Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Commentator slide for JC : Sep 16,2022

Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis

Yunying Feng , Siyu Liang , Yuelun Zhang ,et al.

Journal of the American Medical Informatics Association, 29(8), 2022, 1425–1432

Ekapob Sangariyavanich MD.

Ph.D. student year 2020 in Data Science for Healthcare and Clinical Informatics program

Wisdom of the Land

Outline

- Introduction
- Material & methods
- Results
- Discussion
- Presenter’s comment

Introduction

- The literature screening step can be extremely time-consuming and prevent on-time completion and updates of systematic reviews.
- AI tools are on trial in the highly standardized and repetitive procedures of systematic reviews, such as literature screening, data extraction, and risk-of-bias assessment.

- Few studies reviewing automated literature screening have been found.
- To address this gap in knowledge, we sought to perform a systematic review and meta-analysis on accuracy of AI methods for literature screening in medical evidence synthesis.

Materials & methods

Eligibility criteria

(1) Automatic methods were developed for literature screening for medical systematic reviews

(2) The research question and source of dataset used were reported.

(3) The literature screening results by human investigators were set as the reference standard.

- Exclusion criteria

  - Editorials, commentaries, and narrative review articles were excluded.

Information source and search strategy

- Database : PubMed, Embase, and IEEE Xplore Digital Library.
- Papers published between January 1, 2000 and December 22, 2021.
- No restrictions were set on language.

- Data collection and risk of bias assessment

- The “participants” : original medical studies and literatures
- The index test : AI algorithms used for automatic literature screening.
- The reference standard : traditional literature screening by human investigators.
- The outcomes of meta-analysis : effectiveness of literature screening, as well as labor and time saving

- Recall (sensitivity), precision [positive predictive value (PPV)], specificity, and the work saved over sampling (WSS).

$$\boxed{\mathrm{WSS} = (\mathrm{TN} + \mathrm{FN})/\mathrm{N} - (1.0 - \mathrm{R}),}$$

- Study applying semi-automation and active learning methods were not considered in final meta-analysis.

- All citations and abstracts were independently screened by 2 reviewers.
- The full texts of potentially eligible citations were then reviewed independently by the same 2 reviewers to select the studies for final inclusion.
- Disagreements in both initial screening and final screening were resolved by discussion with a methodologist.

- The detailed information of training sets and validation sets, AI algorithms, and effectiveness and work-saving indices were collected.
- Risk of bias assessment was applied with a revised checklist based on Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).

Statistical analysis

- 1 automation study reported multiple groups of effectiveness and labor-saving indices.
- In this meta-analysis, we mainly focused on 2 groups of effectiveness and labor-saving indices:

