Machine Learning-Enhanced Requirements Engineering: A Systematic Literature Review

Ana-Gabriela Núñez[a], Maria Fernanda Granda[b], Victor Saquicela[c] and Otto Parra[d]
Department of Computer Science, Universidad de Cuenca, Cuenca, Ecuador

Keywords: Requirements Engineering, Machine Learning, Artificial Intelligence, Natural Language Processing.

Abstract: In the software lifecycle, requirements are often subjective and ambiguous, challenging developers to comprehend and implement them accurately and thoroughly. Nevertheless, using techniques and knowledge can help analysts simplify and improve requirements comprehensibility, ensuring that the final product meets the client's expectations and needs. The Requirements Engineering domain and its relationship to Machine Learning have gained momentum recently. Machine Learning algorithms have shown significant progress and superior performance when dealing with functional and non-functional requirements, natural language processing, text-mining, data-mining, and requirements extraction, validation, prioritisation, and classification. This paper presents a Systematic Literature Review identifying novel contributions and advancements from January 2012 to June 2023 related to strategies, technology and tools that use Machine Learning techniques in Requirements Engineering. This process included selecting studies from five databases (Scopus, WoS, IEEE, ACM, and Proquest), from which 74 out of 1219 were selected. Although some successful applications were found, there are still topics to explore, such as analysing requirements using different techniques, combining algorithms to improve strategies, considering other requirements specification formats, extending techniques to larger datasets and other application domains and paying attention to the efficiency of the approaches.

1 INTRODUCTION

Given the subjective and ambiguous nature of requirements, developers encounter difficulties in comprehending and executing them with accuracy and thoroughness. As such, Requirements Engineering (RE) is deemed as the most pivotal phase in the Software Development Life Cycle (SDLC), given that imprecise and incomplete requirements pose challenges for developers to interpret and implement effectively. RE encompasses some tasks associated with extracting, analysing, specifying, validating, and managing requirements, including needs, goals, functionalities, constraints, qualities, behaviours, conditions, capabilities, and more. Two types of requirements are traditionally considered when producing the Software Requirement Specification (SRS) document: functional requirements (FR) and non-functional requirements (NFR). FRs describe how the software interacts with specific inputs and the functionalities provided by such software. NFR represents any requirement for the software product, including how it will be developed, maintained and put under operation (Alashqar, 2022). Machine Learning (ML) is used to improve the efficiency and effectiveness of tasks such as identifying, extracting, and classifying requirements, which are often written in natural language. Integrating RE with ML has the potential to enhance the efficacy of the requirements elicitation process, thereby improving software development quality.

This paper presents a systematic literature review (SLR) that assesses and synthesizes the state-of-the-art concerning ML techniques within the RE domain. The focus is on existing literature showcasing successful ML applications, the characteristic features within RE that utilize ML, and the prevalent tools in RE that incorporate ML techniques. The review also examines quality criteria, including a minimum number of pages, knowledge area or study domain, scope, methods of extraction and classification, and text processing. Key findings of each paper, including datasets, tools, and technology currently being used, are presented.

[a] https://orcid.org/0000-0002-4996-0390
[b] https://orcid.org/0000-0002-5125-8234
[c] https://orcid.org/0000-0002-2438-9220
[d] https://orcid.org/0000-0003-3004-1025

Núñez, A., Granda, M., Saquicela, V. and Parra, O.
Machine Learning-Enhanced Requirements Engineering: A Systematic Literature Review.
DOI: 10.5220/0012688100003687

In *Proceedings of the 19th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2024)*, pages 521-528
ISBN: 978-989-758-696-5; ISSN: 2184-4895

To understand the current state of ML and RE domains, we applied a search in the SCOPUS Digital Library using the keyword “machine learning” in “requirements engineering”. This search returned 4868 papers out of a total of 54231. The integration of ML in RE can enhance process efficiency, reduce errors, improve data quality, and increase stakeholder satisfaction and collaboration among different roles (Pei et al., 2022).

We conducted a systematic literature review in accordance with the guidelines proposed by Kitchenham and Charters (Kitchenham and Charters, 2007) to address the following research questions (RQs):

- RQ1. What existing works demonstrate the successful application of ML techniques in Requirements Engineering?
- RQ2. What are the characteristic features employed in Requirements Engineering that leverage Machine Learning?
- RQ3. What are the tools used in the field of Requirements Engineering that apply Machine Learning?

