Increasing comprehensiveness and reducing workload in a systematic review of complex interventions using automated machine learning

Olalekan A Uthman,[1*] Rachel Court,[1] Jodie Enderby,[1] Lena Al-Khudairy,[1] Chidozie Nduka,[1] Hema Mistry,[1] GJ Melendez-Torres,[2] Sian Taylor-Phillips[1] and Aileen Clarke[1]

[1]Warwick Medical School, University of Warwick, Coventry, UK
[2]Peninsula Technology Assessment Group (PenTAG), College of Medicine and Health, University of Exeter, Exeter, UK

[*]Corresponding author

Disclosure of interests

Full disclosure of interests: Completed ICMJE forms for all authors, including all related interests, are available in the toolkit on the NIHR Journals Library report publication page at https://doi.org/10.3310/UDIR6682.

Primary conflicts of interest: none.

Published November 2022
DOI: 10.3310/UDIR6682

This article should be referenced as follows:

Uthman OA, Court R, Enderby J, Al-Khudairy L, Nduka C, Mistry H, *et al*. Increasing comprehensiveness and reducing workload in a systematic review of complex interventions using automated machine learning. *Health Technol Assess* 2025;**29**(37):45–74. https://doi.org/10.3310/UDIR6682

Abstract

Increasing comprehensiveness and reducing workload in a systematic review of complex interventions using automated machine learning

Olalekan A Uthman,[1*] Rachel Court,[1] Jodie Enderby,[1]
Lena Al-Khudairy,[1] Chidozie Nduka,[1] Hema Mistry,[1]
GJ Melendez-Torres,[2] Sian Taylor-Phillips[1] and Aileen Clarke[1]

[1]Warwick Medical School, University of Warwick, Coventry, UK
[2]Peninsula Technology Assessment Group (PenTAG), College of Medicine and Health, University of Exeter, Exeter, UK

[*]Corresponding author olalekan.uthman@warwick.ac.uk

Background: As part of our ongoing systematic review of complex interventions for the primary prevention of cardiovascular diseases, we have developed and evaluated automated machine-learning classifiers for title and abstract screening. The aim was to develop a high-performing algorithm comparable to human screening.

Methods: We followed a three-phase process to develop and test an automated machine learning-based classifier for screening potential studies on interventions for primary prevention of cardiovascular disease. We labelled a total of 16,611 articles during the first phase of the project. In the second phase, we used the labelled articles to develop a machine learning-based classifier. After that, we examined the performance of the classifiers in correctly labelling the papers. We evaluated the performance of the five deep-learning models [i.e. parallel convolutional neural network (CNN), stacked CNN, parallel-stacked CNN, recurrent neural network (RNN) and CNN–RNN]. The models were evaluated using recall, precision and work saved over sampling at no less than 95% recall.

Results: We labelled a total of 16,611 articles, of which 676 (4.0%) were tagged as 'relevant' and 15,935 (96%) were tagged as 'irrelevant'. The recall ranged from 51.9% to 96.6%. The precision ranged from 64.6% to 99.1%. The work saved over sampling ranged from 8.9% to as high as 92.1%. The best-performing model was parallel CNN, yielding a 96.4% recall, as well as 99.1% precision, and a potential workload reduction of 89.9%.

Future work and limitations: We used words from the title and the abstract only. More work needs to be done to look into possible changes in performance, such as adding features such as full document text. The approach might also not be able to be used for other complex systematic reviews on different topics.

Conclusion: Our study shows that machine learning has the potential to significantly aid the labour-intensive screening of abstracts in systematic reviews of complex interventions. Future research should concentrate on enhancing the classifier system and determining how it can be integrated into the systematic review workflow.

Funding: This project was funded by the National Institute for Health and Care Research (NIHR) Health Technology Assessment programme and will be published in *Health Technology Assessment*. See the NIHR Journals Library website for further project information.

Contents

List of tables

Glossary

Deep learning A type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data.

False negative An outcome where the model incorrectly predicts the negative class.

False positive An outcome where the model incorrectly predicts the positive class.

Precision (also known as positive predictive value) The proportion of relevant instances among retrieved instances.

Recall (also known as sensitivity) The proportion of relevant instances that were retrieved.

Test data set A sample of data that is held back from the training of the model and instead is used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

Training data set The sample of data used to create the model.

True negative An outcome where the model correctly predicts the negative class.

