Journal of Software Engineering Research and Development, 2025, 13:8, doi: 10.5753/jserd.2025.4892

Requirements Engineering for Machine Learning-Based AI Systems: A Tertiary Study

Mariana Crisostomo Martins [**Universidade Federal de Goiás (UFG)** | ***maricrisostomo.martins@gmail.com***]
Lívia Mancine C. Campos [**Instituto Federal Goiano** | ***livia.mancine@ifgoiano.edu.br***]
João Lucas R. Soares [**Universidade Federal de Goiás (UFG)** | ***soares.joao@discente.ufg.br***]
Taciana Novo Kudo [**Universidade Federal de Goiás (UFG)** | ***taciana@ufg.br***]
Renato F. Bulcão-Neto [**Universidade Federal de Goiás (UFG)** | ***rbulcao@ufg.br***]

Abstract Context: In the last decade, machine learning (ML) components have become more and more present in contemporary software systems. A number of secondary literature studies reports challenges impacting on the development of ML-based systems, including those for requirements engineering (RE) activities. **Motivation/Problem**: Synthesizing secondary literature contributes to building knowledge and reaching conclusions about the existing RE approaches for ML-based systems (RE4ML), besides the novelty of a tertiary study on that subject. **Objective**: Through a tertiary study protocol we elaborated on, this paper synthesizes the body of evidence present in secondary studies on RE4ML systems. **Method**: We followed well-accepted guidelines about tertiary study protocols, including automatic search, the snowballing technique, selection and quality criteria, and data extraction and synthesis. **Results**: Nine secondary studies on RE4ML systems were aligned to our tertiary study's goal. We extracted and summarized the requirements elicitation, analysis, specification, validation, and management techniques for ML-based systems as well as the great challenges identified. Finally, we contribute with a nine-item research agenda to direct current and future searches to fill the gaps found. **Conclusions**: We conclude that RE has not been left aside in ML research, however, there are still challenges to be overcome, such as dealing with non-functional requirements, collaboration between stakeholders, and research in an industrial environment.

Keywords: *Requirements Engineering, Machine Learning, AI Systems, Tertiary Study*

1 Introduction

The emerging use of machine learning (ML) components in software segments raises concerns about the quality of components. Factors such as the large amount of data generated daily and increasing computational power and storage have contributed to the intensive use of ML components integrated into traditional software. Examples of systems can be seen in healthcare Jiang et al. (2017), finance Goodell et al. (2021), education Kucak et al. (2018), and others.

The development of ML-based systems represents a paradigm shift compared to traditional software development. ML-based systems present ML models with data-based behavior, while in traditional software the behavior conforms to users' needs and business rules Martínez-Fernández et al. (2022). ML-based systems development presents several challenges from a Software Engineering (SE) perspective Martínez-Fernández et al. (2022), to name a few, new quality attributes, such as fairness and explainability, a lot of experimentation, unrealistic stakeholder expectations, and multidisciplinary teams Lewis et al. (2021); Nahar et al. (2022).

Regarding the Requirements Engineering (RE) process, RE traditional practices are not well defined for ML-based systems development Giray (2021); Zaharia et al. (2018); Hu et al. (2020); Villamizar et al. (2021). Among RE challenges, additional effort for the successful development of ML-enabled systems is required and may contribute to the fact that 87% of ML-based projects never reach production. Due to the communication- and collaboration-intensive nature, as well as the inherent interaction with other development processes, the literature suggests RE can help mitigate most of these challenges when engineering ML-based systems Ahmad et al. (2021); Villamizar et al. (2021); Vogelsang and Borg (2019).

However, the effective establishment of RE practices in ML-based projects is challenging, primarily due to: (i) the lack of practitioners engaged in formal RE activities Alves et al. (2023); and (ii) the scarceness of tailored techniques and tools for data-driven projects, as research in this intersection predominantly focuses on using ML techniques to support RE activities rather than exploring how RE can improve the development of ML-based systems Dalpiaz and Niu (2020). It is not surprising then that recent studies claim that practitioners find RE as the most difficult phase in ML-based projects Ishikawa and Yoshioka (2019); Kuwajima et al. (2020); Nahar et al. (2022).

Hundreds of primary studies on RE and ML usually focus on one or more of the following situations toward addressing the ML-based systems particularities: (i) evaluating to what extent the existing RE approaches can be used, (ii) tailoring the current RE approaches, and (iii) proposing (and experimenting) novel RE approaches. Over the years, the contributions of these primary studies have been summarized in secondary studies on RE for Artificial Intelligence (AI) systems or RE for ML-based systems, including Systematic Literature Review and Mapping (SLR and SLM, respectively) and Literature Surveys. Comparing and contrasting evidence from secondary studies is essential to build a comprehensive understanding and drive conclusions about the empirical sup-

port of a phenomenon Cruzes and Dybå (2011). Therefore, research synthesis in RE for ML-based systems plays a central role in the scientific evolution of the RE discipline.

Those secondary studies identify requirements modeling tools and techniques, addressing aspects such as security and explainability, and map challenges and limitations of RE for ML-based systems. However, no research was found that comprehensively synthesizes the challenges and gaps highlighted in the literature or proposes a research agenda to guide further research. This gap is particularly relevant given the challenges of RE in ML-based projects.

This paper synthesizes the body of evidence present in secondary studies on RE for ML-based systems. We planned and performed a tertiary study protocol, following SE classic approaches Kudo et al. (2020b); Kitchenham et al. (2010); Cruzes and Dybå (2011). Through automatic search, snowballing, and inclusion and exclusion criteria, we selected nine studies to answer our protocol's research questions. We found research gaps and proposals of RE techniques, tools, and metrics for ML-based systems. We propose a nine-item research agenda to direct current and future research endeavors. At the time of this writing, tertiary studies on RE for ML-based systems had not been published.

This paper is organized as follows: Section 2 defines the tertiary study protocol; Section 3 describes the data extraction activity; Section 4 details each secondary study found; Section 5 synthesizes our results; Section 6 presents threats to validity results; Section 7 outlines a research agenda; and Section 8 summarizes our contributions.

2 Tertiary study protocol

This tertiary study aims to consolidate a body of knowledge on RE for ML-based systems, drawing on secondary studies. Figure 1 presents the entire process, from protocol conception to data synthesis, collected from relevant secondary studies. This section describes the tertiary study protocol. Further details about this protocol can be seen elsewhere[1].

2.1 Research questions, pilot search, and search string

According to the objective of this study, we formulated the following research questions (RQ):

RQ1: What is the state of the art about RE for ML-based systems?

Justification: The goal is to identify current RE approaches for ML-based systems, considering methods, techniques, metrics, supporting tools, and stakeholders, to name a few.

RQ2: What are the challenges and gaps highlighted by the RE literature for ML-based systems?

Justification: The goal is to map out challenges and potential research directions on RE for ML-based systems.

We defined a search string according to the tertiary study's goal and the following criteria:

- the search string should include consensual terms for RE activities as described in the Guide to the SE Body of Knowledge v4.0 (SWEBOK v4.0): elicitation, analysis, specification, validation, and management Washizaki (2024).
- the search string should include relevant terms related to the AI and ML fields commonly used in RE research for ML-based systems. In our pilot tests, we observed that some studies returned contained the term "data driven".
- the search string should include terms representing secondary research.

The final search string is as follows:

((“machine learning” OR “artificial intelligence” OR “data-driven”) AND (“requirements engineering” OR “requirements elicitation” OR “requirements analysis” OR “requirements specification” OR “requirements validation” OR “requirements management”) AND (“systematic review” OR “systematic literature review” OR “systematic mapping” OR “systematic literature mapping” OR “literature survey”)).

2.2 Search strategy

The search strategy includes automatic search[2] over six sources, including digital databases and search engines: ACM DL, IEEE Xplore, Scopus, Engineering Village, Wiley, and Web of Science.

In addition, as an attempt not to leave out relevant secondary studies, we also performed backward and forward snowballing Wohlin (2014) over the studies' citations and references resulting from papers' full reading.