The remainder of this paper is structured as follows: Section II with related work; Section III on methodology; Section IV presenting SLR results; Section V discussing validity threats; and Section VI concluding and outlining future work.

2 RELATED WORK

RE’s study domain, especially its evolving relationship with ML, has been widely covered. Significant works include Iqbal et al. (Iqbal et al., 2018), who provided comprehensive insights on ML in RE, covering themes like FR, NFR, prioritisation, and more. They detailed ML model types and datasets used, highlighting ML’s emerging role in RE.

Sonbol et al. (Sonbol et al., 2022) conducted an in-depth study on NLP in RE, distinguishing between ML and rule-based methods, covering techniques like tokenisation and POS tagging. Similarly, Zamani et al. (Zamani et al., 2021) focused on challenges in applying ML to RE, discussing datasets, document types, and evaluation metrics for ML approaches in RE, such as recall and precision. Jindal et al. (Jindal et al., 2021) explored NFR classification, including types of NFRs, ML and NLP techniques, and datasets. Their research adds to understanding ML’s role in predicting FR and NFR.

While some studies have explored specific ML techniques for predicting FR and NFR, there is a shortage of research that systematically reviews empirical studies on this topic. More work is needed to identify the most effective ML practices and challenges in RE.

3 METHODOLOGY

The utilized approach follows the model outlined in the literature review by Kitchenham (2007) (Kitchenham and Charters, 2007) and has been customized to outline a search protocol consisting of three key phases: planning, execution, and results.

3.1 Planning Stage

This stage is essential to defining the basic review procedures and producing a search protocol to support our vertebrate research method. The activities identified in this stage include establishing the research questions, creating the search string, and selecting the information sources.

3.1.1 Establishment of the Research Questions

We have started from these contributions as the primary consultation elements to comprehend the ongoing studies related to empirical studies and formulated our research questions. For this, we consider first-hand that when implementing Machine Learning techniques, it’s essential to conduct an analysis of user requirements. This preliminary assessment provides a foundation for clearly defining the scope of ML within the target domain. In light of this, it’s crucial to determine the specific Artificial Intelligence (AI) strategies and algorithms that are most aligned with the given context. Once these foundational aspects are solidified, the emphasis should shift to understanding the nuanced elements of Requirements Engineering ML, which might enhance that. This involves identifying pertinent data sources, detailing data pre-processing methods, and selecting the most appropriate evaluation metrics. As our exploration draws to a close, we aim to highlight the tools that have proven most effective in bridging ML in RE.

3.1.2 Creation of the Research String

For this study, the research string has been composed of the following terms (”requirement* engineering” OR ”non-functional requirement*” OR nfr OR ”functional requirement*”) AND (”Machine Learning” OR ml OR ”artificial Intelligence” OR ai OR ”data mining” OR ”text mining” OR NLP OR ”Natural language processing”) AND $Year > 2011$. This study aims to analyze different contributions during the past ten years in the domains of RE and ML. Our search primarily relies on information from five pertinent databases: Scopus, WoS, IEEE, ACM, and ProQuest. Until June 2023, 1194 papers were retrieved using the research string.

3.2 Execution Stage

In this stage, we have continued with the review protocol and specified some steps necessary in the SLR, such as the primary studies selection process and the study quality assessment process. For this, we first considered the automatic search and, secondly, the manual search.

3.2.1 Selection of Primary Studies

First, a review was conducted to delete duplicate papers with the support of a script in R Studio[1], obtaining 1049 papers for the next analysis. Then, we handpicked and examined by reviewing the title and the abstract to preserve only those results relevant to our study goal that correspond to the research questions, getting 198 papers. Then, the papers were selected according to the inclusion and exclusion criteria.

3.2.2 Inclusion Criteria

These criteria comprised studies that presented examples or any empirical studies (e.g., study cases, experiments, and developed tools, among others), which included terms such as: *"Requirement Engineering", "Functional Requirement*", "Non-functional Requirements", "Requirement* Elicitation", "Requirement* Validation", "Data-driven Requirement*", "Extraction", Classification", "Prioritisation", "ML", "NLP", "AI", "text-mining", "data-mining"*

3.2.3 Exclusion Criteria

We eliminated papers that were not written in English, were not published in journals or conferences, and had a page count of fewer than six pages (for example, workshop papers and posters).

