Automated Screener Based on Convolutional Neural Network for Randomized Controlled Trials in Chinese Language: A Comparative Study of Different Classification Strategies

Shengkai Chen  
Sichuan University West China College of Stomatology

Bochun Mao  
Peking University School of Stomatology  https://orcid.org/0000-0002-2598-0282

Yu Xie  
Sichuan University West China College of Stomatology

Pan Yao  
Sichuan University

Chunjie Li  
Sichuan University West China College of Stomatology

Sijin Yang  
Southwest Medical University

Li Dong  
The Affiliated Hospital of Southwest Medical University

Bo Li (libo2004051192@163.com)  
Southwest Medical University

Research

Keywords: Evidence-based medicine, Convolutional neural networks, Chinese literature, Citation screening, Meta-analysis.

DOI: https://doi.org/10.21203/rs.3.rs-713655/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Objective: To explore the influence of modified literature classification strategies of Chinese biomedical literature on an automated screener based on conventional algorithm.

Methods: Citations of studies indexed as ‘Oral Science’ published in Chinese between 2014 and 2018 were retrieved from the China National Knowledge Infrastructure. Apart from dividing the studies into 2 categories (RCTs and non-RCTs), 3-category (RCTs, may-be-RCTs, and non-RCTs) and 5-category (RCTs, randomization-unclear controlled trials, non-randomized clinical trials/studies, non-clinical literature, and unclear) classification were also employed. The multi-category strategies took into consideration the diversity of study types and the presence of expression vagueness. Similar to real-world practice, full-text-needed studies included those that certainly concerned RCTs and those that might be RCTs but lacked information in their abstracts. Screening and classification were performed independently by 2 experienced researchers. The classification results after peer discussion and/or senior decision were used for the training of the CNN model. The probability thresholds for the classification of each category were set at a high sensitivity level. The area under the receiver-operator curve (AUC) was calculated when applicable. An isolated sample of citations was used in a prospective comparative trial that compared the sensitivity (SEN) and specificity (SPE) of screening RCTs, may-be-RCTs, and full-text-needed studies by using algorithms with different strategies and manual screening.

Results: In total, 12,166 citations were used for CNN model training. All 3 training strategies performed well in RCTs-screening with AUCs being higher than 0.99. The training exhibited that, when screening for RCTs, the 5- and 3-category strategies can yield better performance than the 2-category strategy. When screening for may-be-RCTs and full text-needed studies, the 5-category model achieved better SENs while the 3-category model achieved higher SPEs. The comparative trial with 1,422 samples presented similar results.

Conclusion: The CNN algorithm has promising results in the automatic screening of Chinese literature. The multi-category training strategies considering different study types and expression vagueness are more suitable for CNN training and can help achieve better screening sensitivity and specificity.

Introduction

Evidence-based medicine (EBM), which is based on the principle that every clinical decision should be made according to the best clinical evidence, has significant value for clinical practice and research. Systematic reviews (SRs) are the foundation of EMB and final clinical guidelines. In general, an SR involves protocol design, literature searching, citation screening, full text obtaining, data extraction, critical appraisal, statistical analysis, and writing. Among these steps, literature searching and citation screening determine how systematic a review can be.

To minimize any potential bias from the design of a study, researchers often need, whenever possible, to screen for randomized controlled trials (RCTs) for SRs that focus on medical interventions. Due to that
SRs must be comprehensive with all available evidence, researchers often adopt search strategies with high sensitivity (> 0.98) and relatively low specificity (< 0.75), which often yield thousands of references to be screened at the title and abstract level. Moreover, due to the ever-growing corpus of published literature, performing abstract-level screening is a task that is increasingly time-consuming. In 2015 alone, an average of 100 manuscripts describing RCTs for medical interventions were published daily. Thus, this task is one of the several bottlenecks in performing SRs that is both time and money consuming. SRs concerning clinical trials are a critical resource supporting policy and clinical decision-making; however, it is challenging to keep them up-to-date. The production and update of SRs is resource-intensive, while the constantly increasing volume of new evidence produced can outpace our ability to keep up. Researchers have found that to complete an SR, an average of 1,781 studies need to be screened with a screen-out rate of 97.1%. In another study, it was revealed that it takes an average of 2 years before the publication of a Cochrane review, which is far from the requirement from up-to-date. It is well known that SRs require highly skilled reviewers for completing a series of very specialized manual and repetitive tasks. Interestingly, a study has indicated that, if not volunteered, the costs of some SRs may reach approximately $141,194.80 to conduct. A fundamental problem is that current screening methods, though rigorous enough, simply are not designed to meet the demands imposed by the voluminous scale of the current evidence base.

