**TU**Delft

Delft University of Technology

# Systematic Mapping Study on the Machine Learning Lifecycle

Xie, Yuanhao; Cruz, Luís; Heck, Petra; Rellermeyer, Jan S.

**Citation (APA)**
Xie, Y., Cruz, L., Heck, P., & Rellermeyer, J. S. (2021). Systematic Mapping Study on the Machine Learning Lifecycle. In L. O'Conner (Ed.), *Proceedings - 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI, WAIN 2021* (pp. 70-73). Article 9474380 (Proceedings - 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI, WAIN 2021). IEEE. https://doi.org/10.1109/WAIN52551.2021.00017

**Important note**
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

# Systematic Mapping Study on the Machine Learning Lifecycle

Yuanhao Xie
*AI for Fintech Research, ING*
Amsterdam, Netherlands
yuanhao.xie@ing.com

Luís Cruz
*Delft University of Technology*
Delft, Netherlands
l.cruz@tudelft.nl

Petra Heck
*Fontys ICT*
Eindhoven, Netherlands
p.heck@fontys.nl

Jan S. Rellermeyer
*Delft University of Technology*
Delft, Netherlands
j.s.rellermeyer@tudelft.nl

*Abstract*—The development of artificial intelligence (AI) has made various industries eager to explore the benefits of AI. There is an increasing amount of research surrounding AI, most of which is centred on the development of new AI algorithms and techniques. However, the advent of AI is bringing an increasing set of practical problems related to AI model lifecycle management that need to be investigated. We address this gap by conducting a systematic mapping study on the lifecycle of AI model. Through quantitative research, we provide an overview of the field, identify research opportunities, and provide suggestions for future research. Our study yields 405 publications published from 2005 to 2020, mapped in 5 different main research topics, and 31 sub-topics. We observe that only a minority of publications focus on data management and model production problems, and that more studies should address the AI lifecycle from a holistic perspective.

*Index Terms*—AI lifecycle management, Artificial Intelligence, Software Engineering, Systematic mapping study

## I. INTRODUCTION

With the development of AI, various industries are eager to discover and reap the benefits of AI, and progressively more AI-related research is being carried out. However, the purpose of most of the research is to develop new AI algorithms or techniques to solve an issue in a given field. There are growing numbers of problems related to AI model life cycle management, such as version control of data and models, the difficulties of model deployment, transparency, model reproducibility, fairness, and so on [1, 2, 3]. Despite this, there is a limited number of studies on AI model life cycle management, and there is no comprehensive study on the life cycle management of AI models. To fill this gap, we have conducted a systematic mapping study of the life cycle management of AI. By summarising the current situation in this field through quantitative and qualitative research we have developed an overview, identified research gaps and provided suggestions for future research directions.

We collect 405 papers and map them into 5 main categories with 31 sub categories. We also analyse the papers in terms of venue, year of publication, type of research, and Moreover, we discuss problems that are being overlooked by the AI research community.

The research questions are defined as: RQ1: In what years, from which countries, affiliations, venue were these research papers published? RQ2: What research approaches do these studies apply? RQ3: Which subtopics of AI model life cycle management have already been investigated?

*Replication Package*

All data (includes reference list of the 405 publications), instructions, necessary to replicate this study are available in the online appendix:

https://luiscruz.github.io/ai-lifecycle-mapping-study/

## II. RELATED WORK

In this section, we describe previous mapping studies that include the lifecycle of AI model applications in their scope. Previous work conducted a systematic literature review to understand the use of software engineering practices in the development of AI/ML systems [4]. The study maps 57 ML studies into 11 different software engineering topics. [5]. It shows that little research has focused on deployment, maintenance, quality, and management. Our work differentiates by mapping the fields not only in terms of software engineering topics, but mostly by focusing on the different stages of the lifecycle of AI artefacts.

Vakkuri and Abrahamsson [6] conceptualised and classified the ethics of AI, and found 37 keywords related to the ethics of AI in 83 papers. Sherin et al. [7] conducted a system mapping study on the testing of machine learning systems. After researching 37 selected articles, they identified the trends in the field of machine learning testing and discovered research opportunities, such as testing machine learning programs using reinforcement learning. Spaargaren [8] conducted a systematic mapping study in the field of machine learning (ML) and software engineering (SE) in the context of financial technology (FinTech). None of the AI engineering related systematic mapping studies above are about AI model life cycle management, and only fragmented work about a specific sub topic of AI model lifecycle management.