True positive An outcome where the model correctly predicts the positive class.

Validation data set A sample of data held back from training of the model that is used to give an estimate of model skill while tuning the model's hyperparameters.

Work saved over sampling Defined as 'the percentage of papers that meet the original search criteria that the reviewers do not have to read (because they have been screened out by the classifier)' (Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc* 2006;**13**:206–19).

List of abbreviations

CNN	convolutional neural network	TP	true positive
CVD	cardiovascular disease	WSS	work saved over sampling
FN	false negative	WSS@95%	work saved over sampling at 95% recall
RNN	recurrent neural network		
TN	true negative		

Introduction

Systematic reviews of biomedical literature are the cornerstones of the creation of evidence-based clinical practice guidelines. Systematic assessments are used not only to assess the comparative effectiveness of medical procedures, but also as additional insight into decisions on reimbursement for technology globally. Systematic reviews require the time-consuming procedure of human reviewers manually screening title and abstract records to determine their importance to the analysis. This method can include reading thousands, or even tens of thousands, of journal abstracts. As the body of articles available continues to expand, this process becomes increasingly difficult and more expensive. In fact, the increasing cost of performing a systematic review of biomedical literature has resulted in the need to reduce the total workload of researchers.[1,2] One way this can be achieved is by applying natural language processing techniques to 'automate' the classification of publications that are potentially applicable to a given topic.[3–7]

Text-mining software and artificial intelligence algorithms help undertake broad reviews, which are increasingly common for improving efficiency and lowering the costs of completing these reviews.[6,8,9] Machine-learning algorithms are used mainly at the screening stage of the systematic review process. This screening stage entails the categorisation of records found from the search into 'relevant' or 'irrelevant' categories and is usually performed by two independent human reviewers with a third reviewer resolving any differences. Artificial intelligence text classification algorithms have proved to be highly effective for the identification of randomised controlled trials.[5,10–13] To our knowledge, no such attempt has been made when performing systematic reviews of complex interventions. Therefore, the aim of this study was to explore the viability and effectiveness of using machine-learning modelling to classify abstracts according to specific exclusion/inclusion criteria, as would be done in the first stage of a systematic review of a complex intervention.

Methods

We followed a three-phase process to develop and test an automated machine learning-based classifier for screening potentially relevant studies on interventions for the primary prevention of cardiovascular disease (CVD). We labelled a total of 16,611 articles during the first phase of the project. In the second phase, we used the labelled articles to develop machine learning-based classifiers. After that, we examined the performance of the classifiers in correctly labelling the papers and applied the best-performing one to the unseen records retrieved by a more sensitive version of the database searches used in the first phase of the project.

Building the text collections

The data set used for this study comprises a corpus of 16,611 titles and abstracts of articles (a subset of the 133,260 articles yielded by the more sensitive search), which were collected and labelled by a pair of human reviewers (with 0 representing a 'not relevant' study and 1 representing a 'relevant' study) from an ongoing systematic review on interventions for the primary prevention of CVD.[14] We evaluated each identified study against the following selection criteria.

Study population

Study populations were adults (≥ 18 years of age) included in population-based studies that may or may not be targeted at moderate/high CVD risk groups (e.g. hypertension, obesity, hyperlipidaemia, type 2 diabetes or a combination of these). As the review was to focus on the primary prevention of CVD, we excluded trials that included people who had experienced a previous myocardial infarction, stroke or revascularisation procedure (coronary artery bypass grafting or percutaneous transluminal coronary angioplasty) and those with angina or angiographically defined coronary heart disease. Studies with mixed populations (i.e. both those with and those without CVD) were included if data for the relevant primary prevention could be extracted.

Intervention

Included was any form of intervention aimed at the primary prevention of CVD, including but not limited to drugs (lipid-lowering medications, blood pressure-lowering medications, antiplatelet agents), diet (nutritional supplements, dietary interventions), physical activity and public health (health promotion programmes, structural and policy interventions).

Comparators

Comparators were other forms of intervention (e.g. minimal intervention, active intervention, concomitant intervention), placebo, usual-care or no-intervention control group, or waiting list control.

Outcome measures

The primary outcome was all-cause mortality. Secondary outcomes were CVD-related mortality, major cardiovascular events (defined as fatal and non-fatal myocardial infarction, sudden cardiac death, revascularisation, fatal and non-fatal stroke, and fatal and non-fatal heart failure), coronary heart disease (fatal and non-fatal myocardial infarction and sudden cardiac death, excluding silent myocardial infarction) and incremental costs per quality-adjusted life-years gained reported alongside a randomised trial.