To tackle this crucial problem, several efforts have been made which can be categorized into 3 main subfields: modified search filters, crowdsourcing, and machine learning (ML). Search filters are based on combinations of text strings and database tags, which are developed by information specialists. Current best-performing filters achieve near-perfect sensitivity, but they provide comparatively low specificity. Crowdsourcing, such as the Cochrane Crowd which is the most well-known crowdsourcing task, has been used among others in RCT-screening, and table and figure selection. However, some serious disadvantages of crowdsourcing have prevented it from becoming popular. These include the unsatisfactory inter-person consistency and the extensive efforts required to perform a crowdsourcing task.

ML refers to the application of artificial intelligence (AI) that enables computer systems to learn and improve from experience, typically from large amounts of training data, without being explicitly programmed. The use of ML in the field of text analyzing is called natural language processing (NLP), referring to the analysis of the human language. Several studies have been focused on methods for semi-automating SRs via NLP. In recent years, a few researchers used various ML-aided software or platforms in real-world SRs to replace conventional counterparts, and the results verified their feasibility. In a series of reviews published in the Journal of Clinical Epidemiology in 2018, the term ‘living systematic review’ was introduced, which called for the fully-automated workflow of SR that can only be achieved with ML. In the past few years, neural network models have dominated NLP generally and text classification specifically. In particular, convolutional neural networks (CNNs), which have been originally used for image classification, have emerged as state-of-the-art models for text categorization.
Up to now, current studies on assisting citation screening of scientific articles in English are thriving with an accompanying body of work. However, none of the existing studies have included Chinese literature. The EBM principle requires SRs to be as inclusive as possible. A potential source of bias in SRs is the publication language of the included studies. It has been pointed out that trials with significant results are more likely to be published in English-language journals. Cochrane Collaboration standards recommend that “whenever possible review authors should attempt to identify and assess for eligibility all possibly relevant reports of trials irrespective of language of publication”. Including studies published in non-English languages not only prevents language bias, but also increases the power of meta-analysis estimates. It has been suggested that all SRs of epidemiology and public health should include literature published in the major languages of the world. This means that the use of regional and non-English bibliographic databases should become routine. In China, there are already 263 research institutions and medical laboratories with an estimated 926,000 researchers, making Chinese medical researchers second in number only to those of the United States in 2006. Moreover, the share of China in the published scientific papers worldwide increased from less than 1% in 1980 to about 12% in 2011, ranking second behind the US in 2013. A study has pointed out that the utilization of Chinese-language databases can significantly increase the number of potentially relevant references for each search. Nevertheless, only a few researchers in the EBM field take this issue seriously. For example, in 2015, a study pointed out that among 8,680 published reviews indexed in the Cochrane Library, only 243 (3%) had included in their searches at least one of the major Chinese databases. If there would be more convenient ways to overcome the language barrier and identify RCTs in Chinese, it would be helpful to increase the inclusion rate of Chinese studies in SRs.

This study aimed at Chinese literature of RCTs. The ambiguity of abstracts and the variety of clinical study designs were taken into consideration when the CNN model was trained in order to improve the screening performance of Chinese RCTs. Although ML has been already used in various studies to assist English literature screening, it cannot be applied directly to Chinese literature without specific training and calibration. This is due to that many Asian languages, such as Chinese and Japanese, are written without using explicit word delimiters. These characteristics of the Chinese language require complex preprocessing before the application of ML, which pose additional difficulties on the automation of Chinese literature screening. Furthermore, although there are guidelines that instruct proper reporting, abstracts can still be ambiguous or lack important methodological description. Consequently, dichotomous classification of abstracts into RCTs and non-RCTs may miss an important body of evidence, i.e., may-be-RCTs. May-be-RCTs refer to those studies that cannot be determined either as RCTs or non-RCTs. In the actual title and abstract screening process, full texts of both RCTs and may-be-RCTs need to be acquired. Dichotomous classification cannot solve this problem. Thus, the sensitivity of the screening results from dichotomous classification can be flawed. In order to further exploit the potential help from clinical epidemiology, this study aims to explore the influence of modified literature classification strategies of Chinese biomedical literature on an automated screener based on CNN. In this
In the classification criteria, the ambiguity of title and abstract expression and the variety of study types was taken into consideration. In the finest version of classification, i.e. classify the literature by its nature regardless of the screeners’ intend to include or exclude the literature in their reviews, literature was divided into 5 different categories: RCTs, randomization-unclear controlled trials (RUCTs), non-randomized clinical trials/studies (NRCTs), non-clinical literature (NC), and Unclear. The RUCTs comprised comparative studies where the group assigning method was not stated clearly enough. The NRCTs included clinical studies other than RCTs and RUCTs, e.g., case-control studies. The NC category included medical reviews, laboratory experiments, and other literature that was not related to medicine at all. In the Unclear category, the literature contained so little information in the titles and abstracts that could not be assigned into any of the aforementioned categories.