  (1) the precision and WSS when achieving the maximized recall

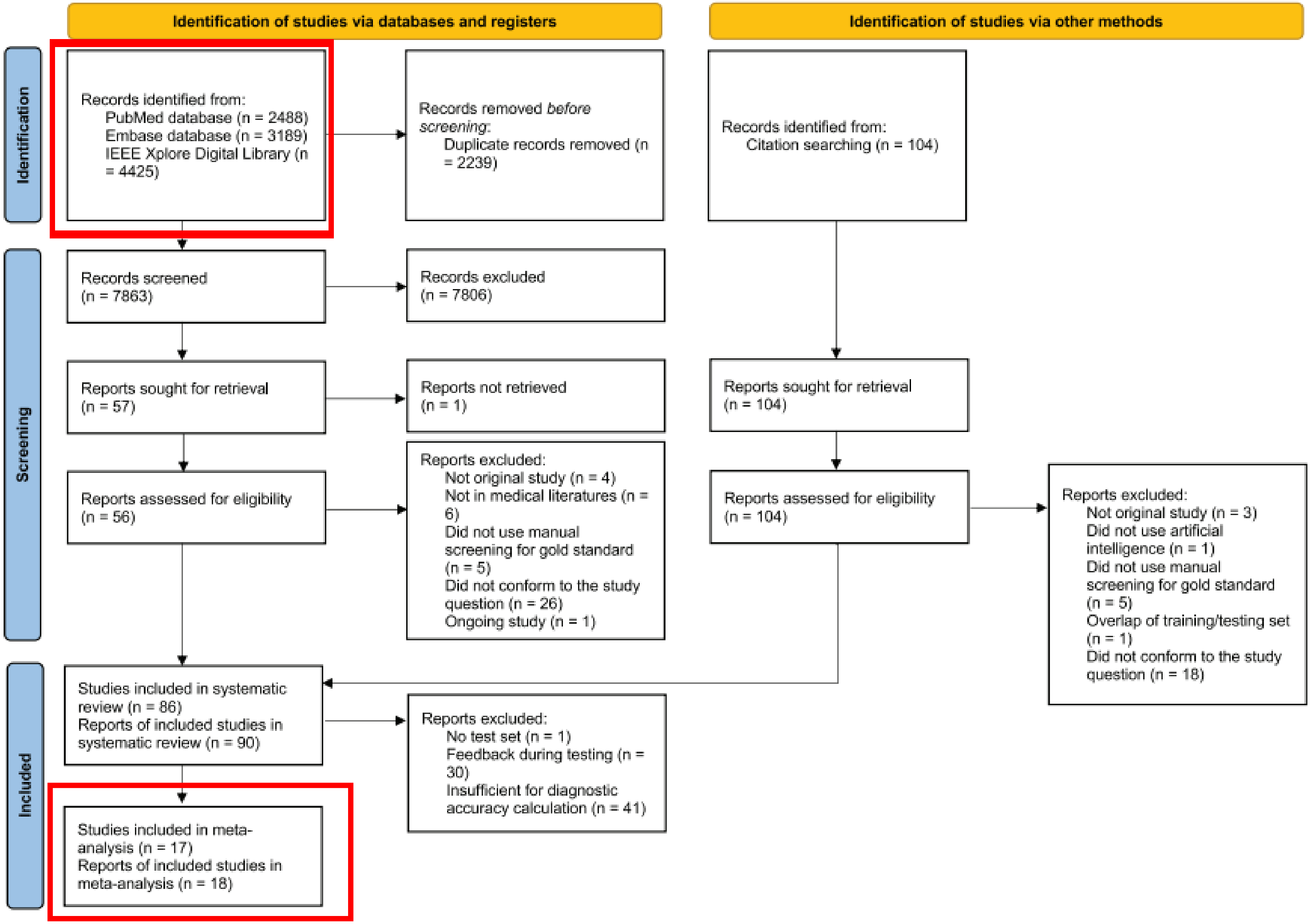
  (2) the recall and WSS when achieving the maximized precision

- Predefined subgroup analyses were conducted according to

  - AI algorithms (SVM VS other algorithms group)
  - Number of screened literatures for model validation
  - Fraction of included literatures

Results

Search and screening



Identification of studies via databases and registers

Identification of studies via other methods

Identification

Records identified from:
PubMed database (n = 2488)
Embase database (n = 3189)
IEEE Xplore Digital Library (n = 4425)

Records removed *before screening*:
Duplicate records removed (n = 2239)

Records identified from:
Citation searching (n = 104)

Screening

Records screened
(n = 7863)

Records excluded
(n = 7806)

Reports sought for retrieval
(n = 57)

Reports not retrieved
(n = 1)

Reports sought for retrieval
(n = 104)

Reports assessed for eligibility
(n = 56)

Reports excluded:
Not original study (n = 4)
Not in medical literatures (n = 6)
Did not use manual screening for gold standard (n = 5)
Did not conform to the study question (n = 26)
Ongoing study (n = 1)

Reports assessed for eligibility
(n = 104)

Reports excluded:
Not original study (n = 3)
Did not use artificial intelligence (n = 1)
Did not use manual screening for gold standard (n = 5)
Overlap of training/testing set (n = 1)
Did not conform to the study question (n = 18)

Included

Studies included in systematic review (n = 86)
Reports of included studies in systematic review (n = 90)

Reports excluded:
No test set (n = 1)
Feedback during testing (n = 30)
Insufficient for diagnostic accuracy calculation (n = 41)

Studies included in meta-analysis (n = 17)
Reports of included studies in meta-analysis (n = 18)