2.3 Studies selection

We formulated the following inclusion and exclusion criteria (IC and EC, respectively) for studies selection:

- EC1: Full text not available for free on the Web or through the CAPES Periodicals Platform.
- EC2: It is not an English-written paper.
- EC3: It is not a secondary study.
- EC4: It does not address requirements engineering for artificial intelligence or machine learning.
- EC5: It is a preliminary or summarized version of another study already included.

A study is excluded when it falls into at least one of the ECs presented. If the study is not excluded, it has to meet IC1, as follows:

- IC1: The secondary study identifies proposals, use, or evaluation of requirements engineering research on machine learning or artificial intelligence.

Figure 2 depicts the whole process performed in our tertiary study, including protocol definition, studies identification through automatic search (**193**), duplicate studies removal (**-89**), studies selection via metadata reading and inclusion and exclusion criteria (**-95**), studies selection based

[1] `https://zenodo.org/records/14617471`

[2] We performed automatic search on October 18, 2023.

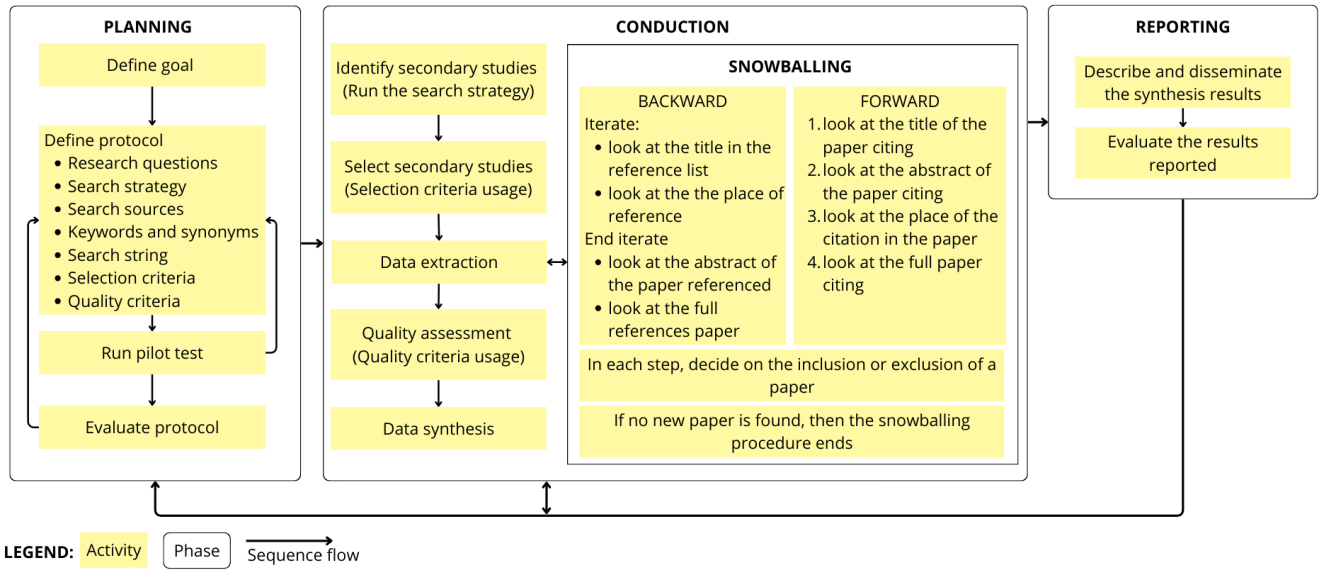

Figure 1. Phases and activities of this tertiary study (adapted from Kudo et al. (2020a); Fabbri et al. (2013); Wohlin (2014)).

on full reading of papers (**-5**), backward (**+5**) and forward (**0**) snowballing, quality assessment, and data synthesis (**9**). Therefore, nine secondary studies represent the state of the art on RE for ML-based systems.

A five-member team carried out the process detailed in Figure 2. **R1** and **R2** are PhD students and professors with experience in planning, conducting, and publishing systematic literature studies. Both performed study selection, data extraction, quality assessment, and analysis and synthesis of results. An undergraduate student (**R3**) also assisted in the study selection as a third reviewer. Finally, two more experienced researchers (**R4** and **R5**), with several publications on RE and systematic literature studies, defined the study protocol, validated the pilot test, resolved conflicts during study selection, and reviewed findings analysis and synthesis.

It is worth describing how we performed the snowballing technique. Seven papers remained after full reading, from which we examined their references (November 2, 2023). From this first round of backward snowballing, we included five new secondary studies, from which we also examined their references in a second round (November 11, 2023), but no new secondary study was found. Next, we looked for papers that cited each of the 12 secondary studies remaining (December 12, 2023), and we did not find new study in the only round of forward snowballing.

After the full reading of the 12 papers identified (seven from automatic search and five from snowballing), three papers were eliminated, and nine remained for data extraction and synthesis. Still regarding studies selection, Table 1 summarizes the number of studies removed as duplicates and those excluded by a particular EC. We cut out a great number of duplicates (243), mainly due to the overlap between study sources and the recursive nature of snowballing. As primary studies represent most of the papers' references and citations during snowballing, hundreds of papers (789) were removed by EC3. Finally, the number of studies excluded by EC4 (155) means that lots of secondary studies exist in related themes but not specifically about RE and ML-based systems.

2.4 Quality Assessment

An important decision when performing a systematic tertiary study is to check the quality of secondary studies. The Centre for Reviews and Dissemination (CDR) maintains a database of systematic reviews in Medicine selected according to pre-established quality criteria Centre for Reviews and Dissemination (2002). These same quality criteria can also be used to assess systematic studies in the SE area Kitchenham et al. (2010); Cruzes and Dybå (2011); Costal et al. (2021).

In this tertiary study, we analyze each secondary study through eight questions — or quality criteria (QC) — based on the CDR criteria. The acceptable responses for each question are Y (Yes), P (Partially), and N (No). The following is the list of quality assessment-related questions.

QC1- Are the inclusion criteria (IC) appropriately described?

Y: Criteria are explicit;
P: Criteria are implicit;
N: Criteria are not defined or not easily identified.

QC2- Are the exclusion criteria (EC) appropriately described?

Y: Criteria are explicit;
P: Criteria are implicit;
N: Criteria are not defined or not easily identified.

QC3- Does the search cover all relevant studies?

Y: It uses 4 or more sources relevant and one additional search strategy;
P: It uses 3 relevant sources, but no extra search strategy;
N: It uses at most 2 sources relevant to the area of interest.

QC4- Is the quality or validity of the included primary studies assessed?

Y: Quality criteria are explicit and associated with each primary study;

Table 1. The number of studies removed as duplicates and applying an exclusion criterion.

	Duplicate	EC1	EC2	EC3	EC4	EC5	Total
Automatic Search	89	2	0	31	62	0	184
1st Backward Snowballing	99	0	0	566	38	0	703
2nd Backward Snowballing	23	0	0	219	29	0	271
Forward Snowballing	31	2	1	73	15	0	122
Total	**242**	**4**	**1**	**889**	**144**	**0**	**1280**