Study design

The included design was a randomised controlled trial with at least 6 months' follow-up. The units of randomisation could be either individuals or clusters (e.g. family or workplace).

About 80% of the full set of data ($n = 13,288$) was taken as a random sample for training, and the other 20% ($n = 3323$) was saved for testing. Before the machine was trained on the data, the data had to be processed. The machine was then put to the test on the testing set, and its predictions were compared with how the data had been labelled by humans. It is important to note that the machine sees the processed abstract as well as the human reviewer's classification during training, but, during testing, the machine only sees the processed abstract and produces its own classifications, which are then compared with those of the human reviewers.

Classifier system

We developed automated machine-learning classifiers using Uber's Ludwig low code deep-learning toolbox.[15] Ludwig is a user-friendly deep-learning toolbox that allows users to train and test deep-learning models for a variety of applications, including text classification. We developed and compared the performance of the five deep-learning models [parallel convolutional neural network (CNN), stacked CNN, parallel-stacked CNN, recurrent neural network (RNN) and CNN–RNN].[16–21] Deep learning refers to 'neural networks with multiple layers of perceptrons inspired by the human brain' and has been shown to bring benefits in text generation, word representation estimation, sentence classification and feature presentation.[22–36] Recently, deep-learning models have been shown to perform better than traditional machine-learning algorithms in many natural language processing text classification applications.[22–36]

Deep-learning methods, in addition to their high performance, do not rely on any hand-engineered features and can instead learn the most appropriate features for a given task. As a result, deep learning is currently the most popular and successful approach to natural language processing. The most common deep-learning architectures are CNN and RNN.[22,23] CNN is a type of deep, feed-forward artificial neural network (where node connections do not form a cycle) that employs a variant of multilayer perceptrons that is designed to require minimal pre-processing.[22,23] A RNN is a type of artificial neural network in which node connections form a directed graph along a sequence. Theoretically, RNN with a sequential architecture should be better suited for sequence modelling tasks (e.g. machine translation, language modelling or speech recognition) because these tasks require the representation of complex context dependencies.[22,23] Similarly, CNN's hierarchical architecture should be more useful for text classification, where detecting representative patterns can be critical to solving the problem.[22,23] In fact, CNN architectures have been shown to outperform state-of-the-art text classification algorithms because they can extract the most informative ngrams (i.e. a contiguous sequence of *n* items from a given sample of text or speech) describing a text.[22,23]

Evaluating the classifiers

We used the conventional information retrieval terminology of recall and precision, which are synonymous with sensitivity and positive predictive value, respectively, to evaluate the classifiers. In the current scenario, the recall statistic is of primary concern, as a means of confirming that eligible study reports are not discarded incorrectly from the systematic review. Precision is also an interesting metric because it can be used to calculate the number of articles that require manual screening by reviewers. We were concerned with the number of irrelevant articles that were incorrectly classified as relevant by machine-learning classifiers (i.e. records with an assigned probability score greater than the identified threshold score), which had to be manually filtered out by the reviewers. Precision was calculated as the proportion of retrieved articles that genuinely report a relevant study. Recall and precision were calculated using a two-by-two table that represented 'relevant'/'irrelevant' articles and whether they were classified correctly or incorrectly (*Table 1*).

TABLE 1 The 2 × 2 table from which precision and recall were calculated

	Relevant (human labelled)	Irrelevant (human labelled)
Machine learning classed as 'relevant'	True positives	False positives
Machine learning labelled as 'irrelevant'	False negatives	True negatives

The following formulas are used to calculate precision and recall:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}. \tag{1}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}. \tag{2}$$

To be useful for systematic reviews, a document classification system must save reviewers the time and effort of reading each paper. Simultaneously, the number of missed papers containing high-quality evidence must be very small. For this study, we assumed that the system needed a recall of ≥ 0.95 to identify an adequate fraction of the positive papers.[37,38] As long as the recall is at least 0.95, precision should be as high as possible.[37,38]