Following the finest classification, the literature was further merged into 3-category and 2-category training. The 3-category training comprised RCTs, may-be-RCTs, and non-RCTs. The may-be-RCTs category included both RUCTs and Unclear, while the non-RCTs category contained the literature classified as NRCTs and NC. Finally, traditional dichotomous classification was performed by further merging the may-be-RCTs and non-RCTs categories.

2.1.2 Literature collection and labelling

Chinese literature published between Jan 1st 2014 and Dec 31th 2018 indexed as ‘Oral Science’ in the China Network Knowledge Infrastructure (CNKI) was searched and exported. Citations were excluded in the cases where key information, such as the abstract, author, or publication information, was missing.

Literature classification was conducted by 2 experienced researchers (S.C. and Y.X.) separately using an online platform that our group developed especially for this study. The platform aims at assisting researchers with the screening and label-checking tasks for Chinese literature citation, and provides the functions of citation management, screening assistance, and automatic label checking. The citations were screened by the 2 researchers independently, according to the 5-category criteria. After the labeling results of the 2 researchers were checked, citations with the same labels were used as the gold standard. Results with different labels were resolved by consensus after peer discussion and/or turning to an experienced senior researcher (C.L.) for final decision. The 3- and 2-category labels were derived from the 5-category labels according to the aforementioned classification criteria.

2.2 AI screener training
2.2.1 CNN model description

In this paper, a customized neural network architecture is proposed that uses a CNN to process text inputs (Fig. 1). The input of the model is the citation of a Chinese study that includes the title and abstract information, while the output is the confidence coefficient of whether the study should be labeled as one of the different categories. The architecture represents words as vectors and the input text is a concatenation of word vectors. During preprocessing, JIEBA, a Chinese text segmentation tool, was employed as the tokenizer for the text segmentation and deleting stop words tasks. Then, a citation was processed into a sequence of words based on the order of words that appeared in the original text. A word vector was randomly initialized, which would be further adjusted during model training.

Next, the $i$-th word corresponded to a $d$-dimensional word vector. A convolution operation involves a filter which was applied to a window of $w$ words in order to produce a new feature. Multiple filters were applied to each possible window of words in the sequence, producing many feature maps. Subsequently, a max-pooling operation was applied over the feature map and the maximum value was set as the feature corresponding to that particular filter. These features that formed the penultimate layer were passed to a fully connected layer. The final output was the probability of the citation being from a specific category, e.g., RCT, which ranged between 0 and 1. The closer the probability to 1, the more likely the citation to be in the category, and vice versa. Cross entropy was the loss function of the model, and the Adam optimizer was used to update the network parameters by backpropagation.

2.2.2 Parameter adjustment

The model adjusted the parameters on the dev set with the learning rate being initially set as 0.0005. The batch size was set to an integer power of 2 in order to adjust the number of filters and the size of each filter convolution window. The pooled results were treated with dropout and the values in the range of parameters (0,1). In addition, in order to prevent the over-fitting of the model, L2 penalty was applied to the model, the parameters of which could be adjusted according to the accuracy rate on the dev set.

2.2.3 Model storage and usage

After each training process with the respective classification strategies, a stable model was obtained. The model parameters and the word vectors were saved. Then, when the model was used, the parameters were loaded, while the fully connected layer ‘dropout_keep_prob’ was set to 1. When these models were used, the input data were processed and the results were predictions of the possibility belonging to each literature category. The results were saved with the relevant citation data.