3.2.4 Quality Assessment

Besides inclusion or exclusion criteria, we considered certain aspects to evaluate the quality of our selected studies. These aspects include the number of pages in the paper (more than five pages), all RQs have been answered, other authors have cited the study, and the study has been published in relevant journals or conferences.

In our work, we have included studies manually. Then, we used the snowballing technique to observe if relevant papers existed that we needed to include. For this, criteria and quality assessment (QA) have been considered, as well as some papers from notifications we received in emails (new articles published in relation to the study domain) and others from Google Scholar. Initially, we did not select this database because we considered twofold issues. On one side, the limited precision of search results, and on the other side, the outcomes of many irrelevant results.

Therefore, first, we collected information for each study, such as year of publication, publication type (journal or conference), publication source (journal or conference name), publication title and abstract, number of citations and authors' names. Second, we compiled the necessary study data to address our research questions by defining a data extraction form. This strategy aimed to ensure a consistent classification of all the primary studies and understand the current state-of-the-art. Finally, we did not consider papers that did not report works on the RE domain and its relationship with ML or provide examples of empirical studies.

3.3 Reporting Stage

This section presents the preliminary results obtained from the SLR process, such as the data extraction and data synthesis processes.

3.3.1 Data Extraction Strategy

Our analysis has identified a crucial set of topics (T) for each pre-defined research question (RQs). These sixteen topics, derived from a comprehensive review of prior works, hold significant implications for RE and ML.

Concerning *RQ1*, the results are classified by topics that include strategies, learning algorithms, study domain, and scope. For this review, we analyzed three *strategies* to support RE: (1) *classification*, (2) *clustering*, (3) *association*.

We found *supervised learning* algorithms, *unsupervised learning* algorithms, requirements obtained from some methods, type of requirement, preprocessing and metrics. We have seen the need to know the works currently exist in which ML is helping in automating different tasks related to requirements, the study domain and their scope.

For *RQ2*, we are interested in knowing the types of dataset repositories used (e.g. open-source), types of

[1]R Studio: https://www.r-studio.com/data-recovery-software/

dataset documents with requirements to be processed with ML (e.g. SRS, textual, review data, book), *the dataset gathered from* (Alashqar, 2022), *variables* treated in the datasets (e.g., continuous and categorical) and the *dataset size*. Regarding *RQ3*, we recognised topics regarding *tools* and *technology* and developed them.

3.3.2 Syntesis Method

We utilized a quantitative analysis that categorised the primary studies according to the research questions. In this regard, we considered the number of studies and the percentages by different categories identified for each topic.

3.3.3 Conducting Stage

The application of the review protocol yielded the following preliminary findings, as outlined in Table 1. Throughout this process, four participants, who are the article's authors, contributed extensively to the entire review. After that, we complemented papers by manually searching 25 papers not considered in the automatic search. The methodology was employed to apply the same criteria as established before, and 17 papers were obtained. Then, 74 selected studies were obtained for analysis. These studies are shown in Appendix A, and the selected studies detailed by inclusion criteria are displayed in Appendix B.

Table 1: Results of conducting stage.

Source	Potential Studies	Removing duplicates	Scanning title and abstract	Selected Studies
Scopus	1036	969	160	46
WoS	48	17	9	2
IEEE	57	35	21	8
ACM	38	21	3	1
ProQuest	15	7	7	0
Manual search	25	25	25	17
Total	1219	1074	223	**74**

4 RESULTS AND FINDINGS

This section presents the results and the findings obtained through the SLR process regarding the three research questions. The results were classified by year to determine the frequency of selected studies and showed that 2018 and 2021 were the years of greatest scientific production regarding empirical studies regarding works about ML techniques in RE (see Figure 1). Figure 2 summarise the extracted information quantitatively (for a detailed synthesis of the selected studies, please refer to Appendix C).

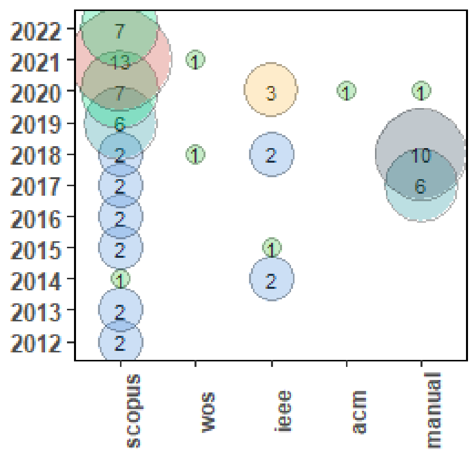

Figure 1: Selected studies per year.