Furthermore, the most important aspect of the system to evaluate is how much future work the reviewers could save in the process of screening abstracts and titles.[38] We defined work saved as the percentage of papers that meet the original search criteria but are not required to be read by the reviewers (because they have been screened out by the classifier). With a 0.95 random sampling of the data, a recall of 0.95 can be obtained, and this process would save the reviewers 5% of the time spent reading the papers. Clearly, for the classifier system to be advantageous, the amount of work saved must be greater than the amount of work saved by simple random sampling. As a result, for a given level of recall, we measure the work saved in addition to the work saved by simple sampling.[38,39] We define work saved over sampling (WSS) as follows:

$$\text{WSS} = \frac{(\text{TN} + \text{FN})}{N} - (1.0 - R) = \frac{\text{TN} + \text{FN}}{N} - 1 + \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

where TP represents the number of true positives identified by the classifier, TN represents the number of true negatives identified by the classifier and FN represents the number of false negatives identified by the classifier. The recall is *R*, and the total number of samples in the test set is *N*. We have fixed recall at 0.95[37,38] for the current work, so WSS at 95% recall (WSS@95%) is:

$$\text{WSS@95\%} = \frac{(\text{TN} + \text{FN})}{N} - 0.05. \tag{4}$$

Results

We 'human' labelled a total of 16,611 articles. Only 676 (4%) articles were labelled as 'relevant', and 15,935 were labelled as 'irrelevant' (96%). *Table 2* shows the performance of the five machine-learning approaches from Uber's Ludwig toolkit. The recall ranged from as low as 51.9% to as high as 96.6%. Only parallel CNN reaches the desired recall (also known as sensitivity) of 95%. The precision ranged from as low as 64.6% to as high as 99.1%. The work saving ranged from 8.9% to as high as 92.1%. The best-performing model (recall 96.64%) was parallel CNN, yielding a 96.4% recall, as well as 99.1% precision, and a potential workload reduction of 92%. Parallel CNN was chosen as the best-performing classifier to be applied to all other unseen citations identified from our searches.

TABLE 2 Performance of the five classifiers developed for identifying studies relevant for systematic review of complex interventions for primary prevention of CVD

Classifier type	Recall	Precision	WSS@95%
Parallel CNN	96.64	99.08	89.9
Stacked CNN	89.26	99.08	52.9
Parallel-stacked CNN	93.02	98.95	92.1
RNN	56.65	64.61	17.8
CNN–RNN	50.89	98.00	8.9

Discussion

Main findings

We developed five separate artificial intelligence models to predict whether or not a paper is applicable for our systematic analysis of complex interventions. The best-performing classifier exhibited excellent performance on the testing set, yielding a recall of 96.4%, as well as a precision of 99.1%, and demonstrating the potential for a workload saving of 92%. This automated approach has the greatest potential for reducing workload in conducting systematic literature reviews of complex interventions.

Comparison with previous studies

To optimise productivity, support vector machine-based natural language processing approaches have been developed to classify important medical literature papers on a variety of topics. In 2005, Aphinyanaphongs *et al.*[40] created the first support vector machine tool to assist in the systematic analysis of literature by defining specific documents in the field of internal medicine. Several similar methods were subsequently suggested, including one introduced by Wallace *et al.*[7] that involves active learning to minimise annotation costs.[41] Wallace *et al.*[42] reduced the number of papers to be checked manually by approximately 50% for a systematic review. Fiszman *et al.*[43] developed a framework that uses symbolic significance analysis to classify potentially important documents for cardiovascular risk factor guidance. The recall of their system was 56% and the accuracy was 91%.[44] While most of the current approaches concentrate on clinical literature, Miwa *et al.*[3] expanded the reach of their approach to social science literature. CNN-based natural language processing methodologies have been introduced for short-text and sentence classification.[19] However, few methodologies have been introduced and evaluated for the classification of medical literature. Using the probability of bias in text classification data sets, Zhang *et al.*[16] developed a CNN model to test the bias of the study design in the literature on randomised clinical trials. The reduction in workload ranged from 64.0% to 75.0%.

