Requirements Engineering for Machine Learning: A Systematic Mapping Study

Hugo Villamizar, Tatiana Escovedo, Marcos Kalinowski
Software Engineering Laboratory, Department of Informatics
Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
Rio de Janeiro, Brazil
Emails:{hvillamizar, tatiana, kalinowski}@inf.puc-rio.br

***Abstract*—Machine learning (ML) has become a core feature for today's real-world applications, making it a trending topic for the software engineering community. Requirements Engineering (RE) is no stranger to this and its main conferences have included workshops aiming at discussing RE in the context of ML. However, current research on the intersection between RE and ML mainly focuses on using ML techniques to support RE activities rather than on exploring how RE can improve the development of ML-based systems. This paper concerns a systematic mapping study aiming at characterizing the publication landscape of RE for ML-based systems, outlining research contributions and contemporary gaps for future research. In total, we identified 35 studies that met our inclusion criteria. We found several different types of contributions, in the form of analyses, approaches, checklists and guidelines, quality models, and taxonomies. We discuss gaps by mapping these contributions against the RE topics to which they were contributing and their type of empirical evaluation. We also identified quality characteristics that are particularly relevant for the ML context (e.g., data quality, explainability, fairness, safety, and transparency). Main reported challenges are related to the lack of validated RE techniques, the fragmented and incomplete understanding of NFRs for ML, and difficulties in handling customer expectations. There is a need for future research on the topic to reveal best practices and to propose and investigate approaches that are suitable to be used in practice.**

***Index Terms*—requirements engineering, machine learning, systematic mapping study**

I. Introduction

Machine Learning (ML) is the study of computer algorithms that improve automatically through experience [42]. Its purpose is to build a model based on sample data, known as training data. This kind of systems, unlike traditional software systems, base its behavior on data from the external world instead of explicitly programming hard rules. ML components are often developed by data scientists who typically lack foundations to build reliable software systems [30]. On the other hand, the paradigm shift that makes data, to some extent, replace code, supposes a change in the way of designing, developing and testing this type of systems.

This is challenging from the point of view of Software Engineering (SE). For example, data should be tested and models be validated just as thoroughly as code, but there is, currently, a lack of best practices on how to do so [5]. In that sense, Requirements Engineering (RE) plays an important role in the development of ML-based systems since a machine-learned model is a specification that is built based on training data, that is, a learned description of how the system shall behave. In line with this argument, Kästner [26] stated that ML corresponds to the RE phase of a project rather than the implementation phase. In addition, he also argued that ML should worry about its validation, typically associated with RE, instead of verification, that is, whether the model has learned the right specification. This means that the learned behavior of a ML-based system might be incorrect, even if the learning algorithm is implemented correctly, a situation in which traditional testing techniques are ineffective.

The intersection of RE and ML has been studied in recent years by the RE community [15] and discussed in renown SE conferences such as RE, REFSQ, ESEC/FSE, ICSE, and ESEM. However, current research on this intersection focuses on using ML techniques to support RE activities rather than on exploring how RE can improve the use of these techniques in the entire software development process [53]. ML can benefit from the RE perspective even from studies suggesting that RE is the most difficult activity for the development of ML-based systems [24], [34]. For instance, RE techniques for ML could help in identifying quality metrics beyond accuracy, to allow better dealing with customer expectations, understanding why models do not fit and for whom they do not fit, which data is missing and how the data is analyzed and generalized.

In response to the importance and benefits that RE can offer to the development of ML-based systems, we contribute a first step in synthesizing existing work on RE for ML. In particular, we report on a systematic mapping study with the research objective of outlining the state of the art of RE for ML-based systems. We aim at characterizing RE contributions in terms of RE topics, quality characteristics, challenges, research directions, research type facets and empirical evaluations.

We found that the main RE contributions are in the form of approaches that address different RE activities, quality models, analysis of unique characteristics of RE for ML, taxonomy of problems, checklists and guidelines to support requirements engineers. We identified that requirements elicitation and requirements analysis are the main RE activities addressed by such contributions. We also identified unique quality characteristics (e.g., explainability, fairness, transparency and ethics) and RE challenges to ML (e.g., how to deal with stakeholder expectations, aligning data with the business goals and how to

978-1-6654-2705-0/21/$31.00 ©2021 IEEE
DOI 10.1109/SEAA53835.2021.00013

properly cover and validate requirements).

The remainder of this paper is organized as follows. Section II provides the background and an overview on related work. Section III describes the mapping study protocol and how it was applied. Section IV presents the mapping study results. Section V discusses the results. Section VI describes the threats to the validity of our study. Finally, Section VII presents the concluding remarks.

II. Background and Related Work

RE research is characterized by the involvement of interdisciplinary stakeholders and uncertainty [54]. Hence, RE is highly volatile and inherently complex by nature [17]. On the other hand, the use of ML-based systems has grown considerably in recent years resulting in increasing demands for high quality for such applications. This tacitly involves RE since it plays a critical role for addressing software quality characteristics. RE for ML-based systems faces unique difficulties as it constitutes a paradigm shift compared to conventional software development. As a response, a couple of software process models for ML have emerged [3], [41], [59]. They all have similar stages, starting with requirements and ending with quality assurance. However, the vision of RE is limited and superficial.

