MMiDaS-AE: Multi-modal Missing Data aware Stacked Autoencoder for Biomedical Abstract Screening

Eric W. Lee[†], Byron C. Wallace[‡], Karla I. Galaviz[†], Joyce C. Ho[†]
[†]Emory University, [‡]Northeastern University
ewlee4@emory.edu,b.wallace@northeastern.edu,
{karla.galaviz,joyce.c.ho}@emory.edu

ABSTRACT

Systematic review (SR) is an essential process to identify, evaluate, and summarize the findings of all relevant individual studies concerning health-related questions. However, conducting a SR is labor-intensive, as identifying relevant studies is a daunting process that entails multiple researchers screening thousands of articles for relevance. In this paper, we propose MMiDaS-AE, a Multi-modal Missing Data aware Stacked Autoencoder, for semi-automating screening for SRs. We use a multi-modal view that exploits three representations, of: 1) documents, 2) topics, and 3) citation networks. Documents that contain similar words will be nearby in the document embedding space. Models can also exploit the relationship between documents and the associated SR MeSH terms to capture article relevancy. Finally, related works will likely share the same citations, and thus closely related articles would, intuitively, be trained to be close to each other in the embedding space. However, using all three learned representations as features directly result in an unwieldy number of parameters. Thus, motivated by recent work on multi-modal auto-encoders, we adopt a multi-modal stacked autoencoder that can learn a shared representation encoding all three representations in a compressed space. However, in practice one or more of these modalities may be missing for an article (e.g., if we cannot recover citation information). Therefore, we propose to learn to impute the shared representation even when specific inputs are missing. We find this new model significantly improves performance on a dataset consisting of 15 SRs compared to existing approaches.

CCS CONCEPTS

• **Applied computing → Health informatics**; • **Information systems →** *Clustering and classification.*

KEYWORDS

Systematic Review, Multi-modal Stacked Autoencoder, Missing Data Imputation

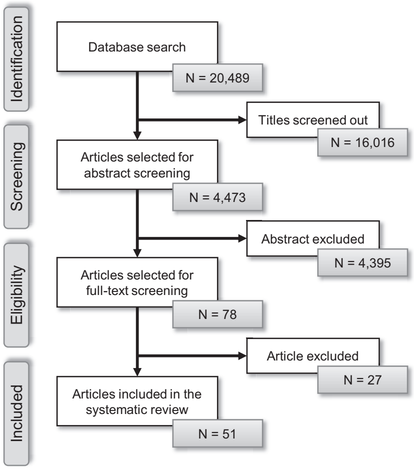


Figure 1: A simplified illustration of the SR screening process by Galaviz *et al.* [17].

ACM Reference Format:
Eric W. Lee[†], Byron C. Wallace[‡], Karla I. Galaviz[†], Joyce C. Ho[†]. 2020. MMiDaS-AE: Multi-modal Missing Data aware Stacked Autoencoder for Biomedical Abstract Screening. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3368555.3384463

1 INTRODUCTION

Systematic reviews (SRs) are essential knowledge translation tools focused on bridging the research-to-practice gap across a wide range of domains. In health research, SRs aim to identify, evaluate, and summarize the findings of all individual studies (which typically describe clinical trial results) relevant to a clinical question, thereby making the available evidence more accessible. SRs (and meta-analyses) in this way provide high-quality evidence that can inform healthcare decision making, support clinical guidelines, and guide health policies [10, 19, 20]. For instance, a SR can be used to synthesize findings from randomized intervention studies to robustly determine which interventions are best supported for a particular condition.

Conducting SRs is a time-consuming and complex task [21]. Established methodologies for performing a SR [11, 35, 37] require a comprehensive search to identify all the relevant studies for inclusion. Indeed, comprehensiveness (so as to avoid bias via 'cherry-picking' of evidence) is a key property of rigorous evidence syntheses. Yet the broad searches necessary to achieve this yield imprecise

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384463

search results including searches that often yield only ∼1% relevant results. Domain experts must wade through these mostly irrelevant articles to identify those that meet the *inclusion criteria*; thus, producing a single review can require thousands of person-hours [2]. Figure 1 provides an example of the laborious citation screening process for a SR based on diabetes prevention interventions. Only 0.38% of the articles were selected for full-text review based on the title and abstract and 0.24% were included (i.e., analyzed and evaluated) in the actual review itself [17]. The exponential growth of biomedical literature has further exacerbated this problem [5].

