dHealth 2024
D. Hayn et al. (Eds.)

doi:10.3233/SHTI240034

Screening Automation in Systematic Reviews: Analysis of Tools and Their Machine Learning Capabilities

Elias SANDNER[a,b,1], Christian GÜTL[b] and Igor JAKOVLJEVIC[a,b] and Andreas WAGNER[a]

[a] *IT-Department, CERN, Geneva, Switzerland*
[b] *CoDiS-Lab ISDS, Graz University of Technology, Graz, Austria*

Abstract. Systematic reviews provide robust evidence but require significant human labor, a challenge that can be mitigated with digital tools. This paper focuses on machine learning (ML) support for the title and abstract screening phase, the most time-intensive aspect of the systematic review process. The existing literature was systematically reviewed and five promising tools were analyzed, focusing on their ability to reduce human workload and their application of ML. This paper details the current state of automation capabilities and highlights significant research findings that point towards further improvements in the field. Directions for future research in this evolving field are outlined, with an emphasis on the need for a cautious application of existing systems.

Keywords. Systematic Review Automation, Title and Abstract Screening, Research Tool, Machine Learning, Text Classification

1. Introduction

A systematic review (SR) aims to evaluate and interpret all existing studies related to a particular field of interest. Annual Review[2] releases SRs in more than 30 disciplines, including computer science and medicine. In medical contexts, SRs address research questions about the frequency of disease occurrence, their expected progression, the dangers involved in diagnosing them, and the strategies for their management, to name a few aspects [1]. SRs have their level of evidence confined to that of the studies they encompass. Nonetheless, by aggregating more data than individual studies, SRs enhance the precision of the overall findings. Consequently, they offer the most reliable evidence to address research questions [2]. However, this evidence comes at the cost of an enormous workload, as it typically takes several months to complete SRs [3,4]. Within the SR process, the screening phase is described as the most difficult and time-consuming aspect of the process and is the most urgent task that requires a reliable support system [5]. Therefore, this survey paper aims to answer the following research question: **How advanced**

[1] Corresponding Author: Elias Sandner, IT-Department, CERN, Geneva, Switzerland and CoDiS-Lab ISDS, Graz University of Technology, Graz, Austria, E-mail: elias.sandner@cern.ch

[2] https://www.annualreviews.org/

is the current level of automation in the title and abstract screening phase, which methods show promise for further advancement, and what key research findings should inform future improvements in this area?

Section 2 emphasizes the need for enhanced automation through an overview of the current SR process. Section 3 describes how machine learning is integrated into existing tools, and Section 4 describes recent research findings. The paper concludes with Section 5 which suggests promising paths for future work.

2. Background and Related Work

Conducting an SR involves multiple stages and support tools to streamline several steps are already in use. For example, in [6] it is demonstrated how such systems helped to conduct an SR within two weeks and the main tools used are listed in Table 1. This paper focuses on the screening phase within the SR process. Subsequently, this task is outlined in detail, followed by an in-depth analysis of the required time effort.

To determine studies that comply with the specified inclusion and exclusion criteria, retrieved studies undergo a thorough screening process. Initially, the title and abstract of each paper are assessed to eliminate the bulk of the nonrelevant papers and then a detailed examination of the full text is executed for comprehensive evaluation. To reduce bias and errors, it is recommended that the screening is conducted independently by a minimum of two reviewers. Discrepancies should be addressed through collaborative discourse or by incorporating the judgment of an additional expert. This approach ensures the comprehensive identification of relevant studies, although it is associated with a high workload. In [6] RobotSearch was utilized during the title and abstract screening phase (TiAb screening). It is capable to filter out documents that are definitely not randomized controlled trials (RCTs), the specific study type to which this SR was limited. However, it is important to consider that the exclusive focus on RCTs notably simplified the screening process. This is a significant point, as such a limitation might not apply to other SRs, where a broader range of study types could lead to a more complex and time-consuming screening effort. Additionally, the SRA Helper was employed in both the TiAb screening and full text screening phases. It offers a user-friendly interface that allows documents to be included or excluded using hotkeys.

SRs yield significant evidence, although the completion of all steps in the outlined process typically spans several months. To highlight the effort required to perform SRs, Table 2 sums up the results of three time analyzes. Given the substantial time needed to conduct SRs, a survival analysis of 100 SRs indicates that 7% of these reviews already showed signs of quantitative or qualitative obsolescence at the time of publication [8]. [5] highlighted that the screening phase is the most time-intensive aspect of the entire

Table 1. Overview of support tools employed in major tasks of the SR process as reported by Clark et al. (2020)[6].