2.3 Prospective validation trial

2.3.1 Sample and model preparation

In this comparative study, a sample containing 1,422 citations that had been isolated from the screening researchers and CNN models were adopted to reveal the performance of the models with different
classification strategies. During model preparation, cutting thresholds of the possibility coefficient given by the CNN models were determined. In our previously published study,\textsuperscript{26} the sensitivity (SEN) and specificity (SPE) of the CNN model could be adjusted by setting different thresholds. The primary goal was to maintain the screening as sensitive as possible. Ideally, the sensitivity should be 1. However, if such SEN could not be reached, 0.95 should be the bottom line. The secondary goal was to maintain the SPE as high as possible. Therefore, an SPE higher than 0.8 was desired but not mandatory. Hereby, the High-sensitivity Threshold was determined as follows: when the SPE can be maintained higher than 0.8, we first seek for SEN = 1. If that threshold does not exist, the standard will be lowered to 1 > SEN > 0.95 when the SPE is still higher than 0.8. If the threshold still cannot meet the subordinate standard, the High-sensitivity Threshold will be set at SEN = 0.95 irrespective of how low the SPE is. Receiver-operating characteristic (ROC) curves and the area under curve (AUC), along with the SEN and SPE with the chosen thresholds, were employed to compare the models obtained by the different classification strategies.

It should be noted that, while the 2-category model had only one High-sensitivity Threshold for the RCTs category, the 3- and 5-category models had respectively 3 and 5 High-sensitivity Thresholds for each category. For example, when trained with 5-category strategy, the CNN model will provide 5 sets of probabilities for a single citation, 5 thresholds were therefore needed to determine to which category did the citation belong. This might lead to the result that a specific citation may be labelled with multiple category tags if the possibility given for those categories were all above the respective thresholds. Therefore, combination of different thresholds was needed. The workflow of the combination of different thresholds is shown in Fig. 2. Since the sensitivity of citation screening was the main aim, no citations were discarded until the positive choosing was completed. The first screening was for RCTs and the positive results were kept. Then, the RUCTs and Unclear categories were screened from the remaining results and kept aside. Finally, screening for NRCTs and NC was performed in the remainder, where the negative results (Ambiguous) were kept and the positive ones were discarded.

2.3.2 Screening performance validation

The 1,422 samples underwent both manual and CNN model screening. The same independent screening process was performed by the 2 experienced researchers. After discussion, the final classification labels were used as the gold standard. The independent results of both researchers prior to discussion were deemed as performance of manual screening. Fully-prepared CNN models performed the screening according to the aforementioned thresholds and workflow. Finally, SEN and SPE were given with 95% confidence interval (95% CI).

Results

3.1 CNN model training

In total, 12,166 citations (2,382 RCTs, 886 RUCTs, 402 Unclear, 3,839 NRCTs, and 4,657 NC) were collected for CNN model training. Among them, 10,266 citations were used as training set, while 1,900
randomly selected citations were used for testing. Screening for RCTs, may-be-RCTs, and full-text-needed citations were 3 principal scenarios. After testing 3 differently trained CNN models, the SEN, SPE, and AUCs for the above 3 scenarios with the High-sensitivity Thresholds were documented (Table 1). The 2-category model was not able to screen for may-be-RCTs and full-text-needed citations. Due to the fact that some screening tasks were performed by threshold combinations, ROC curves and AUCs of those scenarios were not applicable.

| a. Screening for RCTs | b. Screening for may-be-RCTs | c. Screening for full-text-needed |
|-----------------------|----------------------------|---------------------------------|
| **Categories**        | **Threshold**              | **SEN** | **SPE** | **AUC (95% CI)** |
| 2                     | 0.002175                   | 0.9937  | 0.8004  | 0.9924 (0.9882 ~ 0.9966) |
| 3                     | 0.011535                   | 0.9979  | 0.8004  | 0.9951 (0.9929 ~ 0.9972) |
| 5                     | 0.074541                   | 1       | 0.8921  | 0.9978 (0.9966 ~ 0.9990) |
| □                     | N/A                        | N/A     | N/A     | N/A                    |
| □                     | 0.011315                   | 0.9543  | 0.7293  | 0.9458 (0.9309 ~ 0.9607) |
| □                     | RUCTs: 0.012651            | 0.9644  | 0.6588  | N/A                    |
| □                     | Unclear: 0.002711          |         |         |                       |

When screening for RCTs, i.e., those that were certainly RCTs, the 5-category model reached a SEN of 1 and maintained a high SPE of 0.8921, outperforming the other 2 models. When screening for may-be-RCTs, the 5- and 3-category models exhibited a higher SEN and SPE, respectively. Interestingly, when the
threshold of the 3-category model was adjusted to 0.00651, the SPE and SEN were the same as the performance of the combined thresholds of the 5-category model. When it came to finding the full-text-needed citations, the performance of the 5- and 3-category models was almost similar. However, the 5-category model had a slightly higher SEN (missed 2 out of 629 citations) compared to the 3-category model (missed 7 out of 629 citations). Furthermore, the SPE of the 3-category model was higher than that of the 5-category model since it had a workload of 127 citations less.

