Sports, Health & Wellness Insights, e-ISSN: 3107-7617
July-December 2025 Volume: 01, Issue: 02, Page No.: 22-39
https://doi.org/10.64006/shwi/1203

ORIGINAL ARTICLE | Received: 18th September. 2025| **Revised**: 16th October. 2025| **Accepted**: 27th October 2025| **Published**: 24th December. 2025

Design Validity and Quantitative Metric Analysis of Artificial Intelligence for the Proficiency of Systematic Review

Ms. Sadia, MOT[1];
[1]Department of Occupational Therapy, Jamia Hamdard, New Delhi, India
sadiatajuddin@gmail.com

Dr. Deepika Singla, PhD[2*];
[2] Department of Physiotherapy, Jamia Hamdard, New Delhi, India

Ruchi Basista, MPT[3];
[2] Department of Physiotherapy, Jamia Hamdard, New Delhi, India

Corresponding Author:
Dr. Deepika Singla, PhD, Assistant Professor (II), Dept. of physiotherapy, School of Allied Health Sciences and Research, Jamia Hamdard, New Delhi, India
Address:
Dept. of physiotherapy, Exercise Performance Testing Lab, room no 501, fifth floor, central library building, Jamia Hamdard, New Delhi, India
Email ID: *dpkasingla@gmail.com*
Funding/financial disclosure*: Authors received no funding.*

Author contributions

MS: Conceptualization; Data curation; Formal analysis; Methodology; Resources; Writing -original draft, review & editing.
DS: Conceptualization, Methodology; Resources; Supervision; Writing -review & editing.
RB: Methodology; Formal analysis; Resources; Writing - review & editing

ABSTRACT

With emergence of Artificial Intelligence (AI) has created opportunities for automation with natural language processing (NLP) and semantic mapping. AI tools can identify relevant literature, classifying evidence, extract data, and evaluate quality.
To explore methodological validation of design and quantitative metrics analysis of artificial intelligence and machine learning tools used in the automated systematic review process.
Electronic databases were searched from 2015 to September 2025. Investigating methodological validation of design and quantitative metrics analysis of AI tools.
8 studies were included in this review for reporting quantitative metrics analysis and design validity of ML tool automation in the systematic review process, and evaluating model performance components.
The findings showed metrics analysis was done across a wide range of simulations and workload savings vary according to task across the included studies.
Keywords: Machine Learning; Natural Language Processing (NLP), Semantic Mapping

INTRODUCTION

Systematic reviews are a comprehensive approach that is widely used to bring together the findings from multiple studies in a refined way and are often used for evidence synthesis[1]. Regrettably, this could be challenging because of relevant evidence is unstructured, small population size is deficient in the randomization and blinding process, and describing the conduct of such articles is already massive and continues to expand rapidly. Marshall and Wallace (2019) highlighted that the growth of scientific knowledge and an expansion of publications across various subjects create both opportunities and challenges for researchers and reviewers at the same time[2].

Higgins et al (2019) pointed out the process of systematic review entails several explicit searching for studies, describing the sources of studies, search process, search strategies, managing references during the search process, correctly documenting the search process, and extracting data from eligible studies, synthesizing the results[3]. Traditionally, screening of studies and extraction of data from selected studies have been the most laborious process in conduction of systematic reviews. The assessment of methodological qualities, intervention and outcome measure is assessed by a team of reviewers independently, ensuring reliability, but it subsumes and protracts reviewers in cognitive burden. This can challenge the persistence of reviewers' variability, error in recording data, fatigue, and confusion [4].The methodological foundation of systematic review starts forms hours of literature searches, rigid inclusion criteria, ultra careful data extraction, and data synthesis, ensuring the reliability, credibility, and replicability. However, as the literature is expanding exponentially, traditional manual methods of systematic reviews are inadequate to keep up with the new diversified trends and evidence[5].

Watt et al (2008) highlighted that the deadline to complete the review, especially in health-related technological assessments, and the necessity for search to minimize the number of studies screening, even though finding relevant research is a laborious task[6]. Time limit may compromise the quality of the systematic review and produce biased results. Many innovations have occurred in the processes of systematic reviews since the publication of the PRISMA 2009 statement[7]. For instance, technological advances have enabled the use of natural language processing and machine learning to identify relevant evidence. The methodological frameworks of PRISMA 2020 require explicit recording of statistical tools and human error, implying the chances that AI should be augmented along with scholarly judgment[8]. Gusenbauer and Haddaway et al (2020) explains, reviewer searches databases, using keywords and Boolean operators to get relevant reports [9] into a reference manager. Two independent reviewers check the records' titles and abstracts on the basis of inclusion criteria; most of the records are excluded at this phase. A small number of research makes to the full-text screening, which is a significant step in systematic review. De Vries et al. analyzed 10,115 records and excluded 9,847 after title and abstract screening, a drop of more than 95%26, implying how labor-intensive this step is [10]. A recent estimate states that conducting a review takes about 67 weeks from registration to publication [1]. Along with that, current processes are not sustainable for producing current evidence efficiently because they often go out of date quickly once they are published [11].Kaveh G. Shojania et al conducted a systematic review to assess the updates of new evidence that occurred in systematic reviews within 2 years, it was 23% reviews, within 1 year 15% reviews, and before publication7% reviews, which also found several statistically significant predictors signaling for updates [11]. It is recommended by Cochrane to update reviews every 2 year.