Given the importance of SRs for realizing evidence-based practice and the labor that conducting these entails, there is a clear need to expedite tasks necessary for evidence synthesis while maintaining rigor and comprehensiveness. In particular, semi-automation can help speed up the screening process, an extremely time-consuming endeavor due to a large number of citations [36]. The standard methodology for semi-automating the citation screening step of SRs entails training a custom classification model for each new review. Unfortunately, many of the previous approaches assume that they have small labeled batches from the reviewers, and train their model on those batches to predict the rest [15, 27]. Moreover, the existing methods focus primarily only on the text itself including using representations like bag-of-words, or word embeddings [6, 13, 15, 27, 28, 34, 52]. However, there is rich information (i.e., citation relationships between the articles) that can be used to learn more accurate models.

Our goal in this work is to minimize the number of relevant articles (articles included after the full-text screening) excluded by the classifier while reducing the reviewers' workload by excluding the maximum number of irrelevant documents. Our general strategy builds upon a body of work on semi-automating citation screening for evidence syntheses via machine learning [15, 41, 49, 52]. To address the limitations of existing semi-automation SR models, we introduce MMiDaS-AE, a Multi-modal Missing Data aware Stacked AutoEncoder. We adopt the multi-modal stacked autoencoder [9] to encode a variety of information that includes 1) text from the document, 2) Medical Subject Headings (MeSH) terms, and 3) citation networks. In addition to the textual data in the documents, each article in PubMed (a repository of biomedical articles) is associated with MeSH terms, which codify abstract concepts and can be used to learn topic representations. MMiDaS-AE also uses co-citation relations between articles. The intuition is that an unknown article with co-citation relations to an article that passes the SR screening is more likely to be relevant.

However, it is crucial for the model to be robust to missing data representations, especially when learning a shared representation using three different sources of information. Thus, to mitigate the effects of missing data, we extend work for bimodal speech classification [38] to design an imputation technique for multi-modal data in which we intentionally leave one or more representations out while learning to induce a shared representation in a latent space from which we can reconstruct all input modalities. Consequently, this multi-modal stacked autoencoder is robust to missing data. We also introduce a multi-label classification task to improve the prediction result by utilizing whether the article passed the abstract screening and whether it passed the full-text screening. Finally, we utilize a cross-topic learning strategy to utilize existing SRs to pre-train MMiDaS-AE, and then fine-tune the weights of the model to a specific SR topic.

We perform extensive studies on 15 SRs (or topics) related to drug efficacies provided by Cohen *et al.* [15]. Our pre-trained model achieves the best predictive performance (measured using the area under the receiver operating curve) on 11 out of 15 of the topics compared to existing approaches. Moreover, MMiDaS-AE achieves the best performance on 13 of the 15 topics with a small amount of labeled data and can reduce the workload by 13.2% to 69.4% where one of the existing approaches reduces workload by 11.1% to 62.9%. In addition, our ablation study demonstrates the importance of imputation and multi-label classification, as it can reduce the reviewer workload by 11.5% to 67.8%, compared to the previously proposed multi-modal stacked autoencoder [9]. As a result, MMiDaS-AE reduces the reviewers' workload by excluding the maximum number of irrelevant documents.