4.1 RQ1. What Existing Works Demonstrate the Successful Application of ML Techniques in RE?

4.1.1 RQ1 Results

This review analyzed three strategies (classification, clustering and association) to support RE. These strategies were found in 66 of the reviewed papers as follows. Most papers (53 of the 74 papers, that is, 71.6%) report using classification strategy to categorize requirements using different ML *classification* algorithms.

Mostly, the works focused on the supervised classification strategy, indicating an interest in analysing and classifying requirements according to defined criteria (with the existence of a class). Eight relevant works corresponding to 11% were found regarding the application of *clustering strategies*, which seek to find similar requirements through similarity metrics, as for *association rules*, which aim to find rules of the "if ... then" type regarding requirements, four relevant works were found. Finally, eight works were classified as "other" because they did not meet the classification defined in RQ1. In this group some works proposing an *ontology* (S37, S56, S63), and others are based on a *graph theory* (S71) and *question-answering (chatbot)* (S03).

Then, this review examined different ML techniques and considered how these contribute to RE. Regarding the *supervised learning* we have obtained 49 papers, where one was identified as *Linear Regression*, and seven as *logistic Regression*. 27% of the

RQs and topics		Emerged answers	Studies #	Studies %
RQ1.	T1: Strategies	Classification	53	71.6%
		Clustering	8	11%
		Association Rules	4	5%
		Others	8	11%
	T2: Supervised learning algorithms	Linear Regression	1	1%
		Logistic Regression	8	11%
		Decision trees	20	27%
		Random Forest	10	14%
		k-NN	11	15%
		SVM	23	31%
		SVC	1	1%
		Naive Bayes	23	31%
		LSTM	6	6.8%
		Bagging	3	4%
		Deep Learning	5	7%
		CNN	4	5%
		RNN	3	4%
		others	4	5%
	T3: Unsupervised learning algorithms	K-means	4	5%
		LDA	3	4%
		HAC	4	5%
		Deep Learning	1	1%
		Word2vec	4	5%
		others	3	4%
	T4: Study domain	health	8	11%
		industry	12	16%
		education	1	1%
		others	19	26%
	T5: Scope	requirements identification	7	9%
		requirements extraction	6	8%
		requirements classification	21	28%
		requirements clustering	3	4%
		requirements flaws	2	3%
		ontology	5	7%
		others	5	7%

RQs and topics		Emerged answers	Studies #	Studies %
RQ2.	T6: Requirements obtained from	interview	3	4%
		experiment	2	3%
		expert	5	7%
		information not available	8	11%
		others	16	22%
	T7: Type of requirement	FR	23	31%
		NFR	30	41%
		NLP	40	54%
		text-mining	12	16%
		not-provided	1	1%
	T9: Metrics	Accuracy	24	32%
		Precision	35	47%
		Recall	40	54%
		F-measure	27	36%
		others	11	15%
	T10: Dataset repository	open-source	14	19%
		others	6	8%
	T11: Dataset document type	SRS	4	5%
		textual	3	4%
		mobile apps	3	4%
		review data	2	3%
		book	1	1%
	T12: Dataset gathered from	PROMISE	19	26%
		iTrust	2	3%
		PURE	2	3%
		others	22	30%
	T13: Dataset size	dataset less than 10000	33	45%
		dataset varied from 10000 to more than 7 million	6	8%
	T14: Variables	continuous	1	1%
		categorical	5	7%
		others	7	9%
RQ3.	T15: Tools and technologies	open-source	15	20%
		visualization	4	5%
		chatbot	2	3%
	T16: Developed	Python	10	14%
		Java	4	5%
		MySql	1	1%
		Weka	2	3%
		R	1	1%
		Google Colab	1	1%
		others	12	16%

Figure 2: RQs, Topics, and Summary of the Result corresponding (N=74).

papers were reported using *decision trees* (DT) as a method, of which four papers related to the J48 decision tree method (S24, S30, S39, S58), two to C4.5 (S64, S68), and ten to *Random Forest* (RF). In most of the studies related to SLR, 31% corresponding to the SVM method and 31% to *Naive Bayes* (NB).

