The Past, Present, and Future of Research on the Continuous Development of AI

Monika Steidl*, Rudolf Ramler§, and Michael Felderer‡
*‡*University of Innsbruck*, Austria
§*Software Competence Center Hagenberg GmbH*, Austria
‡*German Aerospace Center (DLR), Institute of Software Technology*, Germany
‡*University of Cologne*, Germany
ORCID: *0000-0002-3410-7637, §0000-0001-9903-6107, ‡0000-0003-3818-4442

***Abstract*—Since 2020, 33 literature reviews have systematically synthesized research on the continuous development of AI, also known as Machine Learning Operations (MLOps), reflecting the increasing prevalence of AI models across various fields and the multifaceted challenges in their development, integration, and deployment. Yet, the lack of comprehensive analysis of these literature reviews and their covered topics complicates selecting relevant ones and anticipating future trends and research. In addition, these literature reviews gathered related 1397 primary sources to describe aspects of AI's continuous development, integration, and deployment, posing a hidden gem to gain insights into the past and present work and derive insights into the future of AI's continuous development.**

With this work, we 1) systematically collected and summarised 33 literature reviews via a Multivocal Literature Review (MLR) that focus on the continuous development, deployment, and integration of AI models. 2) Due to minimal overlap between the literature reviews' primary sources, we offer holistic insights into and interrelations of frequently addressed topics. These topics encompass the AI development pipeline, respective Software Engineering (SE) practices, and associated challenges. 3) We discuss future research directions for AI's continuous development, integration, and deployment. Therefore, we base our arguments on identified clusters in the primary sources of literature reviews. This discussion focuses on AI model reliability and resource consumption, emphasizing the interrelation of proposed future work and the effects on the whole pipeline.

***Index Terms*—tertiary study, MLOps, CD4ML, continuous development of AI, lifecycle pipeline, literature review, challenges, SE practices, AI reliability, Green AI**

I. Introduction

Continuous integration (CI) and continuous deployment (CD) are common practices to handle the dynamics and complexity of continuous software development in the DevOps lifecycle. CI/CD practices have also been adopted for the continuous development of Machine Learning (ML), Deep Learning (DL), and Artificial Intelligence (AI) models. In Machine Learning Operations (MLOps), these practices are applied to developing, integrating, testing, and deploying code, data, and the AI model itself [9], [13].

The widespread use and importance of AI have led to a tremendous amount of research, with more than a thousand individual papers (primary sources/studies) published on various aspects related to the continuous development of AI models. Hence, since 2020, numerous literature reviews (secondary studies) have attempted to collect and synthesize this research from different perspectives. Now, however, the increasing number of literature reviews has started to become an inextricable body of literature in itself. For instance, nine new literature reviews were published in 2023 alone, which are closely related to the topics covered by 24 existing reviews from previous years.

Furthermore, varying literature reviews employ different search terms and selection criteria to answer similar research questions, partitioning the primary sources and, thus, the available work in this field for synthesis. We found that there is only minimal overlap in considered primary sources among literature reviews, which introduces a bias and potential limitation for the robustness and reliability of the results and findings of these reviews. A holistic overview needs to incorporate knowledge from the entire set of primary studies collected by the various existing literature reviews (currently 1397 different primary sources) by analyzing the topics and results from available literature reviews. However, up to now, a meta-analysis of all existing literature reviews is not available.

Thus, the goal of our work is to provide a systematic overview of the existing literature reviews to assist researchers and practitioners in selecting appropriate secondary studies on the continuous development of AI, to prevent redundant research efforts in this area, and to help future research endeavors to focus on new advances [22], [23], [44]. By clustering and analyzing the body of primary sources along the timeline, we identify trends and gaps in current research and provide insights into emerging trends, open questions, and new possibilities for future research directions. Therefore, we target the following three research questions (RQs):

- **RQ_1: Which literature reviews regarding the continuous development, integration, and deployment of AI are available?**
- **RQ_2: What are the main topics covered by systematic literature reviews on the continuous development, integration, and deployment of AI?**
- **RQ_3: What are potential future directions for research on the continuous development, integration, and deployment of AI?**

This work was supported by the Austrian Research Promotion Agency (FFG) in the frame of the project ConTest [888127] and the SCCH COMET Competence Center INTEGRATE [892418].

The remainder of the paper is structured as follows: Section II describes the applied methodology, data analysis, and threats to validity. In Section III, we answer RQ_1 by concisely presenting available literature reviews and RQ_2 by summarising the main topics handled in these. Section IV discusses and reflects RQ_3 on potential future directions on the continuous development, integration, and deployment of AI. Section V summarizes the paper's results.

II. Methodology

We applied a Multivocal Literature Review (MLR) to collect literature reviews regarding the continuous development, integration, and deployment of AI. All data and detailed information regarding the data collection, extraction, and analysis steps can be found in our replication package [54].

A. Collection of Literature Reviews

We applied the guidelines established by Garousi et al. [11] to conduct our MLR. Our decision to incorporate peer-reviewed and non-peer-reviewed papers available at ArXiv and TechRxiv aimed to present a comprehensive overview of existing research and ongoing scientific advancements. This is crucial given the ever-evolving nature of this research area. To ensure a thorough quality of the included non-peer-reviewed literature reviews, we assessed their quality based on the proposed assessment guidelines[1] by [11].

TABLE I
Selection criteria for literature reviews

	Inclusion Criteria	Exclusion Criteria
Method	Description of the applied methodology (systematic collection of sources with the primary aim to depict the current state of literature)	No methodology available or goal to validate proposed approach without any systematic source collection process
Relevance	Relevant information to answer research questions (lifecycle/pipeline of the continuous development of AI)	Sources focusing on AI for DevOps/AIOps, culture, specific tasks of the lifecycle (e.g., only data handling)
Language	English	Any other language
Publication Type	Peer-reviewed & non-peer reviewed (ArXiv, TechRxiv) when they fulfilled quality assessment for grey literature[1]	Short papers, master thesis
Access	Full text accessible	Restricted access

Table I describes the **selection criteria** we applied to the literature reviews on the continuous development, integration, and deployment of AI. We executed the search for relevant reviews between February 2023 and the end of December 2023 to retrieve literature reviews published before the end of December 2023. We used Google Scholar, IEEEXplore, ACM Digital Library, Springer Link, and Web of Science. The search terms were derived from established terminology in continuous Software Engineering (SE) [10]. Figure 1 illustrates the search strings formed via various combinations of these terms, such as

```
("Literature Review" OR "Literature Study")
AND ("Artificial Intelligence" OR "AI")
AND "Continuous Integration" OR "CI".
```

The database-specific search strings are documented in our replication package [54]. In the search, we covered title, abstract, keywords, and full text.