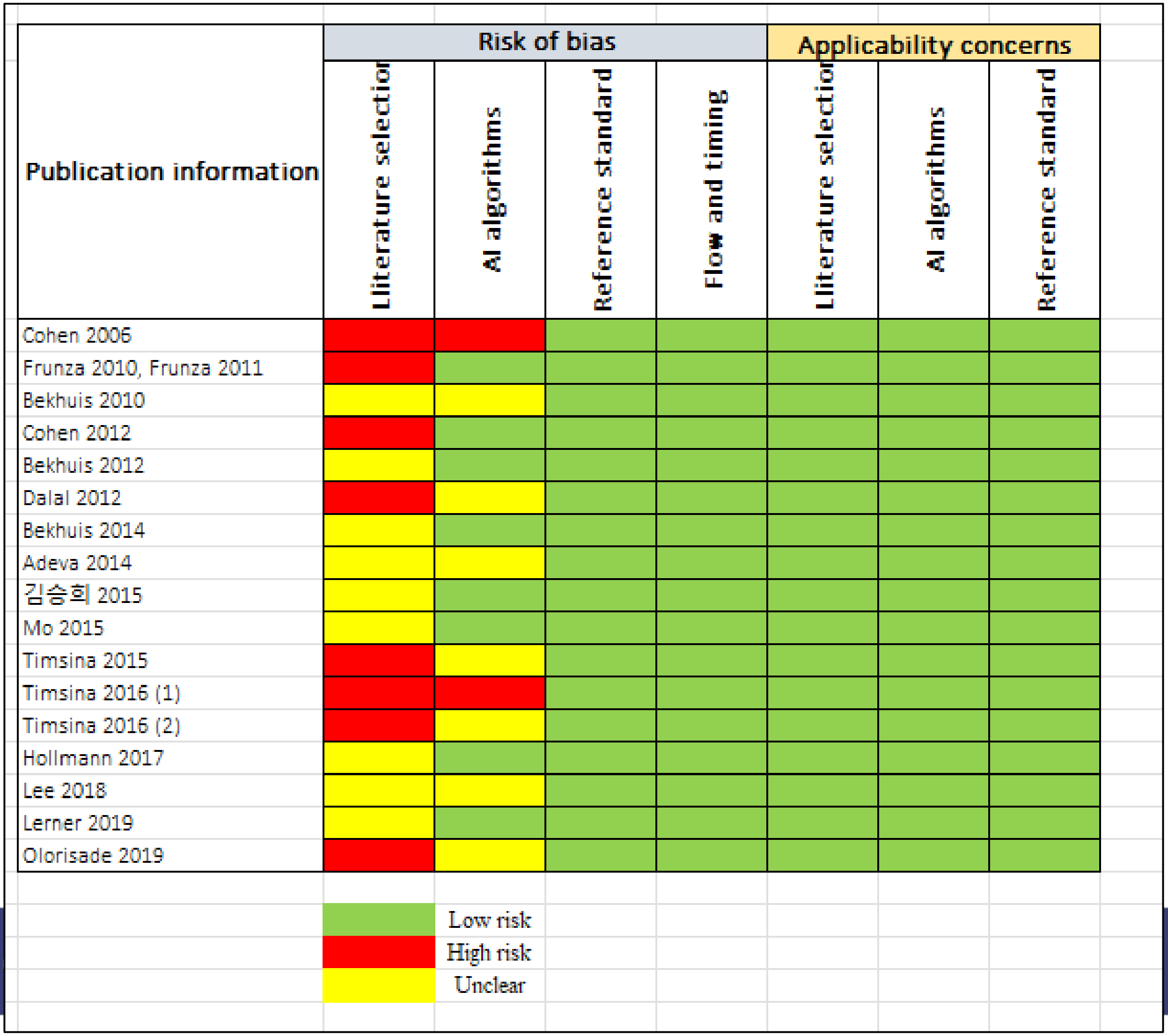
Figure 1. Review flow diagram.

Characteristics of included studies

- The studies included in the systematic review were published between 2006 and 2021.
- SVM was the most commonly used classifier for literature screening.
- NB, kNN, perceptron, random forest, convolutional neural networks, radial basis function kernel, and other algorithms were applied as well.

- Most automation studies used results of literature screening from existing systematic reviews to train and evaluate their classification.
- All studies used article titles, abstracts, and metadata rather than full texts for training or validation.