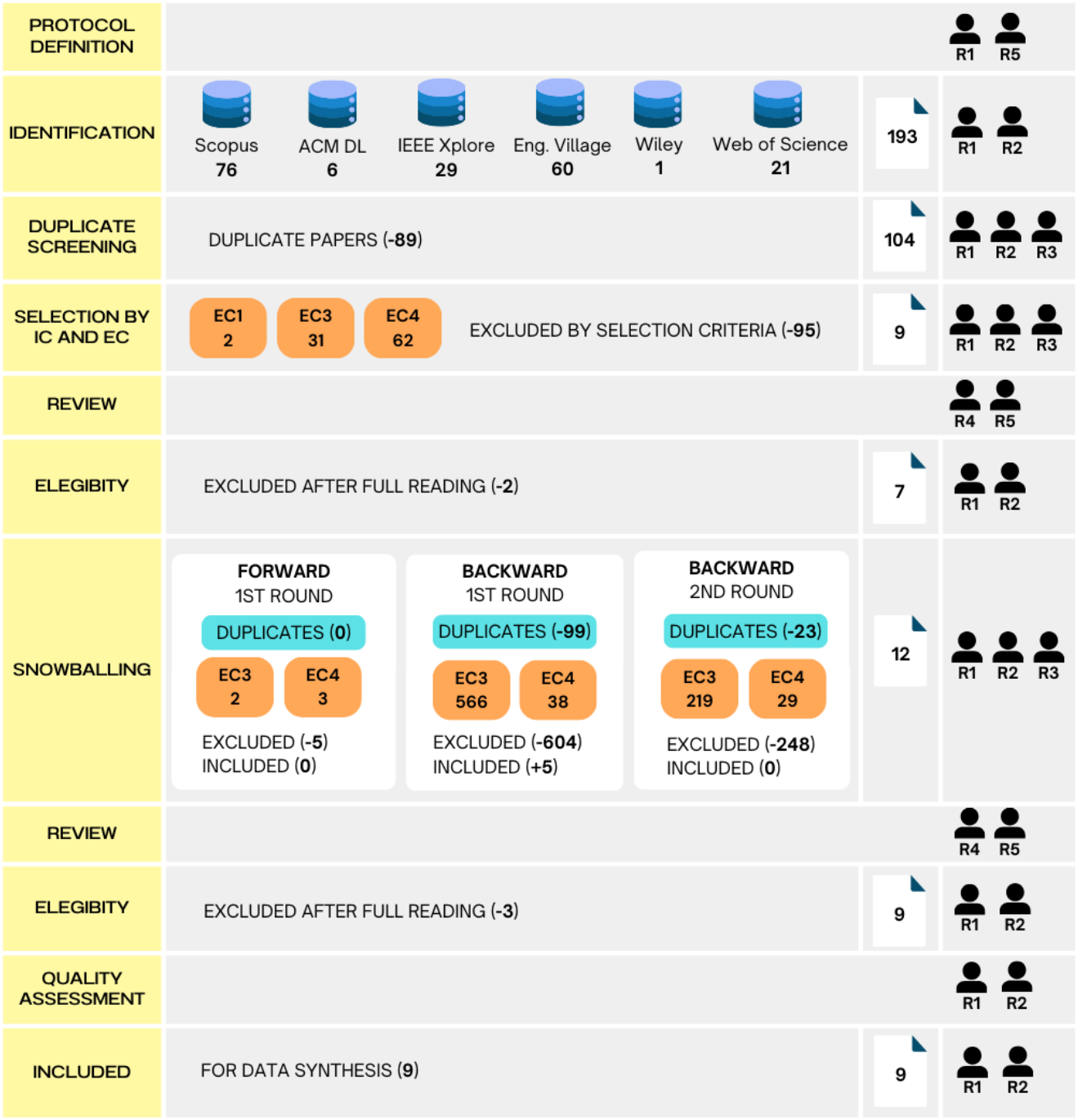

Figure 2. A detailed view of the identification and selection processes of secondary studies.

P: Research questions from the secondary study address the quality of primary studies;
N: Quality assessment not performed or described.

QC5- Are primary studies adequately described?

Y: Details of each primary study are explicit;
P: There is only a summary of each primary study;

N: There is no information on the primary studies analysed.

QC6- Is the justification for the secondary study duly described?

Y: Details of justifications for the study are explicit;
P: There is little information about justifications for the study;
N: There is no justification for the study proposed.

QC7- Is the protocol validation properly described?

Y: Details of the protocol validation are explicit;
P: There is little information about the protocol validation;
N: There is no information about the protocol validation.

QC8- Is data extraction properly described and appropriate?

Y: Details of data extraction are explicit;
P: There is only a summary of the data extraction step;
N: There is no information about data extraction.

Considering the response to each QC, we assign the following score: two for 'Yes', one for 'Partly', and zero for 'No'. The final quality score of each secondary study is the arithmetic mean of the scores for its quality questions. That quality score allows secondary studies to rank through the analysis of their protocol (QC), which is useful for guiding the synthesis activity of the secondary studies. As depicted in Figure 2, researchers **R1** and **R2** run quality assessment during the full reading of papers and data extraction.

3 Data extraction

This section describes the data extraction step through the full reading of the nine remaining papers. We elaborated on a data extraction form based on the RQs, including the following data fields: publication year, type, and title; authors' names and affiliations; number of citations[3]; search source strategies (e.g., automatic, manual, snowballing, or hybrid); number of primary studies analyzed; main objective; search string (if applicable); contributions of RE to AI/ML according to RE activities Washizaki (2024); challenges identified; and future work.

According to publication year, one paper was published in 2019, five in 2021, and three in 2023. Five secondary studies were published in journals, two in the *Journal of Systems and Software* in 2021. One author group contributed twice (**S1** and **S2** in Table 2): Klood Ahmad, Muneero Bano, John Grundy, and Mohamed Abdelrazek, who are affiliated with the School of Information Technology, Deakin University, Geelong, Australia. Two secondary studies (**S4** and **S8**) came from Japanese institutions, namely the Nippon Institute of Technology and Waseda University. Other researchers also contributed to the state-of-the-art RE4ML systems research, coming from Brazil, Germany, USA, Turkey, and Korea.

[3]We made use of Google Scholar to get the number of citations.

Although the first secondary research found dates back to 2019 (five years before this tertiary study), only two (**S4** and **S7**) of the nine studies found have less than 40 citations. It is worth highlighting the works **S9** and **S1** with 213 and 104 citations, respectively. All these numbers demonstrate not only the current state of requirements engineering research for AI or ML-based systems, but also point to a hot research topic.

Still regarding Table 2, eight secondary studies implement a hybrid search strategy combining automatic search and snowballing. Study **S9** goes beyond by adding a manual search step. Conversely, study **S6** performed only an automatic search.

During the data extraction step, researchers R1 and R2 performed the quality assessment based on the quality criteria (1 to 8) presented in the previous section. Table 3 shows the final quality assessment score for each secondary study. The table shows the score for each QC corresponding to each evaluator (R1 or R2). No study obtained a score with a discrepancy greater than 1 for the assessment comparing the score given by each researcher. In case of discrepancy, there would be a review with other researchers (R1, R2, R4 and R5).

The quality assessment was performed independently by researchers R1 and R2. The authors meet weekly and discuss the results of the quality assessment with the other authors. At the end of the readings, we calculated an average score for each question given by the researchers. Finally, we calculated the total average given the averages for each question. For example, M1 = (x1+y1)/2 where x1 is the grade given by researcher 1 for question M and y1 is the grade given by researcher 2 for question M. Then we averaged the questions, for example, total = (M1+N1+...+T1)/8, where M1 is the average of the scores for question M calculated previously.

We noticed that the studies with higher scores had contributions and objectives more aligned with the objectives of this tertiary study, that is, they instigated RE4ML. The other studies that obtained lower scores (**S6** to **S8**) were characterized by investigation of a specific topic, such as explainability or startups. Those studies do not follow the classical protocol definition and disclosure guidelines in software engineering systematic secondary studies Kitchenham (2004); Wohlin (2014); Ampatzoglou et al. (2019).

4 Studies Characterization

This section details each secondary study considered relevant to answer the research questions of this tertiary review.

4.1 About S1

Ahmad et al. (**S1**) conducted a systematic mapping to identify existing empirical assessments, emerging theories, and instances of limitations and challenges in RE for AI systems Ahmad et al. (2023). The Australian authors published the study in 2021 at the *29th International Requirements Engineering Conference*. The authors built a search string of related synonyms for RE and AI. They analysed 43 primary studies through an automatic search in six sources and backward and forward snowballing.

Table 2. Summary of data extracted from secondary studies: J for journal, C for conference, A and M mean automatic and manual search, respectively, and S for snowballing.