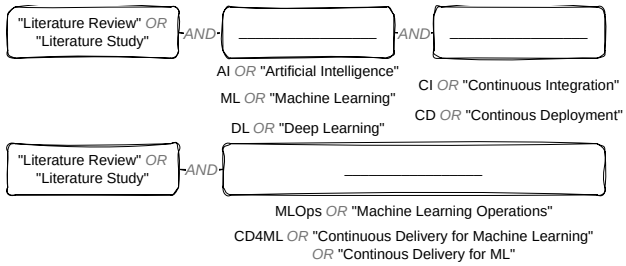

Fig. 1. Search string compilation for the Multivocal Literature Review (MLR), where each text under the box is combined with the other texts of the linked box

B. Data collection, extraction, and analysis

We used a data extraction form (see guidelines by Kitchenham et al. [22], [23]) to systematically collect, extract, and analyze the literature reviews [54]. If a review did not include all the necessary information, we contacted the authors of the papers to ensure we had a complete set of information for our analysis. To ensure consistency in data extraction, any uncertainties were discussed with the author team, and we performed a test-retest procedure to verify consistency [22].

For RQ_1, we extracted publication details such as title, source, author(s), publication date, type and venue, and the number of primary sources. We classified the literature reviews according to research method type, research question, search strings, and topics. We rated how well the literature reviews comply with empirical standards for systematic reviews [2] [3] [4] [43]. Therefore, we employed an equivalent rating scale proposed by Kitchenham et al. [22] with the categories fulfillment (1.0), partial fulfillment (0.5), and lack of coverage (0.0).

For RQ_2, we applied the thematic analysis proposed by Cruzes and Dyba's [3] to identify major themes in the literature reviews. We marked text segments, labeled them, and translated them into themes that refer to major stages in the continuous development lifecycle pipeline, where sub-themes specify SE practices and associated challenges.

[1]Quality assessment checklist for grey literature in software engineering including authority of the producer, methodology, objectivity, date, position w.r.t related sources, novelty, impact, outlet type see Table 7 in [11]

[2]**essential attributes**: identify type of review; replicable search process with search terms & selection criteria; describe data extraction, synthesis, & coding schemes; chain of evidence from data to RQ; provide conclusion & recommendation.

[3]**desirable attributes**: supplementary materials; mitigate sampling & publication bias; sufficiently rigorous search processes; assess primary sources' quality; evaluate coverage; use multiple analysts for reliability; reflect on biases; consolidate results visually; include PRISMA flow diagram; employ appropriate meta-analysis methods; integrate results into prior research; present practical, evidence-based guidelines; distinguish results from interpretations.

[4]**extraordinary attributes**: researchers independently conduct preliminary searches to refine scope & keywords; verify interpretations with primary study authors; apply integrative data analysis.

For RQ_3, we aggregated the 1397 primary sources of the identified literature reviews. To analyze these primary sources, we clustered them based on their abstracts using BERT (Bidirectional Encoder Representations from Transformers). Specifically, we applied BERTopic [14]. This approach assumes that documents containing the same topic are semantically similar by converting sentences and paragraphs to dense vector representations using pre-trained language models.

C. Threats to validity

For **internal validity**, we evaluated the methodological quality of the literature reviews according to Ralph et al. [43] and Garousi et al. [11] to ensure that the papers provide methodological transparency and soundness of the derived results. It has to be noted that while the absence of reporting the details about the applied method does not necessarily imply unmet empirical standards, our evaluation only relied on the available statements the authors made in their reviews. Furthermore, we encountered discrepancies between the stated and our actual retrieved primary sources due to duplicates, unavailable papers, and a missing list of primary sources. Please refer to the replication package [54] for further insights. However, we did not modify the provided white, gray, and $\sum$ amounts from the literature reviews in Table II to prevent inconsistencies.

Regarding **external validity** in RQ_3 on the future of continuous development in AI, the obtained 33 literature reviews are published within a narrow time frame of four years, which may limit their insights on trends over time. To mitigate this, we also considered their primary sources spanning a broader research period (1990 to 2024). Additionally, we integrated recent related work into our discussion to mitigate the inherent lag of literature reviews in capturing the latest research advances.

Construct validity may be affected by the search string used. Although we did not include `"mapping study"` in our search string, we included three mapping studies. However, we potentially may not have found all related mapping studies. Furthermore, we clustered the primary source topics using BERTopic, which assigns a paper's abstract to a single cluster, even though it may belong to multiple ones. Thus, we examined these clusters using hierarchical clustering and an inter-topic distance map to address potential clustering issues.

For **reliability**, we manually collected the title of primary sources either via the information provided in the literature reviews or contacted authors. Then, we automatically extracted the information from the required primary sources from Google Scholar, where we selected peer-reviewed versions over pre-prints when both were available.

III. Results

Our systematic search yielded 33 literature reviews on the continuous development, integration, and deployment of AI for further analysis. First, we provide a detailed overview of these reviews in Section III-A. Second, we explore the main topics covered in these literature reviews in Section III-B.

A. RQ_1: Available Literature Reviews

Table II shows the 33 literature reviews identified in our systematic search process sorted by publication date. Publication dates range from October 2023 to July 2020, with most reviews published in 2022 (14 papers), followed by nine in 2023, six in 2021, and four in 2020. We did not find literature reviews before July 2020.

The literature reviews were published in journals (13), conferences (11), workshops (3), and on ArXiv (5) or TechRxiv (1). Most reviews were either multivocal literature reviews (14) or systematic literature reviews (13), with additional gray literature reviews (3) and systematic mapping studies (3).

The number of primary studies considered in the reviews ranges from 9 to 405 (median 55.5). In total, we were able to extract 1397 primary sources from these literature reviews.

Furthermore, the column *Empirical standards* in Table II indicates which literature reviews fulfilled or exceeded 50% of the standards for systematic reviews[2][3][4] [43] (marked with ✓) and which do not (∼). The identified reviews perform very well in fulfilling essential attributes[2], such as identifying the type of review and presenting a detailed description of the search process, search terms, and clear selection criteria. However, only 9% of the literature reviews demonstrated that their search process was sufficiently rigorous (desirable attribute[3]) by defining control papers during their initial pilot study ([40]) or cross-checking the identified papers by additionally executing a search via Google Scholar ([1], [7]).

Key-Takeaways about available literature reviews on the continuous development of AI (RQ_1):

- 33 literature reviews are available until the end of 2023 (see Table II)
- These literature reviews include 9 to 405 primary sources and collectively report 1397 primary sources
- 67% of the reviews fulfill at least half of the empirical standards for systematic literature reviews[2][3][4] [43]

B. RQ_2: Main topics covered by literature reviews

In the analysis of the literature reviews, we identified 12 different topics which are covered by the literature reviews: Pipelines (covered by 15 reviews), challenges (16), Software Engineering (SE) for AI (17), tools (11), application settings (9), demography (7), architectures (3), definitions (4), maturity models (4), triggers (3), roles and team (4), and requirements (1). Table II shows which reviews cover which of these topics.

In the following, we focus our analysis on the three most frequently covered topics. Figure 2 illustrates the key findings related to these topics and their interrelation. The topic **pipeline** covers information regarding the continuous development lifecycle of AI and its respective four main stages, (`Design Decision`, `Data`, `Model`, `DevOps`) (rectangles in Figure 2). The topic **SE for AI** describes how the tasks in each stage are realized via best practices (in brackets). **Challenges** describe obstacles during the continuous development of AI and are mapped to respective stages (blue and red lines).