Study limitations and strengths

There are some drawbacks to our existing methods. We used only title and abstract terms, MeSH (medical subject headings) and MEDLINE publication styles as possible classification features. Additional work needs to be carried out to examine potential changes in performance, including additional features, such as full document text. Furthermore, it has recently been reported that deep-learning methods surpass conventional machine-learning algorithms in many natural language processing applications.[39,45,46] In addition to the high-precision, deep-learning methods do not rely on hand-engineered features; rather, they are able to learn the most suitable features for a specific task. Another possible limitation is the generalisability of the approach to other complex systematic reviews of other topics. Methodological research is required to assess the validity of developing high-performing classifiers for a wide range of topics. Deep learning, therefore, is potentially the leading and most promising path to natural language processing at the moment. Our reason for using this approach in our review was to efficiently increase the comprehensiveness of our searches and save workload in our update searches. CNN has been shown to be better at classifying text, where finding representative patterns is often the key to solving a text classification problem.[47] In fact, CNN architectures have shown to be better at text classification than the state-of-art algorithms in text classifications because they can find the most informative ngrams that describe a text.[18,47] 'Traditional' algorithms look at data based on words or features, but they do not look at the meaning of a sentence as a whole.[48] CNN, on the other hand, figures out how words are related without using a parser or a vocabulary.[49] We are aware of other methods of potentially increasing the comprehensiveness of searches efficiently, such as forward-citation searching, but the breadth and number of interventions in this review made these less attractive. In addition, we did not conduct error analysis to check the misclassified studies by the best-performing classifier. If the recommended recall threshold was not met, error analysis could have been performed, and a second model could have been built using updated training data.[37]

Patient and public involvement

Drawing on INVOLVE's guidance and support for best practice, we worked closely with three dedicated patient and public involvement advisors, and we invited guidance and support from our advisors during the preparatory phase of the project.

Conclusions

Our study shows that machine learning has the potential to significantly aid the labour-intensive screening of abstracts in systematic reviews of complex interventions. This is an exciting first step in applying machine-learning methods to the systematic review of complex interventions, which, if further developed, has the potential to revolutionise the systematic review process by allowing researchers to better manage the massive number of papers they must read during a systematic review. Future research should concentrate on enhancing the classifier system and determining how it can be integrated into the systematic review workflow. Further research to compare alternative information retrieval approaches with machine learning and to compare 'off-the-shelf' machine-learning applications with the use of these effective architectures would be useful.

Acknowledgements

Contributions of authors

Olalekan A Uthman (https://orcid.org/0000-0002-8567-3081) (Professor, Evidence Synthesis) contributed to the protocol, study selection, data extraction, validity assessments and synthesis of the included studies. He developed the classifiers and undertook the analyses. He also contributed to the interpretation of the results and the writing of the report and had overall responsibility for the project.

Rachel Court (https://orcid.org/0000-0002-4567-2586) (Information Specialist) contributed to the protocol development, developed the search strategies, conducted a range of searches to locate studies and wrote the sections of the report relating to the literature searches. She also contributed to the protocol and interpretation of the results and commented on drafts of the report.

Jodie Enderby (https://orcid.org/0000-0002-1446-7512) (Research Associate) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. She also contributed to the interpretation of the results and the writing of the report.

Lena Al-Khudairy (https://orcid.org/0000-0003-0638-583X) (Associate Professor, Evidence Synthesis) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. She also contributed to the interpretation of the results and the writing of the report.

Chidozie Nduka (https://orcid.org/0000-0001-7031-5444) (Senior Research Fellow, Evidence Synthesis) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. He also contributed to the interpretation of the results and the writing of the report.

Hema Mistry (https://orcid.org/0000-0002-5023-1160) (Associate Professor, Health Economics) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. She also contributed to the interpretation of the results and the writing of the report.

GJ Melendez-Torres (https://orcid.org/0000-0002-9823-4790) (Professor, Evidence Synthesis) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. He also contributed to the interpretation of the results and the writing of the report.

Sian Taylor-Phillips (https://orcid.org/0000-0002-1841-4346) (Professor, Evidence Synthesis) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. She also contributed to the interpretation of the results and the writing of the report.

Aileen Clarke (https://orcid.org/0000-0001-8299-3146) (Professor, Evidence Synthesis) contributed to the protocol, study selection, validity assessments and synthesis of the included studies. She also contributed to the interpretation of the results and the writing of the report.

Ethics statement

This work is a systematic review of accessing, processing and analysing data that has already been published and is available to the public. As a result, no patient data were processed, and patient consent and/or registration through human research ethics committees were therefore not relevant.

Data-sharing statement

No new data have been created in the preparation of this article and therefore there is nothing available for access and further sharing. All queries should be submitted to the corresponding author.

Funding

This project was funded by the National Institute for Health and Care Research (NIHR) Health Technology Assessment programme and will be published in *Health Technology Assessment*. See the NIHR Journals Library website for further project information.