Van De Schoot et al. (2021) claims that Artificial intelligence, natural language processing, and automation are components of systematic review. A large number of tools are available and being used in the process of conducting systematic review and meta-analysis are based on machine learning and natural language processing. From study selection, citation management, risk assessment, certainty analysis, forest plot, and funnel plot are created using software based on machine learning and natural language processing (NLP)[12]. With emergence of Artificial Intelligence (AI), specifically Machine Learning (ML), has created opportunities to automate through natural language processing (NLP) and semantic mapping can identify relevant literature, classifying evidence, extract quantitative data, and even evaluate methodological quality. Recent advances in Artificial Intelligence (AI) use a type of machine learning known as active learning, in which a model can choose the data points. For obtaining records from a systematic search and learns from it and thereby reducing a significant number of records required for manual screening[12]

The machine learning validation process systematically assesses how well an application works. providing researchers to understand and assess components like relevance, occurrence of hallucinatory response, speed of response, and accuracy of responses. Evaluation of machine learning tracks improvements for iterations of prompts, parameters, or retrieval. The constant iteration and modification of prompts, adjusting temperature settings, or refined the retrieving approach. The reliability and effectiveness of machine learning tools, a design and quantitative metrics analysis for the intended area, such as healthcare and research, is a must, focusing on how the machine learning tool is structured, its components, algorithms, interfaces, and how its performance objectivity is measured using quantitative data.

The primary objective of this review is to systematically identify, evaluate, and synthesize evidence by exploring methodological validation of design and quantitative metrics analysis of artificial intelligence and machine learning tools used in the automation systematic review process.

METHOD

This systematic review is based on "The Systematic Methodological Review Design" following PRISMA-ScR and automation-specific guidelines[13]. Machine learning and Artificial intelligence-based tools developed or validated for literature search, screening, or data extraction were included in this review (Table-1). The tools that are manual or statistical tools without ML components were excluded. *Search strategy:* The scientific Databases searched, e.g., PubMed, Scopus, Web of Science, Semantic Scholar, with terms such as 'artificial intelligence', 'AI tool validation', 'machine learning', 'literature search', 'systematic review, 'data extraction', automation, 'Semi-automation', 'quantitative metrics analysis', 'design validation'. *Study selection:* In order to ensure reliability, reviewers screened titles and abstracts prior to screening all sources independently. Two reviewers worked separately to screen for inclusion criteria.

Table 1: AI Tools for Automation of Systematic Review

Step of systematic review	AI functions	Tool name	Advantages
Literature Search [14 15]	semantic retrieval, transformer-based query expansion	Elicit, Rayyan,	Identifying gray literature, synonym expansion, Year, Author
Screening and selection [5,14–16]	Active learning and supervised classification for abstract/title screening	ASReview, Abstrackr, Elicit, Rayyan	Reduces manual screening workload, recall precision analysis
Data Extraction [14,17,18]	NLP-driven extraction models	Robot Reviewer, Elicit, research screener	Reduces manual workload
Risk of Bias Assessment [19]	Automated bias identification	Robot Reviewer	Integrated with the Systematic Review guideline.

Conflicts were resolved through discussion with a third investigator and the process of searching and selecting research according to the PRISMA2020 flow. (Fig. 1). Two reviewers individually extracted data using an inclusion criterion. According to the objectives of this systematic review, categorized information was extracted to meet the purpose of the review (e.g., describe the ML models, simulation performance metrics). Data charting began when a sufficient number of reports was selected, then the reviewer recorded the name of the tool, developer, model, purpose, and application stage in a tabular format for critical appraisal and narrative synthesis.