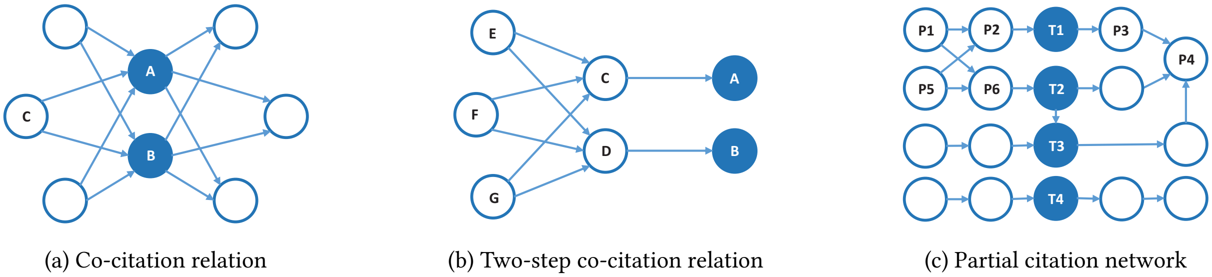
2 RELATED WORK

Methods for semi-automating the citation screening step of SRs have been widely studied; see [41] for a survey of this work. The typical approach is to adopt a supervised learning model – equivalent to training a custom classification model for each new review. The classification models used to discriminate between relevant and irrelevant articles for a given topic include support vector machines (SVMs) [22, 42, 49, 52], generalized linear models [23], Voting Perceptron [15], Random Forest [27], Complement Naive Bayes [33], Decision Tree [6], and k-NN [1]. Note that models can be used either to make 'hard' include/exclude decisions, or can be used to rank citations in order of likely relevance.

Because supervision is expensive for this task, and a new model must be trained for each new review, a common strategy explored is active learning [14, 28, 34, 49, 50] in which the learner starts with a small subset of manually labeled records, which are used to train the initial classifier. After each learning (or annotation) cycle, the newly trained model classifies the remaining unlabelled citations and presents a sample of these records to the reviewer for annotation. This iterative approach may be used to train a model that is used to classify all remaining (unscreened) citations, or can simply be used to prioritize identification of relevant abstracts so that the review team can begin data extraction from these [41].

Several software tools for reducing abstract screening time. Abstrackr [49] is a semi-automated screening tool that uses active learning approaches to reduce the number of relevant and irrelevant labels necessary to learn a robust predictive model. Rayyan [42] is also a semi-automated web and mobile screening application that builds an inclusion/exclusion model based on individual words. EPPI-Reviewer [45] is an online tool for research synthesis that clusters documents to describe the range of studies that have been identified which has been used by hundreds of users for over 200 SR. Finally, SWIFT-Review [23] is an interactive workbench that uses text-mining tools to prioritize the relevant documents.

Most of the semi-automation SR approaches use bag-of-words and their combinations [6, 13, 15, 27, 28, 34, 52]. For example, Cohen *et al.* [13] proposed to use uni-grams and bi-grams to treat each of them as a single word, Bannach-Brown *et al.* [4] used tri-gram and GENIA tagger [48] prior to extracting uni-grams, and Khabsa *et*



(a) Co-citation relation (b) Two-step co-citation relation (c) Partial citation network

Figure 2: A simplified example of the co-citation relations and the partial citation network used to learn representations. Solid nodes denote target articles that we need to classify; empty nodes are the articles used to learn the representation of target articles.

al. [27] used brown clustering on the bi-grams to tackle the data sparsity problem. Some more recent efforts have proposed to learn a paragraph vector using neural models [22, 31].

Some prior work has considered an intra-topic setting [15, 27, 33, 40] which assumes the models have access to small labeled batches from the reviewers and train their model on those batches to predict the rest. On the other hand, Cohen *et al.* [14] proposed a cross-topic (inter-topic) learning strategy by observing that classifiers can learn to exclude a large proportion of articles using any kind of sampling strategy. They used multiple SVM models by subsampling non-topic specific data to prioritize relevant articles. While semi-automation models are evaluated predominantly on a private dataset, many of them also provide direct comparisons to the Cohen dataset [15].

3 FEATURE REPRESENTATIONS

Feature extraction is a crucial component to the success of the classification process. Previous approaches use bag-of-words of titles, abstracts, and MeSH terms [6, 13, 15, 27, 28, 34, 52]. Khabsa *et al.* [27] used co-citation data as a feature to semi-automate the SR. However, unlike previous approaches that deal with each representation separately, we propose to learn a shared representation that encodes different article information. As a result, the model can be robust to missing data and a limited number of samples.

3.1 Document Representation

Natural language processing (NLP) systems typically transform input documents into fixed-dimensional vector representations that can subsequently be used as feature vectors by 'downstream' modules (e.g., logistic regression or a feed-forward neural network). Previous work for semi-automating screening for SRs predominantly represented documents via sparse bag-of-words (BOW) representations [4, 15, 27, 33, 40, 52]. More recent work in NLP has moved towards learning better representations of texts, in particular by mapping high-dimensional and sparse BOW representations into dense, low-dimensional vectors. For example, doc2vec extends word2vec to learn distributed representations of documents (rather than words) [26, 30]. ELMO [44] and BERT [16] were proposed to learn the contextual representations.