TABLE II

OVERVIEW OF LITERATURE STUDIES ON THE CONTINUOUS DEVELOPMENT OF AI (TYPE: ***A**rXiv*, ***C**onference*, ***J**ournal*, ***T**echRxiv*, ***W**orkshop*; METHOD: ***M**ultivocal **L**iterature **R**eview, **S**ystematic **L**iterature **R**eview, **G**ray **L**iterature **R**eview, **S**ystematic **M**apping **S**tudy*; EMPIRICAL STANDARDS: $\sim < 50\%$ OR ✓$\geq 50\%$ *fulfilled*; TOPICS OTHERS: ***A**rchitecture, **D**efinition, **M**aturity Model, **T**riggers, **R**oles/Team, **Re**quirements*)

Paper	**Publication**			**# Sources**					**Topics**						
	Date	**Type**	**Method**	**White**	**Gray**	**Total**	**End data collection**	**Empirical standards**	Pipeline	Challenges	SE for AI	Tools	Apl. Setting	Demography	Others
Faubel et al. [7]	10.2023	J	SLR	69	0	69	05.2022	✓	●	●	●		●		D,M,T
Diaz-de-Arcaya et al. [5]	10.2023	J	SLR	93	0	93	2023	✓		●		●		●	A
Lakha et al. [27]	08.2023	C	SLR	37	0	37	11.2021	✓		●	●		●		
Heiland et al. [17]	07.2023	A	MLR	35	16	51	2022	✓			●				
Alves et al. [1]	07.2023	J	MLR	37	93	130	12.2021	✓	●		●				
Steidl et al. [52]	05.2023	J	MLR	79	72	151	06.2021	✓	●	●	●	●	●		D,T
Moreschini et al. [38]	04.2023	A	MLR	51	203	254	11.2022	✓				●			
Kreuzberger et al. [25]	03.2023	J	SLR	27	0	27	05.2021	$\sim$	●		●	●			D,A,R
Schlegel & Sattler [46]	01.2023	J	MLR	-	-	-	05.2022	$\sim$				●			
Kumara et al. [26]	11.2022	T	GLR	0	58	58	03.2022	$\sim$	●						R,A,Re
Lu et al. [33]	09.2022	A	MLR	205	69	274	07.2022	✓			●				
Recupito et al. [45]	09.2022	C	MLR	6	54	60	05.2020	✓				●			
Shivashankar & Martini [49]	09.2022	C	SLR	56	0	56	01.2022	✓	●	●					
Barrak et al. [2]	09.2022	J	SMS	53	0	53	06.2022	✓		●		●	●	●	
Mboweni et al. [37]	07.2022	C	SLR	60	0	60	2022	$\sim$							D
Testi et al. [56]	06.2022	J	SLR	-	-	-	2022	$\sim$	●	●		●			
Kolltveit & Li [24]	05.2022	W	SLR	24	0	24	09.2021	✓		●		●			
Lakshman & Eisty [28]	05.2022	W	MLR	-	-	33	2021	$\sim$	●		●		●		
Lima et al. [29]	04.2022	C	SLR	30	0	30	07.2021	✓		●	●	●		●	M,R
Martinez-Fernandez et al. [36]	04.2022	J	SMS	-	-	248	03.2020	✓		●	●			●	
Serban & Visser [48]	03.2022	C	MLR	-	-	42	01.2022	✓	●	●					
Warnett & Zdun [59]	03.2022	C	GLR	0	35	35	-	✓			●				T
Warnett & Zdun [60]	03.2022	J	GLR	0	29	29	-	$\sim$			●				
Lorenzoni et al. [32]	11.2021	A	MLR	23	10	33	06.2020	$\sim$	●		●				
Giray [12]	10.2021	J	SLR	141	0	141	12.2019	✓		●	●			●	
John et al. [21]	09.2021	C	MLR	6	15	21	03.2021	✓	●						M,R
Lo et al. [31]	05.2021	J	MLR	-	-	231	12.2019	✓	●				●		
Xie et al. [61]	05.2021	W	SMS	-	-	405	07.2020	✓	●					●	
John et al. [20]	01.2021	C	MLR	13	6	19	08.2020	$\sim$	●	●	●		●		
Nascimento et al. [40]	11.2020	A	SLR	55	0	55	2019	✓		●	●	●		●	
Figalist et al. [8]	11.2020	C	SLR	-	-	-	-	$\sim$	●				●		
Lwakatare et al. [34]	11.2020	C	MLR	3	6	9	-	$\sim$		●					M
Lwakatare et al. [35]	07.2020	J	SLR	72	0	72	-	✓		●	●		●		

1) Design Decision: In this initial stage of the pipeline, *requirements engineering* tasks necessitate extensive collaboration among experts (e.g., data scientists, domain experts) [56]. This task applies SE practices to understand the problem (e.g., user needs and expectations) [1], [12], [27], [40], grasp the current context (e.g., aligning data to the problem, re-engineering processes, assessing AI suitability) [1], [17], [33], specify requirements for data, models, and systems (e.g., defining hypotheses, verifiable (non-)functional requirements) [1], [27], [28], [33], [36], [48], [56], and validate the requirements through verifiable tests (e.g., ethical principles, prototyping, preliminary experiments) [1], [12], [27], [33]. Additionally, in this stage, one decides on appropriate *pipeline iteration* cycles, considering triggers such as data or code updates, concept drifts, scheduled or manual triggers [7], [8], [20], [26], [52], [60]. DevOps practices, including continuous training and tool support for end-to-end development, are also part of this stage [1], [8], [12], [17], [21], [29], [36].

2) Data: In this pipeline stage, data is continuously *acquired* through integration or synthetic generation to prevent performance degradation [1], [12], [20], [21], [29], [31], [33], [35], [48], [49], [52]. Data *preprocessing* tasks require practices to analyze, clean and transform the acquired data [8], [12], [21], [31], [33], [49], [52], [61]. For *data Quality Assurance (QA)*, practices involve observing descriptive statistics and checking feature impact while adhering to data requirement specifications for reliability, accountability, and fairness [8], [27], [33], [35], [52]. Data *tracking* ensures provenance through versioning [1], [21], [29], [52], [59], [61], avoiding data fragmentation over pipeline iterations [32].