Study	Reference	Country	Year	Source	Type	Strategy	Citations
S1	Ahmad et al. (2023)	Australia	2023	Scopus	J	A, S	104
S2	Ahmad et al. (2021)	Australia	2021	Scopus	C	A, S	77
S3	Villamizar et al. (2021)	Brazil	2021	Scopus	C	A, S	71
S4	Yoshioka et al. (2021)	Japan	2021	Scopus	C	A, S	11
S5	Dey and Lee (2021)	Rep. of Korea	2021	Scopus	J	A, S	42
S6	Clement et al. (2023)	Germany	2021	Scopus	J	A	88
S7	Lakha et al. (2023)	USA	2023	Scopus	C	A, S	2
S8	Kumeno (2020)	Japan	2019	IEEE Xplore	J	A, S	97
S9	Giray (2021)	Turkey	2021	ScienceDirect	J	A, S, M	213

Table 3. Quality score of the secondary studies.

Study	QC1		QC2		QC3		QC4		QC5		QC6		QC7		QC8		Score
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	
S1	2	2	2	2	2	2	2	2	2	1	2	2	1	2	0	0	**2,0**
S3	2	2	2	2	2	1	0	0	1	2	2	2	2	2	2	2	**2,0**
S4	2	2	2	2	1	2	0	0	0	0	2	1	2	2	2	2	**2,0**
S5	2	2	2	2	1	2	0	0	0	0	2	1	2	2	2	2	**2,0**
S9	2	2	2	2	2	2	2	2	0	1	2	2	1	1	1	1	**2,0**
S2	0	1	0	1	2	2	1	2	2	1	2	2	2	2	1	1	**1,5**
S7	1	1	0	0	1	2	1	2	0	1	1	2	0	1	0	0	**1,0**
S8	0	2	0	0	0	1	0	0	0	1	2	1	2	1	1	0	**1,0**
S6	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	1	**0,5**

The authors highlighted modeling languages such as SysML, ontoML, activity diagrams, and semi-formal UML models. They utilized the jUCMNav tool for OO requirements modeling. The most common evaluation methods identified were case studies, surveys, and controlled experiments, with the domain of greatest interest being autonomous driving and robotics. The challenges highlighted by the authors include requirements definition, introducing new methods to existing techniques, lack of stakeholder integration, and ethical, explainability, and data issues. The authors identified future work involving the identification of a modeling language and the promotion of a collaboration platform for stakeholders, alongside the evaluation of proposals.

4.2 About S2

Those same authors also conducted a systematic review (**S2**) to investigate current writing and modeling requirement approaches for AI/ML systems Ahmad et al. (2021). This work was published in the *Information and Software Technology* journal in 2023. They borrowed the same search strategy and search string used in **S1**.

The authors analyzed 27 primary studies, from which 18 are empirical evaluations of the use of existing RE techniques when building a system with an AI component, and nine are solution proposals with little or no evaluation concern. The authors also identified requirements modeling languages, such as UML and GORE. The most commonly found application domains were autonomous driving and computer vision. Ethics, trust, and explainability issues were cited as relevant, but are treated only theoretically, with no evaluation. Moreover, they reported that the literature presents challenges when dealing with data requirements and non-functional requirements. The authors pointed out the lack of integration of tools and methods. Aspects that help with communication and documentation still need to be studied and require empirical evaluation. In future work, the authors aim to involve requirements documentation, identify suitable language models, and provide a collaboration platform for stakeholders.

4.3 About S3

The Brazilian authors, Villamizar et al. (**S3**), published a systematic mapping in 2021 at the *47th Euromicro Conference on Software Engineering and Advanced Applications* Villamizar et al. (2021). The objective of the investigation was to outline the state-of-the-art research on RE for ML-based systems. The string used was *(Software OR Applications OR Systems) AND (Machine Learning) AND (Requirements Engineering)*. The authors used an automatic search strategy and backward and forward snowballing applied to two search engines.

From the analysis of 35 primary studies, the authors found that requirements elicitation and analysis were the most investigated RE activities, as well as the brainstorming technique for elicitation purposes. They also highlight that ad hoc methods are employed to elicit and ensure the fulfillment of NFRs. The authors mentioned that it remains unclear which traditional tools and techniques can be applied to RE4ML.

Finally, it was observed that most studies addressing the intersection between RE and ML employ ML in RE activities.

4.4 About S4

The systematic review conducted by Yoshioka et al. (**S4**), in turn, reports current RE techniques and practices for ML-based systems Yoshioka et al. (2021). The authors from Japan published their study in 2021 at the *28th Asia-Pacific Software Engineering Conference Workshops*. They performed an automatic search using a search string combining ML, AI, and RE keywords and synonyms over six sources, followed by the forward snowballing technique.

From the analysis of 32 papers, the authors point out that 62% of these address specific requirements, such as explainability, transparency, accountability, precision, robustness, dataset requirements, fairness, and overfitting. Besides, 12.5% address specific techniques for ML activities, such as data collection and training. Requirements elicitation is covered in 15 primary studies, followed by requirements specification, analysis, modeling, validation, and management. The authors highlight that only one article is dedicated to monitoring, despite runtime model degradation being a concern for ML systems. They also point out that 25 of 32 studies are focused on system requirements. The most cited domains in the studies were automotive, aviation, and health. Ten studies highlight the participation of AI-knowledge stakeholders, while only three deal with domain experts. The authors also note that the most commonly used specification techniques are goal-oriented. As challenges, they point out domain and data understanding and uncertainty issues through explainability and requirements analysis and modeling. As for future work, the authors demonstrated interest in including gray literature, analyzing additional characteristics, extending the databases, and evaluating with professionals.

4.5 About S5

In 2021, the Korean researchers, Dey and Lee, published a literature review (**S5**) in the *Journal of Systems and Software* Dey and Lee (2021), whose goal is to identify the state of the art of safety approaches for ML systems. The authors elaborated on a search string with keywords related to AI, ML, deep learning, and software engineering. Automatic search over six sources and backward and forward snowballing represent the study's search strategy.

The authors raise concerns for each of the RE phases. For elicitation, they point out the need to include data scientists and legal experts, use domain benchmarks, elicit new data sources, identify sensitive characteristics of the data, and analyze situations that require explanation. Regarding analysis, they highlight the importance of discussing user-understandable performance measures and conditions for data pre-processing and cleaning, besides the relevance of obtaining the system automation level required and carrying out goal- and evidence-oriented modeling and analysis. About specification, they highlight the relevance of dealing with data and model requirements, the ML process, and quantitative and measurable markers for requirements writing. The authors also emphasize the importance of specifying explainable, ethical, and legal requirements that address robustness. Concerning validation, the authors cite monitoring data dependence and reliability, the analysis of quantitative markers, evolving dataset documentation, model versioning, and requirements traceability. As challenges, the authors highlight (i) data requirements specification, (ii) collaborative and knowledge-sharing initiatives, and (iii) a lack of an integrated structure to deal with traceability. In future work, the authors describe a methodology to guide safety analysis and perform verification for ML-based systems.

4.6 About S6

A systematic metareview (**S6**) developed by German researchers was published in the *Machine Learning and Knowledge Extraction* journal in 2023 Clement et al. (2023). The goal is to identify and analyze methods and tools used in the explainable AI software systems development process. The search strategy includes only automatic search over multiple sources, using a search string with keywords about explainable AI, resulting in 227 studies for analysis.

Regarding the requirements phase, the authors emphasize the need to specify what needs to be explained and to whom, and the importance of defining relevant stakeholders and users and their different characteristics, such as AI knowledge, attitude toward AI, responsibilities, and skills. They propose a trade-off analysis as part of the requirements analysis, considering that explainability can impact aspects such as ease of use. Furthermore, they highlight the importance of defining what happens if explainability is not possible and how much uncertainty is tolerable, in addition to emphasizing the importance of real-world evaluation. As future work, the authors suggest a proposal for a literature metareview.

4.7 About S7

Of authorship of USA researchers, the paper (**S7**) aims at understanding software engineering practices in software and ML startups Lakha et al. (2023). Published in the *IEEE/ACIS International Conference on Software Engineering, Management, and Applications* in 2023, the study's search strategy includes an automatic search over three databases and forward snowballing, resulting in 37 studies for analysis.