Article history

The research reported in this article was funded by the HTA programme under project number 17/148/05. The contractual start date was in February 2019. This article began editorial review in December 2021 and was accepted for publication in June 2022. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and production house have tried to ensure the accuracy of the authors' article and would like to thank the reviewers for their constructive comments on this article document. However, they do not accept liability for damages or losses arising from material published in this article.

Disclaimer

This article presents independent research funded by the National Institute for Health and Care Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care.

References

1. Borah R, Brown AW, Capers PL, Kaiser KA. 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. https://doi.org/10.1136/bmjopen-2016-012545

2. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp Clin Trials Commun* 2019;**16**:100443. https://doi.org/10.1016/j.conctc.2019.100443

3. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. *J Biomed Inform* 2014;**51**:242–53. https://doi.org/10.1016/j.jbi.2014.06.005

4. 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. https://doi.org/10.1186/s13643-015-0067-6

5. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, *et al.* Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol* 2021;**133**:140–51. https://doi.org/10.1016/j.jclinepi.2020.11.003

6. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev* 2014;**3**:74. https://doi.org/10.1186/2046-4053-3-74

7. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinform* 2010;**11**:55. https://doi.org/10.1186/1471-2105-11-55

8. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015;**4**:5. https://doi.org/10.1186/2046-4053-4-5

9. Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, *et al.* SWIFT-Review: a text-mining workbench for systematic review. *Syst Rev* 2016;**5**:87. https://doi.org/10.1186/s13643-016-0263-z

10. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. *J Am Med Inform Assoc* 2015;**22**:707–17. https://doi.org/10.1093/jamia/ocu025

11. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Res Synth Methods* 2018;**9**:602–14. https://doi.org/10.1002/jrsm.1287

12. Noel-Storr A, Dooley G, Elliott J, Steele E, Shemilt I, Mavergames C, *et al.* An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomised trials. *J Clin Epidemiol* 2021;**133**:130–9. https://doi.org/10.1016/j.jclinepi.2021.01.006

13. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *J Am Med Inform Assoc* 2017;**24**:1165–8. https://doi.org/10.1093/jamia/ocx053

14. Uthman OA, Al-Khudairy L, Nduka CU, Court R, Mistry H, Melendez-Torres GJ, *et al.* Determining optimal strategies for primary prevention of cardiovascular disease: systematic review, cost-effectiveness review and network meta-analysis protocol. *Syst Rev* 2020;**9**:105. https://doi.org/10.1186/s13643-020-01366-x

15. Molino P, Dudin Y, Miryala SS. Ludwig: a type-based declarative deep learning toolbox. *arXiv* 2019;1909.07930.

16. Zhang Y, Marshall I, Wallace BC. *Rationale-Augmented Convolutional Neural Networks for Text Classification*. Paper presented at Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, November 2016. https://doi.org/10.18653/v1/D16-1076

17. Zhang H, Xiao L, Wang Y, Jin Y. *A Generalized Recurrent Neural Architecture for Text Classification with Multi-Task Learning*. Paper presented at Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia, 19–25 August 2017. https://doi.org/10.24963/ijcai.2017/473

18. Wang J, Wang Z, Zhang D, Yan J. *Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification*. Paper presented at Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia, 19–25 August 2017. https://doi.org/10.24963/ijcai.2017/406

19. Kim Y. *Convolutional Neural Networks for Sentence Classification*. Paper presented at Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014. https://doi.org/10.3115/v1/D14-1181

20. Conneau A, Schwenk H, Barrault L, Lecun Y. *Very Deep Convolutional Networks for Text Classification*. Paper presented at Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, 2017. https://doi.org/10.18653/v1/E17-1104

21. Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. *Trans Assoc Comput Linguistics* 2016;**4**:357–70. https://doi.org/10.1162/tacl_a_00104

22. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, *et al.* A guide to deep learning in healthcare. *Nat Med* 2019;**25**:24–9. https://doi.org/10.1038/s41591-018-0316-z

23. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 2021;**8**:53. https://doi.org/10.1186/s40537-021-00444-8

24. Balakrishnan V, Shi Z, Law CL, Lim R, Teh LL, Fan Y. A deep learning approach in predicting products' sentiment ratings: a comparative analysis. *J Supercomput* 2022;**78**:7206–26. https://doi.org/10.1007/s11227-021-04169-6

25. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, *et al.* A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019;**19**:1. https://doi.org/10.1186/s12911-018-0723-6