Risk of bias



Publication information	Risk of bias				Applicability concerns		
	Lliterature selection	AI algorithms	Reference standard	Flow and timing	Lliterature selection	AI algorithms	Reference standard
Cohen 2006	🟥	🟥	🟩	🟩	🟩	🟩	🟩
Frunza 2010, Frunza 2011	🟥	🟩	🟩	🟩	🟩	🟩	🟩
Bekhuis 2010	🟨	🟨	🟩	🟩	🟩	🟩	🟩
Cohen 2012	🟥	🟩	🟩	🟩	🟩	🟩	🟩
Bekhuis 2012	🟨	🟩	🟩	🟩	🟩	🟩	🟩
Dalal 2012	🟥	🟨	🟩	🟩	🟩	🟩	🟩
Bekhuis 2014	🟨	🟩	🟩	🟩	🟩	🟩	🟩
Adeva 2014	🟨	🟨	🟩	🟩	🟩	🟩	🟩
김승희 2015	🟨	🟩	🟩	🟩	🟩	🟩	🟩
Mo 2015	🟨	🟩	🟩	🟩	🟩	🟩	🟩
Timsina 2015	🟥	🟨	🟩	🟩	🟩	🟩	🟩
Timsina 2016 (1)	🟥	🟥	🟩	🟩	🟩	🟩	🟩
Timsina 2016 (2)	🟥	🟨	🟩	🟩	🟩	🟩	🟩
Hollmann 2017	🟨	🟩	🟩	🟩	🟩	🟩	🟩
Lee 2018	🟨	🟨	🟩	🟩	🟩	🟩	🟩
Lerner 2019	🟨	🟩	🟩	🟩	🟩	🟩	🟩
Olorisade 2019	🟥	🟨	🟩	🟩	🟩	🟩	🟩

🟩	Low risk
🟥	High risk
🟨	Unclear

Effectiveness and labor-saving indices

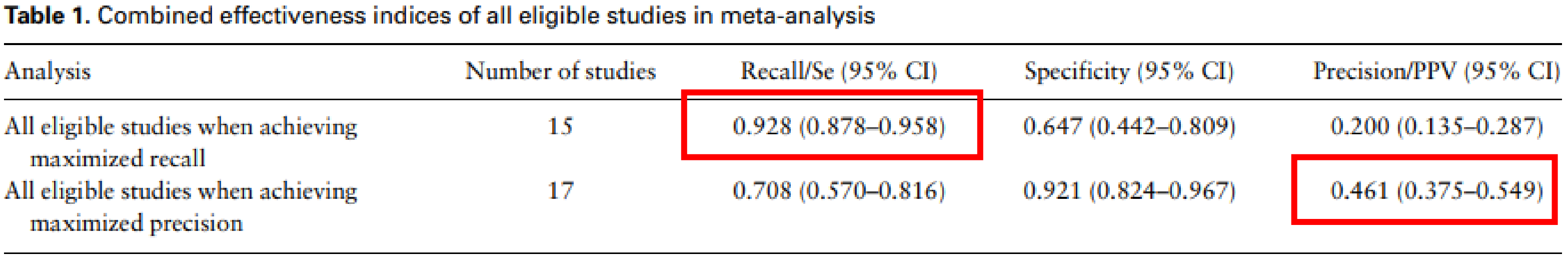


Table 1. Combined effectiveness indices of all eligible studies in meta-analysis

Analysis	Number of studies	Recall/Se (95% CI)	Specificity (95% CI)	Precision/PPV (95% CI)	WSS
All eligible studies when achieving maximized recall	15	0.928 (0.878–0.958)	0.647 (0.442–0.809)	0.200 (0.135–0.287)	-0.003 - 0.897
All eligible studies when achieving maximized precision	17	0.708 (0.570–0.816)	0.921 (0.824–0.967)	0.461 (0.375–0.549)	0.095 – 0.841