3.2 CNN model performance validation

In the validation trial, 1,422 citations (381 RCTs, 194 RUCTs, 8 Unclear, 474 NRCTs, and 365 NC) were included. The screening was performed by both humans and CNN models. The SEN and SPE with 95% CIs are shown in Table 2. The validation results revealed that the sensitivities of the CNN models were close to or even better than those of manual screening, especially when screening for may-be-RCTs. Nevertheless, in all 3 scenarios, manual screening had much higher specificities than all CNN models. This can be partially attributed to that the high-sensitivity workflow sacrifices model specificity. The performance of the CNN models in the validation trial was close to or somewhat better than that during training.
Table 2
Performance of manual and CNN screening in validation trial.

| Method | SEN (95%CI)          | SPE (95%CI)          |
|--------|----------------------|----------------------|
| Manual | 0.9960 (0.9875 ~ 0.9989) | 0.9942 (0.9896 ~ 0.9968) |
| 2-cat  | 0.9921 (0.9751 ~ 0.9979) | 0.8117 (0.7863 ~ 0.8347) |
| 3-cat  | 0.9947 (0.9790 ~ 0.9990) | 0.7608 (0.7334 ~ 0.7861) |
| 5-cat  | 1 (0.9875 ~ 1)         | 0.8799 (0.8582 ~ 0.8987) |

| Method | SEN (95%CI)          | SPE (95%CI)          |
|--------|----------------------|----------------------|
| Manual | 0.9455 (0.9174 ~ 0.9647) | 0.9922 (0.9876 ~ 0.9951) |
| 2-cat  | N/A                  | N/A                  |
| 3-cat  | 0.9653 (0.9269 ~ 0.9847) | 0.7688 (0.7439 ~ 0.792) |
| 5-cat  | 0.9900 (0.9608 ~ 0.9982) | 0.7475 (0.7219 ~ 0.7715) |

| Method | SEN (95%CI)          | SPE (95%CI)          |
|--------|----------------------|----------------------|
| Manual | 0.9854 (0.9762 ~ 0.9912) | 0.9862 (0.9791 ~ 0.991) |
| 2-cat  | N/A                  | N/A                  |
| 3-cat  | 0.9862 (0.9720 ~ 0.9936) | 0.6972 (0.6647 ~ 0.7279) |
| 5-cat  | 1 (0.9918 ~ 1)       | 0.6591 (0.6257 ~ 0.6909) |

By comparing the CNN models, it was found that, in RCTs screening, the 5-category model had an astonishing SEN of 1 with the highest SPE. In the other 2 scenarios, although the 95% CIs overlapped, the 5-category model tended to be more sensitive, while the 3-category model tended to be more specific. The actual benefits and costs were that the 5-category model rescued 5/202 may-be-RCTs at the cost of 26 more false positive citations than the 3-category model. As for the full-text-needed citations, 8/583 citations were rescued at the cost of 32 false positive citations. Subsequently, the choice between these 2 models depends on the balance between the cost and benefit of SEN and SPE required by the researchers according to their screening requirement.

Discussion

In general, SRs are time- and resource-intensive, requiring an average of 5 researchers and approximately 41 weeks to be submitted to a journal. An average-sized SR search cites 1,781 references (range 27 ~
92,020), requires an abstract-level screening of 1,286 references (range 14~77,910), full-text screening of 63 studies (range 0~4,385), and final inclusion of 15 studies (range 0~291). Proper methods are required to solve this long-existed problem. This study focused on the automation of Chinese medical literature screening and employed different training strategies to improve the screening results.