26. Short RG, Bralich J, Bogaty D, Befera NT. Comprehensive word-level classification of screening mammography reports using a neural network sequence labeling approach. *J Digit Imaging* 2019;**32**:685–92. https://doi.org/10.1007/s10278-018-0141-4

27. Hernandez V, Suzuki T, Venture G. Convolutional and recurrent neural network for human activity recognition: application on American sign language. *PLOS ONE* 2020;**15**:e0228869. https://doi.org/10.1371/journal.pone.0228869

28. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, *et al.* Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020;**27**:457–70. https://doi.org/10.1093/jamia/ocz200

29. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci* 2021;**2**:420. https://doi.org/10.1007/s42979-021-00815-1

30. Bangyal WH, Qasim R, Rehman NU, Ahmad Z, Dar H, Rukhsar L, *et al.* Detection of fake news text classification on COVID-19 using deep learning approaches. *Comput Math Methods Med* 2021;**2021**:5514220. https://doi.org/10.1155/2021/5514220

31. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, *et al.* Entity recognition from clinical texts via recurrent neural network. *BMC Med Inform Decis Mak* 2017;**17**:67. https://doi.org/10.1186/s12911-017-0468-7

32. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *J Am Med Inform Assoc* 2019;**26**:1247–54. https://doi.org/10.1093/jamia/ocz149

33. Obeid JS, Heider PM, Weeda ER, Matuskowitz AJ, Carr CM, Gagnon K, *et al.* Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Stud Health Technol Inform* 2019;**264**:283–7. https://doi.org/10.3233/SHTI190228

34. Prabhakar SK, Won DO. Medical text classification using hybrid deep learning models with multihead attention. *Comput Intell Neurosci* 2021;**2021**:9425655. https://doi.org/10.1155/2021/9425655

35. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019;**7**:e12239. https://doi.org/10.2196/12239

36. Zhang Z, Cheng H, Yang T. A recurrent neural network framework for flexible and adaptive decision making based on sequence learning. *PLOS Comput Biol* 2020;**16**:e1008342. https://doi.org/10.1371/journal.pcbi.1008342

37. Bannach-Brown A, Przybyła P, Thomas J, Rice ASC, Ananiadou S, Liao J, Macleod MR. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst Rev* 2019;**8**:23. https://doi.org/10.1186/s13643-019-0942-7

38. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc* 2006;**13**:206–19. https://doi.org/10.1197/jamia.M1929

39. Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Med Inform Decis Mak* 2012;**12**:33. https://doi.org/10.1186/1472-6947-12-33

40. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform Assoc* 2005;**12**:207–16. https://doi.org/10.1197/jamia.M1641

41. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. *Int J Comput Biol Drug Des* 2013;**6**:5–17. https://doi.org/10.1504/IJCBDD.2013.052198

42. Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, *et al.* Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genet Med* 2012;**14**:663–9. https://doi.org/10.1038/gim.2012.7

43. Fiszman M, Bray BE, Shin D, Kilicoglu H, Bennett GC, Bodenreider O, *et al.* Combining relevance assignment with quality of the evidence to support guideline development. *Stud Health Technol Inform* 2010;**160**:709–13.

44. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, *et al.* Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Res Synth Methods* 2014;**5**:31–49. https://doi.org/10.1002/jrsm.1093

45. Che W, Zhang Y. Deep learning in lexical analysis and parsing. In Deng L, Liu Y, editors. *Deep Learning in Natural Language Processing*. Singapore: Springer; 2018. pp. 79–116. https://doi.org/10.1007/978-981-10-5209-5_4

46. Artetxe M, Labaka G, Agirre E. *Unsupervised Statistical Machine Translation*. Paper presented at Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October–November 2018. https://doi.org/10.18653/v1/D18-1399

47. Segura-Bedmar I, Colón-Ruíz C, Tejedor-Alonso MÁ, Moro-Moro M. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *J Biomed Inform* 2018;**87**:50–9. https://doi.org/10.1016/j.jbi.2018.09.012

48. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review* 2020;**53**:5455–516. https://doi.org/10.1007/s10462-020-09825-6

49. Kalchbrenner N, Grefenstette E, Blunsom P. *A Convolutional Neural Network for Modelling Sentences*. Paper presented at Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Volume 1 Long Papers, Baltimore, MD, USA, June 2014. https://doi.org/10.3115/v1/P14-1062