Some research has pointed out problems and challenges related to RE for ML [9], [24], [34]. Others have proposed approaches, methods and frameworks to address different concerns [23], [49], [58]. Despite the important contributions in the field so far, the synergy of RE for ML has not been so comprehensively studied [21], [53]. For instance, researchers would benefit from more studies about how RE is addressed in practice in order to propose new contributions. Practitioners, in turn, would benefit from those contributions in order to improve the development of ML-based systems.

To the best of our knowledge, we are aware of only two mapping studies that somehow relate to RE for ML. Schuh *et al.* [51] conducted a systematic literature review aiming at identifying design patterns, data model requirements, and technology potentials for ML systems in manufacturing companies. The authors found data characteristics such as quantity, quality and dimensionality. On the other hand, Borg *et al.* [10] conducted a review of verification and validation for ML in the automotive industry. The authors identified challenges such as transparency and requirements specification. However, we consider that the scope of these studies is significantly different from ours since it only addresses RE partially and limits its scope to one specific industrial sector.

Nevertheless, in order to broaden our vision, we also covered secondary studies concerning SE for ML that shed some light on RE as part of our related work. Kumeno *et al.* [32] and Lorenzoni *et. al* [38] surveyed the literature in order to outline the SE challenges that emerge during the development of ML systems. Regarding RE, the authors found that software requirements activities for ML applications involve activities such as data and feasibility analysis, requirements elicitation, requirement specification, and validation and performance evaluation of ML-models. We know that pointing out these insights is important, however, we consider the coverage of these studies insufficient from the RE perspective since there are other interesting aspects to review in the literature, such as proposed scientific contributions and their evaluation.

Nascimento *et al.* [46] conducted a systematic literature review in order to investigate how SE has been applied in the development of ML systems and identified challenges and practices that are applicable. Somehow related to RE, they identified which practices data scientists use to improve the quality of project data such as cross-validation, checking data distribution, and checking implicit constraints.

In summary, these studies agree that RE is a challenging task in SE for ML, but do not focus on synthesizing existing work on RE for ML. To address this gap, our systematic mapping study aims at providing an overview of research contributions regarding RE for ML.

III. Systematic Mapping Protocol

Systematic Mapping (SM) studies are designed to provide a wide overview of a research area, to establish if research evidence exists on a topic and provide an indication of the quantity of the evidence [31].

The SM study was performed following the guidelines proposed by Kitchenham and Charters [31] and the SM-specific guidelines by Petersen *et al.* [48]. After identifying the need for the review (*cf.* Section II), we define the research questions, search strategy and inclusion/exclusion criteria.

A. Research Objectives and Questions

The main research objective is **to outline the state of the art of RE for ML-based systems**. The following research questions were derived from the objective in order to further characterize the RE contributions.

RQ1. What RE contributions have emerged to support the software development of ML-based systems? This question aims at providing a general overview of RE contributions (e.g., approaches, quality models, checklists) that have been proposed for ML.

RQ2. What RE topics do the contributions address? The aim of this question is to identify the specific RE topics that were the focus of the contributions (*cf.* Table III), helping to further understand their purpose.

RQ3. What quality characteristics do the RE contributions consider for ML-based systems? There is a consensus in the SE community that the quality of ML-based systems must go beyond metrics such as accuracy [29]. This question aims at pointing out concerns about quality in ML systems.

RQ4. What are the reported challenges and research directions on the interplay between RE and ML-based systems? This question aims at identifying open challenges. One of the main reasons to conduct a SM is supporting the planning of new research. Thus, this question seeks to indicate the aspects that may be studied by other researchers.

RQ5. What are the research type facets of the contributions? The purpose of this question is to classify the papers according to their research type facets. We adopt the classification scheme by Wieringa *et al.* [57].

RQ6. Which kind of empirical evaluations have been performed to assess the contributions? The purpose of this question is to identify what types of empirical studies have been conducted, focusing on the research type facets of evaluation and validation research from the previous question. Obtaining this information allows to get a first idea on the scientific rigour of the evidence reported in the field.

While questions RQ1-RQ4 aim at structuring the publication landscape in a conceptual manner, the last two shall provide insights into the nature of the current reported evidence.

B. Search Strategy

The mapping study employed a hybrid search strategy [43] that involves conducting a search string-based database search on a specific digital library (Scopus) and then complementing the set of identified papers with iterative backward and forward snowballing (using Google Scholar) following the guidelines by Wohlin [60]. We intentionally refrained from using various specific libraries, given that the chosen hybrid strategies typically achieve an appropriate balance of precision and recall [43]. Between the different hybrid strategies (sequential, parallel, and iterative) [43], we chose the more complete iterative snowballing for maximizing the recall, even though it would imply in analyzing more papers [61]. Iterative backward and forward snowballing concerns applying backward and forward snowballing on each new included paper.

We chose Scopus because it claims to be the largest database of titles and abstracts [31], which would allow us to identify a representative and unbiased seed set [43]. It is, however, backward and forward snowballing via Google Scholar which we used as an effective way to complement the identification of the broader population of studies [43], [60].