For our task, we restrict our document to titles and abstracts due to potential copyright issues inherent to full-text articles. As a result, each article's input is relatively short (an average of 118 words after simple preprocessing). For short texts, averaging the embeddings of all words in the text can serve as the document representation [26]. Therefore, we adopt PMCVec [18], a pre-trained word2vec embedding, and learn the document representations of all articles by averaging embeddings in the title and the abstract of each document. PMCVec was trained on titles and abstracts from ∼27 million documents indexed in the PubMed database. We explored SciBert [7], a deeper representation, but this did not yield better predictive power as demonstrated in our empirical results.

3.2 Topic Representation

In the "Identification" step of SRs depicted in Figure 1, a combination of MeSH terms that represent the SR topic is used for database search to retrieve the initial articles list. Using these MeSH terms, we can compute the distance between the article and the MeSH terms. Thus, we learn a topic representation of an article by using the relationship between MeSH terms and the article. This can be done in two steps. First, we learn the representation of all MeSH terms of the topic. Second, we subtract the document representation we learned from the previous section from the MeSH term representation. Thus, this topic representation captures the relationship between the article and the MeSH terms used in the SR search. This has the added benefit of distinguishing articles that are in multiple SRs.

Because we are learning the relationships between documents and associated MeSH terms of the topic by subtracting their representations, both representations should be learned from the same embedding space. One benefit of PMCVec [18] is that it learns representations of both single words and multi-words from PubMed abstracts as technical phrases in biomedical texts such as diseases or symptoms are multi-words phrases. Thus for MeSH terms, instead of using the composition of single words, multi-words MeSH terms also appear in the embedding space, and we can directly use them to compute the MeSH terms embedding.

3.3 Citation Network Representation

Most existing SR screening methods primarily rely on text features derived from titles and abstracts. This ignores the rich citation structure (e.g., the study is cited by other studies) available for each article. Figure 2(a) depicts a simple citation network (a network that in which articles are nodes and citations are edges), and *C*

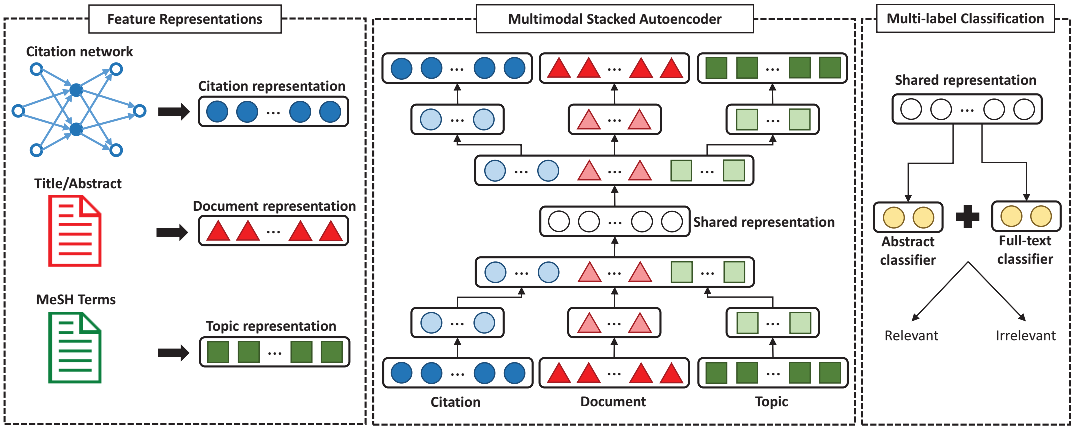


Figure 3: Framework overview of the MMiDaS-AE.

and A implies A is cited by C (or C is a citation of A). From the citation network, co-citations (two articles cited together by the other articles) might be used to find related studies. For example, in 2(a), A and B are co-citations as there exists an article C that cites both A and B. This is motivated by the intuition that if one article is included, then co-cited articles are more likely to be included as well. Using features consisting of just the bag-of-words of uni-grams and co-citations, Khabsa *et al.* [27] showed that their machine learning model could achieve good recall. Yet, co-citation only captures one perspective of the article. There exist cases in which two articles do not have a co-cited article, but their citations having co-cited articles. For example, in Figure 2(b), A and B do not share any citations directly, however, their citations, C and D are co-citations as they share E, F, and G as citations. In such examples, directly looking at the co-citations of the first two articles, A and B, do not help to find this relationship. Therefore, we propose to construct a citation network and learn a representation (low-dimensional projection) of each article.