Other systematic mapping studies in the field of AI focus on the intersection between AI and other specific fields – for instance, network security. Wiafe et al. [9].

## III. METHODOLOGY

Our methodology, as described in Fig. 1, is based on the guidelines proposed by Petersen et al. to conduct systematic mapping studies in software engineering [10, 11]. We pinpoint each step below.
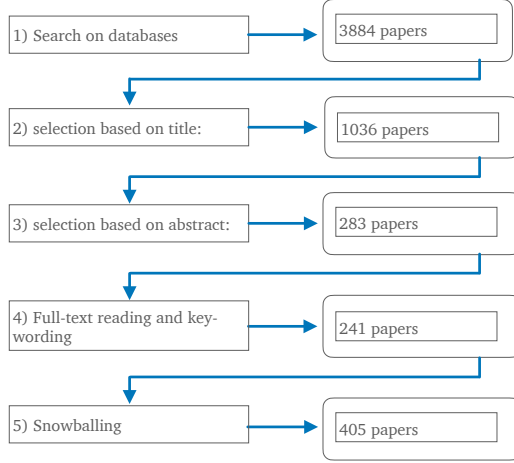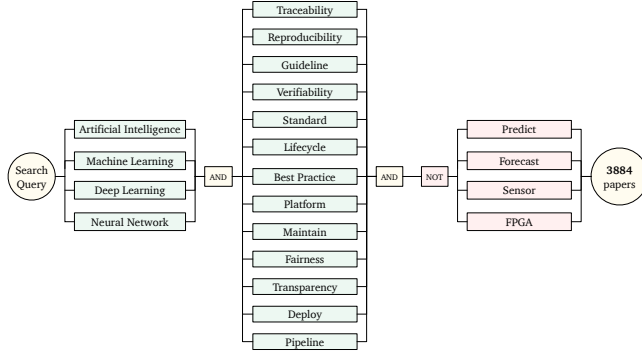
Fig. 1. Methodology used to collect publications.



Fig. 2. Diagram of the query used to retrieve papers from the DBLP and Scopus databases.

*1) Search on Databases:* The search query we use to retrieve publications from the databases is illustrated in Fig. 2. We use the Population, Intervention, Comparison, and Outcome (PICO) strategy to derive the search terms.

AI-related keywords includes **artificial intelligence (AI), machine learning (ML), deep learning (DL), neural network.**

We define keywords related to the lifecycle of AI based on previous work that pinpoints its different steps [12, 13, 14, 15]. A preliminary analysis yielded several keywords that cover three main aspects: model-oriented works (e.g., *Pipeline*), code meets data (e.g., data *Traceability*, *Reproducibility*), and DevOps works (e.g., *Deploy*). After analyzing the papers from the initial result set, we have iteratively improved the query by including, for example, *Standard*, *Best Practice*, and *Platform*.

*2) Apply selection criteria based on title:* This step was undertaken by the first two authors in parallel. The papers were only excluded when both researchers agreed in that decision. In the end, we select 1036 papers, The papers were excluded according to the following criteria:

- Studies not presented in English.
- Duplicates.

| | | Researcher 1 | | |
|---|---|---|---|---|
| | | Include | Uncertain | Exclude |
| Researcher 2 | Include | 195 | 67 | 71 |
| | Uncertain | 129 | 85 | 212 |
| | Exclude | 119 | 158 | **2848** |

## TABLE II
### AGREEMENT TABLE OF THE SELECTION BASED ON ABSTRACT.

| | | Researcher 1 | | |
|---|---|---|---|---|
| | | Include | Uncertain | Exclude |
| Researcher 2 | Include | 253 | 6 | 38 |
| | Uncertain | 31 | 21 | **6** |
| | Exclude | 57 | **12** | **610** |

- Studies not accessible in full-text.
- Books and grey literature.
- Studies that discussed applications of AI.
- Studies about the design of new AI algorithms unrelated to AI model lifecycle management.