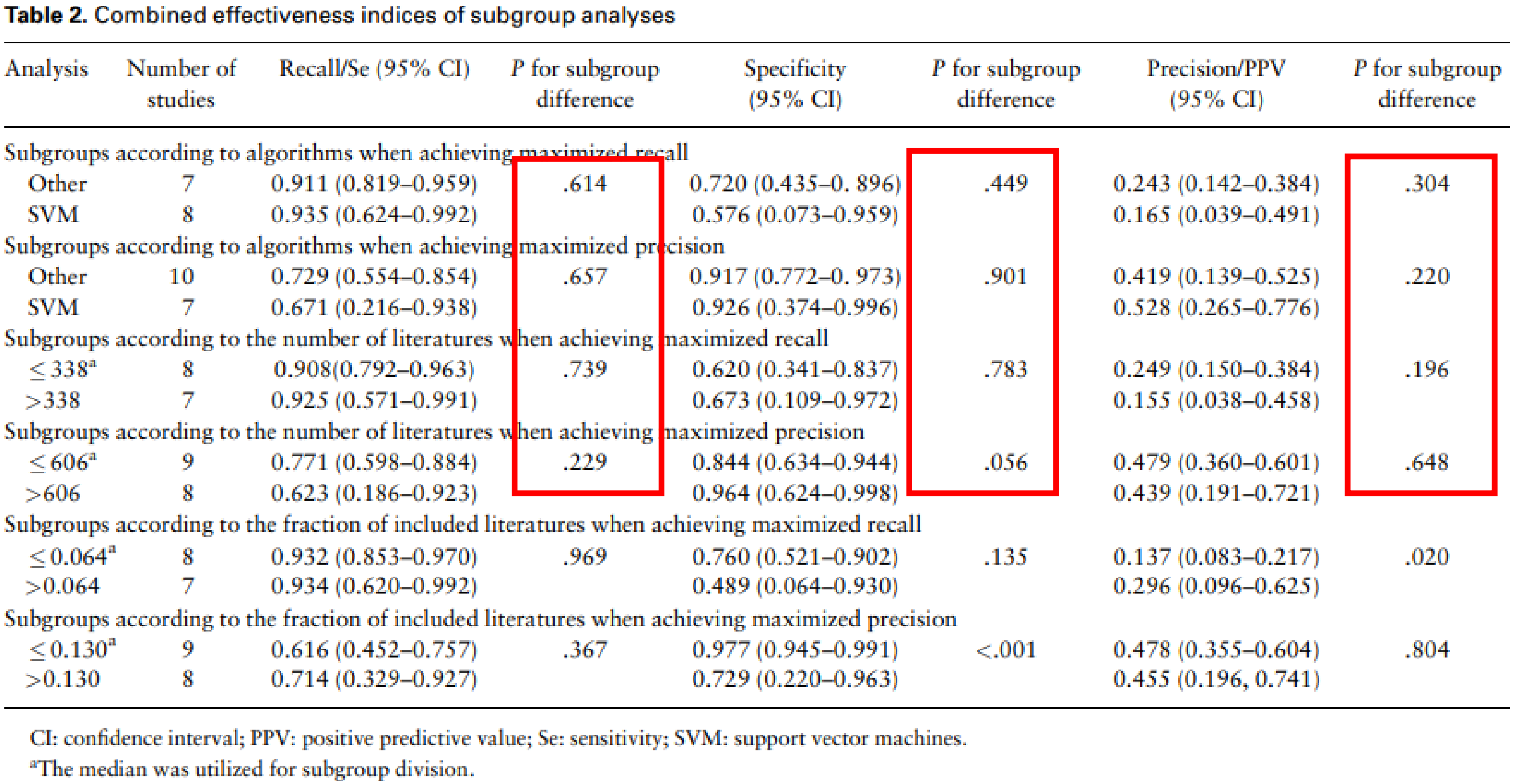


Table 2. Combined effectiveness indices of subgroup analyses

Analysis	Number of studies	Recall/Se (95% CI)	*P* for subgroup difference	Specificity (95% CI)	*P* for subgroup difference	Precision/PPV (95% CI)	*P* for subgroup difference
Subgroups according to algorithms when achieving maximized recall							
Other	7	0.911 (0.819–0.959)	.614	0.720 (0.435–0. 896)	.449	0.243 (0.142–0.384)	.304
SVM	8	0.935 (0.624–0.992)		0.576 (0.073–0.959)		0.165 (0.039–0.491)	
Subgroups according to algorithms when achieving maximized precision							
Other	10	0.729 (0.554–0.854)	.657	0.917 (0.772–0. 973)	.901	0.419 (0.139–0.525)	.220
SVM	7	0.671 (0.216–0.938)		0.926 (0.374–0.996)		0.528 (0.265–0.776)	
Subgroups according to the number of literatures when achieving maximized recall							
≤338[a]	8	0.908(0.792–0.963)	.739	0.620 (0.341–0.837)	.783	0.249 (0.150–0.384)	.196
>338	7	0.925 (0.571–0.991)		0.673 (0.109–0.972)		0.155 (0.038–0.458)	
Subgroups according to the number of literatures when achieving maximized precision							
≤606[a]	9	0.771 (0.598–0.884)	.229	0.844 (0.634–0.944)	.056	0.479 (0.360–0.601)	.648
>606	8	0.623 (0.186–0.923)		0.964 (0.624–0.998)		0.439 (0.191–0.721)	
Subgroups according to the fraction of included literatures when achieving maximized recall							
≤0.064[a]	8	0.932 (0.853–0.970)	.969	0.760 (0.521–0.902)	.135	0.137 (0.083–0.217)	.020
>0.064	7	0.934 (0.620–0.992)		0.489 (0.064–0.930)		0.296 (0.096–0.625)	
Subgroups according to the fraction of included literatures when achieving maximized precision							
≤0.130[a]	9	0.616 (0.452–0.757)	.367	0.977 (0.945–0.991)	<.001	0.478 (0.355–0.604)	.804
>0.130	8	0.714 (0.329–0.927)		0.729 (0.220–0.963)		0.455 (0.196, 0.741)	

CI: confidence interval; PPV: positive predictive value; Se: sensitivity; SVM: support vector machines.

[a] The median was utilized for subgroup division.

Discussion

- This is the first systematic review and meta-analysis in the area of automatic literature screening aimed to quantitatively evaluate the performance of AI methods and provide recommendations based on evidence.
- Literature screening is an imbalanced classification task, for the total number of screened literatures is large while fraction of included literatures is usually very low.

- The combined recall was 0.928 when achieving the maximized recall by optimizing the AI model. However, this value was only 0.708 when achieving the maximized precision, indicating that more literatures might be missed if the automation model focused on precision.