This study was designed to face the problem concerning the absence of a Chinese medical literature screener based on ML able to classify abstracts according to study type. Additional efforts were made to improve the screening performance by adopting different training strategies. In the scenario of screening for citations that certainly concern RCTs, the 5-category strategy yielded the best performance. The SEN and SPE of the 5-category model were higher than those of the other 2 models with the SEN reaching the perfect standard of 1. Although the results of perfect SEN might be due to sampling error, the excellent performance and the superiority over the 3- and 2-category models could be guaranteed. As for the scenarios of screening for may-be-RCTs, the 2-category model was inherently not applicable. The remaining 2 models exhibited relative advantages over each other; i.e., the 3-category model had higher SPE, while the 5-category model had higher SEN, while no statistical significance was found. The screening performance for full-text-needed studies (i.e., certain RCTs and may-be-RCTs) was similar. Despite that no statistical significance was found, again the 5-category model achieved a perfect SEN of 1, while the 3-category model had a higher SPE than its counterpart (0.6972 vs. 0.6591). The reason why the 5-category model had a perfect SEN in screening full-text-needed studies while it missed some may-be-RCTs was that all the missed ones were labeled as RCTs. In real-world practice, the choice between 5- and 3-category models depends on whether the user demands a higher SEN or SPE.

The earliest study that used ML to emulate the inclusion decisions for SRs was that of Cohen et al. In 2015, Cohen et al. used a linear-kernel support vector machine (LKSVM) to assist citation screening, and the results demonstrated that the area under the ROC curve was 0.965–0.969. Most of the current researches on ML-assisted screening have reached sensitivities over 90% and specificities of around 30–70%; however, their use on Chinese literature remains rare. In addition, by comparing the performance of the models proposed in this study with other models reported to classify literature according to study type, a relatively high level was reached. Marshall et al. trained several ML models with a 2-category strategy to screen for RCTs, and different models with different combinations were tested. The best performance achieved by their model was an AUC of 0.987 (95% CI 0.984~0.989), which was lower than the 3 AUCs obtained in this study. Moreover, the corresponding SEN was 0.985 (95% CI 0.978~0.990) with a SPE of 0.840 (95% CI 0.836~0.843). In an earlier study, Wallace et al. used LKSVM and trained it with a 2-category strategy, obtaining a SEN of 0.992 with a SPE of 0.774. The superiority of the SEN, SPE, and AUC achieved by the 5-category model in this study suggests that apart from the combination of different ML model types, a training strategy with multiple study type categories might be another means of improving screening performance. It should be noted that since multiple-category labels are usually not available or inaccurate, extra efforts of manual labelling are needed for this way of model training. In most studies, readily classified data sets are adopted, such as the MEDLINE publication type label or the Cochrane Crowd RCT screening label, which are mostly dichotomous. A recent study used
citations labeled simply with the results from a search filter as the training set for CNN. Due to the significant noise of the labels, the screening results exhibited a sensitivity of 96.9% with a low specificity of 34.6%. In the present study, a rigorous screening and labeling process was followed to generate the database of Chinese literature citations, ensuring the quality and sufficient labeling of the training set at the cost of extra manual labor. However, it should be put out clear that, since the corpus of different studies varied, the sensitivities and specificities can’t be directly compared cross studies. Thus, the results of this study only implied a potential tendency of influence of different literature classification strategies on automatic screening results. Further study is needed in order to expand the conclusion of universality.

As mentioned previously, there are 2 other subfields that can be used to improve literature classification: modified search filters and crowdsourcing. Modification of search filters has never stopped. Nowadays, search filters contain dozens of text strings and database tags, whose sensitivity for RCTs can reach a level as high as 99.5%, but at the cost of a specificity as low as 30 ~ 50%. The search filter recommended by Cochrane Collaboration was developed by Glanville et al. It provides a better specificity of 72.7%, but still leaves a large amount of studies for citation screening. Although manual work is still required, search filters can still save researchers’ time when identifying RCTs. Moreover, this can help in the data preparation process for ML. As for the Chinese database, currently there are no developed search filters. On the other hand, crowdsourcing can be defined as the act of taking a job usually performed by a designated agent and outsourcing it to an undefined, generally large group of people in the form of an open call. In this field, Cochrane Crowd is one of the most well-known projects. This project has been remarkable in its success, since over 1,600,000 articles have already been labeled as RCTs or clinical controlled trials. An evaluation of its output against double assessment by experienced researchers revealed that the sensitivity and specificity exceeded 99%. Crowdsourcing may facilitate the update of previously published reviews, or contribute to real-time up-to-date online SRs, i.e., “living SRs”. However, the quality of the crowd’s work and the large effort required to begin a reliable crowdsourcing task remain the top 2 barriers hindering its wide application.

