benedikt Weber. Dynamic data management for continuous retraining. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, pages 359–366, 2022.
- [PS196] Jing Peng, Shiliang Zheng, Yutao Li, and Zhe Shuai. From source code to model service: A framework’s perspective. In *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 1355–1362. Springer, 2022.
- [PS197] Markus Borg. Agility in software 2.0—notebook interfaces and mlops with buttresses and rebars. In *International Conference on Lean and Agile Software Development*, pages 3–16. Springer, 2022.
- [PS198] Eyad Kannout, Michal Grodzki, and Marek Grzegorowski. Considering various aspects of models’ quality in the ml pipeline-application in the logistics sector. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 403–412. IEEE, 2022.

Appendix B BACKGROUND

B.1 CI/CD practices in traditional software engineering and DevOps

Continuous integration (CI) practices within agile development approaches [6] have widespread usage in today’s software development community. CI refers to frequent integration of software artifacts (e.g., code base, dependencies) during development [31]. The benefits of CI include reducing release cycles by keeping software always ready for release, increasing team productivity, finding bugs as early as possible, and improving software quality [30, 31]. On the other hand, continuous deployment (CD) indicates the automatic and frequent delivery of software to the customer or production environment [19, 29, 31]. This practice allows for the customer to receive early feedback on the functionality delivered of the software, helping to identify further improvements [9]. CI/CD contributes to reducing risks and improving software quality [13].

DevOps (Dev for development and Ops for operations) is a natural evolution of agile and CI/CD practices. It is a paradigm of continuous and collaborative software engineering [19]. It aims to deliver software more frequently using an automated, repeatable, and continuous process flow. Furthermore, DevOps involves practices, tools, and sociotechnical aspects that facilitate easier collaboration between development and operations teams [PS42].

B.2 MLOps

CI/CD practices have been very useful in the development of ML-enabled systems, much like their success in traditional software development. Based on its benefits in traditional software engineering, the ML development community adopts CI / CD practices in ML development [22]. The experimental and probabilistic nature of AI/ML systems requires frequent repetition of development

and deployment flow. Therefore, the ML development and deployment process demands practices similar to those of CI / CD [PS3].

MLOps concept has emerged with the goal of establishing a set of practices for more efficient ML development and deployment flow by mimicking the DevOps practices [PS1]. MLOps introduces additional practices specific to ML, emphasizing the automation of development and monitoring at all stages of ML, and deployment [PS1, PS10].

One important practice of MLOps is monitoring. Monitoring helps to track data and model performance shifts to ensure the reliability and quality of system over time [PS7], and enables continuous feedback loop between development, operations, and production environments. At this point, along with CI and CD, MLOps incorporates the continuous training (CT) concept, because the continuous delivery of an ML system is not independent from data collection, analysis, and model training phases [PS78].

Further, MLOps introduces new practices regarding the roles, responsibilities and collaboration between people with different backgrounds. MLOps engineer is a new role introduced with MLOps, referring to the professionals responsible for maintaining the operational stability of ML pipelines and the entire ML system [26, PS106]. Moreover, new collaboration practices, i.e., collaboration points [PS151], have been introduced to improve the efficiency of MLOps projects by enabling better collaboration between data science and software engineering teams.

Here we also mention DataOps and ModelOps which are two concepts that share similarities with MLOps. However, our work focuses on MLOps. Thus, we only highlight differences between concepts and only cover MLOps in our RQs.

DataOps is coined to refer to the automation of data analytics life-cycle [14, PS42]. DataOps establishes a set of practices for data-centric systems mimicking DevOps practices and aims to ship data frequently in an automated [33].

ModelOps is defined as continuous practices for AI-enabled software systems [PS33]. The key activities revolve around continuous model training, model metadata versioning, deployment, and management of AI pipelines.

While both DataOps and ModelOps are related to MLOps, these three terms refer to different concepts. DataOps centralizes data analytics, management, and the shipment of data. ModelOps centralizes building and managing various versions of models and serving them in AI systems. On the other hand, MLOps is primarily concerned with productionalization of ML systems and automating the whole ML pipeline from data collection through model training, deployment, and monitoring. Further, the skills required by these three concepts vary; DataOps mainly involves data science, data analysis, and data engineering, and ModelOps involves data science. MLOps utilizes a broader range of skills such as data and ML engineering, data science, and MLOps engineering (see Section 4.3).