We formulated the search string to conduct the initial database search on Scopus using the PICO (*Population, Intervention, Comparison, Outcome*) criteria [35] strategy. Our study focuses on ML-based systems (*population*) and aims at identifying RE contributions for such systems (*intervention*). As our study concerns a mapping study, there was no specific *comparison* nor the need of limiting the search space regarding *outcomes*. Therefore, we needed keywords for ML and RE. The defined search string, to be applied on titles, abstracts and keywords was: *"(Software OR Applications OR Systems) AND (Machine Learning) AND (Requirements Engineering)"*.

C. Study Selection

Following suggestions by Mendes *et al.* [40], we initially defined a well limited intended timeframe for our mapping study, comprising papers published by the end of 2020. The primary inclusion criteria was on papers that describe RE contributions in the context of ML. When several papers reported the same study, only the most recent one was included. When multiple studies were reported in the same paper, each study was considered separately. The exclusion criteria applied for filtering the papers are shown in Table I.

TABLE I
EXCLUSION CRITERIA.

Criteria	Description
EC1	Papers that do not meet the inclusion criteria
EC2	Papers about the use of ML techniques for improving RE activities
EC3	Papers not written in English
EC4	Grey literature, including blogs, white papers, theses, and papers that were not peer reviewed
EC5	Papers that are only available in the form of abstracts/posters and presentations

Fig. 1 shows all the steps performed in the paper selection process. The database search results on Scopus, filters, and backward and forward snowballing are detailed below.

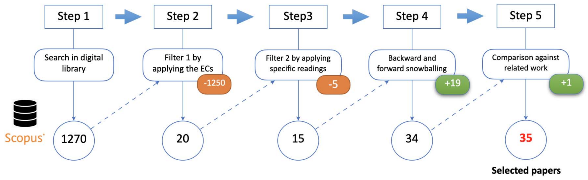
Step 1
Step 2
Step3
Step 4
Step 5
Search in digital library
Filter 1 by applying the ECs
Filter 2 by applying specific readings
Backward and forward snowballing
Comparison against related work
-1250
-5
+19
+1
Scopus
1270
20
15
34
35
Selected papers

Fig. 1. Papers selection process.

The first step consisted of searching for papers using the search string in the digital library selected for this study. The search string was applied on titles, abstracts and keywords in Scopus in January 2021, and returned 1270 papers. In the second step (Filter 1), the first filtering took place. In this step, we applied the exclusion criteria. Regarding EC1, at this step we excluded the papers that clearly didn't have information on RE for ML-based systems. We identified that a substantial number of papers (175) concern EC2. This confirms that the intersection between RE and ML is predominant for papers that use ML to support RE activities. As a result, we reduced our set of candidate papers to 20.

In the third step, we applied a second filter (Filter 2), filtering papers by reading the titles, abstracts and selected paper parts (when necessary) while applying the inclusion and exclusion criteria. This step left us with 15 papers representing the result of the search on Scopus. All exclusions and the final set of included papers were peer reviewed by an independent researcher. In case of divergence, a third researcher was involved and a discussion was held to reach consensus.

In the next step, during the month of February, we applied backward and forward snowballing iteratively following the snowballing guidelines in [60]. In total, four backward snowballing (BS) and two forward snowballing (FS) iterations were applied (order: BS1, BS2, FS1, BS3, BS4, FS2) until reaching our final set of papers. The four backward snowballing iterations involved analyzing 624 (259 + 200 + 131 + 34) papers (including duplicates) and allowed identifying twelve additional papers to be included. The two forward snowballing iterations involved analyzing 304 (290 + 14) papers (including duplicates) and allowed identifying seven additional papers to be included. The whole snowballing process was peer reviewed. It is noteworthy that the first forward snowballing iteration retrieved a paper accepted for publication in 2020, but published January 1st 2021 [45]. As this paper was on the borderline of our scoped time frame, but represents a valuable contribution, we decided to also include it in our mapping. This explains the single paper from 2021.

Finally, in the fifth step, we compared our results against the results provided by the related work (*cf.* Section II). We found only one paper [8] that was not identified by our search strategy, because this paper didn't cite any of the remaining studies on the topic. Hence, in total, 35 papers were included in the SM study, where 15 papers came from Scopus, 19 papers came from snowballing and one paper came from analyzing related work as shown in Fig. 1. The selected papers are shown in Table III. A spreadsheet with all details on the filtering and snowballing process, documenting each iteration, can be found in our online open science repository [1].

D. Data Extraction and Classification Scheme

The information extracted from each of the selected papers and the classification schemes describing the different categories are presented in Table II. The complete extracted data is also available in our online open science repository.

TABLE II
DATA EXTRACTION FORM.

Information	Description
Study Metadata	Includes the paper title and information such as venue, type of venue and year of publication.
RE contribution (RQ1)	Description of the RE contribution for ML.
RE topics (RQ2)	RE topics that the contribution addresses. These topics were coded based on typical RE activities (*e.g.*, elicitation, analysis, modeling, specification, validation, verification, and management) or other RE aspects (*e.g.*, data quality requirements, requirements assurance) that were the focus of the contributions.
Quality characteristic (RQ3)	Characteristic that influences the quality of ML-based systems. It often refers to Non-Functional Requirements (NFRs) (*e.g, safety, explainability, performance*).
RE problems (RQ4)	Fact or situation that requires an action by RE researchers.
Research directions (RQ4)	RE topics that previous studies point out to research.
Research type facet (RQ5)	Classification of research type facets according to Wieringa *et al.* [57], including the following categories: *evaluation research, solution proposal, philosophical paper, opinion paper, or experience paper*.
Empirical evaluation (RQ6)	Classification of the empirical strategy, according to Wohlin *et al.* [62], including the following categories: *experiment, case study, survey*.