- The recall of model reflects the ability to correctly identify eligible literatures. More eligible literatures containing quality evidence would be missed by the low-recall models, introducing significant selection bias to systematic review.
- A low precision model would mistakenly identify many irrelevant literatures, leading to more paper-reading load in the follow-up manual screening.

- For medical evidence synthesis, the introduction of bias is unacceptable. Thus, in practice, a high level of recall should be prioritized to make sure the automatic screening process includes as many eligible literatures as possible.
- Current models would miss 7.2% of literatures on average when achieving maximized recall (combined recall: 0.928) (miss 29.2% when achieving maximized precision).
- We therefore recommend that recall should take priority over precision and other indices.

- Cohen et al assumed that a recall of 0.95 or greater might be required for the system to identify an adequate fraction of the relevant literatures, though no further evidence was given.
- Our findings provide direct evidence proving that a large number of studies failed to achieve the recall of 0.95 even using a high-recall strategy in the model training.

- We therefore propose that 0.95 is still an important benchmark of recall for future screening automation to hold.
- When a high recall is achieved, the secondary goal of training is to improve precision or specificity to decrease the false negative identification, as well as to save the work to review every literature.

- According to our results, the combined specificity and precision were 0.647 (95% CI, 0.442–0.809) and 0.200 (95% CI, 0.135– 0.287) when achieving maximized recall.
- The low ranges of specificity and precision indicate that more newly adjusted algorithms are required for efficiency improvement in literature screening.
- The results of the automation studies included in this review have limited generalizability given that the training datasets applied by these studies were mostly MEDLINE.

- The AI algorithms were divided into SVM and other algorithms, as current evidence showed that SVM classifiers performed well for text classification.
- The algorithms and the number of screened literatures were not found to affect the accuracy of automated literature screening indicating a relatively homogeneous effectiveness.