Task	Support system
Project proposal design	template
Systematic search	SRA word frequency analyzer, SRA polgot search translator
Eligibility Screening	RobotSearch, SRA Helper
Data extraction	digital spreadsheet

Table 2. Time demands of SRs based on three analyses. Data sources and metrics are derived from the studies conducted by Beller et al. (2013)[9], Demetres et al. (2023)[3], and Borah et al. (2017)[4]; All values represented in Days.

	Beller et al. (2013)	**Demetres et al. (2023)**	**Borah et al. (2017)**
Data origin	Medline	Weill Cornell Medicine	PROSPERO
Data quantity	300	101	195
Time of SR conduction	2009-2011	2011-2021	before July 2014
Time measured from/to	last search / publication	Requesting librarian support / submission	Registration / publication
Min	0	42	42
Max	1314	930	1302
Median	153	N/A	461
Average	N/A	295	473

process. Additionally, this phase was recognized as the most challenging and the one that most urgently necessitates a dependable support system. As indicated in [4] most of the literature is removed during the TiAb screening with a median reduction in citations of 95% at this stage and 3.7% during full text screening. Furthermore, [3] outlined that abandoned SRs, most likely occur in the TiAb screening phase.

Data filtration carried out by highly compensated experts, a process uncommon in most fields, is currently a consistent component of the SR process. Therefore, tools designed to increase expert efficiency are commonly used to reduce human workload.

3. Current Level of Screening Automation

Despite the current inability to fully automate the screening process, numerous software tools significantly aid human experts. [10] analyzed 16 tools based on 21 features. The five tools that rank highest (Table 3) offer a stable and supported release, comprehensive documentation, active customer support and features for multiple users, importing and allocating references, and the inclusion/exclusion of references with labeled reasons for exclusion, and resolution of discrepancies. While the top four tools support the distinction between TiAb and full text screening, this is the only mandatory feature not fulfilled by Rayyan[3]. However, users can simply export relevant citations from the TiAb screening and import them into a new review. Therefore, we do not interpret this as a major drawback. Furthermore, the study (published in 2019) indicates that Covidence[4] does not offer any ML automation feature, but the current version does. As a result, 5 out of the 16 analyzed tools now provide some form of ML support, as summarized in Table 3. The typical workflow for using one of these tools during the TiAb screening phase is illustrated in Figure 1 and is subsequently described.

Search results, retrieved from various databases, are imported into the selected review management tool, typically in RIS, RevMan, or PubMed format. Initially, integrated solutions for deduplication are utilized. Subsequently, the user manually screens citations using the interface's hotkeys for including and excluding reasons. Once a number of decisions are made, the ML system uses labeled documents to estimate inclusion probabilities of unscreened citations and reorders the citation queue accordingly, prioritizing those with the highest likelihood of inclusion. With more labeled data, the rank-

[3] https://www.rayyan.ai/

[4] https://www.covidence.org/

Table 3. Feature analysis and AI support of selected review management tools. (*2 according to [10] but machine learning features are available in the current version.)

	Feature Analysis according to [10]			Machine learning support			
Tool	Mandatory features (out of 9)	Desirable features (out of 9)	Optional features (out of 3)	Deduplication support system	Relevance ranking	RCT classifier	Custom classifier
DistillerSR	9	8	3	yes	yes	No	Yes
EPPI-Reviewer	9	7	3	yes	yes	Yes	Yes
SWIFT Active Screener	9	7	2	yes	yes	No	No
Covidence	9	5	3*	yes	yes	Yes	No
Rayyan	8	6	2	yes	yes	No	Yes