IV. SYSTEMATIC MAPPING RESULTS

This section presents the results of the SM study. First, we provide an overview of the included papers. Overall, we identified 35 papers. Regarding the years of publication, the papers range from 2018 to 2021. Most of the publications (31) are conference and workshop papers and only 4 papers have been published in journals. The venues in which the topic has been addressed comprise premier international SE conferences and journals such as FSE, ICSE, RE, ESEM, REJ, and TSE. This gives an idea of the relevance and interest on this topic on behalf of the SE community in the last years.

A. RQ.1 What RE contributions have emerged to support the development of ML-based systems?

Similar to our previous mapping study in the field of RE [52], we followed open coding guidelines [50] with the aim of characterizing the papers by the type of contribution. We coded the following different main contribution types for the papers: analyses (e.g., analyzing some RE aspects for ML), approaches (e.g., methods, methodologies, processes, and conceptual frameworks), checklists and guidelines (C & G), quality models (QM), and taxonomies (T). Table III shows an overview of these contributions by contribution type.

TABLE III
IDENTIFIED CONTRIBUTIONS.

Type	Id	Short description
Analyses	P3 [29]	Teaching Software Engineering for AI-Enabled Systems
	P4 [37]	Emerging and changing tasks for developing ML systems
	P11 [53]	Perspectives from data scientists
	P14 [21]	Challenges and new directions of NFRs for ML
	P16 [33]	How to adapt SQuaRE for ML-based AI systems
	P17 [24]	How engineers perceive difficulties in engineering ML systems
	P20 [55]	How does ML change software development practices?
	P21 [56]	Studying SE patterns for designing ML systems
	P24 [7]	Approaches for requirements assurance in traditional safety-related software
	P25 [12]	The importance of requirements for DL and the wisdom of requirements quality
	P28 [47]	Challenges of ML applied to safety-critical cyber-physical systems
	P33 [39]	Using conceptual modeling to support ML
	P34 [14]	NFRs orienting the development of socially responsible software
Approaches	P2 [23]	Approach for evidence-driven RE to handle uncertainty in ML
	P6 [2]	Conceptual framework for ML model lifecycle management
	P8 [11]	Approach to software metrics for ML systems
	P9 [13]	Method for identifying stakeholders needs
	P12 [49]	Approach to improve requirements specification in ML systems
	P15 [1]	Methodology to guide the development of ML systems
	P18 [22]	Approach for specifying and testing requirements for robustness based on human perception
	P22 [18]	Method for understanding XAI requirements in ML systems
	P23 [36]	Method that contains a XAI question bank to bridge the spaces of user needs for AI explainability
	P26 [25]	Method for dataset augmentation to improve neural networks by satisfying the customer's requirements
	P27 [45]	Conceptual framework for requirements elicitation, design, and development of ML solutions.
	P29 [58]	Methodology for security requirements elicitation in ML systems
	P31 [4]	Methodology for the evaluation of non-functional properties in ML
	P32 [20]	Process for developing data-driven applications
	P35 [8]	Methodology for bridging the gap between ML and business goals
C & G	P7 [19]	Guidelines for quality assurance of ML-based AI
	P13 [16]	Checklist to support business modeling
	P19 [3]	Best practices with ML in SE that could be seen as requirements
	P30 [6]	Ethical guidelines for developing AI systems
QM	P1 [44]	Determining quality characteristics and measurements for ML
	P7 [19]	Guidelines for quality assurance of ML-based AI
T	P5 [34]	Engineering problems in ML systems
	P10 [9]	RE Challenges in Building AI-Based Complex Systems

B. RQ2. Which RE topics do the contributions address?

The majority of the selected papers concern requirements elicitation practices (14 out of 35), where authors consider problems such as defining business goals and problems of understanding. Furthermore, we found five papers about requirements analysis, more specifically addressing customer

[1] https://doi.org/10.5281/zenodo.4682374

expectations and requirements prioritization. We also identified contributions regarding data related requirements in five papers. Other contributions regard requirements specification, assurance (typically related to quality model based software product requirements evaluation), modeling, verification of documented requirements, and validation. Fig. 2 shows the distribution of the papers by the covered RE topics.

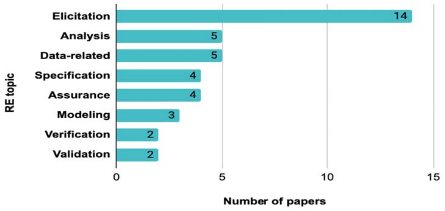

Fig. 2. Distribution of papers per RE topics.

C. RQ3. What quality characteristics do the RE contributions consider for ML-based systems?

During the analysis of the contributions, it was possible to identify several quality requirements aka non-functional requirements (NFRs) that authors consider in their research. Table IV shows the quality characteristics that were considered in the papers with their frequencies. Note that one paper can address one or more quality characteristics.

TABLE IV
FREQUENCY OF QUALITY CHARACTERISTICS.