Limitations

- Due to diverse recall levels as well as missing reported WSS in many studies, we were unable to further analyze the work savings in this task.
- There was significant heterogeneity in literature topics for investigating the screening performance of different AI algorithms, which limits the generalizability of the findings.
- The reference standard is actually imperfect, since human investigators may still miss eligible literatures during screening.

Conclusion

- Workload reduction in automated medical literature screening has been acceptable, but the recall level of current automation studies still needs to be improved.
- Our findings suggest that a recall of 0.95 should be prioritized in the model training.
- We recommend to report recall and other indices separately rather than report average form such as F-score in automated medical literature screening.

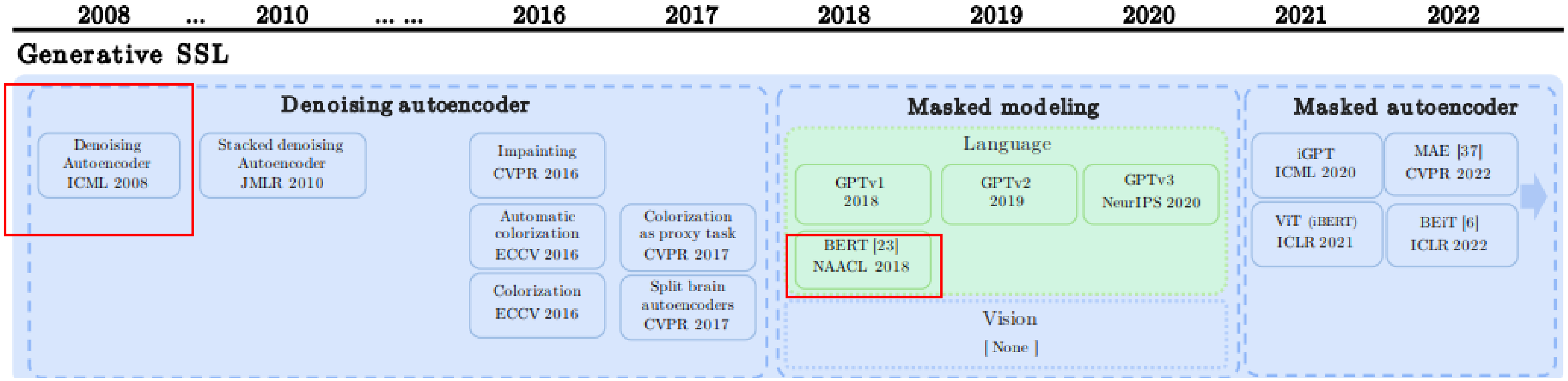
Comment for

Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews

Georgios Kontonatsios[a,*], Sally Spencer[b], Peter Matthew[a], Ioannis Korkontzelos[a]

[a] *Department of Computer Science, Edge Hill University, United Kingdom*
[b] *Faculty of Health and Social Care, Edge Hill University, United Kingdom*

1. Feature extraction method



Zhang C, Zhang C, Song J, Yi JS, Zhang K, Kweon IS. A Survey on Masked Autoencoder for Self-supervised Learning in Vision and Beyond. arXiv preprint arXiv:2208.00173. 2022 Jul 30.

BOW → statistical / no semantic meaning

“ I hate cat but love dog “ VS “ I hate dog but love cat” → same vector

BERT → contextual word embedding / semantic meaning

- BERT

Aum, S., & Choe, S. (2021). srBERT: automatic article classification model for systematic review using BERT. Systematic reviews, 10(1), 1-8.