The authors cite that there are initiatives to understand the problem and define the goal, but there are no additional concerns regarding the lack of data. They also highlight that data requirements must be unbiased, sufficient, consistent, robust, and correct, as well as the difficulty of communicating with customers with high expectations and do not adequately understand the data metrics. The authors are concerned about requirements affecting data management, model construction, quality assurance, and deployment. Startups do not have a well-defined process, but good practices show improvements in performance, quality of work, and stakeholder satisfaction. As a challenge, the authors mention the pressure on delivery time and the struggle in handling data requirements. In future work, the authors seek to combine case studies and interviews with startup participants to highlight the use of practices and add gray literature to this study.

4.8 About S8

The systematic review (**S8**) conducted by Kumeno (2020) aims to clarify software engineering challenges for ML applications. Published in the *Intelligent Decision Technologies* journal in 2019, this research from Japan implements a search strategy combining automatic search and iterative forward and backward snowballing using the Google search website.

The author reveals as challenges the need to capture requirements, make changes, and adapt the existing process, beyond the need to improve support for domain experts by providing approaches to capture domain, data, and business requirements. Developing languages and tools to support RE activities was also highlighted in this study, as well as the need for further investigation on security, fairness, and privacy.

4.9 About S9

At last, the study (**S9**) identifies, analyzes, summarizes, and synthesizes the state-of-the-art research in software engineering for ML-based systems Giray (2021). This study was developed by an author affiliated with a Turkish corporation and published in the *Journal of Systems and Software* in 2021. The search string combines software engineering and machine and deep learning keywords and synonyms, applied to six databases. Besides automatic search, primary studies were also collected from manual search and snowballing, resulting in 141 papers.

The author calls attention to the struggles in managing stakeholders' expectations, deciding who is responsible for each need, convincing people about the value of real and possible resources, dealing with requirements that depend on data, and managing the uncertainty level. The author also points out the need to extend current specification practices with quantitative measures concerned with quality and concerns, such as performance, fairness, and explainability. As a challenge, he highlights the importance of requirements specification in different ways, possibly using a hypothesis-based approach and integrating quality aspects into the requirements.

The importance of addressing GDPR is still highlighted, as well as process changes with adapted practices. As future work, the author highlights the importance of rethinking system development with greater collaboration with industry. The author also suggests that he seeks to perform a multivocal literature review in the industry focused on RE.

5 Data synthesis

This section presents a synthesis of the data extracted from the **S1** to **S9** studies to answer the research questions of this tertiary review.

5.1 About RQ1

Regarding the research question: "What is the state-of-the-art of RE4ML systems?", the goal is to report current RE approaches for ML-based systems. The results below are evaluated according to the elicitation, analysis, specification, validation and management activities described in SWEBOK v4.0 Committee et al. (2022). Only study S8 does not present a clear contribution regarding those RE activities. The studies remaining conduct discussions about RE4ML on the RE activities, with an emphasis on elicitation, analysis, and specification.

5.1.1 Elicitation

For the elicitation phase, where users' needs are discovered, study **S1** presents the RE4ML framework. Study **S2** describes concerns with data requirements and non-functional requirements (NFR) in their primary studies analyzed, which is also found in studies **S4**, **S5**, **S7**, and **S9**. Study **S3** shows that the most commonly used elicitation techniques are brainstorming and ad-hoc methods.

Highlighted among the studies are the following NFR: ethics, trust, and explainability (**S1**, **S2**), precision (**S4**), reliability (**S7**), robustness and legal requirements (**S5**), safety (**S5**, **S8**), security, interpretability, justice, and privacy (**S8**, **S9**). Study **S2** also points out the challenges in enabling communication and collaboration between stakeholders, which is corroborated by studies **S3**, **S5**, and **S7** to **S9**. Study **S3** points out the importance of dealing with uncertainty and managing customer and user expectations in ML-based AI systems projects, as also found in studies **S4**, **S6**, and **S9**. Study **S5** also refers to the relevance of eliciting new data sources, identifying data-sensitive characteristics, and obtaining quantitative markers for the quality of this type of system.

Studies **S5** and **S6**, which focus on safety/security and explainability, emphasize the importance of identifying situations that require explainability, as well as defining stakeholders and user characteristics. Study **S5** contradicts **S7** by stating that the definition of the problem and the domain have received less attention in the primary studies analyzed.

5.1.2 Analysis

As for the analysis activity, the prioritization and classification of the elicited needs are reported. Studies **S2**, **S5** and **S9** identified that the most used modeling languages for this type of system are UML and GORE-based. Study S1 also mentions that others were also cited in primary studies, such as SysML, ontoML, activity diagram, and semi-formal UML models.

Additionally, the study **S1** presents modeling tools, including jUCMNav and two others developed by researchers named Rius and Sirius. Furthermore, study **S1** also highlights frameworks for analysis, holistic DevOps and Ethics-Aware. Study **S5** emphasizes the importance of discussing performance measures in connection with NFR. Study **S6** also addresses the need to perform trade-off analysis, e.g., balancing fairness versus precision or performance versus explainability.

5.1.3 Specification

Studies **S1** and **S2** found specification languages, such as signal temporal logic (STL) and traffic signal control (TSC).

Study **S4** indicated other languages used for specification: i*, KAOS, UML and safety-case.

Studies **S5** and **S7** highlight the importance of specifying data, models, and processes requirements. Study **S5** also presents the importance dataset documentation, model versioning, and traceability management. Study **S6** outlines the importance of identifying the essential information that must be conveyed and to whom, establishing a correlation with the characteristics of the users identified during the elicitation activity. Given the inherent characteristics of ML-based AI systems, study **S9** shows that the specification could be declared as a hypothesis to tested as an experiment.

5.1.4 Validation and Management

Study **S3** was the only one that presented four verification and validation studies. Surprisingly, the management phase was not mentioned in any of the secondary studies analyzed.

The frames below synthesize the key information extracted from the analyzed studies, covering core topics in RE such as elicitation, analysis, specification, validation, and management. This consolidated view highlights the main contributions and challenges, providing a comprehensive overview of the the state-of-the-art in RE4ML systems.

S1

Elicitation: NFR (e.g., ethics, trust, and explainability).
Analysis: UML, GORE, sysML, ontoML, activity diagram, semi-formal UML model, jUCMNav, Rius, Sirius; frameworks for analysis, holistic DevOps, and Ethics-Aware.
Challenges: Defining, treating, and evaluating NFR, such as ethics, explainability and data requirements. Integrating stakeholders and development team. Need for new methodologies.

S2

Elicitation: NFR (e.g., ethics, trust, and explainability) and data requirements.
Analysis: UML and GORE.
Challenges: Enabling communication and collaboration between stakeholders. Defining, treating, and evaluating NFR, such as ethics, explainability, and data requirements. Need for new methodologies.

S3

Elicitation: Brainstorming and ad-hoc methods.
Specification: Defining business metrics.
Validation: Ad-hoc methods for NFR assurance.
Management: Presenting a framework for management.
Challenges: Integrating stakeholders and development team. Comprehend domain, problem, and uncertainty nature. Need for new methodologies and evaluated approaches.

S4

Elicitation: NFR and data requirements.
Challenges: Defining, treating, and evaluating NFR, such as ethics, explainability, and data requirements. Comprehend domain, problem, and uncertainty nature.

S5

Elicitation: NFR (e.g., safety and explainability) and data requirements.
Analysis: UML and GORE; discussing performance measures aligned with NFR.
Specification: Specifying data, model, and process requirements; dataset documentation, model versioning, and traceability management.
Challenges: Defining, treating, and evaluating NFR, such as ethics and explainability. Integrating stakeholders and development team. Addressing traceability and specifying metrics and acceptance criteria.

S6

Elicitation: NFR (e.g., security and explainability).
Analysis: Trade-off analysis (fairness vs. precision; performance vs. explainability).
Specification: Identifying the essential information that must be conveyed and to whom.

S7

Elicitation: NFR (e.g., ethics, trust, and explainability) and data requirements.
Specification: Specifying data, model, and process requirements.
Challenges: Lack of a process.