3) Model: This stage involves *planning* decisions for developing the AI model, addressing issues like selecting an appropriate AI framework and ethical considerations [1], [17], [33], [40]. The *training* task incorporates strategies that align

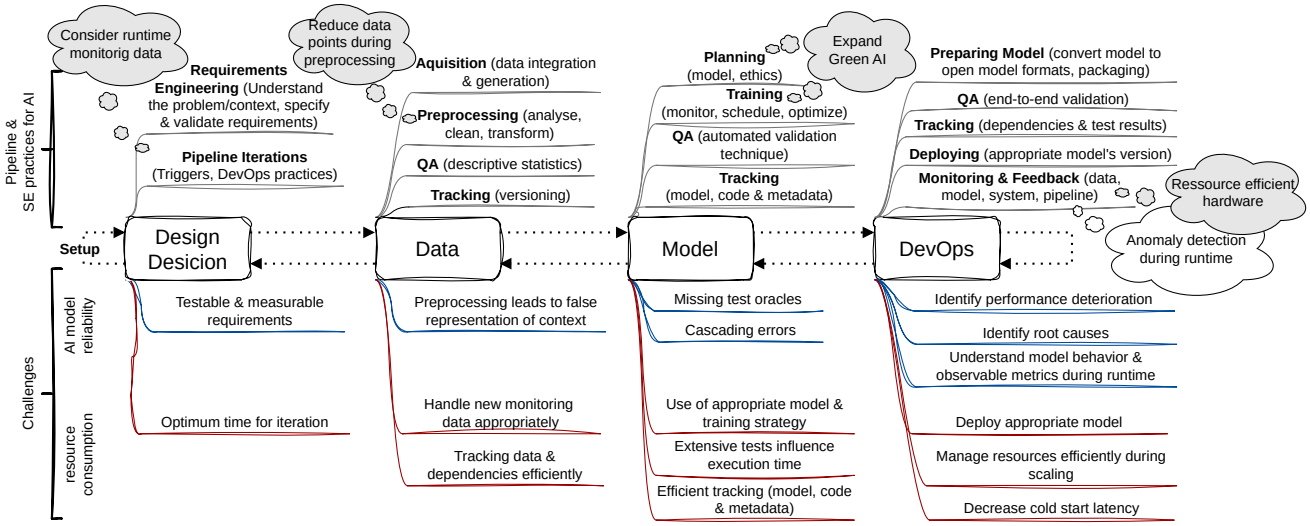

Fig. 2. Holistic overview of literature reviews on the pipeline for the continuous development of AI, encompassing four stages (rectangles), respective tasks (arrows), SE practices (in brackets), challenges (blue: AI model reliability; red: resource consumption), & potential future research directions (white cloud: AI model reliability & gray cloud: resource consumption)

with available computing resources, time frames, and AI algorithms, [1], [35], [48], [52]. Identified SE practices monitor and schedule training activities [2], [12], [20], [26], [56]. Furthermore, practices such as trial and error or advanced approaches like AutoML optimize the AI model's performance [1], [12], [49], [52], [61]. Model *QA* practices focus on test automation, management, and validation techniques (e.g., performance, quality, robustness) [8], [12], [35], [52]. Model *tracking* needs to version not only trained models but also the respective code and metadata (e.g., configurations, dependencies) [21], [26], [33], [35], [45], [48], [52], [59].

4) DevOps: This stage focuses on *preparing* models for deployment by converting them to open model formats [26], [52], and packaging using containerization practices [12], [20], [38], [52], [59]. *QA* validates the end-to-end system before pushing it to production [8], [40], [52]. Dependencies and test results are *tracked* for reproducibility [26], [28], [52]. Once successfully validated, the appropriate model's version is *deployed* [52], [59], with costs to be minimized [20]. Since AI model performance tends to decrease over time [56], the data [8], [52], model [1], [8], [12], [29], [33], [52], [59], system [27], [52], and the pipeline itself [8], [45] need to be continuously *monitored*.

5) Tasks without a stage: Tasks that cannot be assigned to a specific pipeline stage belong to the *setup* of the AI model lifecycle pipeline [1], [8], [17], [40], [52], [59], [61]. These pipelines should prevent decay due to fast and efficient re-deployment of AI models [1], [12], [20], [35], [36]. Furthermore, the pipelines should be extensible and portable to use preferred tools, libraries, and programming languages [2], [26], [45], [60].

6) Challenges: The literature reviews usually discuss task-specific challenges. However, these challenges should not be discussed independently because they are interrelated and affect each other. Hence, we consolidate them into two overarching challenges: (a) AI model reliability and (b) resource consumption. We use the previously presented holistic view in Figure 2 to map the identified challenges (AI model reliability depicted with a blue line and resource consumption depicted with a red line) to the pipeline stages. We discuss potential future work in RQ_3 on these two challenges concerning the entire lifecycle for the continuous development of AI.

When evaluating the **AI model reliability**, the stage `Design Decision` is essential to define correct, testable, and measurable functional and non-functional requirements [2], [27], [29], [40], [48], [56]. Otherwise, the stage `Data` might use preprocessing strategies that lead to a false representation of the context, misaligning with the actual requirements and leading to unexpected consequences [36], [48], [49]. Consequently, the stage `Model` is challenged to specify test oracles, define correct behavior, and identify the root cause of performance deviations due to cascading errors from previous tasks [12], [35], [36], [40], [49]. Furthermore, models might exhibit strong performance during the testing phases. Yet, their performance in a production environment may deteriorate [7], [20], [34], [49], [52] or errors can cascade to other software parts when deployed during the stage `DevOps` [48]. Thus, the root cause of errors must be identified. However, the impact of handled tasks on the AI model in production is poorly understood, such as how reducing the dataset during preprocessing affects the model's quality [52]. In addition, handling high-dimensional data, complex model architectures, and executing interrelated tasks throughout the pipeline complicates identifying the root cause of errors or behavior anomalies. Thus, domain-specific knowledge about the whole pipeline or knowledge about error characteristics is essential [34], [40],

[48]. Understanding how AI models normally behave during runtime is challenging, especially when scalability is involved, which makes it rare to have observable metrics that are easy to interpret [2], [24], [40], [52].

Furthermore, challenges associated with resource consumption are often discussed, highlighting the unclear impact of individual tasks on the total resource consumption of the entire pipeline and AI model in production. For instance, one key challenge during the `Design Decision` stage is finding the optimum time to rerun the pipeline [12], [35], [36], [40], [48] without relying on fixed intervals [20], [52] or constant manual monitoring [34], [52]. However, a data-triggered rerun may also not be optimal to avoid unnecessary pipeline iterations, as more data may not necessarily enhance model reliability. While it might reduce reruns, it still requires significant resources for storing and preprocessing the additional monitored data [5], [36], [40], [52]. When rerunning the pipeline, it is essential to appropriately handle new monitoring data and track data dependencies in the stage `Data` to ensure explainability, reproducibility, and consistency [20], [48], [52]. Deciding when, what, and where to store data with unstable dependencies is challenging because changes can arbitrarily harm the model and are costly to diagnose [5], [40].

For the stage `Model`, challenges often arise during training, such as selecting the appropriate model for the specific domain [40], and choosing an appropriate training strategy (e.g., federated or transfer learning) to meet defined requirements while considering resource constraints [35], [49]. The challenge lies in ensuring model performance, which is uncertain and stochastic, leading to extensive tests that impact pipeline execution time and increase its resource consumption [48], [52]. Efficiently tracking multiple versions of AI models, including their dependencies, configurations, and test results, is crucial to ensure deploying models, similar to the aforementioned data tracking challenges [12], [34], [35], [40], [48], [52].

During the stage `DevOps`, determining which model should be deployed is a challenging decision, where not only the model performance [52] but also resource consumption (e.g., carbon footprint) [56], reasonable scalability [5], [20], [40] and memory consumption during cold start latency needs to be considered [2], [24], [35].