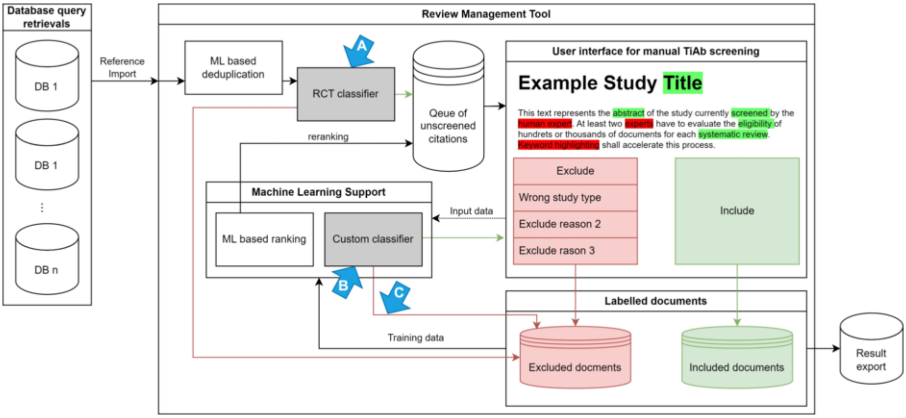
Database query retrievals
DB 1
DB 1
⋮
DB n
Reference Import
Review Management Tool
ML based deduplication
RCT classifier
A
Queue of unscreened citations
reranking
Machine Learning Support
ML based ranking
Custom classifier
B
C
Input data
Training data
User interface for manual TiAb screening
Example Study Title
This text represents the abstract of the study currently screened by the human expert. At least two experts have to evaluate the eligibility of hundrets or thousands of documents for each systematic review. Keyword highlighting shall accelerate this process.
Exclude
Wrong study type
Exclude reason 2
Exclude rason 3
Include
Labelled documents
Excluded docments
Included documents
Result export

Figure 1. Title and abstract screening with support tool; A: only applicable to Covidence and EPPI-Reviewer[11,7]; B: only applicable to DistillerSR, EPPI-Reviewer and Rayyan; C: Only applicable to DistillerSR and EPPI-Reviewer as Rayyan provides suggestions without making autonomous decisions. [13,15,14]

ing's accuracy improves. For example, EPPI-Reviewer[5] updates its ranking every 25 citations. Swift-Active-Screener[6] also estimates the number of relevant citations left in the unscreened document list.

However, Swift Active Screener also stands out as the only one that does not use ML classification. As mentioned in Section 2, some SRs are confined to RCTs. The RCT classifier component (Figure 1 A) autonomously excludes citations with different study designs, considerably lessening the need for manual screening. The Cochrane RCT classifier[12] achieves a recall rate of 0.99. Consequently, Covidence has integrated this original classifier. EPPI-Reviewer also provides a solution with a recall rate of 0.99, based on data manually labeled by the Cochrane crowd, as explained in [7].

DistillerSR[7], Eppi-Reviewer and Rayyan furthermore provide classifiers that can be trained based on the initial manual decisions as illustrated with the custom classifier component (Figure 1 B). DistillerSR offers integrated classification software based on a statistical approach. Based on the application in three SRs, it is claimed to reduce human workload by 57.4%. The false negative rate is 1. 17%, and the recall is not disclosed.

[5] https://eppi.ioe.ac.uk/
[6] https://www.sciome.com/swift-activescreener/
[7] https://www.distillersr.com/

Furthermore, only a minimum of 10% of the citations were manually reviewed, leaving the number of relevant papers missing unclear. [13] Rayyan employs a random forest ensemble model, evaluated on 15 pre-labeled SRs. The best performance showed a 0.986 recall and reduced workload by 46.9%, but when applied to a different review, the recall dropped to 0.75 and workload reduction to 3%. In both instances, 50% of pre-labeled data was used for training and the rest for testing. Once enough training data is gathered, Rayyan activates its prediction model, offering suggestions for undecided studies. However, to maintain review quality, the final decision always rests with a human expert. [14] EPPI-Reviewer offers the capability to manually create bespoke classifiers using previously screened citations. This tool utilizes the scikit-learn Python library[8] and is adept at binary classification of new records. Additionally, it provides statistical insights before applying these created classifiers to new sets of citations.