S8

Challenges: Enabling communication and collaboration between stakeholders. Defining, treating, and evaluating security, safety, justice, interpretability, and privacy. Need to investigate the use of ad-hoc techniques.

S9

Elicitation: NFR and data requirements.
Analysis: UML and GORE.
Specification: Specification could be declared as a hypothesis to be tested through an experiment.
Challenges: Enabling communication and collaboration between stakeholders. Defining, treating, and evaluating the privacy requirement. Addressing traceability and specifying metrics and acceptance criteria. Lack of process. Need to evaluate proposals in different scenarios and in partnerships with industry.

5.1.5 Miscellaneous

According to studies **S1** and **S2**, the most investigated areas deal with autonomous driving and computer vision. However, study **S4** describes that the most investigated areas were autonomous driving, aviation, and healthcare. Studies **S1** to **S3** present the most commonly used empirical evaluation methods in RE research for ML-based AI systems. Study **S3** also indicates a lack of RE techniques and tools for ML-based AI systems, as well as a lack of evaluation to determine whether existing methods and tools are suitable for this type of system.

On the other hand, study **S7** claims that ad-hoc methods used in startups improve performance, delivery quality, and developer satisfaction. Unfortunately, those methods are not presented in the paper. Study **S8** highlights the need to adapt the RE process for ML-based AI systems. Studies **S1** and **S5** broght significant contributions to the elicitation, analysis and specification, highlighting structures, tools, and needs of these phases. Study **S9** points out the importance of considering trade-offs and explaining them in a hypothesis format.

Finally, studies **S3** and **S4** identified that the most investigated activity was elicitation, with brainstorming being the most used technique. These studies highlight uncertainty about which existing approaches and tools are suitable for this type of system, suggesting that empirical assessments applied to different scenarios could provide clarity.

In Figure 3 we highlight the main contributions pointed out by the secondary studies according to RE activity Committee et al. (2022).

> **Findings**: the elicitation and analysis phases received the most contributions, with a focus on non-functional and data requirements. Key non-functional requirements include ethics, trust, explainability, precision, reliability, robustness, legal aspects, security, interpretability, fairness, and privacy. Autonomous driving and computer vision were the most studied domains, and brainstorming was the most common elicitation technique. The suitability of existing techniques and tools for requirements engineering in ML-based AI systems remains unclear. Ad hoc methods seem insufficient but require further investigation. Secondary studies emphasize stakeholder communication, clear documentation, and the need for more empirical research.

5.2 About RQ2

Regarding the research question: "What are the challenges and gaps highlighted by the RE literature for ML?", the goal is to collect challenges and future research directions on RE for ML-based systems.

5.2.1 Challenges

Defining, treating, and evaluating NFR, such as ethics, explainability (**S1**, **S2**, **S4**, **S5**), security, justice, interpretability (**S8**), privacy (**S8**, **S9**), and data requirements (**S1**, **S2**, **S4**), are very discussed challenges in the secondary studies.

Studies **S1**, **S3**, and **S5** also highlight the struggles of integrating the development team and stakeholders and ensuring a consensual understanding. Comprehending the domain and navigating uncertainty are pointed out by **S3** and **S4**, enabling customers and users to understand and engage in the process with appropriate expectations.

Studies **S1** and **S2** indicate the need to introduce new methodologies, while **S3** indicates a scarcity of such proposals and shows how the absence of formal evaluation affects the certainty that ad hoc approaches are adequate for defining requirements in ML-based AI systems. Study **S5** presents the challenge of addressing traceability and specifying metrics and acceptance criteria, which is corroborated by **S9**.

Studies **S7** and **S9** present challenges related to the lack of an RE process and the pressure of delivery time that can affect the adherence of RE with ad hoc methods for this type of system. This contribution is complemented by **S8**, which argues that requirements in ML-based AI systems change rapidly and therefore ad hoc techniques may not be appropriate. Study **S9** shows that it is difficult to work together with industry and academics and assess different situations, which is corroborated by **S2** and **S3**, which additionally present the challenge of evaluating different scenarios.

5.2.2 Future work

Studies **S4**, **S6**, **S7**, and **S9** propose to carry out bibliographic study initiatives in their future research. Study **S4** aims to extend its study by adding gray literature and feedback from practitioners to its analysis. Study **S6** indicates a meta-review as future work. Besides gray literature, **S7** also promises to extend the research with case studies and interviews with professionals. Finally, **S9** cites a multifocal review as future work.

S1 seeks to evaluate techniques in different areas, while **S2** aims to develop documentation and modeling language proposals and a platform for stakeholder collaboration and communication. Study **S5** proposes a methodology to guide the work of security teams. Only studies **S3** and **S8** do not present future work proposals.

Figure 4 depicts the main challenges identified in our analysis. Regarding requirements elicitation, we highlight collaborative communication between stakeholders (including data scientists) and stakeholders' expectations management due to the uncertainty of AI-based solutions, and data requirements and NFRs (e.g., security, explainability, ethics, trust, fairness, privacy, legal requirements, etc.).

For the requirements analysis activity, we give special importance to discussing performance measures with stakeholders and performing trade-off analysis, especially aligned with NFRs. Concerning the requirements specification, we point out the significance of specifying data, model, and process requirements in a versionable and traceable manner, and the alignment between specifications and the type of stakeholder.

The analysis of secondary studies calls attention to the low number of studies concerned with requirements validation and management. Challenges related are the definition and specification of performance measures and acceptance criteria for these requirements activities and the establishment of

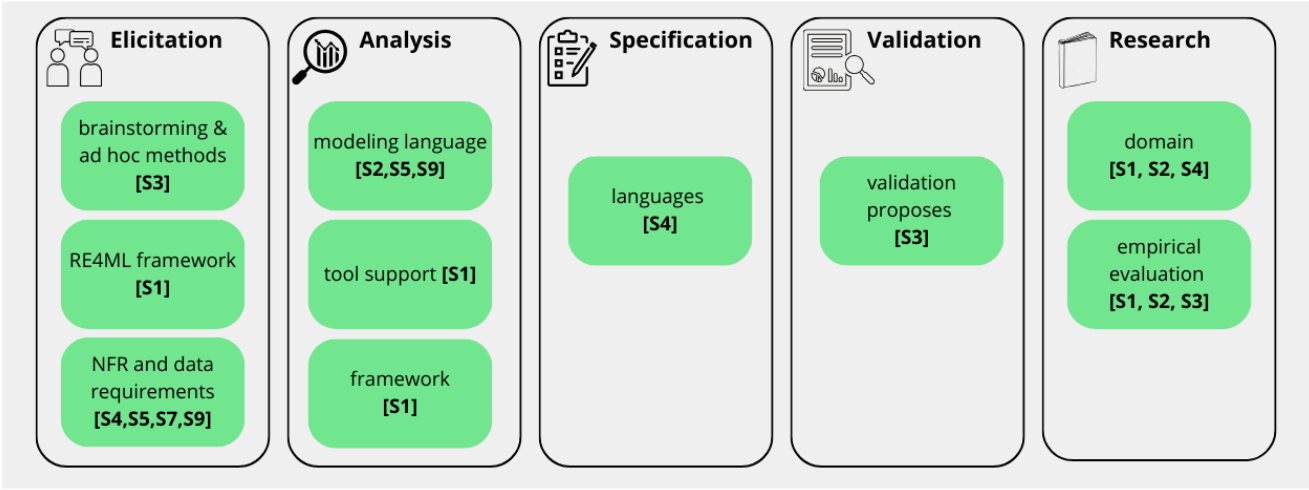

Figure 3. Synthesis of main contributions found in secondary studies.

traceability between requirements artifacts. Finally, as challenges for future RE4ML research, studies refer to a tailored RE process for data-driven systems development, the proposal of new RE-oriented techniques and tools, and experimental use of new and existing RE approaches regarding their suitability for the ML domain.