3.Prisma flowchart

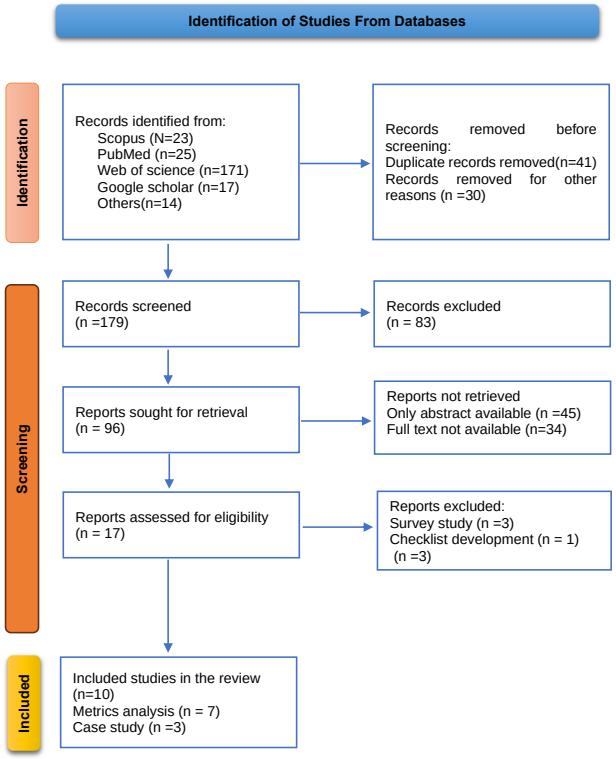

Fig.1. Study flow diagram of steps followed throughout the review process.

4.Results

4.1Collating, summarizing, and reporting

To answer the research question, results were presented in tables summaries in tables, including title, author, application stage, validity, advantages and limitations, and developers of included studies mentioned in (Table 2).

Table 2: Characteristics Summary of Included Studies

Author	AI/ML Tool-task	Developed by	ML Type/model	Application stage	Access	Validation Study	Strength/ limitation
Van De Schoot et al. (2021) [12]	ASReview LAB	Utrecht University, The Netherlands	Active learning (supervised)	Screening	open-source	An open-source machine learning framework for efficient and transparent systematic reviews	ASReview with state-of-the-art systems across a wide range of real-world systematic reviewing applications. According to the tests conducted, this tool offers default settings for its parameters, which demonstrated promising performance.
Rathbone et al. (2015) [16]	Abstrackr	Tufts Brown University.	Supervised classifier	Screening	Web-based, open access	Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers	Saves screening workload, but performance depends strongly on review complexity and dataset balance; caution is needed because some relevant items can be missed
Marshall IJ et al,(2016) [17]	Robot Reviewer	University of Bristol	NLP (BERT variants)	Risk of Bias	open-access	Robot Reviewer: evaluation of a system for automatically assessing bias in clinical trials	Risk of bias assessment may be automated with reasonable accuracy. Substantially reduces workload.
N. Bernard et al.,(2025)[14]	Elicit	Ought (an applied ML research lab)	Transformer (LLM)	Search & Extraction	Not fully open source	Using artificial intelligence for systematic review: the example of elicit	A valuable complementary tool for researchers in the conduct of systematic reviews.

Sophie Elizabeth Park 2018[20]	EPPI-Reviewer	EPPI-Centre (Social Science Research Unit, UCL Institute of Education, University College London, UK	Hybrid supervised ML	Full review pipeline	not fully open-source	Evidence synthesis software	The EPPI-Reviewer is a text-mining automation tool for searching relevant literature
A Hookway et al (2019) [19]	Robot Analyst	National Centre for Text Mining (NaCTeM)	NLP, transformer-based classification (BERT variants)	risk-of-bias detection and text summarization	Free to access	Could machine learning aid the production of evidence reviews? A retrospective test of Robot Analyst	Robot Analyst can be used as a second screener, especially on larger reference sets when the human resource demands of duplicate screening are considerable.
Amir Valizadah et al (2020) [18]	Rayyan	Qatar Computing Research Institute (QCRI)	NLP classifier	Screening	free access	Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews	A tool for excluding ineligible records, but it was not very reliable for finding eligible records.
Chai et al (2021) [15]	Research Screener	Curtin University	deep learning and natural language processing	abstract screening	Web based, open access	Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews	Reduce the burden of researchers in conduct a comprehensive systematic review with scientific rigor.

4.2 Search results and characteristics: About 15 AI tools were identified claiming automation features for systematic review processes. According to those AI tools, a search strategy was built using controlled vocabulary, Boolean operators, and MeSH strategy to search studies of reliability and validity of design and quantitative metric analysis of AI tools for automation of systematic review. The initial database search yielded 458 articles. After removing duplicates and initial screening of titles and abstracts, 52 full text articles were eligible for screening. 406 articles were excluded for not coinciding the eligibility criteria. 13 articles were included in this review. A PRISMA flow diagram shows each stage of the selection studies phase (Figure 1). In the 8 studies were included quantitative metrics analysis and design validity of AI tools [5,14,21,22,17,17,23,16,24,12,15,15] articles were published across a variety of peer-reviewed journals.