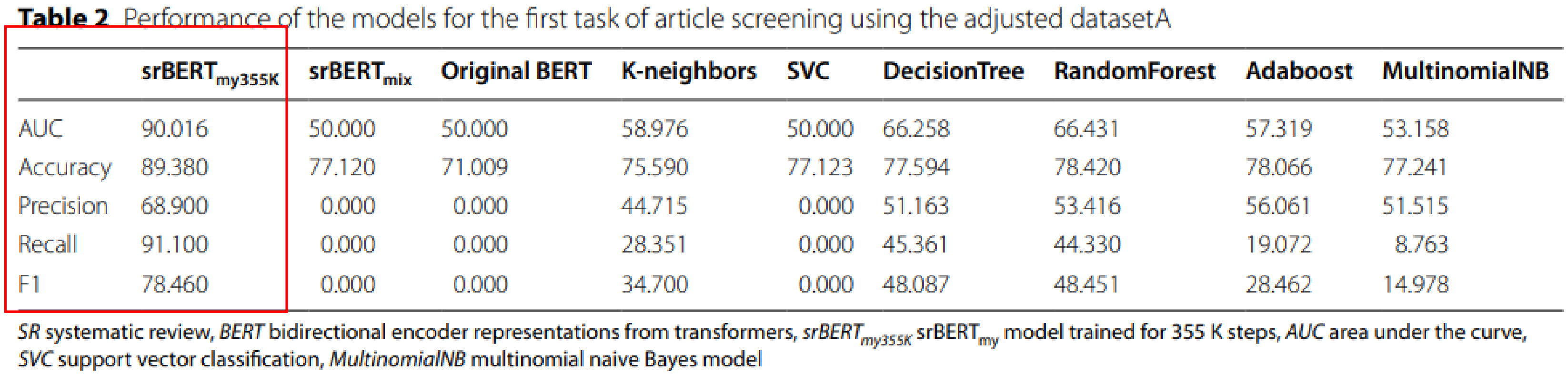


Table 2 Performance of the models for the first task of article screening using the adjusted datasetA

	srBERT$_{my355K}$	srBERT$_{mix}$	Original BERT	K-neighbors	SVC	DecisionTree	RandomForest	Adaboost	MultinomialNB
AUC	90.016	50.000	50.000	58.976	50.000	66.258	66.431	57.319	53.158
Accuracy	89.380	77.120	71.009	75.590	77.123	77.594	78.420	78.066	77.241
Precision	68.900	0.000	0.000	44.715	0.000	51.163	53.416	56.061	51.515
Recall	91.100	0.000	0.000	28.351	0.000	45.361	44.330	19.072	8.763
F1	78.460	0.000	0.000	34.700	0.000	48.087	48.451	28.462	14.978

SR systematic review, *BERT* bidirectional encoder representations from transformers, *srBERT*$_{my355K}$ srBERT$_{my}$ model trained for 355 K steps, *AUC* area under the curve, *SVC* support vector classification, *MultinomialNB* multinomial naive Bayes model

2. Model performance

Study reported only the WSS@95% metric –> avg. 0.564 (0.095 -0.848)

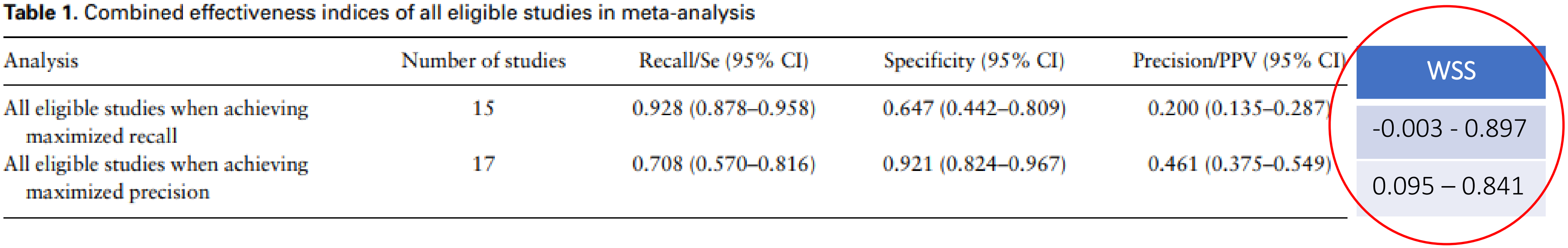


Table 1. Combined effectiveness indices of all eligible studies in meta-analysis

Analysis	Number of studies	Recall/Se (95% CI)	Specificity (95% CI)	Precision/PPV (95% CI)	WSS
All eligible studies when achieving maximized recall	15	0.928 (0.878–0.958)	0.647 (0.442–0.809)	0.200 (0.135–0.287)	-0.003 - 0.897
All eligible studies when achieving maximized precision	17	0.708 (0.570–0.816)	0.921 (0.824–0.967)	0.461 (0.375–0.549)	0.095 – 0.841

- WSS@95% metric   --- appropriate
- unable to directly compare the model performance (recall , precision ,specificity) with other studies.

THE END