Assessment of the Relevance of Best Practices in the Development of Medical R&D Projects Based on Machine Learning

Jan Cychnerski
Department of Computer Architecture, Gdańsk University of Technology
Gdańsk, Poland *jan.cychnerski@pg.edu.pl*

Tomasz Dziubich
Department of Computer Architecture, Gdańsk University of Technology
Gdańsk, Poland *tomasz.dziubich@pg.edu.pl*

Abstract

Machine learning has emerged as a fundamental tool for numerous endeavors within health informatics, bioinformatics, and medicine. However, novices among biomedical researchers and IT developers frequently lack the requisite experience to effectively execute a machine learning project, thereby increasing the likelihood of adopting erroneous practices that may result in common pitfalls or overly optimistic predictions. The paper presents an assessment of the significance of best practices in the implementation of R&D projects supporting the medical diagnostic process. Based on the literature and authors' experiences, 27 good practices influencing three fundamental stages of project implementation were identified. The evaluation was based on the Analytic Hierarchy Process, which relies on subjective assessments from experts, whose credibility is expressed through the consensus of assessment. Initially focusing on DevOps methodology, research integration, interdisciplinary information sharing were prioritized over automation. Furthermore, annotation tools and data / model quality control were identified as of significant importance.

Keywords: medical machine learning projects, guidelines, best practices, DevOps, MLOps

1. Introduction

Research and Development (R&D) projects represent a distinct category of initiatives, significantly divergent from commercial projects. Their primary aim is the pursuit of innovative solutions, highly consequential for advancing progress within the within the specific research area, thereby fostering economic growth [2]. Consequently, both their motivations, objectives, and execution methods markedly differ from generally construed commercial projects, which primarily focus on application, are market-driven, ensuring revenue for the executing companies, and providing various benefits for potential clients [12]. Unlike commercial projects, R&D projects are divided into two parts: research-oriented, aimed at acquiring new knowledge and creating new solutions, and development-oriented, focusing on building prototypes and laying the groundwork for the potential introduction of a new product to the market. The first stage is characterized by a much higher risk of ultimate failure [24]. This paper focuses on a specific type of R&D IT projects related to supporting medical diagnostics using machine learning methods, defined in this work as the Medical Machine Learning (MedML) project. MedML projects require a complex executive structure due to the interdisciplinary nature of the work and the need for extended data collection, high costs of specialists, interdisciplinary collaboration, research-oriented model building, and client involvement. Ensuring proper organization,

Work partially funded by the *Cloud Artificial Intelligence Service Engineering (CAISE)* project No. KPOD.05.10-IW.10-0005/24, KPO, European Programme IPCEI-CIS

adequate financial resources, and effective communication among teams is crucial for success. During the practical implementation and deployment of projects, it is crucial to perform tasks appropriately as this significantly impacts the final outcome of the project. This manner of execution is commonly defined in methodologies through a set of general and detailed recommendations and rules of conduct referred to as *best practices* or guidelines [18].

The contributions of this paper are as follows: (1) a comprehensive list of best practices for conducting MedML projects, specifically those that apply machine learning to support medical diagnostics is presented; (2) the use of the Analytic Hierarchy Process (AHP) for the relative evaluation of these best practices is introduced. This approach distinguishes this research from others in this field, which typically rely only on literature review or surveys; (3) the significance of each best practice in the context of conducting MedML projects is assessed. To the best of the authors' knowledge, there are no publications that parallel this study.

The paper is structured as follows. Section 2 include a detailed description of the selected guidelines and best practices, as well as their sources. Subsequently, in Section 3, method of experimental verification of these guidelines is proposed, and in Section 4 results of conducted verification, in the form of best practices' importance measurement, is presented and discussed. At the end, final remarks are presented in Section 5.

2. Identification of best practices in MedML

Best practices heavily depend on the field, and their initial origins stem from the experience of practitioners and experts. The areas that have been developing dynamically in recent years include, among others, machine learning and medical diagnostics. Hence the need for adaptation or defining new best practices in this area. Adopting best practices in medical projects involving machine learning require the accuracy, reliability, and ethical use of predictive models, which are critical for patient safety and effective treatment. These features distinguish them apart from other interdisciplinary projects. This includes maintaining high-quality data standards, ensuring transparency in model development, and continuously monitoring and validating model performance in real-world settings. Treating both fields separately, numerous publications on best practices can be found in the literature, e.g. [9, 21]. It is worth noting that the most popular methodology used in machine learning projects is MLOps, where authors adopt practices from agile methodologies in it. The accepted practice for defining practices is discussion based on experience.