4. Current Approaches Towards Further Automation

Building on the analysis of existing tools, this section highlights further relevant research findings. Custom classifiers, integrated in existing tools are especially suitable for living SRs as they require labeled data. For their effective use in living SRs, it is essential to maintain consistent scope, unchanged field terminology, and an original SR large enough to supply adequate training data. [15] explored this use case employing EPPI-Reviewer's classifier function, specifically using the stochastic gradient descent (SGD) classifier[9] with logistic regression in both instances. They first evaluated ML classifiers' performance in recall and screening reduction. Classifiers assigned relevance scores from 0 to 99 for each citation, using a threshold of 10 to filter out low-relevance citations. Recalls ranged from 92% to 100%, and screening reduction, measured by papers left for manual review, ranged from 40% to 74%. The study demonstrated improved classifier performance when supplemented with specific exclusion and inclusion criteria, rather than solely binary labels. They also highlighted that text preparation might impact classification performance more significantly than the choice of algorithm. Based on those findings, they continued to apply a custom classifier for the update search. EPPI-Reviewer's classification tool provides insights into the relevance score distribution of citation records, aiding in estimating the classifier's reliability for a particular citation set. This information assists in determining the classifier's integration into the screening workflow. Citations with a relevance score greater than 20 were manually screened. Citations scoring between 13 and 20 were batch screened in sets of 500, and if two consecutive batches lacked relevance, all remaining citations were deemed irrelevant. Citations ranking below 13 were automatically discarded. Implementing these rules resulted in a 61% decrease in screening efforts. 98% of the relevant references were identified in the top 21% of the citations, with relevance scores ranging from 20 to 99. Notably, a highly relevant study with a lower score of 14 was also included. This approach is estimated to save around 25 hours of screening time, considering an average 7-second review time for less relevant records.

While [15] applied a statistic approach, [16] applied a large language model to build a custom classifier for update search. Three models based on Bidirectional Encoder Rep-

[8] https://scikit-learn.org/

[9] https://scikit-learn.org/stable/modules/sgd.html

resentations from Transformers (BERT)[17] were tested. Model one utilizes text from Wikipedia and books for both vocabulary development and pre-training. The second model builds its vocabulary using text from Wikipedia and books and pre-trains with abstracts from selected articles. Meanwhile, the third model constructs its vocabulary from abstracts of articles acquired and uses abstracts from included articles for its pre-training phase. Each model underwent fine-tuning with the titles of included articles. The third model excelled in nearly all performance metrics, notably achieving an AUC above 90, surpassing other models which did not exceed 67. This implies that enriching language representation with domain-specific data boosts performance. To address performance distortion from imbalanced class composition, a common problem in applying ML classification for screening, dummy data were generated by altering keywords in excluded citations. This adjustment led to an increase of recall from 0.55 to 0.91.

Developing classifiers focues on standard eligibility criteria provides a promising alternative, applicable across various SRs, unlike those designed for specific SRs. The importance of study design in these criteria is notable, and [18] developed a classifier to categorize COVID-19 literature into one of 22 study designs. Five classifiers, based on both general and domain-specific corpora, were trained using manually annotated data records. The five classifiers were subsequently combined into an ensemble model. In the evaluated ensemble model, a voting strategy is employed, while another possibility is aggregating class probabilities. Study designs were classified into classes and subclasses. The ensemble model outperformed all standalone models, registering an AUC-ROC of 94.33 at the class level and 94.77 for specific study designs, compared to 91.77 and 92.07, respectively, by the best standalone model.

5. Conclusion and Discussion

Systematic review is the research methodology that provides the most evidence. The associated workload justifies the demand for efficient automation tools. This paper offers a detailed investigation of current tools and recent research aimed at automating the title and abstract screening process, particularly emphasizing the role of machine learning. Custom classifiers, informed by initial human decisions, are integrated into the available tools, but their accuracy for specific SRs often remains uncertain. Consequently, there is a need to integrate and report predefined rules transparently when using these systems. Furthermore, focusing on text preparation rather than selecting specific algorithms could lead to further improvements. Considering specific inclusion and exclusion reasoning instead of binary labels to train custom classifiers not only increases transparency but also performance. Although large language models significantly influenced other fields, this was not observed. However, for further research in this direction, the relevance of domain-specific vocabulary must be considered. Additional research on classification based on specific eligibility criteria should be preferred over focusing on specific reviews, as this enables the collection of more training data and expands its applicability to a broader range of reviews. This approach is already effective for SRs focused on RCTs, and efforts to classify other study types are in progress. It is proposed to identify additional common eligibility criteria and use ML classifiers to address them. In conclusion, the research underscores the need to advance automation in SRs, highlighting the potential and limitations of current solutions. It emphasizes the importance of continuous innovation and the cautious application of existing systems.

Acknowledgement

The World Health Organisation, in particular Alice Simniceanu and Luca Fontana, are acknowledged for providing invaluable advice and guidance, dedicating time to discuss existing solutions and current challenges in systematic review automation. The study was done as part of the joint CERN and WHO ”CAiMIRA” project, that is funding the PhD project, in the context of which this paper was written.