Key-Takeaways about summary of main topics in literature reviews (RQ_2):

- Three main topics (pipeline, SE for AI, challenges) mapped to four stages `Design Decision`, `Data`, `Model` & `DevOps` as depicted in Figure 2
- AI model reliability & resource consumption summarise interrelated challenges throughout the pipeline

IV. Discussion

Based on the results of RQ_2, we identified trends by computing the relative occurrence of tasks covered in the reviews per year, as shown in Figure 3. Furthermore, we use identified clusters (Figure III) from the primary sources over time to predict potential future directions for addressing challenges in the continuous AI development (specifically AI model reliability and resource consumption) to answer RQ_3.

TABLE III
Identified clusters of primary sources via BERTopic

ID	Cluster Name	Description
2	*safety_system_autonomous_vehicle*	Covers research on safety-critical aspects and ensuring reliability for autonomous driving
3	*explanation_model_machine_interpretable*	Covers explainability of black-box models to improve trust and transparency by enhancing the interpretability of predictions
4	*adversarial_attack_example_model*	Covers potential security threats, such as adversarial attacks and perturbation, and methods to provide robustness of AI models through defense strategies
5	*fairness_discrimination_bias_decision*	From 2017 onward, there has been increased research interest in improving AI model fairness, reducing bias, and mitigating discrimination for ethical decision-making
6	*testing_dl_dnn_test*	Covers related topics to DL and neural network, where testing techniques have been increasingly discussed since 2017
7	*edge_iot_computing_cloud*	Covers data safety, privacy, response time, and resource management (e.g., battery life constraints) for edge and IoT devices, with TinyML identified as a main keyword from 2021 onwards
8	*serverless_cloud_cost_computing*	Covers challenges regarding compliance with service level agreements (latency, inference, auto-scaling to mitigate provisioning latency) and resources and cost-effectiveness
9	*data_science_software_process*	Covers data science processes, including data-driven development, analysis of big data, and workflow supported by tools, DevOps pipelines, and platforms
10	*testing_test_bug_machine*	Covers challenges in identifying failures, emphasizes the importance of testing and validating AI models' quality using techniques like test oracles, metamorphic testing, or mutation testing. This research increased from 2017 onwards
11	*ai_ethical_intelligence_artificial*	Ethical guidelines and principles for software development (specifically healthcare)
12	*dl_deep_model_learning*	Covers deep neural networks (DNNs), focusing on interpreting black-box models to identify unreasonable predictions
Others	0 *ml_machine_data_learning*: software engineering for AI & ML 1 *federated_learning_data_privacy*: federated learning *13 recommendation_user_online_system*: recommender systems	

A. AI model reliability

Future research on AI model reliability is paramount because half of the clusters (see Figure III) can be mapped to the AI models' reliability, such as safety in autonomous vehicles (2), adversarial attacks (4), explanation (3), fairness and bias (5), ethics (11), QA for DL (6,12) and AI (10). Because there is a surge in research interest in monitoring the

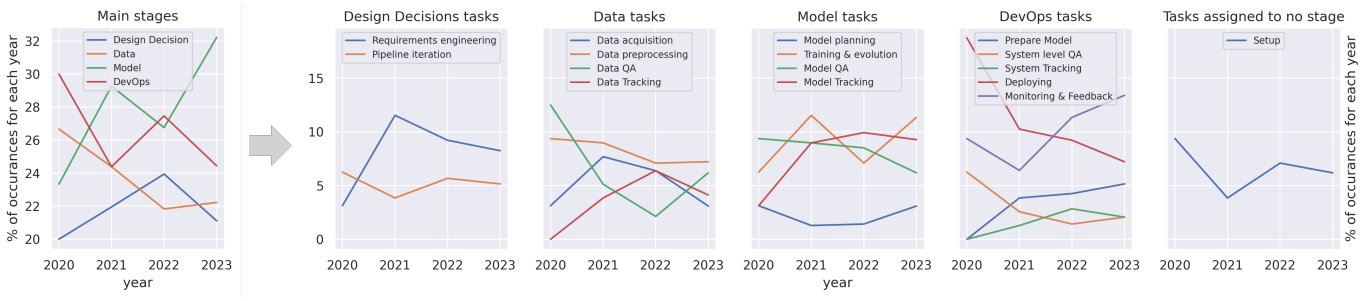

Fig. 3. Percentage of occurrences for the four stages per year broken down to the percentage of occurrences for each stage's tasks per year

system behavior in the stage `DevOps` (Figure 3), we propose to leverage this monitored runtime data to ensure reliable AI models in production. Therefore, we suggest providing automated solutions to analyze this monitored runtime data to overcome identified challenges in the stage `Data` (preprocessing leads to false representation), `Model` (missing test oracles, cascading errors), or `DevOps` (performance deterioration and root cause analysis). Efficient automated monitoring combined with QA techniques is essential to identify unanticipated behavior or performance variance in time to improve reliability and quality [16], [21], [41], [45], [53], especially important when adversarial attacks happen. So far, previous research advances detect anomalies by identifying deviations in monitoring data that consist of logs, traces (the invocation path of microservices), and metrics data (e.g., CPU, GPU, resources, latency, etc.). By applying this already available knowledge to systems with AI models' monitoring data, it will be possible to pinpoint failures or anomalous behavior and identify root causes of deviations of the AI model performance [30], [51], [63]. In addition, future research should look into enhancing available monitoring data by using AI specific monitoring data, such as inference data, inference latency, and accuracy, that give insights into AI faults and anomalies [15], [18], [19], [39], [55], [64].

B. Resource consumption

As the challenges in Figure 2 indicate, efficient resource consumption is of high interest, where decisions taken in one stage can highly influence the resource consumption of the whole pipeline (e.g., an optimum time for pipeline iteration, appropriate algorithms, data, and model tracking). This knowledge would help one identified cluster focusing on edge devices (7) because they require efficient resource management (e.g., tinyML) for battery life. Related work also identified that 93% of survey participants consider resource consumption when setting up their pipeline [53], focusing on the carbon footprint when developing AI [6], [47], [58].

Thus, we propose that future work prioritizes energy consumption and carbon footprint, integrating them into the comprehensive pipeline rather than studying them in isolation. For instance, individual strategies presented in the following must justify their additional resource consumption for savings in the other tasks. So far, Verdecchia et al. [58] identified in their literature review that only three out of 98 papers consider the whole pipeline to ensure Green AI. For instance, evaluating already collected runtime monitoring data, including information on the model's quality and the associated technique to ensure reliability, might suggest whether another pipeline iteration in the stage `Design Decision` is required.

In addition, data preprocessing has a significant effect on energy consumption [58]. Although data preprocessing was slightly less often discussed in identified literature studies (see Figure 3), a major cluster in the primary sources focuses on data science processes (9), including feature selection and sub-sampling techniques, such as removing redundant data points. Thus, future research should investigate the interrelation between resources needed for the preprocessing and saved energy due to fewer storage requirements [4], [57], [58]. However, as highlighted in Section III-B, preprocessing may distort representations, complicating training and testing. Hence, there is a need to balance a potential decrease in resource consumption with the reliability of the AI model.