We also found that several works have used *K-means* (K-NN) method (15%), LSTM (6.8%), CNN (5%), Bagging (4%), *Deep Learning* (7%), and *Recurrent Neural Network* (RNN) (4%), and *Support Vector Classification* (SVC) (1%). Several works have combined different techniques in their work. For instance, Jindal et al. (S30) used 8 ML techniques developed for the classification of NFR descriptions, including *Logitboost* (LB), *Adaboost* (AB), *Multi-Layer Perceptron* (MLP), *Radial Basis Function network* (RBF), and four more that are classified previous such as Bagging, NB, DT and RF. Yaseen et al. (S71) present the *Analytical Hierarchical Process* (AHP) and *spanning tree* combination to prioritise and implement FRs. In addition, Younas et al. (S73) mention employing supervised learning but do not specify which algorithm was used in their work.

In connection with *unsupervised learning* algorithms, we identified nineteen studies where four works corresponding to K-NN, four works Hierarchical agglomerative clustering (HAC), three works LDA, one work to DL, four works to *Word2vec*, one work to *cluster-boosted regression (CBR)* (S59). Muhairat et al. (S45) presented in their work *Apriori* and *FP-Growth* algorithms based on association rule analysis to perform experiments to improve the accuracy and the completeness of the gathered requirements. Elhassan et al. (S18) described the requirements conflict detection automation model based on the unsupervised ML model.

Different study domains were presented, with *health* (11%), *industry* (16%), *education* (1%) and among other areas (26%)

Regarding the scope of the study, most of the papers reported that its research assumes *requirements classification*, (S06, S10, S11, S18, S20, S25, S30, S31, S33, S35, S36, S40, S46, S50, S53, S54, S55, S56, S70, S73) (28%), *requirements identification* (S12, S50, S52, S58, S70, S72) (9%), *requirements extraction* (S02, S03, S12, S26, S55, S59) (8%), *requirements clustering*, (S19, S50, S59) (4%), *requirements flaws* (S34, S49) (3%), *ontology* (S34, S37, S56, S61, S63) (7%), and among others.

4.1.2 RQ1 Findings

These results suggest that identifying and classifying requirements is an important step towards automating the analysis of requirements written in natural language. However, applying ML techniques to classify requirements revealed that a supervised classification strategy is mostly used.

Additionally, these studies show the tendency to use ML mainly to classify NFR. This trend suggests that this activity is critical and necessary since inappropriate NFR management could increase software development and maintenance costs and impact its quality.

The studies highlight the potential of ML techniques to improve RE activities, including user validation, elicitation and prioritization of requirements. Results suggest that no works related to regression techniques were found since we focused explicitly on papers that analyze requirements expressed in natural language. Therefore, it is necessary to perform preprocessing in the requirements using natural language processing or text mining techniques before applying ML techniques. Also, classification and extraction algorithms can be improved or combined with other state-of-the-art algorithms to improve the strategies (i.e. classification, clustering and association rules). The analysis of the papers revealed that they apply their approaches to different domains, such as industry, health, education, and others. These domains generate sets of free and private requirements data, which are used to apply ML algorithms.

4.2 RQ2. What Are the Characteristic Features Employed in RE that Leverage ML?

4.2.1 RQ2 Results

In this SLR, 54 papers covered ML techniques for extracting or classifying requirements, as defined in RQ1. These techniques were applied to both FR and NFR. Additionally, it was identified that requirements were obtained through different methods, such as interviews, experimentation, and experts. It is important to note that some works do not explicitly specify the type of requirement analyzed or its source. Conversely, the application of ML techniques in RE is divided into two main groups: supervised learning and unsupervised learning.

Concerning requirements obtained from different ways, we recognized some got the *interview* (4%), another through *experts* (7%), and the *experiment* (3%). The rest of the papers regard other specifications from where was obtained the requirements (22%), and in some papers, the information was not available (S03, S05, S32, S37, S38, S43, S45, S47) (11%).