*3) Apply selection criteria based on abstract.:* We exclude all the papers in which one of the researchers proposed to exclude and the other agreed or was uncertain – yielding 628 papers, as stated in the **red cells highlighted in bold**.

*4) Full-text reading and Keywording.:* In this step, the first author goes through the contents of the 283 papers to remove articles that were out of scope. As a result, 241 papers are remaining.

*5) Snowballing:* The first step is to browse the reference list, read the abstracts of the articles in the list, exclude papers that do not meet the basic criteria (such as language, publication year, and publication type), and then delete duplicate articles. After this step, 164 articles were obtained.

## IV. RESULTS

In this section, we present all the results collected in this mapping study. We map 405 papers according to year, countries, affiliations, venue and lifecycle topic. We also provide a fine-grained mapping in each subtopic of the lifecycle.

*A. RQ1: In what years, from which countries, affiliations, venue were these research papers published?*

*1) Publication by country:* The country is determined according to the institution where first author is located. The top five countries are the United States(49.4%), Germany(8.4%), the United Kingdom(6.2%), China(5.9%), and Canada(4.2%), while other countries accounted for less than 3%. Publications from the United States account for about half of the total. This trend is not uncommon [16], but it emphasises the lack of diversity and the importance of having more countries shaping the future of AI.

*2) Publication by year:* The number of studies on the lifecycle management of AI models has increased over time. Especially from 2016 to 2019, the number of publications each year is double that of the previous year. The publications in this mapping study were collected in the first half of 2020, so the 2020 publication information is not complete.

TABLE III
DISTRIBUTION OF PAPERS ACROSS RESEARCH TYPES.

| Research Type | Count | Percentage (%) | |
|---|---|---|---|
| Solution | 186 | 45.9 | ████████ |
| Evaluation | 74 | 18.3 | ███ |
| Philosophical | 73 | 18.0 | ███ |
| Validation | 56 | 13.8 | ██ |
| Opinion | 14 | 3.5 | ▪ |
| Experience | 2 | 0.5 | ı |

TABLE IV
DISTRIBUTION OF PAPERS AMONGST TOPIC.

| TOPIC/Subtopic | Count | % | |
|---|---|---|---|
| - RISK MANAGEMENT | 143 | 35.3 | ██████ |
| Security | 61 | 15.1 | ██ |
| Fairness | 44 | 10.9 | ██ |
| Transparency | 14 | 3.5 | ▪ |
| Privacy | 9 | 2.2 | ▪ |
| Reproducibility | 7 | 1.7 | ı |
| Ethics | 6 | 1.5 | ı |
| Risk management | 2 | 0.5 | ı |
| - MODEL MANAGEMENT | 138 | 34.1 | ██████ |
| Explainability | 49 | 12.1 | ██ |
| Interpretation | 24 | 5.9 | ▪ |
| Visualization | 18 | 4.4 | ▪ |
| Development | 16 | 4.0 | ▪ |
| Evaluation | 9 | 2.2 | ▪ |
| AutoML | 4 | 1.0 | ı |
| Experiment Management | 4 | 1.0 | ı |
| Model Management | 4 | 1.0 | ı |
| Hyperparameter Optimization/Management | 3 | 0.7 | ı |
| Sharing | 2 | 0.5 | ı |
| Model Traceability/Versioning | 2 | 0.5 | ı |
| Training Management | 2 | 0.5 | ı |
| Technical Debt | 1 | 0.2 | ı |
| - PRODUCTION | 59 | 14.6 | ███ |
| Deployment | 31 | 7.7 | ▪ |
| Testing | 22 | 5.4 | ▪ |
| Make use of AI | 6 | 1.5 | ı |
| - LIFECYCLE MANAGEMENT | 46 | 11.4 | ██ |
| Lifecycle Management | 46 | 11.4 | ▪ |
| - DATA MANAGEMENT | 19 | 4.7 | ▪ |
| Data preprocessing/preparation | 7 | 1.7 | ı |
| Data Cleaning | 3 | 0.7 | ı |
| Data Management | 3 | 0.7 | ı |
| Annotation | 2 | 0.5 | ı |
| Data Traceability/Versioning | 2 | 0.5 | ı |
| Extraction, Transform, Load | 1 | 0.2 | ı |
| Quality Assessment | 1 | 0.2 | ı |

*3) Publication by affiliations:* 270 different companies, universities or organizations have published research on the life cycle management of artificial intelligence models. Eight of the universities/companies produced more than five publications. IBM has produced 22 articles, half of them are about risk management. Google produced 11 articles, five main themes are all included. University of Maryland, Stanford University, Carnegie Mellon University, University of California, and University of Oxford produced less than 10 related publications.