Drawing inspiration from discussions surrounding optimal methodologies for machine learning across various domains such as manufacturing [7, 10], physical systems [14], and health informatics [6, 20, 19], we have opted to introduce best practices tailored for conducting projects centered on the analysis of medical data for ML projects. Our aim is to circumvent prevalent errors and challenges observed in numerous studies within this field. While certain studies in the computer science and medical literature have already introduced guidelines for applying machine learning techniques to medical image analysis [22, 4, 13], these resources primarily focus on research projects (not development phase). Nevertheless, numerous fields necessitate a more methodical approach to prevent harm to the environment and human health, as well as to minimize risks throughout the process. Hence, our research method was based on the comparative AHP method, unlike other studies which rely on literature reviews or surveys. In the first step, a set of proposed practices was selected based on a literature review. Next, experts participating in the same 4 MedML projects were selected for the comparative evaluation. The knowledge gained from them formed the basis for preparing this paper.

Based on a literature review conducted utilizing the IEEE, Scopus and ACM databases, focusing on the keywords "best practices" and "MLOps", 59 papers (research papers and surveys) were identified (after merging and removing duplicate results). It is noteworthy that these publications were produced after the year 2021 (without date filtering), unequivocally indicating the

freshness of the subject matter. Zinkevich assembled a set of optimal techniques for machine learning that are employed within Google [25]. Serban et al. identified 29 optimal practices of software engineering for machine learning through a systematic literature review prior to conducting a survey among practitioners to explore the adoption and effects of these practices [17]. They divided good practices into 6 categories: data gathering, training, coding, deployment, collaboration and policy compliance. John et al. conducted a similar literature review on the software engineering lifecycle of machine learning models, highlighting the challenges and best practices at each of the 7 stages of the lifecycle [8]. SE4ML group discovered 46 optimal practices for engineering reliable machine learning applications grouped into 6 categories [16]. In total, 50 practices were identified, partially overlapping in content. No guidelines pertaining to healthcare and medical diagnosis could be located. Thus, in this study, the number of categories were limited from the 6-7 proposed in the literature to 3. The practices presented therein can be categorized into one of three main areas related to application development process: (1) data gathering, (2) end-to-end pipeline for model development, (3) practices associated with the classical DevOps process facilitating prototype development.

Drawing from existing methodologies, publications, and experts' professional experience, 27 practices were selected for further analysis, deemed most relevant from the perspective of MedML projects. In following sections these rules are briefly described, followed by an assessment of their significance and utilization in projects using the AHP method.

1. **Data gathering.** A set of best practices regarding the entire data collection process.

   D1 Ensuring Interdisciplinary Work to Understand Project Goals

   D2 Formulating Conclusions Based on the Results of Each Project Iteration

   D3 Defining Relevant Patient Features for Selection from HIS/PACS hospital systems.

   D4 Determining the Types of Collected Data including their modalities, measurement devices, required metadata, types of annotations

   D5 Defining Quality and Quantity Requirements

   D6 Ensuring Good Organization and Planning Data Access Procedures to raw data in HIS/PACS hospital systems.

   D7 Utilizing Effective Tools for Patient Selection and Raw Data Export from HIS/PACS hospital systems, considering technical, legal and privacy requirements.

   D8 Establishing Proper Data Annotation Procedures

   D9 Utilizing Effective Tools to Support Annotation

   D10 Conducting Ongoing Quality Control of Data

2. **Machine Learning Model Development.** A set of practices regarding the process of machine learning models construction.