In recent years, it has been inspiring to see studies using ML-aided methods, improved search filters, as well as crowdsourcing projects to replace conventional manual screening during SR researches. However, the ultimate goal should be a comprehensive database containing well-structured and indexed information concerning the study type, research details, outcome data, etc. Various methods have been employed to identify such sets of information from published studies. If the authors, publishers, and databases could scientifically tag and index the studies prior to publication, it would constitute a further step closer to this ultimate goal.

This study has several limitations that should be mentioned. First, the training set included 12,166 citations. However, ML relies heavily on the amount of training samples to generate better results. More efforts will have to be made to obtain training sets of larger scale, in order to further improve the performance of CNNs. Second, an imbalance of sample distribution was observed, i.e., the Unclear category comprised only a small proportion due to the nature of this category. Three sets of different
approaches could be applied to deal with this imbalance: the pre-processing data approach, the algorithmic approach with cost-sensitive classification, and the set of ensemble methods. Third, it was noticed that a research published in 2020 revealed that compared to CNNs, other novel ML techniques, such as multilayer perceptron, bidirectional long short term memory networks (biLSTM), and CNN–biLSTM can be more efficient during text classification tasks. However, this current study mainly focused on the influence of modified literature classification strategies of Chinese biomedical literature on screening performance, so only conventional CNN was involved. In future studies, carefully combined advanced models with multi-category training strategies might further enhance the screening performance.

Conclusion

Automated screening in Chinese literature was successfully achieved in this study. The proposed CNN algorithm trained with multiple-category strategies in automatic screening of Chinese medical literature outperform the model trained with dichotomous data. They are not only more sensitive and specific when screening for RCTs, but also capable of picking out literature that may also be RCTs. The choice of screening model during real-world practice should depend on the users’ specific needs for higher sensitivity or specificity. Further study is needed in order to expand the conclusion of universality.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: All authors of this manuscript approved the publication.

Availability of data and materials: Not applicable.

Competing interests: The authors declare no conflicts of interest, either directly or indirectly, in the information or products listed in the paper.

Funding: This research was funded by the National Traditional Chinese Medicine Clinical Research Program (NO. 2018-131) and National College Student Innovation and Entrepreneurship Training Program (NO. 202010611329).

Authors’ contributions: Shengkai Chen: Conceptualization, Methodology, Writing- Original draft preparation. Bochun Mao: Conceptualization, Methodology, Software. Yu Xie: Writing- Reviewing and Editing. Pan Yao: Methodology, Software. Chunjie Li: Validation, Supervision. Sijin Yang: Writing- Reviewing and Editing, Supervision. Li Dong: Writing- Reviewing and Editing, Supervision. Bo Li: Writing- Reviewing and Editing, Supervision.

Acknowledgement: We would like to thank Zhonghua Yu and Li Chen, who are both from College of Computer Science, Sichuan University, for their help to this study.
References

1. Deeks JJ, Higgins J, Altman DG, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011). The Cochrane Collaboration 2011: 2.

2. Nye B, Li JJ, Patel R, Yang YF, Marshall IJ, Nenkova A, et al: A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018.

3. Garner P, Hopewell S, Chandler J, MacLehose H, Schunemann HJ, Akl EA, et al. When and how to update systematic reviews: consensus and checklist. *Bmj-British Medical Journal* 2016; 354.

4. 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 (2).

5. Smith R, Tovey D, Bero L, Busuttil G, Farquhar C, Koehlmoos TP, et al. The Cochrane Library Oversight Committee. *Cochrane Database Syst Rev* 2011(2).

6. 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. *Contemporary clinical trials communications* 2019: 100443.

7. 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 (6): 1165-8.

8. 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 (1): 193-201.

9. Rodrigo EG, Aledo JA, Gamez JA. Machine learning from crowds: A systematic review of its applications. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery* 2019; 9 (2).

10. Scott AM, Clark J, Del Mar C, Glasziou P. Increased fluid intake to prevent urinary tract infections: systematic review and meta-analysis. *Br J Gen Pract* 2020; 70 (692): E200-E7.

11. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction-the why, what, when, and how. *J Clin Epidemiol* 2017; 91: 23-30.