Characteristic	Frequency	Characteristic	Frequency
Security	6	Testability	2
Explainability	6	Accountability	2
Privacy	6	Ethics	2
Data quality	5	Accuracy	2
Fairness	5	Suitability	1
Transparency	5	Uncertainty	1
Reliability	4	Autonomy	1
Safety	4	Robustness	1
Performance	3	Modularity	1
Maintainability	3	Scalability	1
Legal requirements	2	Usability	1

D. RQ4. What are the reported challenges and research directions on the interplay between RE and ML-based systems?

Some papers explicitly report challenges from the point of view of RE when developing ML-based systems. We grouped them in order to provide a better understanding and then outline the challenges. A brief overview is summarized below.

Lack of validated techniques. Developing ML-components mainly relies on applying techniques to achieve an objective. However, there seems to be a lack of validated techniques for some important aspects of RE for ML. For instance, several studies (e.g., [P5][P8][P12][P13][P14]) state that ML researchers and users currently lack an ML-specific way to express and specify requirements for ML, including targets and trade-offs, and the influence of domain context. Other studies, such as [P1][P3][P5], mention that measuring quality beyond traditional metrics, such as accuracy and precision, may be complicated since identifying quality attributes is often difficult. The authors of [P13] also outline that identifying business metrics is not trivial since customers want to have policies to improve their business, but do not understand what metrics and data are required to do so. This represents a challenge for requirements engineers of ML-based systems. In addition, in [P28][P35] the authors raise issues on how to properly cover and validate requirements for ML systems and how to deal with testing and verification activities.

Knowledge regarding NFRs. In [P14] the authors state that the understanding of NFRs for ML is fragmented and incomplete, including how to define and refine NFRs in ML-specific contexts. Quality attributes such as explainability ([P22][P23][P33]), safety ([P3]), security ([P18]), fairness ([P3]), robustness ([P5]), and transparency ([P19]) are pointed out as challenging by researchers.

Handling customer expectations. Organizations did not realize that ML models are mainly probabilistic models that commonly have to learn patterns from messy data. This reflects difficulties customers have to understand potential limitations of ML systems. Papers such as [P7][P8][P13][P17] reveal that customers commonly expect to see magic coming out of data.

Furthermore, we wanted to know what research directions are encouraged by the authors. After analyzing the papers, we identified that the authors are mainly asking the community to conduct more empirical studies to uncover more insights on best practices and to propose and investigate approaches that are suitable to be used in practice. The authors also mention other research directions, such as:

- Address transparency, explainability and safety for ML.
- Develop tools to support requirements specification
- Create guidelines related to ML-based system requirements (training data, specification documents, test design, runtime monitoring, and maintenance).
- How to operationalize ethics, security, and privacy.
- How to verify ML requirements and validate ML models.
- Extend the ML quality characteristics.
- Develop standard quality models for ML-based systems.
- Survey ML literature and/or ML experts on NFRs.
- Understand how ML systems integrate with typical software from a quality perspective.
- Provide guidance to non-technical stakeholders about what is possible and what is not.

E. RQ5. What are the research type facets of the approaches?

Fig. 3 shows the distribution of the research type facets of the papers per year. It is possible to observe that most of the papers (16 out of 35) concern evaluation research papers. In the next question we address the types of empirical evaluations that were conducted. Opinion papers, with ten studies, significantly contribute to the account. We also identified that solution proposals are still scarce in this field. This contrasts the identified lack of techniques and the absence of tools supporting RE activities for ML-based systems, further motivating research directions pointed out by the authors.

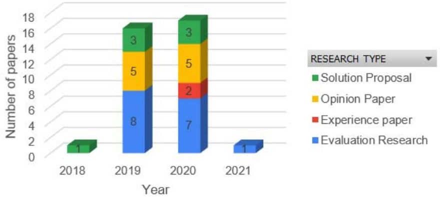

Fig. 3. Distribution of research type per year.

F. RQ6. Which kind of empirical evaluations have been performed?

When analyzing the empirical evaluations conducted within the studies (Fig. 4), it was possible to identify 16 papers out of 35 that have performed empirical evaluations (twelve case studies, three surveys and one experiment). Note that five papers provided a proof of concept, i.e., a realization of a certain method or idea in order to demonstrate its feasibility. This is not considered as an empirical evaluation by Wohlin *et al.* [62], therefore they were not classificated as evaluation research. In fact, 14 studies did not contain any type of empirical evaluation or even a proof of concept. Most of these concern opinion and experience papers. The most applied empirical evaluation strategy in the analyzed studies was case study (12 papers) in contrast with survey (three papers) and experiment (one paper).

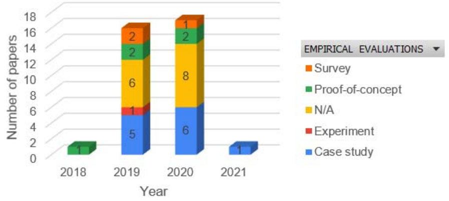

Fig. 4. Distribution of empirical evaluation type per year.

V. Discussion

Several different research contributions have been proposed recently with regard to RE for ML (*cf.* Table III). These contributions comprise analyses, approaches, checklists, guidelines, quality models, and taxonomies. Fig. 5 presents a bubble plot mapping the identified contribution types against the covered RE topics and the type of empirical evaluation that has been conducted. It is evident that there are still relevant gaps to cover.