> **Findings**: Handling non-functional and data requirements is challenging, with the lack of a structured process affecting traceability, uncertainty in ML-based AI systems, and customer expectations. There is also a gap in suitable RE techniques and tools, and uncertainty about existing solutions. Key issues include domain understanding, communication, and stakeholder collaboration. Secondary studies suggest integrating gray literature and professional feedback, while few emphasize evaluation as a future work.

6 Threats to validity

Reporting systematic literature studies presents intrinsic difficulties, e.g., findings are often relevant on a specific topic. We adopted the checklist presented by Ampatzglou et al. Ampatzoglou et al. (2019) to list the actions we took to mitigate threats to the validity of this study.

The tertiary study protocol was proposed by three researchers (R1 to R3) and reviewed by two more experienced ones (R4 and R5). Further information about the tertiary study protocol not available in this paper can be found in the supplementary artifacts.

To mitigate search string-related threats, we selected consensual knowledge about RE, as defined in SWEBOK v4.0, and terms related to AI/ML validated through multiple pilot searches. Inclusion and exclusion criteria were discussed among the research team of this tertiary study to obtain a common understanding of RE-related concepts in AI/ML-based systems. To identify relevant SLRs and mitigate search and selection biases, we searched papers through six sources and also did a recursive search on the references and citations (i.e., snowballing) of relevant papers. Besides, each paper was reviewed by three researchers (R1, R2, and R3). In case of disagreements, we resolved them through discussions and reconsiderations with the most experienced researchers (R4 and R5).

To ensure that the theoretical concepts were correctly represented during extraction and synthesis, we used the RE activities described in SWEBOK v.4.0. This decision aligns the observation of the extracted data with a reference literature. We also defined a data extraction form to ensure consistency in extracting relevant information, and we evaluated the data according to the research questions. In addition, we had at least three researchers who extracted the data independently. The more experienced researchers (R4 and R5) dealt with disagreements and divergences during the process.

Concerning the quality of the nine secondary studies found, we elaborated on quality criteria based on the CDR approach. To mitigate risks, two researchers carried out the quality evaluation process (R1 and R2), and two more experienced ones reviewed it. We also synthesized the results following the RE activities based on the SWEBOK v4.0 definitions.

Finally, despite all the effort spent to search for as many secondary studies on RE for AI/ML as possible, we are aware that some secondary research may not have been retrieved, which would restrict the generalizability of our results.

7 Research agenda

By synthesizing our findings, we identified key areas that require further investigation. Based on these insights, we propose a research agenda encompassing nine critical topics that address the challenges in ML-based systems. The Figure 5 presents the main points that will be discussed throughout this section.

A fundamental aspect of ML systems lies in their intrinsic characteristics, including **data properties**, **model features**, **ML processes**, and **infrastructure**. Defining data requirements is crucial to ensure completeness, representativeness, ethical considerations, privacy compliance, and fair distribution of samples. Equally important is the specification of model requirements, such as explainability constraints, per-

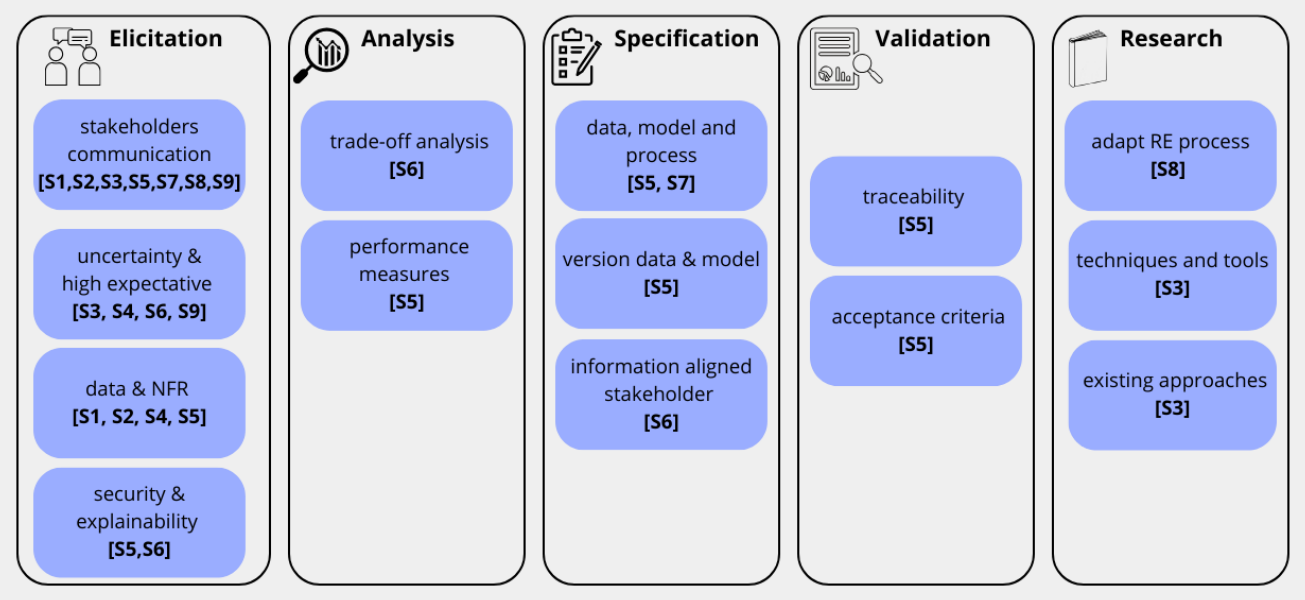

Figure 4. Synthesis of challenges discussed in secondary studies.

formance limitations affecting algorithm selection, relevant performance metrics, execution and learning times, and concerns related to model complexity and degradation. Infrastructure considerations should also be clearly defined, covering aspects such as model provisioning, artifact storage, monitoring, integration needs, and cost management.

Another critical issue is **explainability**. Different stakeholders have varying levels of understanding about the ML-based system and, consequently, distinct needs regarding the interpretability of ML models. When documenting requirements, it is essential to specify what needs to be explained, to whom, and in what context. While explainability focuses on making model decisions comprehensible to different users, interpretability deals with understanding how the model arrives at its conclusions.

NFRs and **trade-offs** also play a significant role in the development of intelligent systems. Different applications may demand specific considerations related to ethics, security, uncertainty, and robustness. Ensuring that data is unbiased, that the model consistently produces reliable results, and that a certain tolerance for errors in predictions is established are all fundamental aspects that must be addressed. Security, in particular, requires careful attention, as it influences and is influenced by other requirements. This is especially relevant in high-stakes scenarios, such as healthcare, where ML models can impact automated decisions not supervised by humans.

Collaboration among stakeholders is another crucial factor for the success of ML-based solutions. Establishing effective communication and expectation management processes is essential for capturing requirements and analyzing problems from multiple perspectives. **Tools** such as structured frameworks and collaborative platforms can facilitate interaction, helping stakeholders align their understanding of the development process and the resulting ML-based solutions. Improving **traceability** mechanisms is also critical to maintaining quality throughout development. Artifacts should support iterative refinement and allow for continuous revisiting and modification of documented information from the goal and problem identification phase through solution delivery.

To ensure that ML models meet business objectives, it is crucial to define measurable **metrics** and **acceptance criteria**. Additionally, refining RE processes for ML systems is necessary to develop or adapt methodologies for requirement elicitation, analysis, specification, validation, and management. Aligning these processes with the ML-based system lifecycle enables better treatment of critical issues and trade-offs while also facilitating acceptance criteria specification and artifacts traceability.

Finally, **empirical studies** in real-world projects are essential for evidence of the effectiveness of RE approaches in ML-based systems development. Conducting controlled experiments and case studies can offer valuable insights into how RE practices influence the success of AI-driven solutions in industry settings.

By addressing these interconnected areas, we aim to advance research and practice in RE for ML-based systems, fostering the development of more transparent, secure, and efficient AI solutions.

8 Conclusions

This paper synthesized the knowledge from secondary studies on RE for ML-based systems through a tertiary study protocol that is available in the supplementary material. From this tertiary study, we reached the following conclusions.