Table 3: Quantitative metrics analysis of AI/ML Tools

Author	AI/ML Tool	No of simulation	Data set	Performance metrics	Results
Van De Schoot et al. (2021) [12]	AS Review LAB	1. viral metagenomic next-generation sequencing 2. Systematic review of studies on fault prediction in software engineering 3. self-reported symptoms of post-traumatic stress assessed after trauma exposure 4. The efficacy of angiotensin-converting enzyme inhibitors	First simulation records-4.84% of 4368 Second simulation records 1.2% of 8911 Third simulation records 0.66% of 5782 Fourth simulation records 1.6% of 2544	WSS@100% (Work saved over sampling) 95%, indicating the work reduction (RRF10%) 10% of the relevant references found	2-user experience (UX) test systematic UX test
Rathbone et al. (2015) [16]	Abstrackr	1.Atypical hemolytic uremic syndrome (aHUS) 1415 records 2.Dietary fiber 517 records 3.Diagnostic accuracy review of echocardiography (ECHO) 1735 records 4.Rituximab 1042 records	1.Atypical hemolytic uremic syndrome 1415,63 were found to be relevant, giving a precision of 16.8 %. 2.Dietary fiber 517 records and 47 were found to be relevant, giving a precision of 24.7 % 3.Diagnostic accuracy review of echocardiography 1735 records potentially relevant, and 181 were	**(aHUS)**: The precision is 16.8%, with a false negative rate of 10% following the review of 18% of the records. **Dietary fiber**: The precision is 24.7%, and the false negative rate is 14.5% after reviewing 23% of the records. **ECHO**: The precision achieved is 29.2%, with a false negative rate of 4.7% after examining 7% of the records. **Rituximab**:	Sensitivity and Specificity rates were not reported.

			found to be relevant giving a precision of 29.2 % 4.Rituximab 1042 records, potentially relevant, and 372 were found to be relevant giving a precision of 45.5 %.	The precision stands at 45.5%, and the false negative rate is 2.4% after screening 12% of the records.	
A. Gates et al (2018) [25]	Abstrackr	**simulation no.1-** Antipsychotics 12,763 records, **simulation no.2-** Bronchiolitis 5893 records **Simulation No. 3-**Child Health SRs 5243 records **Simulation no.4**-Diabetes 47,385 records	1. Diabetes (0.7 (0.4) % of the 47,385 records) 2. Bronchiolitis (10.3 (5.8) % of the 5893 records. 3.Antipsychotics and Child Health SRs was 2.2 (0.3) % (of 12,763 records) and 4.0 (0.2) % (of 5243 records	Abstrackr demonstrated the highest mean (SD) **precision** for Child Health SRs at 64.7 (2.0) %, followed by Bronchiolitis at 38.1 (2.6) %, Antipsychotics at 15.1 (2.6) %, and lastly Diabetes at 14.8 (2.6) %. **The false negative rate** was greatest for Antipsychotics at 21.2 (8.3) %, then for Diabetes at 17.9 (2.3) %, followed by Bronchiolitis at 7.3 (2.2) %, and Child Health SRs at 3.5 (1.4) %. **Simulation no.1-** Specificity of 69% and Sensitivity of 79% after screening 2.2% of records. **Simulation no.2-** Specificity of 85% and Specificity of 92% after screening 10.3% of records. **Simulation no.3-**Specificity of 90% and Specificity of 82%	reliability and workload savings abilities vary according to task.

				after screening 0.7% of records. **Simulation no.4**-Specificity of 19% and Specificity of 96% after screening 4% of records.	
Marshall IJ et al,(2016) [17]	Robot Reviewer	1.Trial included in one Cochrane review 2.Trial included 2 or more Cochrane reviews 3.trial labelled Risk of bias assessment 4.trial labelled using Risk of bias assessment	1: articles matching full title 2: articles with exact author combination with matching publication year 3: articles with matching journal name, volume, issue, and page number	Document-Level Results-significance of differences across the three models **Sentence-Level Results** the top3 model was rated as more relevant than text from the CDSR overall, and in individual domains, but differences were not statistically significant.	total of 1731 judgments from 20 experts from 371 trials.
N. Bernard et al.,(2025)[14]	Elicit	year-based searches (2005, 2010, 2015, 2016, 2017, 2018, 2019, 2020, and 2021)	1. Trial 246 2. Trial 169 3. Trial 172	**Repeatability** was assessed with the search method described above was replicated three times at different moments, the search method described above was replicated three times at different moments. **Accuracy** measured with a mixed method was used to assess accuracy. **Reliability** was assessed by comparing the number of publications by elicit and classical screening	The repeatability test found 246,169 results and 172 results
A Hookway et al (2019) [19]	Robot Analyst	Retrospective testing using 3 previously completed evidence review.	For each test, references were uploaded into Robot	sensitivity 90% observed in reference test set one (n = 500 references)	Robot Analyst could potentially be