12. Thomas J, Noel-Storr A, Marshall F, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol* 2017; 91: 31-7.

13. Simmonds M, Salanti G, McKenzie J, Elliott J. Living systematic reviews: 3. Statistical methods for updating meta-analyses. *J Clin Epidemiol* 2017; 91: 38-46.

14. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schunemann HJ. Living systematic reviews: 4. Living guideline recommendations. *J Clin Epidemiol* 2017; 91: 47-53.

15. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research* 2016; 57: 345-420.

16. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner’s guide. *Research Synthesis Methods*
17. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol 2002; 31 (1): 115-23.

18. Pham B, Klassen TP, Lawson ML, Moher D. Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. J Clin Epidemiol 2005; 58 (8): 769-76.

19. Egger E, ZellwegerZahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997; 350 (9074): 326-9.

20. Fung IC: Seek, and ye shall find: accessing the global epidemiological literature in different languages. Springer, 2008.

21. Dyer G. China overtakes Japan on R&D. Financial Times 2006; 3: 2006.

22. Fu J, Frietsch R, Tagscherer U: Publication activity in the Science Citation Index Expanded (SCIE) database in the context of Chinese science and technology policy from 1977 to 2012. Fraunhofer ISI Discussion Papers-Innovation Systems and Policy Analysis, 2013.

23. Adams D, Wu T, Yasui Y, Aung S, Vohra S. Systematic reviews of TCM trials: how does inclusion of Chinese trials affect outcome? J Evid Based Med 2012; 5 (2): 89-97.

24. Cohen JF, Korevaar DA, Wang JF, Spijker R, Bossuyt PM. Should we search Chinese biomedical databases when performing systematic reviews? Systematic Reviews 2015; 4.

25. Peng KH, Liou LH, Chang CS, Lee DS: Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts. 2015 24th Wireless and Optical Communication Conference. 2015.

26. Mao B, Chen S, Xie Y, Yao P, Li C. Exploration of classical deep learning algorithm in intelligent classification of Chinese randomized controlled trials. Chinese Journal of Evidence-Based Medicine 2019; 19 (11): 1262-7.

27. 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 (2): 206-19.

28. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, et al. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc 2015; 22 (3): 707-17.

29. 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. Systematic Reviews 2015; 4.

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

31. Del Fiol G,Michelson M, Iorio A, Cotoi C, Haynes RB. A Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature: Comparative Analytic Study. J Med Internet Res 2018; 20 (6).
32. McKibbon KA, Wilczynski NL, Haynes RB. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Information and Libraries Journal 2009; 26 (3): 187-202.

33. Glanville J. How to identify randomized controlled trials in MEDLINE: ten years on - Authors' response. J Med Libr Assoc 2007; 95 (2): 119-20.

34. Howe J. The rise of crowdsourcing. Wired magazine 2006; 14 (6): 1-4.

35. Collaboration C: The EMBASE screening project: six months old and going strong.

36. Collaboration C: Embase Project Update. 2014.

37. Badgett RG, Vindhyal M, Stirnman JT, Gibson CM, Halaby R. A Living Systematic Review of Nebulized Hypertonic Saline for Acute Bronchiolitis in Infants. Jama Pediatrics 2015; 169 (8): 788-9.

38. Nama N, Sampson M, Barrowman N, Sandarage R, Menon K, Macartney G, et al. Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study. J Med Internet Res 2019; 21 (4).

39. Karagiannis T, Andreadis P, Manolopoulos A, Malandris K, Avgerinos I, Karagianni A, et al. Decision aids for people with Type 2 diabetes mellitus: an effectiveness rapid review and meta-analysis. Diabet Med 2019; 36 (5): 557-68.

40. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. How to complete a full systematic review in 2 weeks: processes, facilitators and barriers. J Clin Epidemiol 2020.

41. Whittle SL, Johnston RV, McDonald S, Paterson KL, Buchbinder R. Autologous blood product injections including platelet-rich plasma for osteoarthritis of the knee. Cochrane Database Syst Rev 2019(5).

42. Longadge R, Dongre S. Class imbalance problem in data mining review. arXiv preprint arXiv:13051707 2013.

43. Varghese A, Agyeman-Badu G, Cawley M. Deep learning in automated text classification: a case study using toxicological abstracts. Environment Systems and Decisions 2020: 1-15.

**Figures**
Figure 1

The proposed CNN architecture.
Figure 2

Workflow of the combination of different thresholds for the 3- and 5-category models.