- The traditional RE practices, techniques, and methods are unsuitable for AI/ML systems. These RE approaches must be adapted and validated through empirical studies for AI/ML-based systems (**S3**, **S8**).
- Dealing with NFRs (**S2** to **S5**, **S8**, **S9**) and data requirements (**S1**, **S5**, **S7**, **S8**) in AI/ML-based systems is a significant challenge. For now, there is no complete understanding of how to approach these requirement types.

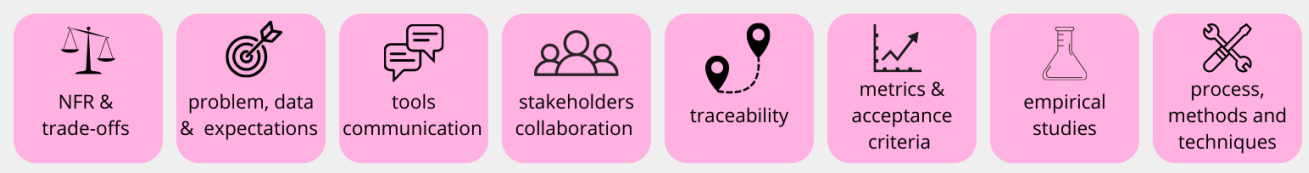

Figure 5. The research agenda's main topics aligned with the challenges collected from secondary studies.

- Ethical issues (**S1**, **S2**) and explainable AI (**S1**, **S2**, **S4**, **S6**, **S9**) in RE for ML-based AI systems still require in-depth investigation.
- Collaboration between multiple types of stakeholders during the AI/ML-based systems development is still a poorly understood point. The lack of collaboration and clarity in the roles of stakeholders can lead to project failures (**S1** to **S3**, **S5**, **S9**).
- Research centers and industry ought to establish a collaboration network with universities to leverage the evidence obtained in academic settings.
- The absence of a custom RE process for ML-based AI systems prevents practical initiatives due to the intrinsic characteristics of this type of system (**S4**, **S5**, **S9**).

Acknowledgment

This study was financed in part by the Coordination of Superior Level Staff Improvement - Brazil (CAPES).

References

Ahmad, K., Abdelrazek, M., Arora, C., Bano, M., and Grundy, J. (2023). Requirements engineering for artificial intelligence systems: A systematic mapping study. *Information and Software Technology*, 159:107176.

Ahmad, K., Bano, M., Abdelrazek, M., Arora, C., and Grundy, J. (2021). What's up with requirements engineering for artificial intelligence systems? In *Proceedings of the 29th IEEE International Requirements Engineering Conference (RE 2021)*, pages 1–12. IEEE.

Alves, A., Kalinowski, M., Giray, G., Méndez Fernández, D., Lavesson, N., Azevedo, K., Villamizar, H., Escovedo, T., Lopes, H., Biffl, S., Musil, J., Felderer, M., Wagner, S., Baldassarre, T., and Gorschek, T. (2023). Status quo and problems of requirements engineering for machine learning: Results from an international survey. In *Product-Focused Software Process Improvement: 24th International Conference, PROFES 2023, Dornbirn, Austria, December 10–13, 2023, Proceedings, Part I*, page 159–174, Berlin, Heidelberg. Springer-Verlag.

Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., and Chatzigeorgiou, A. (2019). Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology*, 106:201–230.

Centre for Reviews and Dissemination (2002). The database of abstracts of reviews of effects (dare). *Effectiveness Matters*, 6(2):1–4.

Clement, T., Kemmerzell, N., Abdelaal, M., and Amberg, M. (2023). XAIR: A systematic metareview of explainable AI (XAI) aligned to the software development process. *Machine Learning and Knowledge Extraction*, 5(1):78–108.

Committee, I. C. S. P. P. et al. (2022). Swebok: Guide to the software engineering body of knowledge, 2022 version beta. *IEEE Computer Society*.

Costal, D., Farré, C., Franch, X., and Quer, C. (2021). How tertiary studies perform quality assessment of secondary studies in software engineering. In *Proceedings of the XXIV Iberoamerican Conference on Software Engineering (ESELAW@CIbSE)*, pages 1–14.

Cruzes, D. S. and Dybå, T. (2011). Research synthesis in software engineering: A tertiary study. *Information and Software Technology*, 53(5):440–455.

Dalpiaz, F. and Niu, N. (2020). Requirements engineering in the days of artificial intelligence. *IEEE Software*, 37(4):7–10.

Dey, S. and Lee, S.-W. (2021). Multilayered review of safety approaches for machine learning-based systems in the days of AI. *Journal of Systems and Software*, 176:110941.

Fabbri, S., Felizardo, K., Ferrari, F., Hernandes, E., Octaviano, F., Nakagawa, E., and Maldonado, J. (2013). Externalising tacit knowledge of the systematic review process. *Software, IET*, 7:298–307.

Giray, G. (2021). A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software*, 180:111031.

Goodell, J. W., Kumar, S., Lim, W. M., and Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32:100577.

Hu, B. C., Salay, R., Czarnecki, K., Rahimi, M., Selim, G. M. K., and Chechik, M. (2020). Towards requirements specification for machine-learned perception based on human performance. In *2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, pages 48–51. IEEE.

Ishikawa, F. and Yoshioka, N. (2019). How do engineers perceive difficulties in engineering of machine-learning systems?-questionnaire survey. In *2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)*, pages 2–9. IEEE.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE-0401, Keele University, UK.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., and Linkman, S. (2010). Systematic literature reviews in software engineering – a tertiary study. *Information and Software Technology*, 52(8):792–805.

Kucak, D., Juricic, V., and Dambic, G. (2018). Machine learning in education: A survey of current research trends. In *Proceedings of the 29th DAAAM International Symposium on Intelligente Manufacturing and Automation*, volume 29, pages 406–410. DAAAM International Vienna.

Kudo, T. N., Bulcão-Neto, R. F., and Vincenzi, A. M. (2020a). Requirement patterns: a tertiary study and a research agenda. *IET Software*, 14(1):18–26.

Kudo, T. N., Bulcão-Neto, R. F., and Vincenzi, A. M. (2020b). Requirement patterns: a tertiary study and a research agenda. *IET Software*, 14(1):18–26.

Kumeno, F. (2020). Software engineering challenges for machine learning applications: A literature review. *Intelligent Decision Technologies*, 13(4):463–476.

Kuwajima, H., Yasuoka, H., and Nakae, T. (2020). Engineering problems in machine learning systems. *Machine Learning*, 109(5):1103–1126.

Lakha, B., Bhetwal, K., and Eisty, N. U. (2023). Analysis of software engineering practices in general software and machine learning startups. In *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE.

Lewis, G. A., Bellomo, S., and Ozkaya, I. (2021). Characterizing and detecting mismatch in machine-learning-enabled systems. In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 133–140. IEEE.

Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A. M., and Wagner, S. (2022). Software engineering for ai-based systems: A survey. *ACM Trans. Softw. Eng. Methodol.*, 31(2).

Nahar, N., Zhou, S., Lewis, G., and Kästner, C. (2022). Collaboration challenges in building ML-enabled systems: communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, page 413–425, New York, NY, USA. Association for Computing Machinery.

Villamizar, H., Escovedo, T., and Kalinowski, M. (2021). Requirements engineering for machine learning: A systematic mapping study. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 29–36. IEEE.

Vogelsang, A. and Borg, M. (2019). Requirements engineering for machine learning: Perspectives from data scientists. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 245–251.

Washizaki, H. (2024). Guide to the software engineering body of knowledge (swebok guide), version 4.0. *IEEE Computer Society, Waseda University, Japan*.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, New York, NY, USA. Association for Computing Machinery.

Yoshioka, N., Husen, J. H., Tun, H. T., Chen, Z., Washizaki, H., and Fukazawa, Y. (2021). Landscape of requirements engineering for machine learning-based AI systems. In *Proceedings of the 28th Asia-Pacific Software Engineering Conference Workshops (APSEC Workshops 2021)*, pages 5–8. IEEE.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., et al. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Engineering Bulletin*, 41(4):39–45.