			Analyst and the decisions made by the original reviewers were input in blocks of 25 to form a training set. The "update predictions" function generated a predicted inclusion decision for the remaining references at each test point, and these were compared to the original review decisions.	sensitivity increased from 51% at the start of testing to 91%	utilised as a second screener for evidence review.
Chai et al (2021) [18]	Research Screener	9 systematic reviews and 2 scoping reviews Sedentary- overweight Low back pain- lifting Low back pain- fear Lung cancer — preop Lung cancer — exercise Falls Acute pain Team reflexivity Sexual health	The mean number of rounds of 50 articles required to find all eligible articles was 9.4, with rounds completed ranging from 1 to 36 (i.e., 50–1800 articles screened)	The relevant articles needed to be screened ranged from 5 to 35%. The WSS was computed at both WSS@95 and WSS@100. The mean time saved using Research Screener was 88.7%	Rates of Sensitivity and Specificity were not reported
Amir Valizadah et al (2020) [15]	Rayyan	1.applied machine learning algorithms on cerebral structural magnetic resonance imaging (sMRI) 2. applied machine learning algorithms on cerebral resting-state functional	2054 records screened manually, of 379 sMRI 122 records fMRI 193 EEG 64 112 reports sMRI 25, fMRI 64 EEG 23	SEN of 21-88% after screening 25% of records. SEN of 86-98% after screening 50% of records. SEN of 89-100% after screening 75% of records	diagnostic test accuracy reviews

		magnetic resonance imaging (rs-fMRI) 3. applied machine learning algorithms on electroencephalogram (EEG)			

5.DISCUSSION

An Artificial Intelligence and machine learning model is validated by evaluating its prediction performance. Model performance indicates how well a machine learning and artificial intelligence performs the assigned task, based on various metrics. Measuring model performance is essential for optimizing an AI tool before releasing it and enhancing it after deployment. Without proper optimization, AI could produce inaccurate or unreliable predictions and suffer from inefficiencies, leading to poor performance. This review explored the metrics analysis of machine learningtools used in systematic reviews, published in peer-reviewed journals. The findings indicate that metrics analysis was done across a wide range of simulations with a large data set. A total of 8 studies were selected in this review to report the metrics analysis[2,5,14–17,19,26], because metrics allow to quantify the performance of an ML model[27].