Many works focused on *NFR* (41%), and others in *FR* (31%), where features or capabilities are the input to applied ML techniques. Most of the studies indicated that applied preprocessing such as NLP (54%), text-ming (16%), and one paper in which the authors did not provide this information (S42). We recognised some metrics among which we have most papers applied *recall* (54%), followed by *precision* (47%), *F-measure* (36%), and *accuracy* (32%). Also, we identified *k-fold cross* (S21), *validity degree* (S15), *silhouette score* (S53), *the curve of Receiver Operating Characteristic* (ROC) (S26), *Area Under ROC Curve (AUC)* (S30), *Flesch Kincaid Reading Ease* (fkre), *Flesch Kincaid Grade Level* (fkgl), *Coleman Liau Index* (cli) (S24), and others (S09, S18, S60, S61).

The dataset is relevant to train and evaluate ML techniques. We identified fourteen studies that used datasets based on open-source repositories (S03, S07, S24, S32, S43, S44, S45, S46, S47, S61, S62, S68, S70). EzzatiKarami and Madhavji (S20) use a public requirements document (PURE dataset). Reahimi et al. (S54) specify that the dataset was obtained in a previous real-world project. Gu et al. (S22) selected a dataset of an experiment concerning 91 effective cases of KLK2 elevator, among others.

Considering the type of document of the dataset applied for research purposes, we found some papers based on *SRS* (S09, S10, S20, S31) (5%), *textual* (S44, S61, S62) (4%), *mobile apps* (S04, S12, S46) (4%), *review data* (S25, S70) (3%), and *book* (S60) (1%). Once the dataset is gathered, it has to be processed to prepare data and use ML algorithms. Most studies use the *PROMISE* dataset (26%), among others include *iTrust* (S24, S65), *PURE* (S02, S20), *SQuAD* (S03), *Aurora 2* (S47).

Regarding the dataset size, several works to group sizes between 14 to 10.000 (S04, S06, S07, S11, S12, S18, S19, S20, S21, S22, S24, S25, S28, S30, S31, S34, S35, S36, S43, S44, S50, S52, S53, S54, S55, S61, S62, S63, S64, S66, S68, S70, S73) (45%), and other works varied between 10.000 to 7 million and more (S02, S03, S11, S39, S45, S57) (8%). Some methods consider continuous and categorical variables, five papers corresponding to *categorical* (S22, S24, S36, S44, S45), and one to *continuous* (S22).

4.2.2 RQ2 Findings

In research on the classification of requirements using ML techniques, evaluating the results is an essential part of the process. In this regard, various metrics widely used in the analyzed works have been identified: Accuracy, Precision, Recall, and F-measure. This confirms that these four metrics are the most commonly used as they are considered the most representative for evaluating an ML model.

According to the results table, most papers use requirement repositories that are freely accessible. Regarding the size of the datasets, most papers comprise less than 10.000 requirements. Although, some papers with a minimal amount of requirements are mainly used to perform a proof of concept of the proposed approach. Only a small number of papers have datasets with more than 10.000 requirements, which is why deep learning algorithms are used to a lesser extent. This situation may be due to the lack of free access datasets with large amounts of requirements.

These results suggest that the investigation of the application of ML techniques to automate the activities of RE has several open avenues, such as (1) considering other types of requirements specification formats, (2) replicating these works to confirm their results and provide benchmarks of different approaches, (3) extend the existing techniques through the use of large data sets and other application domains; and (4) pay attention to the efficiency of the approaches.

4.3 RQ3. What Are the Tools Used in the Field of RE that Apply ML?

4.3.1 RQ3 Results

Some important topics in using ML algorithms applied to software requirements are related to the technology and tools used. Several tools and technology have been applied in different studies involving ML and RE, being that 20% of the works have used *open-source* (S12, S13, S24, S43, S44, S45, S46, S47, S61, S62, S64, S68), whereas 5% indicated visualization tools for presenting the results of ML analyses in a user-friendly (e.g., dashboards) (S13, S12, S61, S62), and 3% specified *chatbot* (S03, S12). Dabrowski et al. (S12) do not mention using visualization or chatbot technologies; rather, they focus on evaluating techniques for opinion mining and searching for feature-related reviews in app reviews. Some authors indicated that they gave a name to their proposals, for example, *Review with Categorized Requirements* (ReCaRe) (S49), *Heuristic Requirements Assistant* (HeRA) (S62), *retraining Bidirectional Encoder Representations from Transformers* (REBERT4RE) (S02), ReqVec (S66), NOMEN (S26), among others.