   M1 Ensuring Interdisciplinary Work to Understand Project Goals

   M2 Formulating Conclusions Based on the Results of Each Project Iteration

   M3 Defining the Target Functionality of Machine Learning Models

   M4 Defining Quality Metrics and Evaluation Procedures Adequate to Project Goals

   M5 Utilizing Advanced Tools Supporting Data Processing Methods

   M6 Correct Execution of Raw Data and Annotation Unification

   M7 Proper Data Splitting into Training, Validation, and Test Sets

   M8 Use of Existing Machine Learning Architectures and Models

M9 Adjustment and Definition of New Architectures

M10 Extensive Hyperparameter Optimization

M11 Methodologically Correct Measurement of Defined Quality Metrics

M12 Optimal Selection of Final ML Models

3. **Development of a prototype IT system and supporting tools.** A set of practices regarding process of implementation and deployment of IT system, applications, and tools.

P1 Ensuring Project Progress Monitoring

P2 Automation of Software Development Process

P3 Information Sharing during Development Process

P4 Adherence to Appropriate Work Culture

P5 Incorporation of R&D Requirements

Each single best practice has been assigned an identifier (from $P1$ to $M12$), consistent with a detailed description which can be found in [3]. The first two practices (D1-D2, M1-M2) relate to activities associated with the overall project implementation, interdisciplinary collaboration, and exchange of experiences among all processes in the project development methodology. The practices (D3-D10) summarize practices defined in a manner consistent with principles applied in the literature (particularly in the *CRISP-MED-DM* methodology [11] and *MLOps* [23]). The (M3-M12) apply to development, training and testing of machine learning models. The (P1-P4) recommended practices and success factors in project implementation formulated in the scientific literature on the DevOps methodology [1, 26], which is one of the recommended methodologies for implementing the prototype IT system in MedML projects. As DevOps is only a part of MedML projects, and is well researched in the literature, the most detailed DevOps recommendations layer was omitted, and only general categories of recommendations were taken for analysis. On the other hand, due to the R&D nature of the considered projects, point *P5* has been added to this list by the authors, which addresses R&D aspects related to building a prototype IT system to support medical diagnostics.

3. Evaluation Methodology of Best Practices

In order to assess the significance of individual best practices in the development of a MedML project, the AHP method [15] was employed. AHP is a multicriteria method of hierarchical analysis of decision problems, allowing for the decomposition of a complex decision problem into sub-problems, the adoption of criteria for their evaluation, and the forming of rankings of the considered alternatives based on the level of fulfillment of these criteria. In this work, the definitions of concepts, notations, and the implementation of the AHP available in [5] were utilized, where the details of this method, as well as mathematical formulas and corresponding algorithms, are described.

To determine the level of significance of best practices, a three-level hierarchy was composed based on them, as depicted in Fig. 1. This hierarchy was constructed in accordance with the principles of DevOps and MLOps methodologies, particularly by distinguishing three main categories of best practices related to the three primary processes of MedML project development: IT system construction, data collection, and ML model generation. At Level 1 of the hierarchy lies the overall assessment of the level of employment of best practices throughout the project; at Level 2, the evaluation of best practices for the three main processes of MedML project implementation; at Level 3, individual best practices pertaining to the execution of these processes.

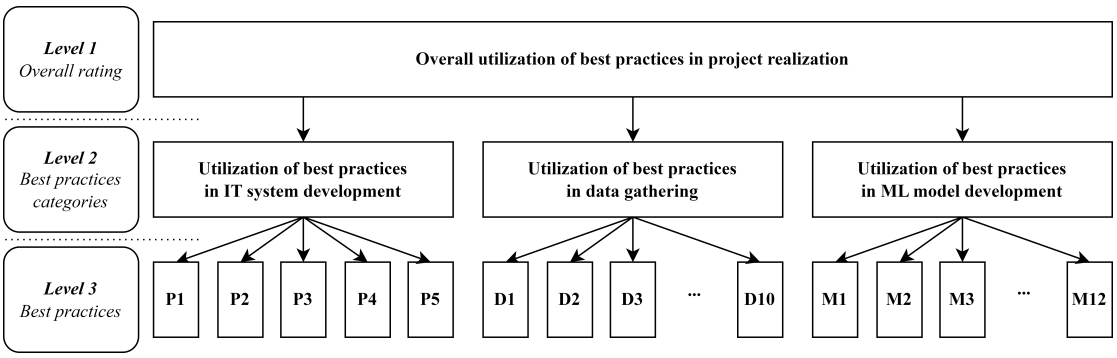

Fig. 1. Hierarchy of assessing the significance of best practices in MedML projects.