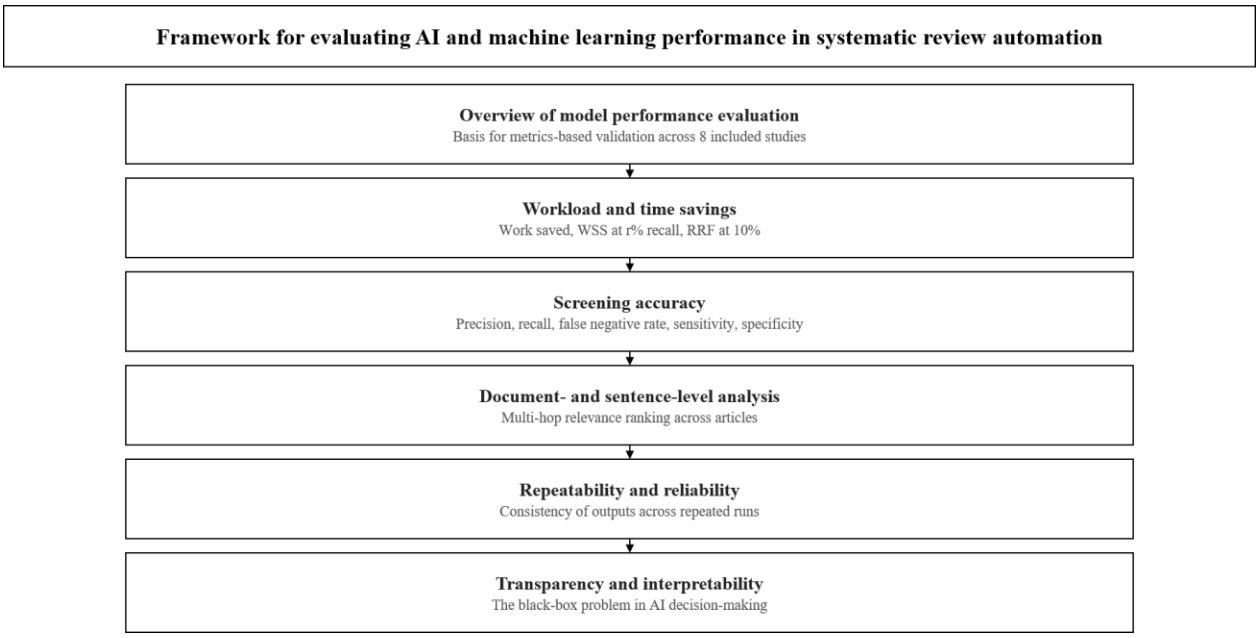

Fig.2 Conceptual framework for performance evaluation of machine learning tools in evidence synthesis.

5.1Workload and Time savings

The fundamental matrix across included studies was workload reduction. van de Schoot et al evaluation of *AS Review LAB* indicated 95% work reduction [12], while Chai 2021 found the *Research Screener* mean time saved was 88.7% evaluated WSS@100% [18]. Gates et al. identified that the workload savings from Abstrackr were most significant for Diabetes, at 88.4 (2.7)%, followed by Antipsychotics at 64.5 (9.8)%, Bronchiolitis at 70.0 (3.7)%, and Child Health systematic reviews at 9.5 (3.5)%. [25]. The work saved over sampling at *r*% recall (WSS@r%) is the most commonly used custom evaluation measure in citation screening. WSS@r% is introduced by Cohen et al as the percentage of papers according to the original search criteria that reviewers don't need to read, as that has been screened by a classifier[28,29]. Norman et al 2020 describe that WSS is straightforward for the interpretation of automation of systematic review, inclined to have large variance and highly influenced by random effects[29].

5.2 Screening Accuracy and Classification Performance

Relevant references found @10% (RRF@10%) in the metrics analysis of AS Review LAB[5]. Reflects the proportion of relevant articles a tool identifies after screening only 10% of records. Mostly, RRF is presented graphical format with relevant references along the *Y* axis and the number of articles screened on the *X* axis [30].

Precision and recall were widely reported. Gates et al evaluation of Abstrackr, precision ranged from 15 to 65% and was lower for Antipsychotics and Diabetes compared to Bronchiolitis and Child Health SRs [25]. Precision measures retrieval specificity, defined as the proportion of retrieved articles that are judged by the researcher as being relevant; this measure penalizes system retrieval of irrelevant articles (false positives) but does not penalize failures by the system to retrieve articles that the user considers to be relevant (false negatives). Recall measures retrieval coverage, defined as the proportion of the set of relevant research articles that is retrieved by the system, and therefore penalizes false negatives but not false positives [31]. In the tasks of classification, anomaly detection, and predictive modeling, the confusion matrix serves as a crucial instrument for assessing model performance, consisting of four elements: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). False negatives arise when positive instances are mistakenly categorized as negative, indicating the presence of the model fails to detect positive articles [32]

The False Negative Rate (FNR) measures the number of positive responses that a classifier wrongly labels as negative [33]. Also varied by review topic in Gates et al evaluation of Abstrackr, the rate of false negatives was greatest for Antipsychotics at 21.2 (8.3)%, followed by Diabetes at 17.9 (2.3)%, Bronchiolitis at 7.3 (2.2)%, and Child Health SRs at 3.5 (1.4)%[25]. Rathbone et al similarly reported topic dependent precision and FNR in a metrics analysis of the Abstrackr tool that Atypical hemolytic uremic syndrome had a precision of 16.8% and FNR of 10% after screening 18% of records. Dietary fiber achieved a Precision of 24.7% and FNR of 14.5% after screening 23% records. ECHO achieved a precision rate of 29.2% and a false negative rate of 4.7% after reviewing 7% of the records. Rituximab demonstrated a precision rate of 45.5% and a false negative rate of 2.4% after reviewing 12% of the records (18).

Sensitivity and specificity were reported together in in Gates et al evaluation of Abstrackr with substantial variation across topics e.g., Antipsychotics 69% Specificity and 79% Sensitivity after screening 2.2% of records. Bronchiolitis had 85% Specificity and 92% Specificity after screening 10.3% records, Child Health Systematic reviews had 90% Specificity and 82% Specificity after screening 0.7% of records and Diabetes had 19% Specificity and 96% Specificity after screening 4% records. Together these measures offers a more nuanced picture of modal discrimination then accuracy alone through interpreting them requires understanding the context in which the modal is applied, since this determine which metrics should be prioritized[34].