However, constructing a complete citation network is infeasible. Instead, we use a partial citation network that contains co-citation information by limiting the network to contain only articles at most two citations away. Figure 2(c) provides an example of a partial citation network that is used. The partial citation network contains citation information of 5 articles which is shown as $P1 \longrightarrow P2 \longrightarrow T1 \longrightarrow P3 \longrightarrow P4$ in Figure 2(c) where *T1* is the article that we are considering about. To account for co-citation information, we adopt LINE [46], a network embedding method that takes into account both first- and second-order proximity. The first-order proximity permits the model to learn the direct link between nodes such as *T2* and *T3* in Figure 2(c), while second-order proximity is determined by the similarity of the "neighbors" (co-citation) of two nodes such as *P2* and *P6*. Overall, from Figure 2(c), after the training, *T2* and *T3* are close to each other in the embedding space because of the citation relation (first-order proximity), and *P2* and *P6* are close to each other by sharing neighbor (second-order proximity). Also, *T1* and *T2* should be close to each other because *P2* and *P6* are already close to each other, but also there is a link between *P2* and *T1*, and *P6* and *T2*. However, *T4* would not be considered while training *T1*, *T2*, and *T3* because there is no citation relationship. This is an important factor because the main goal of our model is to correctly classify articles *T1*, *T2*, *T3*, and *T4* using the co-citation information ultimately encoded into the learned representation.

4 MMIDAS-AE DESIGN

MMiDaS-AE adopts a multi-modal stacked autoencoder [9] which takes multiple input representations and learns a shared representation that encodes all of these modalities. This avoids the unwieldy number of parameters that are introduced with a simple concatenation of each input representation. Also, compressing the feature representations into a shared representation makes it easier to apply any matrix manipulation technique that can not be done in the input space because of the difference in dimensions. However, the existing work was insufficient to deal with missing data representations. Thus we introduce a new learning strategy by using an augmented dataset. Finally, we propose a multi-label classification task to improve the prediction results. An illustration of our framework is shown in Figure 3.

4.1 Multi-modal Stacked Autoencoder

Autoencoders are unsupervised models that learn compressed representations of inputs. The objective for the autoencoder is to reconstruct inputs faithfully from this learned representation with minimal error [43]. Cadena *et al.* [9] proposed multi-modal stacked autoencoders for the task of robotics scene understanding to support different input modalities simultaneously (i.e., RGB image, scene depth, and semantic information). Each input representation was passed through an autoencoder. The three independent autoencoders were concatenated together using their respective hidden layers and then passed to another autoencoder, thus inducing a shared representation from which to reconstruct the original (concatenated) inputs. One may view this approach as a means of learning *disentangled* representations [24, 32] in which we have explicit low-dimensional encodings of the respective input modalities. We found empirically that the best performance was obtained when we unified the length of the independent hidden layer prior to concatenation. For example, if we have 256, 200 and 200 dimensions

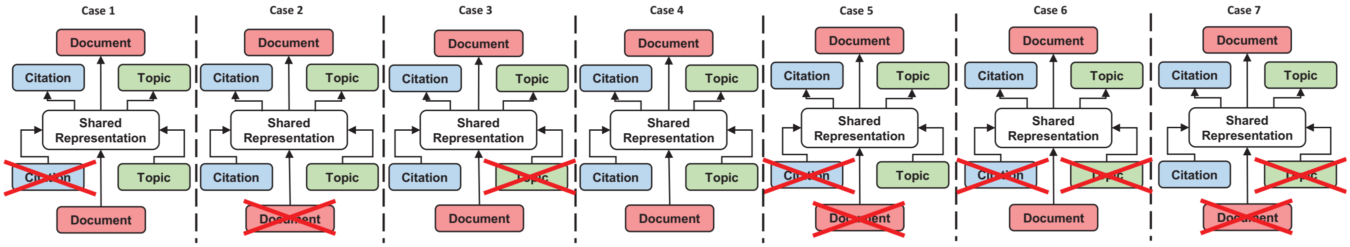


Figure 4: The process of imputation on multi-modal stacked autoencoder to deal with missing data.

as an input for each representation, the best performance we get was when we use unified (e.g. 100) dimensions for the independent hidden layer.