*4) Publication by type of venue:* 65.4% of all publications are conference papers, 17.5% are journal articles, 14.6% are informative and other publications, and the remaining 10 publications are parts in books or collections.

### B. RQ2: What research approaches do these studies apply?

*1) Publication by type of research type:* We map papers according to its research type using the taxonomy proposed in [17]. It includes six different types: Solution, Evaluation, Philosophical, Validation, Opinion, Experience. The distribution is presented in Table III.

Solution papers were by far the most common research type, comprising 45.9% of the publications. This means that most works focus on proposing a novel solution technique with a proof of concept consisting of a small example or a sound argument. Evaluation research follows with 18.3%, comprising papers that investigate a problem or the implementation of a technique in practice. Not so significant yet representative are Philosophical papers with 18.0% and Validation papers with 13.8%.

### C. RQ3: Which subtopics of AI model life cycle management have already been investigated?

*1) Publication by research topic:* We first summarised the keywords of each article, and after reading all the articles, the keyword database is obtained. Then we read all the articles again, and labelled each article by selecting the suitable keywords from the keyword database. All the articles are grouped into 5 overarching research topics and 31 themes (sub-topics) as presented in Table IV. Amongst which, "Risk Management" works that address governance and management of risks associated with AI. This is a broad topic, but here we exclusively focus on works that study risk management from the perspective of the lifecycle of the AI system. "Lifecycle management" is a separate category with no subcategories, containing all articles discussing artificial intelligence management from a holistic perspective.

## V. DISCUSSION

The majority of papers collected in this study tackle a particular stage of the AI model lifecycle management. Albeit relevant, we argue that it is important look into the lifecycle of AI model from a holistic perspective – only 46 papers (11.4%) in our study took such perspective. On the other hand, this suggests that there are still many problems to be solved in each particular stage. Our study also shows that some subcategories have been overlooked by the research community – e.g., *Data Traceability/Versioning*, *Quality Asessment*.

In terms of research type, a considerable portion of the publications (45.9%) revolve around proposing a *Solution*. The nature of the problems related to the lifecycle management of AI projects is mostly related to how to manage AI projects in practice, in a real setting. Thus, we argue that more *Evaluation* papers are necessary for this topic. Thus, strong collaborations between academia and the industry are key to improve the state of the art of the lifecycle management of AI applications.

*Opinion* and *Experience* publications were the least found research types – 3.5% and 0.5%, respectively. This is expected, as academic venues tend to be strict with these types of works that are subject to author bias. However, there are no standards on the AI lifecycle and engineers are figuring out

how to apply it in practice. Hence, we argue that *Opinion* and *Experience* papers are also very important to the community of AI engineers. This is a concern being raised by the research community of empirical software engineering, which highlights grey literature as a valuable data source for research [18].

We also noticed that when referring to specific topics, many articles are confused about topics with similar meanings. For example there are usually two kinds of papers about explanation: using mathematical models to explain ML models, and explaining model features. In the process of reading the literature, we found that many documents confuse explainability and interpretability. This may be due to the lack of a clear and authoritative definition of those topics, because the current research on the life cycle of artificial intelligence is still in its infancy.

## VI. THREATS TO VALIDITY

### A. Theoretical validity

*1) Study Identification/Sampling, Data extraction and classification:* The search query is based on common keywords that may appear in the title or abstract. While we made extensive efforts to iteratively improve the search query, relevant papers may have been missed. This is a common threat in mapping studies [19]. To mitigate this threat, we complement the search with backward snowball sampling. Furthermore, to reduce the researcher bias, two authors performed the paper selection. The threat is not fully mitigated since only the first author classified the selected papers in the last step of selection.

### B. Generalisability

Generalisability is not a major threat since this mapping study cover a wide range of topics. Yet, from 2009 to 2019, the number of papers is increasing almost exponentially every year. Thus, new publications in subsequent years may substantially affect the quantitative results of this mapping study. We argue that this is a common threat in any mapping study done in a fast-changing field, but does not hinder their usefulness.