5.3 Document and Sentence-Level Analysis

Marshall IJ et al's evaluation of Robot Reviewer assessed performance at both Document and sentence level, finding the top three modal rated more relevant than overall text. Document level analysis requires multi-hop reasoning to identify relationships between articles across sentences or paragraph[17]. This reflects a broader shift towards transformer encoded language modelling methods, pre-train big models on a massive text corpus, focusing mainly on learning the representation of contextualized words. Drawing inspiration from human language processing, this explores the improvement of sentence-level representations of pre-trained models by borrowing ideas from predictive coding theory. The predictive mechanism helps to improve the quality of the representations [35].

5.4 Repeatability and Reliability

N. Bernad et al. metrics evaluation of Elicit AI with the Repeatability of the search method described and was replicated three times at different moments [14].Repeatability is the ability to consistently produce similar outputs given identical inputs under controlled conditions. Which can be undermined by AI system relying on stochastic reasoning, may manifest unexpected variability during output[36]. Reliability was assessed by comparing the number of publications by elicit and classical screening, reflecting ability of a system or process to perform its intended function over a specified period of time and under defined conditions with consistency. Reliability assesses factors such as design, manufacturing, maintenance, and operational conditions, and can be assessed through various methods, including testing, modeling, and analysis. Therefore, it measures how dependable and trustworthy a system can be, and the probability of the correct functioning of the system without failure. Reliability takes into account factors such

as design, manufacturing, maintenance, and operational conditions, and can be assessed through various methods, including testing, modeling, and analysis [37].

5.5 *Transparency and Interpretability*

The rapid advancement of methods and techniques has led to the emergence of a new field of artificial intelligence and machine learning. This not only deliver precise predictions but offers clear and interpretable justifications for their decisions and actions, thereby increasing their reliability for human users. And several studies raised the importance of modal transparency. AI algorithms can suffer from a lack of transparency in situations where a system fails to provide any reasoning or adequate explanation for its decisions, commonly known as "the black-box problem". The opaque nature of black-box models makes them difficult for humans to understand. Relying on a black-box model for critical decisions creates an essential requirement for AI algorithms to be transparent regarding their decision-making processes.

6.Conclusion

It was found that the metrics analysis was done across a wide range of simulations with a large dataset, reporting sensitivity, precision, and true positives. Reliability and workload saving vary according to task across the included studies. The relevant research articles that a machine learning technique can identify after screening just 10% of the research articles. The sentence-level results, repeatability of the method to identify relationships between research articles, and the capability to produce similar articles.

7.Conflicts of Interest

The authors declare that no conflicts of interest.

8.Key findings

Conducting a systematic review is a time-consuming process, but utilizing Machine learning tools and human intellect can improve time efficiency and optimization by The complexity of research can be comprehended [38]. Currently, AI tools are being used in the systematic review process to assist researchers with specific tasks, such as formulating research questions, as discussed in this review for screening and data extraction.

9.What study contributes

The machine learning tools highlight the necessity to follow certain principles to uphold the methodological integrity of systematic reviews. The primary crucial aspect to consider is that AI tools lack evidence regarding design and the analysis of quantitative metrics. AI tools should only be utilized at specific stages in the systematic review process, rather than automating the entire workflow. For transparency and reproducibility, it is vital to detail in the methodology section which AI tools were employed. Adopting the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-AI) reporting guidelines is beneficial as it offers a framework for incorporating AI into the systematic review process.

10.References

1. Borah R, Brown AW, Capers PL, et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017; 7: e012545.

2. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews* 2019; 8: 163.

3. Higgins J, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Newark: John Wiley & Sons, Incorporated, 2019.

4. Systematic review automation technologies | Systematic Reviews | Full Text, https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046-4053-3-74 (accessed 29 October 2025).

6. Watt A, Cameron A, Sturm L, et al. Rapid reviews versus full systematic reviews: An inventory of current methods and practice in health technology assessment. *International Journal of Technology Assessment in Health Care* 2008; 24: 133–139.

7. Ganann R, Ciliska D, Thomas H. Expediting systematic reviews: methods and implications of rapid reviews. *Implementation Sci* 2010; 5: 56.

8. Page MJ, McKenzie J, Bossuyt P, et al. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. Epub ahead of print 14 September 2020. DOI: 10.31222/osf.io/jb4dx.

9. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods* 2020; 11: 181–217.

10. De Vries H, Bekkers V, Tummers L. INNOVATION IN THE PUBLIC SECTOR: A SYSTEMATIC REVIEW AND FUTURE RESEARCH AGENDA. *Public Administration* 2016; 94: 146–166.