4.2 Missing Data Imputation in Autoencoder

One advantage of multi-modal auto-encoders is their potential to combine the available modalities to impute representations in the case that one of these is missing [9, 38]. However, robustness to missing data is crucial when learning the shared representation of multi-modal inputs; otherwise, missing inputs may yield poor shared representations. There are three possible cases of missingness (for different modalities):

1. Citation network representation: lacking citation information.
2. Document representation: missing abstract.
3. Topic representation: missing document representation as topic representation is computed by using document representation.

Ngiam *et al.* [38] proposed the use of an augmented dataset in the bimodal autoencoder that has a single modality as an input (the other input is set to zero values) that reconstructs two modalities as an output. However, naively extending this to the multi-modal scenario does not yield desirable results. We introduce a strategy that generalizes this work [38] for multi-modal in which we intentionally leave *one or more* representations out (or 'empty') while learning to induce a shared representation in a latent space from which we reconstruct all input representations. Figure 4 illustrates our proposed imputation process to construct the augmented dataset. In particular, for the inputs with no missing values, we purposely use an empty representation for each input and try to reconstruct the output with the clean representations.

For illustration purpose, we demonstrate our process on a simple 2-dimensional example. Suppose we have 3 representations, $c = [1, 2]$, $d = [3, 4]$, and $t = [5, 6]$. Then we train our encoder with all cases shown in Figure 4. For each case, the inputs are

- Case 1: $c = [0, 0]$, $d = [3, 4]$, and $t = [5, 6]$
- Case 2: $c = [1, 2]$, $d = [0, 0]$, and $t = [5, 6]$
- Case 3: $c = [1, 2]$, $d = [3, 4]$, and $t = [0, 0]$
- Case 4: $c = [1, 2]$, $d = [3, 4]$, and $t = [5, 6]$
- Case 5: $c = [0, 0]$, $d = [0, 0]$, and $t = [5, 6]$
- Case 6: $c = [0, 0]$, $d = [3, 4]$, and $t = [0, 0]$
- Case 7: $c = [1, 2]$, $d = [0, 0]$, and $t = [0, 0]$

and the reconstructed output is $c = [1, 2]$, $d = [3, 4]$, and $t = [5, 6]$ for all cases. Therefore, we intentionally leave one or two representations out using an *empty* representation (vector of zeroes) but still require the multi-modal autoencoder to reconstruct all representations. Using this process we can handle missing input representations because the model is forced to learn a robust shared representation from all possible combinations of the inputs.

4.3 Multi-label Classification Task

The objective of the MMiDaS-AE is to minimize the number of relevant articles (articles after the full-text screening) that are excluded while minimizing the number of irrelevant citations that need to be screened by domain experts. Thus, the model must make a binary prediction for each instance which indicates whether or not it should be screened by a human reviewer. Since SRs are intended to be *comprehensive* assessments of the relevant evidence, achieving high recall (i.e., sensitivity to the relevant citations) is imperative. This is challenging in practice because there is severe *class imbalance* [25, 51], that is, there are far fewer relevant than irrelevant citations. Consider Figure 1: Here we have 20,489 articles in total, but only 51 (0.24%) of these pass full-text screening.

To ensure the identification of relevant articles, we propose a multi-label classification task to use the results of abstract screening as the labels are less imbalanced than the full-text. In the literature identification phase of SRs, there are two steps that are typically performed: title/abstract screening, which is followed by full-text screening. We posit that documents that pass the title/abstract screening are more likely to be "relevant" than those that are discarded. In other words, we were interested in ordering each document into three ordered categories: completely irrelevant, inclusion in the full-text screening, and inclusion in the SR. By including an abstract classifier, we can encode additional information that may help our model distinguish completely irrelevant articles. Thus, MMiDaS-AE uses two classifiers, an abstract classifier, and a full-text classifier. Then, as proposed by Niu *et al.* [39], we sum the prediction probability of the true (relevant) class for each classifier and use this to evaluate the performance of MMiDaS-AE. For example, if the article is predicted as irrelevant by the abstract classifier, it will have a low probability (and be unlikely to meet the final threshold). Thus, MMiDaS-AE will only detect articles that have high probabilities for both the abstract and full-text classifier.