References

[1] P. Glasziou, L. Irwig, C. Bain, and G. Colditz, Systematic Reviews in Health Care: A Practical Guide. Cambridge University Press, 2001.

[2] K. Coleman, S. Norris, A. Weston, K. Grimmer-Somers, S. Hillier, T. Merlin, P. Middleton, R. Tooher, and J. Salisbury, ”NHMRC Additional Levels of Evidence and Grades for Recommendations for Developers of Guidelines,” National Health and Medical Research Council (NHMRC), Canberra, 2009.

[3] M. R. Demetres, D. N. Wright, A. Hickner, C. Jedlicka, and D. Delgado, ”A decade of systematic reviews: an assessment of Weill Cornell Medicine’s systematic review service,” Journal of the Medical Library Association: JMLA, vol. 111, no. 3, p. 728, 2023, Medical Library Association.

[4] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, ”Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry,” BMJ Open, vol. 7, no. 2, p. e012545, 2017, British Medical Journal Publishing Group.

[5] J. C. Carver, E. Hassler, E. Hernandes, and N. A. Kraft, ”Identifying barriers to the systematic literature review process,” in 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2013, pp. 203-212.

[6] J. Clark, P. Glasziou, C. Del Mar, A. Bannach-Brown, P. Stehlik, and A. M. Scott, ”A full systematic review was completed in 2 weeks using automation tools: a case study,” Journal of Clinical Epidemiology, vol. 121, pp. 81-90, 2020.

[7] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, and B. C. Wallace, ”Machine learning for identifying randomized controlled trials: An evaluation and practitioner’s guide,” Research Synthesis Methods, vol. 9, no. 4, pp. 602-614, 2018.

[8] K. G. Shojania, M. Sampson, M. T. Ansari, J. Ji, S. Doucette, and D. Moher, ”How quickly do systematic reviews go out of date? A survival analysis,” Annals of Internal Medicine, vol. 147, no. 4, pp. 224-233, 2007, American College of Physicians.

[9] E. M. Beller, J. K.-H. Chen, U. L.-H. Wang, and P. P. Glasziou, ”Are systematic reviews up-to-date at the time of publication?,” Systematic Reviews, vol. 2, no. 1, pp. 1-6, 2013.

[10] S. Van der Mierden, K. Tsaioun, A. Bleich, C. H. C. Leenaars, et al., ”Software tools for literature screening in systematic reviews in biomedical research,” Altex, vol. 36, no. 3, pp. 508–517, 2019. [Online]. Available: Springer International Publishing AG.

[11] Covidence, ”Automation using the Cochrane RCT classifier,” Covidence Knowledge Base, 2024. [Online]. Available: https://support.covidence.org/help/automatically-tag-studies-not-reporting-on-rcts. [Accessed: Mar. 8, 2024].

[12] J. Thomas, S. McDonald, A. Noel-Storr, I. Shemilt, J. Elliott, C. Mavergames, and I. J. Marshall, ”Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews,” Journal of Clinical Epidemiology, vol. 133, pp. 140-151, 2021, Elsevier.

[13] S. K. Venkata, S. Velicheti, V. Jamdade, S. Ranganathan, M. Achra, K. K. Banerjee, C. Dutta Gupta, M. Happich, and A. Barrett, ”MSR99 Application of Artificial Intelligence in Literature Reviews,” in Value in Health, vol. 26, no. 12, pp. S412, 2023, Elsevier. [Poster presentation].

[14] M. Khabsa, A. Elmagarmid, I. Ilyas, H. Hammady, and M. Ouzzani, ”Learning to identify relevant studies for systematic reviews using random forest and external information,” Machine Learning, vol. 102, pp. 465-482, 2016, Springer.

[15] C. Stansfield, G. Stokes, and J. Thomas, ”Applying machine classifiers to update searches: Analysis from two case studies,” Research Synthesis Methods, vol. 13, no. 1, pp. 121–133, 2022.

[16] S. Aum and S. Choe, ”srBERT: automatic article classification model for systematic review using BERT,” Systematic Reviews, vol. 10, no. 1, pp. 1–8, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[18] J. Knafou, Q. Haas, N. Borissov, M. Counotte, N. Low, H. Imeri, A. M. Ipekci, D. Buitrago-Garcia, L. Heron, P. Amini, et al., ”Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature,” Systematic Reviews, vol. 12, no. 1, p. 94, 2023.