The main identified challenges concern the lack of validated techniques to address the identified gaps, a lack of knowledge regarding specific NFRs that are particularly relevant in this context (e.g., explainability, fairness, safety, transparency), and difficulties in handling customer expectations. We listed several examples of research directions to handle these challenges and other opportunities reported by the authors.

With respect to the kind of research and empirical evidence, we observed that there is still a limited amount of solution proposals and that 14 out of 35 (40%) papers did not contain any type of evaluation (not even a proof of concept). This could be related to the fact that it is a recent topic, which is still to acquire maturity and to move towards more rigorously assessed empirical evidence. Nevertheless, we have relatively little evidence about the feasibility of the presented approaches, which represents a problem for practitioners and an opportunity for researchers.

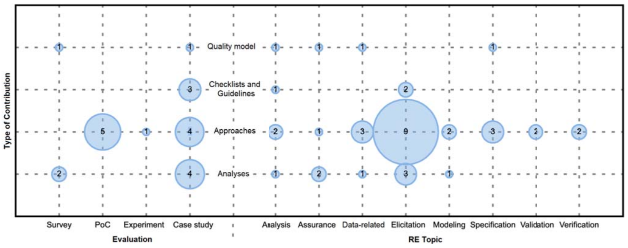

Fig. 5. Contribution types, RE topics, and empirical evaluation types.

An intriguing fact is that almost all papers, except [P25][P26], have the conviction that their proposed contributions are rationally applicable for all kind of ML approaches (e.g., supervised and unsupervised learning, reinforcement learning) and models (e.g., artificial neural networks, decision trees, support vector machines). This gives an idea that within the ML field there are no particularities that may affect the generalization of the studies. Based on our experiences developing ML-based systems [27], [28], we believe that this might not hold and should be further investigated.

Another interesting fact is that only two papers [P13][P21] investigate what processes are used in practice. In [P13] the authors identified that problem understanding in the observed RE for ML context involved studying available data and meeting with customers. For requirements elicitation and specification they found that practitioners typically conduct brainstorming sessions and define business metrics. In [P21] the authors identified that practitioners typically use ad-hoc methods to address requirements elicitation and NFRs assurance.

The importance of RE to design and successfully deliver a ML-based system is clear. However, this work shows that this area faces several problems that need to be addressed both in industry and research. For instance, today it is not clear what tools and methods are used in practice to address requirements for ML-based systems. It is not clear if traditional requirements engineering tools and techniques work for ML, and even worse, it is not known how requirements should be handled in the ML context. We strongly believe that these problems represent interesting research opportunities.

VI. Threats to Validity

Internal validity: We used a hybrid search strategy combining a database search on a single database (Scopus) with iterative backward and forward snowballing (using Google Scholar), and precisely documented each step. One could argue that the search string was confined to a small set of keywords. These keywords were objectively selected using the PICO strategy and are directly related to our research goal. It is important to remember that the database search was used to reveal an unbiased and representative seed set, as starting point for iterative forward and backward snowballing, and that this search strategy has been effective for secondary studies [43].

External validity: We systematically applied a search strategy that has shown good results regarding recall [43] and validated it by comparing it against related work. Still, there is a possibility of having missed studies. Nevertheless, we were unable to manually find any additional study to be included and are confident that we have an unbiased and representative sample. The claims made in our paper are related to the findings reported in the primary studies. While all of them were peer reviewed, and many of them were published in venues that have a rigorous selection process, we did not assess their quality. Such quality assessment is typically not part of mapping studies, and could be part of a systematic review extension. The complete information concerning the process, the extracted data and coding is available in our online repository and is publicly auditable.

Reliability: In order to reduce the bias when selecting relevant studies, it was decided to examine the selected papers in pairs. Hence, two researchers evaluated the selected studies, extracted data and coding in a peer-reviewed manner.

VII. Concluding Remarks

This paper presents the results of a SM study on RE in the context of ML-based systems. We applied a hybrid search strategy, complementing a database search on Scopus with iterative backward and forward snowballing. Our search strategy allowed identifying a total of 35 studies.

We identified several proposed research contributions, some published in premier software engineering conferences and journals. These contributions comprise analyses, approaches, checklists and guidelines, quality models, and taxonomies. We identified research gaps by relating these contributions to RE investigation topics. We also highlighted quality characteristics considered within the papers and reported on challenges and potentially promising research directions.

Hence, the main contributions of this research are twofold: (i) mapping relevant knowledge about the current state of RE for ML, a subject that is not yet widely explored by researchers and confused by practitioners; and (ii) helping to identify points that still require further investigation. As far as we know, this paper is the first systematic mapping study that organizes evidence to provide a comprehensive overview on contributions related to RE for developing ML-based systems.

References

[1] B. Ahmed, T. Dannhauser, and N. Philip, "A lean design thinking methodology (ldtm) for machine learning and modern data projects," in *2018 10th Computer Science and Electronic Engineering (CEEC)*, 2018, pp. 11–14.

[2] R. Akkiraju, V. Sinha, A. Xu, J. Mahmud, P. Gundecha, Z. Liu, X. Liu, and J. Schumacher, "Characterizing machine learning processes: A maturity framework," in *International Conference on Business Process Management (BPM)*, 2020, pp. 17–31.