11. How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis | Annals of Internal Medicine, https://www.acpjournals.org/doi/10.7326/0003-4819-147-4-200708210-00179?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed (accessed 30 October 2025).

12. Schoot R van de, Bruin J de, Schram R, et al. Open Source Software for Efficient and Transparent Reviews. *Nat Mach Intell* 2021; 3: 125–133.

13. Beller EM, Glasziou PP, Altman DG, et al. PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts. *PLOS Medicine* 2013; 10: e1001419.

14. Bernard N, Sagawa Jr Y, Bier N, et al. Using artificial intelligence for systematic review: the example of elicit. *BMC Medical Research Methodology* 2025; 25: 75.

15. Valizadeh A, Moassefi M, Nakhostin-Ansari A, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Medical Research Methodology* 2022; 22: 160.

16. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Syst Rev* 2015; 4: 80.

17. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016; 23: 193–201.

18. Chai KEK, Lines RLJ, Gucciardi DF, et al. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews* 2021; 10: 93.

19. Hookway A, Price S, Knight T. Could machine learning aid the production of evidence reviews? A retrospective test of RobotAnalyst. *Eur J Public Health* 2019; 29: ckz185.095.

20. update b - EPPI-Reviewer Jan 02.indd, https://discovery.ucl.ac.uk/id/eprint/10003909/1/Thomas2007EPPI-Reviewer(Poster).pdf (accessed 10 November 2025).

21. Jain SJ, Sibbu K, Kuri R. Conducting Effective Research using SciSpace: A Practical Approach, https://www.authorea.com/users/703515/articles/689172-conducting-effective-research-using-scispace-a-practical-approach?commit=1063f1660ff682dfbf987b76276bca4fd3a01159 (accessed 31 October 2025).

22. Jeyaraman MM, Rabbani R, Al-Yousif N, et al. Inter-rater reliability and concurrent validity of ROBINS-I: protocol for a cross-sectional study. *Systematic Reviews* 2020; 9: 12.

23. Minozzi S, Dwan K, Borrelli F, et al. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. *Journal of Clinical Epidemiology* 2022; 141: 99–105.

24. Santos PA, Chuquisengo E, Vasquez AE. Design and validation of an instrument to measure the use of the Julius AI tool in Peruvian university students. *Revista Espacios* 2025; 46: 204–212.

25. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews* 2018; 7: 45.

26. Evidence synthesis software | BMJ Evidence-Based Medicine, https://ebm.bmj.com/content/23/4/140.long (accessed 4 November 2025).

27. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep* 2024; 14: 6086.

28. Cohen AM. Optimizing Feature Representation for Automated Systematic Review Work Prioritization. *AMIA Annu Symp Proc* 2008; 2008: 121–125.

29. Kusa W, Lipani A, Knoth P, et al. An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intelligent Systems with Applications* 2023; 18: 200193.

30. Saeidmehr A, Steel PDG, Samavati FF. Systematic review using a spiral approach with machine learning. *Syst Rev* 2024; 13: 32.

31. (PDF) An exact analytical relation among recall, precision, and classification accuracy in information retrieval. *ResearchGate*, https://www.researchgate.net/publication/228906049_An_exact_analytical_relation_among_recall_precision_and_classification_accuracy_in_information_retrieval (accessed 12 November 2025).

32. Raposio E, Baldelli I. Reliability and Accuracy of Generative Artificial Intelligence Tools in Providing General Information on Migraine Surgery. *Plastic and Reconstructive Surgery – Global Open* 2025; 13: e7176.

33. False Negative - an overview | ScienceDirect Topics, https://www.sciencedirect.com/topics/computer-science/false-negative (accessed 11 November 2025).

34. Ting KM. Sensitivity and Specificity. In: *Encyclopedia of Machine Learning*. Springer, Boston, MA, pp. 901–902.

35. Araujo V, Moens M-F, Soto A. Learning Sentence-Level Representations with Predictive Coding. *Machine Learning and Knowledge Extraction* 2023; 5: 59–77.

36. Tian S, Lam AWY, Sung JJ-Y, et al. A six-tiered framework for evaluating AI models from repeatability to replaceability. *Trends in Biotechnology*. Epub ahead of print 31 July 2025. DOI: 10.1016/j.tibtech.2025.07.015.

37. Tamascelli N, Campari A, Parhizkar T, et al. Artificial Intelligence for safety and reliability: A descriptive, bibliometric and interpretative review on machine learning. *Journal of Loss Prevention in the Process Industries* 2024; 90: 105343.

38. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, et al. Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open* 2023; 13: e072254.