Therefore, MMiDaS-AE consists of the following steps (as illustrated in Figure 3). We first train each feature representation, citation network, title/abstract, and MeSH terms into the citation, document, and topic representations. Then, we train a multi-modal

Table 1: Statistics of all datasets used. Abs and Full refer to the number of abstract triage and article triage statuses, respectively. % shows the percentage of the articles that are included after the full-text screening (abstract screening for the last 4 SRs). First 15 SRs are Cohen [15] dataset.

SR	Abs	Full	Total	%
ACEInhibitors	183	41	2544	1.61
ADHD	84	20	851	2.35
Antihistamines	92	16	310	5.16
AtypicalAntipsychotics	363	146	1120	13.04
BetaBlockers	302	42	2072	2.03
CalciumChannelBlocker	279	100	1218	8.21
Estrogens	80	80	368	21.74
NSAIDs	88	41	393	10.43
Opioids	48	15	1915	0.78
OralHypoglycemics	139	136	503	27.04
ProtonPumpInhibitors	238	51	1333	3.83
SkeletalMuscleRelaxants	34	9	1643	0.55
Statins	173	85	3465	2.45
Triptans	218	24	671	3.58
UrinaryIncontinence	78	40	327	12.23
Anemia	653	-	5653	11.55
COPD	196	-	1606	12.20
Clopidogrel	771	-	8291	9.30
Proton Beam	243	-	4751	5.11

stacked autoencoder to learn the shared representations that encode all three representations. While training the multi-modal stack autoencoder, we apply our proposed missing data imputation technique discussed in Section 4.2. Once the shared representation is learned, we use two softmax classifiers, an abstract classifier and a full-text classifier which is trained separately. The prediction probability of true (relevant) classes is then the sum of these two classifiers.

5 EXPERIMENT SETUP

5.1 Dataset

For ease of comparison with previous works, we evaluate our model on the publicly available dataset provided by Cohen *et al.* [15]. The dataset includes 15 SRs (or topics) concerning different drug efficacies.[1] The 15 systematic reviews were performed by members of evidence-based practice centers (EPCs). Each systematic review contains a PubMed identifier (PMID), abstract triage status, and article triage status. The PMID allows us to identify which article was included in the systematic review process. Abstract and article triage status indicates whether the article passed the title/abstract screening and full-text screening stages, respectively.

While the Cohen dataset is used for training and testing (14 SRs used as training and 1 SR as testing), we used 4 additional datasets to use as a validation set for hyperparameter tuning. The other 4 datasets are: COPD [12], proton-beam [47], anemia [29], and clopidogrel [3]. These 4 datasets also contain PMID but only the status for the abstract screening is included.

Table 1 reports the distribution of articles in each topic. The first 15 SRs are Cohen dataset and the last 4 SRs are the datasets used as a validation set. '-' in the Full column denotes that the dataset lacks the full-text screening result. As shown in the table, the number of articles included after the full-text screening varies from 0.55% to 27.04%, demonstrating a relatively large degree of imbalance.

5.2 Data Preprocessing

5.2.1 *Titles and abstracts extraction.*

The Entrez API[2] was used to retrieve the title and abstract of each article using the PMID. In total, 37,149 unique articles were extracted using the API. There are 1,885 duplicate articles between the datasets, and there were 4,548 articles with a missing abstract. The title and abstract of each article are concatenated together and pre-processed using the `nltk` library [8] in Python to remove stopwords, punctuations, and numbers. Each remaining word is then converted to a 200-dimensional vector representation using PMCVec[3] [18]. The individual word representations are then averaged to obtain the final document representation.

Note that we also evaluated the results using a larger pre-trained language model, SciBert [7] and compared the results with PMCVec. The results will be discussed in Section 6.3.

5.2.2 *MeSH terms extraction.*

In the normal SR process, the initial list of articles is retrieved by the combination of MeSH terms. However, all the datasets do not contain the MeSH terms of each SR. Thus, we manually selected the MesH terms that describe each SR the best using the following process. For each SR, we obtained all the MeSH terms (information available from PubMed) that appear in the articles with their associated frequency. Then using the top 50 most frequent terms from this list, we manually searched and selected the MeSH terms that exist on the Wikipedia page associated with the topic (i.e., ACE Inhibitors). We also accounted for the number of times the term appears in the overall corpus to avoid "uninformative" terms such as "Humans", "Male", "Female", and "adult'. After excluding terms that exist in the top 50 for all SRs, each SR contains unique terms.