Concerning how the proposals were developed, the results indicate that ten works use *Python* (S04, S08, S18, S20, S28, S46, S54, S72), four use *Java* (S22, S24, S34, S63) two use *Weka* (S04, S56). The rest of the papers, corresponding to *MySql* (S22), R (S52), *Google Colab* (S06), *wiki* (S61), *Semi-supervised approach for Feature Extraction* (SAFE), *Group MAsking* (GuMa), *Relation-based Unsupervised Summarization* (ReUS) and *Mining App Review using Association Rule Mining* (MARAM) (S12), and *APIs of Apple App Store* (S70).

4.3.2 RQ3 Findings

After the analysis, it was identified that most approaches use open-source technology. Several of them assign names to the generated tools, varying from visual tools to chatbots. Regarding programming languages, the ones used in data science in general, such as Python, R, and Java, among others, are mostly used. Visual frameworks, such as Weka, are also used. It is important to note that several papers use cloud software for experimentation execution.

These findings indicate that there is still the opportunity to develop tools that help automate ML algorithms and that can integrate other AI solutions with this purpose (e.g. ChatGPT).

5 THREATS TO VALIDITY

The completeness of the primary studies enlisted in this SLR hinges heavily on the choice of keywords and the limitations of the digital libraries and search engines employed. Objective search terms were utilised to mitigate risks associated with subjective search terms. However, the syntax and standards across different engines and libraries may have inadvertently excluded relevant studies. A broad search string encompassing various synonyms for each term was crafted to address this concern. Five databases (Scopus, WoS, IEEE, ACM, and ProQuest) were selected to maximize the potential pool of papers. Inclusion and exclusion criteria assessed the relevance and quality of titles and abstracts. The researcher also checked the references of selected papers to identify any additional pertinent works. However, the decision to exclude papers in languages other than English might have overlooked relevant contributions.

Furthermore, the data selection process was rigorously aligned with the research questions to prevent biases or inaccuracies in data extraction based on the researcher's interests. Instances of uncertainty or conflicts regarding the inclusion of a paper were resolved through discussions among the authors.

6 CONCLUSIONS AND FUTURE WORK

The SLR, spanning from January 2012 to June 2023, evaluates the application of ML techniques in software requirements engineering, identifying 74 significant studies. This review highlights ML's potential in automating software RE processes, though complete automation remains a future goal. Key improvements through ML include user validation, requirement elicitation, and prioritization. Emerging technologies like ChatGPT and Google Bard show promise in requirement analysis and representation. Preprocessing in ML for natural language requirements is essential, and enhancing classification and extraction methods by integrating advanced techniques like clustering and association rules is suggested.

ML's application extends across diverse domains, including industry, health, and education, emphasizing the need for accurate requirement classification metrics. Most studies use open-source tools, programming languages like Python, R, and Java, and visual frameworks like Weka. Future research opportunities include varying requirement specification formats, applying ML to larger datasets and new domains, and enhancing efficiency.

ACKNOWLEDGEMENTS

This work was supported by *"Vicerrectorado de Investigación e Innovación (VIUC)"* at the University of Cuenca.

REFERENCES

Alashqar, A. M. (2022). Studying the commonalities, mappings and relationships between non-functional requirements using machine learning. *Science of Computer Programming*, 218:102806.

Iqbal, T., Elahidoost, P., and Lucio, L. (2018). A bird's eye view on requirements engineering and machine learning. In *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*, pages 11–20. IEEE.

Jindal, R., Malhotra, R., Jain, A., and Bansal, A. (2021). Mining non-functional requirements using machine learning techniques. *e-Informatica Software Engineering Journal*, 15(1).

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Keele University*, 2.3.

Pei, Z., Liu, L., Wang, C., and Wang, J. (2022). Requirements engineering for machine learning: A review and reflection. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pages 166–175. IEEE.

Sonbol, R., Rebdawi, G., and Ghneim, N. (2022). The use of nlp-based text representation techniques to support requirement engineering tasks: A systematic mapping review. *IEEE Access*.

Zamani, K., Zowghi, D., and Arora, C. (2021). Machine learning in requirements engineering: A mapping study. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 116–125. IEEE.

APPENDIX

Appendices A and B list selected papers and inclusion criteria for the SLR, while Appendix C offers a detailed summary. The link to access these appendices is available: SLR-Appendices-Resource.