The created hierarchy was evaluated according to the AHP method. Initially, the relative relevance of individual best practices within 3 categories was assessed pairwise, followed by the evaluation of the relative importance of entire categories. These assessments were made using the original 9-point linear scale proposed by [15]. Subsequently, individual preference matrices, the Consistency Index (CI), and the Consistency Ratio (CR) were determined. The inclusion criterion for individual assessments was a consistency ratio $CR \leq 10\%$. Based on the individual preference matrices, a consolidated decision matrix was computed along with its consistency CR and consensus CS (where a value below 50% indicates very low consensus, 50-65% low consensus, 65-75% moderate consensus, 75-85% high consensus, and above 85% very high consensus, meaning a high agreement of individual assessments in the overall evaluation of the significance of best practices). In the final stage, the significance of each best practice was calculated—both the global priority PG and the local priority PL, for each MedML project part (PL_P for prototype IT system, PL_D for data gathering, PL_M for ML model construction). Based on the sorted decreasing global priorities, a ranking of the most significant best practices in MedML projects was created, where for each practice, its global rank RG was determined, representing its position in the ranking (rank 1 denotes the most significant best practice). The sum of all global priorities PG for all practices equals 100%.

4. Results and discussion

The utilization of best practices related to the realization of individual processes and stages of the MedML project has a significant impact on the ability to achieve the intended effects of these processes, as well as on the attainment of the overall objectives of the entire project. Using the AHP method described in the previous section, the level of significance of all defined best practices was determined according to the adopted hierarchy of evaluation. This significance is expressed in terms of local priorities (PL) and global priorities (PG), as well as ranks (RG) for each analyzed best practice. Priorities of three best practices categories (level 2 in AHP hierarchy) are shown in Fig. 2. All obtained priorities and ranks of every best practice (level 3 in AHP hierarchy) are presented in Tab. 1. The overall consensus of the evaluators reached a high level ($CS = 82\%$).

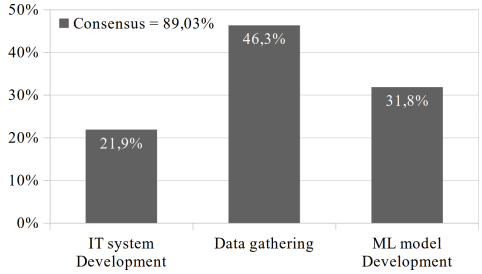

Fig. 2. The significance of best practice categories in relation to the whole MedML project.

Table 1. Significance of best practices in MedML projects. Consecutively, calculated values of local priorities PL (prototype development PL_P, data gathering PL_D, ML model construction PL_M), consolidated global priorities PG, and global ranks RG are provided. A higher priority (lower rank) signifies greater importance of a given good practice. The total consensus of assessors amounted to $CS = 82\%$ (high). Practices are sorted by their global rank RG.

Prototype development				Data gathering				ML models			
Practice	PL_P	PG	RG	Prac.	PL_D	PG	RG	Prac.	PL_M	PG	RG
P3	26,4%	5,8%	3	D1	17,5%	8,1%	1	M1	15,6%	5,0%	6
P5	23,2%	5,1%	5	D2	13,2%	6,1%	2	M2	12,9%	4,1%	10
P4	18,7%	4,1%	11	D9	11,4%	5,3%	4	M4	11,1%	3,5%	14
P1	17,9%	3,9%	13	D10	10,2%	4,7%	7	M10	9,0%	2,9%	18
P2	13,9%	3,0%	17	D8	10,0%	4,6%	8	M3	8,1%	2,6%	19
				D5	9,8%	4,6%	9	M11	8,1%	2,6%	20
				D4	8,8%	4,1%	12	M5	7,2%	2,3%	22
				D7	7,1%	3,3%	15	M12	6,3%	2,0%	23
				D6	6,6%	3,1%	16	M7	6,2%	2,0%	24
				D3	5,4%	2,5%	21	M6	5,6%	1,8%	25
								M9	5,3%	1,7%	26
								M8	4,6%	1,5%	27
$CS = 65,8\%$				$CS = 76,0\%$				$CS = 80,1\%$			

Firstly, the assessment focused on best practices concerning the process of developing a prototype system and supporting tools, corresponding to practices in the *DevOps* methodology. The obtained local priorities PL_P ranged from 13.9% to 26.4%, with the two most important practices being P3 (*information sharing in the development process*) and P5 (*incorporation of research and development requirements*). This assessment deviates from the typical evaluation of the *DevOps* methodology in deployment-oriented IT projects [26], where *work culture* and *automation* are identified as the most crucial factors. The obtained evaluation confirms the assumptions of MedML projects, where research and development aspects, particularly those related to interdisciplinary collaboration and exchange of experiences, are considered more significant than project implementation aspects associated with automation, deployment, and system maintenance.