[3] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 2019, pp. 291–300.

[4] M. Anisetti, C. A. Ardagna, E. Damiani, and P. G. Panero, "A methodology for non-functional property evaluation of machine learning models," in *Proceedings of the 12th International Conference on Management of Digital EcoSystems (MEDES)*, 2020, pp. 38–45.

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

[6] N. Balasubramaniam, M. Kauppinen, S. Kujala, and K. Hiekkanen, "Ethical guidelines for solving ethical issues and developing ai systems," in *International Conference on Product-Focused Software Process Improvement (PROFES)*, 2020, pp. 331–346.

[7] A. Banks and R. Ashmore, "Requirements assurance in machine learning." in *The AAAI's Workshop on Artificial Intelligence Safety (SafeAI)*, 2019.

[8] G. Barash, E. Farchi, I. Jayaraman, O. Raz, R. Tzoref-Brill, and M. Zalmanovici, "Bridging the gap between ml solutions and their business requirements using feature interactions," in *Proceedings of the Joint Meeting of the European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2019, pp. 1048–1058.

[9] H. Belani, M. Vukovic, and Ž. Car, "Requirements engineering challenges in building ai-based complex systems," in *International Requirements Engineering Conference Workshops (REW)*, 2019, pp. 252–255.

[10] M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist, "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *Journal of Automotive Software Engineering*, vol. 1, pp. 1–19, 2019. [Online]. Available: https://doi.org/10.2991/jase.d.190131.001

[11] N. Caporusso, T. Helms, and P. Zhang, "A meta-language approach for machine learning," in *International Conference on Applied Human Factors and Ergonomics*, 2019, pp. 192–201.

[12] H. Challa, N. Niu, and R. Johnson, "Faulty requirements made valuable: On the role of data quality in deep learning," in *7th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, 2020, pp. 61–69.

[13] D. Cirqueira, D. Nedbal, M. Helfert, and M. Bezbradica, "Scenario-based requirements elicitation for user-centric explainable ai," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2020, pp. 321–341.

[14] L. M. Cysneiros and J. C. S. do Prado Leite, "Non-functional requirements orienting the development of socially responsible software," in *Enterprise, Business-Process and Information Systems Modeling*. Springer, 2020, pp. 335–342.

[15] F. Dalpiaz and N. Niu, "Requirements engineering in the days of artificial intelligence," *IEEE Software*, vol. 37, no. 4, pp. 7–10, 2020.

[16] E. de Souza Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, and T. Conte, "Understanding development process of machine learning systems: Challenges and solutions," in *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2019, pp. 1–6.

[17] D. M. Fernández, S. Wagner, M. Kalinowski, M. Felderer, P. Mafra, A. Vetrò, T. Conte, M.-T. Christiansson, D. Greer, C. Lassenius *et al.*, "Naming the pain in requirements engineering," *Empirical software engineering*, vol. 22, no. 5, pp. 2298–2338, 2017.

[18] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, and A. Preece, "A systematic method to understand requirements for explainable ai (xai) systems," in *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China*, 2019.

[19] K. Hamada, F. Ishikawa, S. Masuda, M. Matsuya, and Y. Ujita, "Guidelines for quality assurance of machine learning-based artificial intelligence," in *International Conference on Software Engineering & Knowledge Engineering (SEKE)*, 2020, pp. 335–341.

[20] M. Hesenius, N. Schwenzfeier, O. Meyer, W. Koop, and V. Gruhn, "Towards a software engineering process for developing data-driven applications," in *International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*, 2019, pp. 35–41.

[21] J. Horkoff, "Non-functional requirements for machine learning: Challenges and new directions," in *International Requirements Engineering Conference (RE)*, 2019, pp. 386–391.

[22] B. C. Hu, R. Salay, K. Czarnecki, M. Rahimi, G. Selim, and M. Chechik, "Towards requirements specification for machine-learned perception based on human performance," in *7th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, 2020, pp. 48–51.

[23] F. Ishikawa and Y. Matsuno, "Evidence-driven requirements engineering for uncertainty of machine learning-based systems," in *International Requirements Engineering Conference (RE)*, 2020, pp. 346–351.

[24] F. Ishikawa and N. Yoshioka, "How do engineers perceive difficulties in engineering of machine-learning systems?-questionnaire survey," in *International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)*, 2019, pp. 2–9.

[25] B. Jahic, N. Guelfi, and B. Ries, "Specifying key-properties to improve the recognition skills of neural networks," 2020.

[26] C. Kaestner, "Machine learning is requirements engineering—on the role of bugs, verification, and validation in machine learning," *Medium post, Accessed April*, vol. 25, 2020.

[27] M. Kalinowski, S. T. Batista, H. Lopes *et al.*, "Towards lean r&d: an agile research and development approach for digital transformation," in *Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2020, pp. 132–136.

[28] M. Kalinowski, H. Lopes, A. F. Teixeira *et al.*, "Lean r&d: An agile research and development approach for digital transformation," in *Product-Focused Software Process Improvement (PROFES) 2020, Turin, Italy, November 25-27, 2020. Proceedings*, 2020, pp. 106–124.

[29] C. Kästner and E. Kang, "Teaching software engineering for al-enabled systems," in *International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, 2020, pp. 45–48.