The stage `Model` is the most often discussed stage in the literature reviews in 2023, where the task training and model evolution are also highly discussed (Figure 3). Verdecchia et al. [58] found that green AI primarily addresses model training to counter the trend of developing higher-performing state-of-the-art models due to increased data and powerful hardware [58], [62]. For instance, an experiment showed that model training within the continuous development of AI requires up to 98% of GPU utilization independent of the number of model layers or parameters [66]. Thus, research must expand Green AI to not only focus on making training more environmentally friendly but also identify methods to determine if retraining would lead to better performance, which would then justify an increased resource consumption [50], [65] or whether it is worth reducing data points or parameters during preprocessing. Furthermore, an identified cluster of primary sources provides insights into serverless computing and its resources and cost-effectiveness (8), which could lead to future research directions to provide serverless pipelines or training that profit from optimized scalability and resource consumption.

Tasks like data and model tracking and monitoring have gained popularity over the years (see Figure 3), contributing to a rise in storage capacities. If comprehensive tracking is

unavoidable or runtime monitoring data is used to ensure the AI model reliability, future research should pinpoint suitable hardware and data centers, as highlighted in a study showing that the carbon footprint of Deep Neural Networks (DNN) varied based on the hardware and data center utilized [42].

Key-Takeaways about potential future research directions (RQ_3):

- AI model reliability: explore runtime anomaly detection to use the information covered in data, model, system, and pipeline monitoring data
- Resource consumption: optimize triggers, reduce data points during preprocessing, expand Green AI, resource-efficient storage using hardware & data centers

V. Conclusion

The role of AI is increasing, posing research advances on the continuous development, integration, and deployment of these models. Since 2020, 33 systematic literature reviews have synthesized this research, but a comprehensive overview of these reviews is missing, making it challenging to select the best-fitting one. Different literature reviews use various search terms and selection criteria to answer similar research questions, limiting the primary sources found and hindering the synthesis of available work. Additionally, so far, the primary sources of the literature reviews have not been quantitatively assessed to provide insights into research trends and gaps.

Thus, this study 1) summarized 33 literature reviews regarding the continuous development, integration, and deployment of AI via a Multivocal Literature Review (MLR). Because the identified literature reviews' primary sources vary extensively, we further provide 2) a holistic overview of the three main topics by mapping SE practices and challenges to a continuous development pipeline for AI. Ultimately, we discuss 3) potential future research directions focusing on AI model reliability and resource consumption throughout the summarised pipeline. Our argument is based on identified clusters extracted from 1397 primary sources of literature reviews.

References

[1] Alves, I., Leite, L.A.F., Meirelles, P., Kon, F., Aguiar, C.S.R.: Practices for Managing Machine Learning Products: A Multivocal Literature Review. IEEE Transactions on Engineering Management pp. 1–31 (2023). https://doi.org/10.1109/TEM.2023.3287759, https://ieeexplore.ieee.org/document/10175022/

[2] Barrak, A., Petrillo, F., Jaafar, F.: Serverless on Machine Learning: A Systematic Mapping Study. IEEE Access **10**, 99337–99352 (2022). https://doi.org/10.1109/ACCESS.2022.3206366

[3] Cruzes, D.S., Dybå, T.: Recommended steps for thematic synthesis in software engineering. International Symposium on Empirical Software Engineering and Measurement pp. 275–284 (2011). https://doi.org/10.1109/ESEM.2011.36

[4] Dhabe, P., Mirani, P., Chugwani, R., Gandewar, S.: Data set reduction to improve computing efficiency and energy consumption in healthcare domain. Digital Literacy and Socio-Cultural Acceptance of ICT in Developing Countries pp. 53–64 (jul 2021). https://doi.org/10.1007/978-3-030-61089-0_4/COVER, https://link.springer.com/chapter/10.1007/978-3-030-61089-0_4

[5] Diaz-De-Arcaya, J., Torre-Bastida, A.I., Zárate, G., Miñón, R., Almeida, A.: A Joint Study of the Challenges, Opportunities, and Roadmap of MLOps and AIOps: A Systematic Survey. ACM Computing Surveys **56**(4), 84 (oct 2023). https://doi.org/10.1145/3625289, https://dl.acm.org/doi/10.1145/3625289

[6] Duda, S., Hofmann, P., Urbach, N., Völter, F., Zwickel, A.: The impact of resource allocation on the machine learning lifecycle. Business & Information Systems Engineering 2023 pp. 1–17 (11 2023). https://doi.org/10.1007/S12599-023-00842-7, https://link.springer.com/article/10.1007/s12599-023-00842-7

[7] Faubel, L., Schmid, K., Eichelberger, H.: MLOps Challenges in Industry 4.0. SN Computer Science 2023 4:6 **4**, 1–11 (10 2023). https://doi.org/10.1007/S42979-023-02282-2, https://link.springer.com/article/10.1007/s42979-023-02282-2

[8] Figalist, I., Elsner, C., Bosch, J., Olsson, H.H.: An End-to-End Framework for Productive Use of Machine Learning in Software Analytics and Business Intelligence Solutions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12562 LNCS**, 217–233 (2020). https://doi.org/10.1007/978-3-030-64148-1_14/FIGURES/3, https://link.springer.com/chapter/10.1007/978-3-030-64148-1_14

[9] Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., Brunner, D., Kumar, M., Moser, B.: AI System Engineering—Key Challenges and Lessons Learned. Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 56-83 **3**(1), 56–83 (dec 2020). https://doi.org/10.3390/MAKE3010004, https://www.mdpi.com/2504-4990/3/1/4/htmhttps://www.mdpi.com/2504-4990/3/1/4

[10] Fitzgerald, B., Stol, K.J.: Continuous software engineering: A roadmap and agenda. Journal of Systems and Software **123**, 176–189 (2017). https://doi.org/10.1016/j.jss.2015.06.063, https://www.sciencedirect.com/science/article/pii/S0164121215001430

[11] Garousi, V., Felderer, M., Mäntylä, M.V.: Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. Information and Software Technology **106**, 101–121 (feb 2019). https://doi.org/10.1016/J.INFSOF.2018.09.006

[12] Giray, G.: A software engineering perspective on engineering machine learning systems: State of the art and challenges. Journal of Systems and Software **180**, 111031 (oct 2021). https://doi.org/10.1016/J.JSS.2021.111031

[13] Granlund, T., Kopponen, A., Stirbu, V., Myllyaho, L., Mikkonen, T.: MLOps Challenges in Multi-Organization Setup: Experiences from Two Real-World Cases, http://arxiv.org/pdf/2103.08937v1

[14] Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure (mar 2022), https://arxiv.org/abs/2203.05794v1

[15] Guo, Q., Chen, S., Xie, X., Ma, L., Hu, Q., Liu, H., Liu, Y., Zhao, J., Li, X.: An empirical study towards characterizing deep learning development and deployment across different frameworks and platforms. Proceedings - 2019 34th IEEE/ACM International Conference on Automated Software Engineering, ASE 2019 pp. 810–822 (nov 2019). https://doi.org/10.1109/ASE.2019.00080

[16] Haakman, M., Cruz, L., Huijgens, H., van Deursen, A.: AI lifecycle models need to be revised: An exploratory study in Fintech. Empirical Software Engineering **26**, 1–29 (9 2021). https://doi.org/10.1007/S10664-021-09993-1/FIGURES/8, https://link.springer.com/article/10.1007/s10664-021-09993-1