In the accomplishment of the data gathering process, the most significant best practices were also identified as interdisciplinary work, data understanding (D1, $PL_D = 17.5\%$), and acquiring new knowledge / formulating conclusions based on it for further stages of project implementation (D2, $PL_D = 13.2\%$). This indicates the importance of active information exchange and collaborative work among teams carrying out individual project processes (data gathering, building ML models, prototype development). This aligns with the recommendations of frequent iterations and interweaving of all concurrently executed MedML project implementation processes described in [3]. The next most important best practices were identified as the use of annotation support tools (D9, $PL_D = 11.4\%$) and data quality control (D10, $PL_D = 10.2\%$). The cumulative priority of all data gathering tools (D9+D7) amounted to 18.5%, indicating the very high significance of these tools.

Among the best practices of the machine learning model building, the most significant practices were also those related to interdisciplinary collaboration (M1, $PL_M = 15.6\%$) and acquiring new knowledge (M2, $PL_M = 12.9\%$). Among the next most important practices, adequate determination of metrics and quality assessment procedures for ML models (M4, $PL_M = 11.1\%$) was identified as the most significant, although the differences in the assessment of significance between individual practices were minor (4.6%-11.1%). Among the practices, the use of supporting tools (M5) stood out, with a priority of 7.2% in the process of building ML models, i.e., as 7 out of 12 recommended best practices.

5. Conclusions

We conducted a study to investigate the development of software solutions incorporating machine learning components in medical diagnosis support system. To achieve this, we conducted a review of both academic and grey literature and constructed a compendium consisting of 27 software engineering best practices for MedML projects, categorized into 3 distinct groups. Subsequently, we determined the level of significance of selected practices using AHP method. We posit that adherence to our 27 delineated guidelines can substantially enhance the efficacy of any machine learning practitioner in conducting successful projects within the domain of medicine and its associated fields. Initially focusing on *DevOps* methodologies, we prioritized research integration and information sharing over automation. In machine learning model development, metrics determination and tool usage highlighted. Important guidelines regarded data gathering, especially encouraging using good annotation tools and performing data quality control. In all guideline categories, interdisciplinary work and knowledge acquisition were paramount. This emphasizes the importance of collaboration, research integration, and meticulous data handling in MedML projects. At the team or organizational level, these findings can be utilized to critically evaluate the current utilization of practices and prioritize their adoption based on desired outcomes. For example, a team with a strong emphasis on agility but low adoption of associated practices may develop strategies to enhance the adoption of these practices.

As a subsequent research phase, the influence of implementing the suggested guidelines on the ultimate success of the endeavors will be investigated. The examination will focus on evaluating how their adoption and progression status of the primary MedML processes — namely, prototype IT system development, data gathering, and machine learning model construction — affects the fulfillment of the overarching objectives within the MedML projects.

References

[1] Akbar, M. A., Mahmood, S., Shafiq, M., Alsanad, A., Alsanad, A. A. A., and Gumaei, A.: Identification and prioritization of DevOps success factors using fuzzy-AHP approach. In: *Soft Computing* 27.4 (2023), pp. 1907–1931.

[2] *Commission Regulation (EU) No 1217/2010 of 14 December 2010 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to certain categories of research and development (CELEX: 32010R1217).*

[3] Cychnerski, J.: "The methodology for medical diagnosis support systems implementation using data collection and machine learning tools". PhD thesis. Gdańsk University of Technology, 2023, p. 180.

[4] Cychnerski, J. and Dziubich, T.: Process of Medical Dataset Construction for Machine Learning - Multifield Study and Guidelines. In: *New Trends in Database and Information Systems*. Cham: Springer International Publishing, 2021, pp. 217–229.