### C. Interpretive validity

The conclusions drawn from the data collected are prone to researcher bias. We mitigate this threat by having four researchers discussing and reviewing the interpretation of data.

### D. Reproducibility and Repeatability

We report the systematic mapping process followed. Furthermore, all the scripts and data collected are delivered the online appendix of this mapping study (cf. Section I).

## VII. CONCLUSION

In this paper, we map 405 research publications published from 2005 to 2020 in 5 different main research topics and 31 sub topics related to the lifecycle management of AI systems. All the data and analysis is available online in the replication package (cf. Section I).

As future work, we plan to extend this research to publications in 2020 and 2021, and perform a systematic literature review.

## REFERENCES

[1] M. Haakman, L. Cruz, H. Huijgens, and A. van Deursen, "AI lifecycle models need to be revised. an exploratory study in fintech," *arXiv preprint arXiv:2010.02716*, 2020.

[2] F. Khomh, B. Adams, J. Cheng, M. Fokaefs, and G. Antoniol, "Software engineering for machine-learning applications: The road ahead," *IEEE Software*, vol. 35, no. 5, pp. 81–84, 2018.

[3] A. Arpteg, B. Brinne, L. Crnkovic-Friis, and J. Bosch, "Software engineering challenges of deep learning," in *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2018, pp. 50–59.

[4] E. Nascimento, A. Nguyen-Duc, I. Sundbø, and T. Conte, "Software engineering for artificial intelligence and machine learning software: A systematic literature review," *arXiv preprint arXiv:2011.03751*, 2020.

[5] P. Bourque, R. E. Fairley *et al.*, *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0.* IEEE Computer Society Press, 2014.

[6] V. Vakkuri and P. Abrahamsson, "The key concepts of ethics of artificial intelligence," in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE, 2018, pp. 1–6.

[7] S. Sherin, M. Z. Iqbal *et al.*, "A systematic mapping study on testing of machine learning programs," *arXiv preprint arXiv:1907.09427*, 2019.

[8] W. Spaargaren, "Systematic literature reviews: A case study in fintech and automated tool support," 2020.

[9] I. Wiafe, F. N. Koranteng, E. N. Obeng, N. Assyne, A. Wiafe, and S. R. Gulliver, "Artificial intelligence for cybersecurity: a systematic mapping of literature," *IEEE Access*, vol. 8, pp. 146 598–146 612, 2020.

[10] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, vol. 64, pp. 1–18, 2015.

[11] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, 2008, pp. 1–10.

[12] RWortmanMorris, "Introduction to MLOps and ML lifecycle - learn." [Online]. Available: https://docs.microsoft.com/en-us/learn/modules/start-ml-lifecycle-mlops/2-mlops-introduction

[13] H. Miao, A. Li, L. S. Davis, and A. Deshpande, "Modelhub: Deep learning lifecycle management," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 2017, pp. 1393–1394.

[14] F. Kumeno, "Sofware engneering challenges for machine learning applications: A literature review," *Intelligent Decision Technologies*, vol. 13, no. 4, pp. 463–476, 2019.

[15] "Machine learning life cycle: Datarobot artificial intelligence wiki." [Online]. Available: https://www.datarobot.com/wiki/machine-learning-life-cycle/

[16] B. R. Barricelli, F. Cassano, D. Fogli, and A. Piccinno, "End-user development, end-user programming and end-user software engineering: A systematic mapping study," *Journal of Systems and Software*, vol. 149, pp. 101–137, 2019.

[17] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: a proposal and a discussion," *Requirements engineering*, vol. 11, no. 1, pp. 102–107, 2006.

[18] V. Garousi, M. Felderer, M. V. Mäntylä, and A. Rainer, "Benefitting from the grey literature in software engineering research," in *Contemporary Empirical Methods in Software Engineering*. Springer, 2020, pp. 385–413.

[19] C. Wohlin, P. Runeson, P. A. d. M. S. Neto, E. Engström, I. do Carmo Machado, and E. S. De Almeida, "On the reliability of mapping studies in software engineering," *Journal of Systems and Software*, vol. 86, no. 10, pp. 2594–2610, 2013.