[17] Heiland, L., Hauser, M., Bogner, J.: Design Patterns for AI-based Systems: A Multivocal Literature Review and Pattern Repository (mar 2023). https://doi.org/10.5281/zenodo.7568062, https://arxiv.org/abs/2303.13173v1

[18] Islam, M.J., Nguyen, G., Pan, R., Rajan, H.: A comprehensive study on deep learning bug characteristics. ESEC/FSE 2019 - Proceedings of the 2019 27th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering pp. 510–520 (aug 2019). https://doi.org/10.1145/3338906.3338955

[19] Jia, L., Zhong, H., Wang, X., Huang, L., Lu, X.: The symptoms, causes, and repairs of bugs inside a deep learning library. Journal of Systems and Software **177**, 110935 (jul 2021). https://doi.org/10.1016/J.JSS.2021.110935

[20] John, M.M., Holmström Olsson, H., Bosch, J.: Architecting AI Deployment: A Systematic Review of State-of-the-Art and State-of-Practice Literature. Lecture Notes in Business Information Processing **407**, 14–29 (2021). https://doi.org/10.1007/978-3-030-67292-8_2/COVER, https://link.springer.com/chapter/10.1007/978-3-030-67292-8_2

[21] John, M.M., Olsson, H.H., Bosch, J.: Towards MLOps: A Framework and Maturity Model. Proceedings - 2021 47th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2021 pp. 334–341 (sep 2021). https://doi.org/10.1109/SEAA53835.2021.00050

[22] Kitchenham, B., Charters, S., Budgen, D., Brereton, P., Turner, M., Linkman, S., Jørgensen, M., Mendes, E., Visaggio, G.: Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report (2007)

[23] Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O.P., Turner, M., Niazi, M., Linkman, S.: Systematic literature reviews in software engineering – A tertiary study. Information and Software Technology **52**(8), 792–805 (aug 2010). https://doi.org/10.1016/J.INFSOF.2010.03.006

[24] Kolltveit, A.B., Li, J.: Operationalizing Machine Learning Models - A Systematic Literature Review. Proceedings - Workshop on Software Engineering for Responsible AI, SE4RAI 2022 pp. 1–8 (2022). https://doi.org/10.1145/3526073.3527584

[25] Kreuzberger, D., Kuhl, N., Hirschl, S.: Machine Learning Operations (MLOps): Overview, Definition, and Architecture. IEEE Access **11**, 31866–31879 (2023). https://doi.org/10.1109/ACCESS.2023.3262138

[26] Kumara, C., Nucci, D., Jan Van Den, W., Andrew, D.: Requirements and Reference Architecture for MLOps:Insights from Industry pp. 0–9 (2022). https://doi.org/10.36227/techrxiv.21397413.v1, https://doi.org/10.36227/techrxiv.21397413.v1

[27] Lakha, B., Bhetwal, K., Eisty, N.U.: Analysis of Software Engineering Practices in General Software and Machine Learning Startups. Proceedings - 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications, SERA 2023 pp. 39–46 (2023). https://doi.org/10.1109/SERA57763.2023.10197836

[28] Lakshman, S.B., Eisty, N.U.: Software Engineering Approaches for TinyML based IoT Embedded Vision: A Systematic Literature Review. Proceedings - 4th International Workshop on Software Engineering Research and Practice for the IoT, SERP4IoT 2022 pp. 33–40 (2022). https://doi.org/10.1145/3528227.3528569

[29] Lima, A., Monteiro, L., Furtado, A.P.: MLOps: Practices, Maturity Models, Roles, Tools, and Challenges-A Systematic Literature Review . https://doi.org/10.5220/0010997300003179, https://orcid.org/0000-0002-5439-5314

[30] Liu, D., He, C., Peng, X., Lin, F., Zhang, C., Gong, S., Li, Z., Ou, J., Wu, Z.: MicroHECL: High-efficient root cause localization in large-scale microservice systems. Proceedings - International Conference on Software Engineering pp. 338–347 (may 2021). https://doi.org/10.1109/ICSE-SEIP52600.2021.00043

[31] Lo, S.K., Lu, Q., Wang, C., Paik, H.Y., Zhu, L.: A Systematic Literature Review on Federated Machine Learning. ACM Computing Surveys (CSUR) **54**(5) (may 2021). https://doi.org/10.1145/3450288, https://dl.acm.org/doi/10.1145/3450288

[32] Lorenzoni, G., Alencar, P., Nascimento, N., Cowan, D.: Machine Learning Model Development from a Software Engineering Perspective: A Systematic Literature Review (feb 2021). https://doi.org/10.48550/arxiv.2102.07574, https://arxiv.org/abs/2102.07574v1

[33] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., Jacquet, A.: Responsible AI Pattern Catalogue: A Multivocal Literature Review (sep 2022). https://doi.org/10.48550/arxiv.2209.04963, https://arxiv.org/abs/2209.04963v3

[34] Lwakatare, L.E., Crnkovic, I., Rånge, E., Bosch, J.: From a Data Science Driven Process to a Continuous Delivery Process for Machine Learning Systems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12562 LNCS**, 185–201 (2020). https://doi.org/10.1007/978-3-030-64148-1_12/TABLES/3, https://link.springer.com/chapter/10.1007/978-3-030-64148-1_12

[35] Lwakatare, L.E., Raj, A., Crnkovic, I., Bosch, J., Olsson, H.H.: Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. Information and Software Technology **127**, 106368 (nov 2020). https://doi.org/10.1016/J.INFSOF.2020.106368

[36] Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A.M., Wagner, S.: Software Engineering for AI-Based Systems: A Survey. ACM Transactions on Software Engineering and Methodology (TOSEM) **31**(2) (apr 2022). https://doi.org/10.1145/3487043, https://dl.acm.org/doi/10.1145/3487043

[37] Mboweni, T., Masombuka, T., Dongmo, C.: A Systematic Review of Machine Learning DevOps. International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022 (2022). https://doi.org/10.1109/ICECET55527.2022.9872968

[38] Moreschini, S., Recupito, G., Lenarduzzi, V., Palomba, F., Hästbacka, D., Taibi, D.: Toward End-to-End MLOps Tools Map: A Preliminary Study based on a Multivocal Literature Review (apr 2023), https://arxiv.org/abs/2304.03254v1

[39] Narayanan, N., Chen, Z., Fang, B., Li, G., Pattabiraman, K., Debardeleben, N.: Fault Injection for TensorFlow Applications. IEEE Transactions on Dependable and Secure Computing (2022). https://doi.org/10.1109/TDSC.2022.3175930

[40] Nascimento, E., Nguyen-Duc, A., Sundbø, I., Conte, T.: Software engineering for artificial intelligence and machine learning software: A systematic literature review (nov 2020). https://doi.org/10.48550/arxiv.2011.03751, https://arxiv.org/abs/2011.03751v1

[41] Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in Deploying Machine Learning: A Survey of Case Studies. ACM Computing Surveys **55**(6) (dec 2022). https://doi.org/10.1145/3533378, https://doi.org/10.1145/3533378

[42] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon Emissions and Large Neural Network Training (apr 2021), https://arxiv.org/abs/2104.10350v3