[5] Goepel, K.: Implementation of an Online software tool for the Analytic Hierarchy Process (AHP-OS). In: *International Journal of the Analytic Hierarchy Process* 10.3 (2018).

[6] Granlund, T., Stirbu, V., and Mikkonen, T.: Towards Regulatory-Compliant MLOps: Oravizio's Journey from a Machine Learning Experiment to a Deployed Certified Medical Product. In: *SN Computer Science* 2.5 (June 2021), p. 342.

[7] Heymann, H., Kies, A. D., Frye, M., Schmitt, R. H., and Boza, A.: Guideline for Deployment of Machine Learning Models for Predictive Quality in Production. In: *Procedia CIRP* 107 (2022). Leading manufacturing systems transformation – Proceedings of the 55th CIRP Conference on Manufacturing Systems 2022, pp. 815–820.

[8] John, M. M., Holmström Olsson, H., and Bosch, J.: Architecting AI deployment: A systematic review of state-of-the-art and state-of-practice literature. In: *Software Business: 11th International Conference, ICSOB 2020, Karlskrona, Sweden, November 16–18, 2020, Proceedings 11*. Springer. 2021, pp. 14–29.

[9] Karamitsos, I., Albarhami, S., and Apostolopoulos, C.: Applying DevOps Practices of Continuous Automation for Machine Learning. In: *Information* 11.7 (2020).

[10] Lima, A., Monteiro, L., and Furtado, A.: MLOps: Practices, Maturity Models, Roles, Tools, and Challenges – A Systematic Literature Review. In: *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 2: ICEIS,* INSTICC. SciTePress, 2022, pp. 308–320.

[11] Niakšu, O.: CRISP Data Mining Methodology Extension for Medical Domain. In: *Baltic J. Modern Computing* 3.2 (2015), pp. 92–109.

[12] Przybylek, A.: A business-oriented approach to requirements elicitation. In: *2014 9th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)*. 2014, pp. 1–12.

[13] Rana, M. and Bhushan, M.: Machine learning and deep learning approach for medical image analysis: diagnosis to detection. en. In: *Multimed Tools Appl* (Dec. 2022), p. 1.

[14] Ruf, P., Madan, M., Reich, C., and Ould-Abdeslam, D.: Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools. In: *Applied Sciences* 11.19 (2021).

[15] Saaty, T. L.: "What is the Analytic Hierarchy Process?" In: *Mathematical Models for Decision Support*. 1988, pp. 109–121.

[16] *SE-ML Engineering best practices for Machine Learning*. `https://se-ml.github.io/practices/`. Accessed: 2024-02-13.

[17] Serban, A., Blom, K. van der, Hoos, H. H., and Visser, J.: Adoption and Effects of Software Engineering Best Practices in Machine Learning. In: *CoRR* abs/2007.14130 (2020). arXiv: `2007.14130`.

[18] Sommerville, I.: Software engineering (10th edition). 2016.

[19] Stirbu, V., Granlund, T., and Mikkonen, T.: Continuous design control for machine learning in certified medical systems. In: *Software Quality Journal* 31.2 (June 2023), p. 307.

[20] Testi, M., Ballabio, M., Frontoni, E., Iannello, G., Moccia, S., Soda, P., and Vessio, G.: MLOps: A Taxonomy and a Methodology. In: *IEEE Access* 10 (2022), pp. 63606–63618.

[21] Ueda, D. et al.: Fairness of artificial intelligence in healthcare: review and recommendations. In: *Japanese Journal of Radiology* 42.1 (Jan. 2024), pp. 3–15.

[22] Varoquaux, G. and Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. In: *npj Digital Medicine* 5.1 (Apr. 2022).

[23] Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., and Plöd, M.: *Three Levels of ML Software*.

[24] Wang, J., Lin, W., and Huang, Y. H.: A performance-oriented risk management framework for innovative R&D projects. In: *Technovation* 30.11-12 (Nov. 2010), pp. 601–611.

[25] Zinkevich, M.: Rules of machine learning: Best practices for ML engineering. In: *URL: https://developers. google. com/machine-learning/guides/rules-of-ml* (2017).

[26] Zohaib, M.: Towards Sustainable DevOps: A Decision Making Framework. In: (Mar. 2023). arXiv: `2303.11121`.