[30] M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "Data scientists in software teams: State of the art and challenges," *IEEE Transactions on Software Engineering*, vol. 44, no. 11, pp. 1024–1038, 2017.

[31] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *Keele University and Durham University Joint Report, Technical Report EBSE 2007-001*, 2007.

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

[33] H. Kuwajima and F. Ishikawa, "Adapting square for quality assessment of artificial intelligence systems," in *International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 2019, pp. 13–18.

[34] H. Kuwajima, H. Yasuoka, and T. Nakae, "Engineering problems in machine learning systems," *Machine Learning*, vol. 109, no. 5, pp. 1103–1126, 2020.

[35] R. Leonardo, "Pico: Model for clinical questions," *Evidence Based Medicine and Practice*, vol. 3, no. 115, p. 2, 2018.

[36] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.

[37] H. Liu, S. Eksmo, J. Risberg, and R. Hebig, "Emerging and changing tasks in the development process for machine learning systems," in *Proceedings of the International Conference on Software and System Processes (ICSSP)*, 2020, pp. 125–134.

[38] G. Lorenzoni, P. Alencar, N. Nascimento, and D. Cowan, "Machine learning model development from a software engineering perspective: A systematic literature review," *arXiv preprint arXiv:2102.07574*, 2021.

[39] R. Lukyanenko, A. Castellanos, J. Parsons, M. C. Tremblay, and V. C. Storey, "Using conceptual modeling to support machine learning," in *International Conference on Advanced Information Systems Engineering (CAISE)*, 2019, pp. 170–181.

[40] E. Mendes, C. Wohlin, K. Felizardo, and M. Kalinowski, "When to update systematic literature reviews in software engineering," *Journal of Systems and Software*, vol. 167, p. 110607, 2020.

[41] Microsoft, "The team data science process," https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/, accessed: 2020-12-27.

[42] T. M. Mitchell *et al.*, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[43] E. Mourao, J. F. Pimentel, L. Murta, M. Kalinowski, E. Mendes, and C. Wohlin, "On the performance of hybrid search strategies for systematic literature reviews in software engineering," *Information and Software Technology*, vol. 123, pp. 106 294:1–12, 2020.

[44] K. Nakamichi, K. Ohashi, I. Namba, R. Yamamoto, M. Aoyama, L. Joeckel, J. Siebert, and J. Heidrich, "Requirements-driven method to determine quality characteristics and measurements for machine learning software and its evaluation," in *International Requirements Engineering Conference (RE)*, 2020, pp. 260–270.

[45] S. Nalchigar, E. Yu, and K. Keshavjee, "Modeling machine learning requirements from three perspectives: a case report from the healthcare domain," *Requirements Engineering*, pp. 1–18, 2021.

[46] 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.

[47] A. Pereira and C. Thomas, "Challenges of machine learning applied to safety-critical cyber-physical systems," *Machine Learning and Knowledge Extraction*, vol. 2, no. 4, pp. 579–602, 2020.

[48] 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.

[49] M. Rahimi, J. L. Guo, S. Kokaly, and M. Chechik, "Toward requirements specification for machine-learned components," in *International Requirements Engineering Conference Workshops (REW)*, 2019, pp. 241–244.

[50] J. Saldaña, *The coding manual for qualitative researchers, 4th Edition*. SAGE Publications Limited, 2021.

[51] G. Schuh, P. Scholz, T. Leich, and R. May, "Identifying and analyzing data model requirements and technology potentials of machine learning systems in the manufacturing industry of the future," in *International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, 2020, pp. 1–10.

[52] H. Villamizar, M. Kalinowski, M. Viana, and D. M. Fernández, "A systematic mapping study on security in agile requirements engineering," in *Euromicro conference on software engineering and advanced applications (SEAA)*, 2018, pp. 454–461.

[53] A. Vogelsang and M. Borg, "Requirements engineering for machine learning: Perspectives from data scientists," in *International Requirements Engineering Conference Workshops (REW)*, 2019, pp. 245–251.

[54] S. Wagner, D. M. Fernández, M. Felderer, A. Vetrò, M. Kalinowski, R. Wieringa, D. Pfahl, T. Conte, M.-T. Christiansson, D. Greer *et al.*, "Status quo in requirements engineering: A theory and a global family of surveys," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 28, no. 2, pp. 1–48, 2019.

[55] Z. Wan, X. Xia, D. Lo, and G. C. Murphy, "How does machine learning change software development practices?" *IEEE Transactions on Software Engineering*, 2019.

[56] H. Washizaki, H. Uchida, F. Khomh, and Y.-G. Guéhéneuc, "Studying software engineering patterns for designing machine learning systems," in *2019 10th International Workshop on Empirical Software Engineering in Practice (IWESEP)*, 2019, pp. 49–495.

[57] 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.

[58] C. Wilhjelm and A. A. Younis, "A threat analysis methodology for security requirements elicitation in machine learning based systems," in *International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2020, pp. 426–433.

[59] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the international conference on the practical applications of knowledge discovery and data mining*, vol. 1, 2000.

[60] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2014, p. 38.

[61] C. Wohlin, E. Mendes, K. R. Felizardo, and M. Kalinowski, "Guidelines for the search strategy to update systematic literature reviews in software engineering," *Information and Software Technology*, vol. 127, p. 106366, 2020.

[62] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.