[43] Ralph, P., bin Ali, N., Baltes, S., Bianculli, D., Diaz, J., Dittrich, Y., Ernst, N., Felderer, M., Feldt, R., Filieri, A., de França, B.B.N., Furia, C.A., Gay, G., Gold, N., Graziotin, D., He, P., Hoda, R., Juristo, N., Kitchenham, B., Lenarduzzi, V., Martínez, J., Melegati, J., Mendez, D., Menzies, T., Molleri, J., Pfahl, D., Robbes, R., Russo, D., Saarimäki, N., Sarro, F., Taibi, D., Siegmund, J., Spinellis, D., Staron, M., Stol, K., Storey, M.A., Taibi, D., Tamburri, D., Torchiano, M., Treude, C., Turhan, B., Wang, X., Vegas, S.: Empirical Standards for Software Engineering Research (oct 2020), https://arxiv.org/abs/2010.03525v2

[44] Ralph, P., Baltes, S.: Paving the way for mature secondary research: the seven types of literature review. ESEC/FSE 2022 - Proceedings of the 30th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering pp. 1632–1636 (nov 2022). https://doi.org/10.1145/3540250.3560877, https://dl.acm.org/doi/10.1145/3540250.3560877

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

[46] Schlegel, M., Sattler, K.U.: Management of Machine Learning Lifecycle Artifacts. ACM SIGMOD Record **51**(4), 18–35 (jan 2023). https://doi.org/10.1145/3582302.3582306, https://dl.acm.org/doi/10.1145/3582302.3582306

[47] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green AI. Communications of the ACM **63**(12), 54–63 (nov 2020). https://doi.org/10.1145/3381831, https://dl.acm.org/doi/10.1145/3381831

[48] Serban, A., Visser, J.: Adapting Software Architectures to Machine Learning Challenges. Proceedings - 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2022 pp. 152–163 (2022). https://doi.org/10.1109/SANER53432.2022.00029

[49] Shivashankar, K., Martini, A.: Maintainability Challenges in ML: A Systematic Literature Review. Proceedings - 48th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2022 pp. 60–67 (2022). https://doi.org/10.1109/SEAA56994.2022.00018

[50] Siemers, W., Sallou, J., Cruz, L.: The Two Faces of AI in Green Mobile Computing: A Literature Review. 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) pp. 301–309 (sep 2023). https://doi.org/10.1109/SEAA60479.2023.00053, https://ieeexplore.ieee.org/document/10371497/

[51] Soldani, J., Brogi, A.: Anomaly Detection and Failure Root Cause Analysis in (Micro) Service-Based Cloud Applications: A Survey. ACM Computing Surveys (CSUR) **55**, 39 (feb 2022). https://doi.org/10.1145/3501297, https://dl.acm.org/doi/abs/10.1145/3501297

[52] Steidl, M., Felderer, M., Ramler, R.: The pipeline for the continuous development of artificial intelligence models—Current state of research and practice. Journal of Systems and Software **199**, 111615 (may 2023). https://doi.org/10.1016/J.JSS.2023.111615

[53] Steidl, M., Golendukhina, V., Felderer, M., Ramler, R.: Automation and Development Effort in Continuous AI Development: A Practitioners' Survey. 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) pp. 120–127 (sep 2023). https://doi.org/10.1109/SEAA60479.2023.00027, https://ieeexplore.ieee.org/document/10371413/

[54] Steidl, M., Ramler, R., Felderer, M.: Evaluation of literature reviews & primary sources thereof regarding the continuous development of AI (may 2024). https://doi.org/10.5281/zenodo.10173594, https://zenodo.org/doi/10.5281/zenodo.10173594

[55] Sun, X., Zhou, T., Li, G., Hu, J., Yang, H., Li, B.: An Empirical Study on Real Bugs for Machine Learning Programs. Proceedings - Asia-Pacific Software Engineering Conference, APSEC **2017-Decem**, 348–357 (mar 2018). https://doi.org/10.1109/APSEC.2017.41

[56] Testi, M., Ballabio, M., Frontoni, E., Iannello, G., Moccia, S., Soda, P., Vessio, G.: MLOps: A Taxonomy and a Methodology. IEEE Access **10**, 63606–63618 (2022). https://doi.org/10.1109/ACCESS.2022.3181730

[57] Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., Hotellier, E.: Data-Centric Green AI An Exploratory Empirical Study. Proceedings - 2022 International Conference on ICT for Sustainability, ICT4S 2022 (June), 35–45 (2022). https://doi.org/10.1109/ICT4S55073.2022.00015

[58] Verdecchia, R., Sallou, J., Cruz, L.: A systematic review of Green AI. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **13**(4), e1507 (jul 2023). https://doi.org/10.1002/WIDM.1507, https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1507https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1507https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1507

[59] Warnett, S.J., Zdun, U.: Architectural Design Decisions for Machine Learning Deployment. Proceedings - IEEE 19th International Conference on Software Architecture, ICSA 2022 pp. 90–100 (2022). https://doi.org/10.1109/ICSA53651.2022.00017

[60] Warnett, S.J., Zdun, U.: Architectural Design Decisions for the Machine Learning Workflow. Computer **55**(3), 40–51 (mar 2022). https://doi.org/10.1109/MC.2021.3134800

[61] Xie, Y., Cruz, L., Heck, P., Rellermeyer, J.S.: Systematic Mapping Study on the Machine Learning Lifecycle, http://arxiv.org/pdf/2103.10248v1

[62] Xu, Y., Martínez-Fernández, S., Martinez, M., Franch, X.: Energy Efficiency of Training Neural Network Architectures: An Empirical Study. Proceedings of the Annual Hawaii International Conference on System Sciences **2023-January**, 781–790 (feb 2023), https://arxiv.org/abs/2302.00967v1

[63] Zhang, C., Peng, X., Sha, C., Zhang, K., Fu, Z., Wu, X., Lin, Q., Zhang, D.: DeepTraLog: Trace-Log Combined Microservice Anomaly Detection through Graph-based Deep Learning. Proceedings - International Conference on Software Engineering **2022-May**, 623–634 (2022). https://doi.org/10.1145/3510003.3510180, https://dl.acm.org/doi/10.1145/3510003.3510180

[64] Zhang, R., Xiao, W., Zhang, H., Liu, Y., Lin, H., Yang, M.: An empirical study on program failures of deep learning jobs. Proceedings - International Conference on Software Engineering **12**, 1159–1170 (jun 2020). https://doi.org/10.1145/3377811.3380362, https://doi.org/10.1145/3377811.3380362

[65] Zheng, J., Li, K., Mhaisen, N., Ni, W., Tovar, E., Guizani, M.: Exploring Deep-Reinforcement-Learning-Assisted Federated Learning for Online Resource Allocation in Privacy-Preserving EdgeIoT. IEEE Internet of Things Journal **9**(21), 21099–21110 (nov 2022). https://doi.org/10.1109/JIOT.2022.3176739

[66] Zhou, Y., Yu, Y., Ding, B.: Towards MLOps: A Case Study of ML Pipeline Platform. Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2020 pp. 494–500 (oct 2020). https://doi.org/10.1109/ICAICE51518.2020.00102