Once the MeSH terms of each SR are selected, we compute the MeSH terms representations using PMCVec. As discussed in Section 3.2, it is necessary to use the same pre-trained word embedding because we are learning the relationships between documents and associated MeSH terms of the topic. MeSH terms representations are computed by taking the average of the representations of the MeSH term itself or each word. Then MeSH terms representations are subtracted from the document representation of each article to compute the topic representation. This will distinguish the same article in multiple SRs to have different topic representations. Same as document representation, a 200-dimensional vector representation is used for topic representation.

5.2.3 *Constructing the citation network.*

Using the Entrez API, we can also extract citation data of the article, however, many of the articles contain an insufficient number of citations or none. Thus, we use Semantic Scholar database[4] as our additional resource to

[1] This dataset was later extended to include 24 systematic reviews [14], however, only 15 systematic reviews have been made publicly available.

[2] https://www.ncbi.nlm.nih.gov/books/NBK25501/

[3] Since PMCVec is pre-trained on PubMed abstracts, there was no case where a word did not have a vector representation.

[4] https://api.semanticscholar.org/

extract citation data of the article using PMID. As introduced in Section 3.3, we could not use the entire citation network because of the limitation of computational and memory footprint, thus we used a partial citation network[5]. Starting from the articles that are in the dataset, we looked backward and forward from the citation links by following the semantic scholar identifier (SSID) to construct the citation network. Not all of the articles in the semantic scholar database have a PMID, instead, they have SSID. Therefore, the citation network is constructed by SSID, but preserve the mapping with the PMID.

For learning the citation network representations using LINE, we used 3,158,195 vertices (articles) and 139,270,829 edges (citation links) which is extracted from all 19 SRs and their citations. For both first- and second-order proximity, we use 128 as the dimension of each representation, and as LINE [46] proposed, concatenated the first- and second-order proximity, resulting in 256 dimensions for the citation network representation.

5.3 Evaluation Metrics

Cohen *et al.* [15] introduced a new measure *work saved over sampling* (WSS). WSS measures the work saved over random sampling for a given level of recall. WSS is defined as

$$WSS = (TN + FN)/N - (1.0 - R) \quad (1)$$

where TN denotes true negatives, FN false negatives, N the total number of articles, and R the recall. Cohen *et al.* used the special modification of the WSS called WSS@95% which means WSS for recall at 95%. Note that in some cases, the models may not achieve exactly 95% recall. Thus, to calculate WSS@95%, we compute WSS with the highest recall no less than 95%. In addition to WSS@95%, some works reported area under the receiver operating curve (AUC); we use this as an additional evaluation metric.

5.4 Experimental Design

5.4.1 Inter-topic setting. As the Cohen dataset has 15 SR topics, we evaluate MMiDaS-AE with non-topic specific settings. Specifically, for model training, 14 SR topics are used to classify the one leftover SR topic to evaluate the workload saved. We compare the result with two existing works that used the same inter-topic settings.

- **Norman**: Norman *et al.* [40] constructs a ranker by extracting bag-of-n-grams in titles, abstracts using TF-IDF and binary features. Also, article metadata such as keywords, journal name, and publication types are used as features.
- **Cohen (2008)**: Cohen *et al.* [13] studies the performance of Support Vector Machine (SVM) classifier using both textual (unigram and bigram terms of titles and abstract) and conceptual (MeSH terms) features.

5.4.2 Intra-topic setting. Intra-topic is a topic-specific setting that only uses training data within the same topic. Intra-topic assumes that reviewers labeled small batches of articles. Previous works used 5 × 2 cross-validation within each SR topics to evaluate intra-topic. Under 5 × 2 cross-validation, each SR topic is divided into two parts – one split is used for training and the other as testing. Then the roles of each half are switched. This entire process is then is repeated 5 times. 5 × 2 cross-validation results in 10 experiments and the final score is the average of the 10 experiment scores. We compare the result with four existing works that use the intra-topic setting.

- **Cohen (2006)**: Cohen *et al.* [15] uses a voting perceptron algorithm with varying learning weights using bag-of-words, MeSH terms, and publication type as their features.
- **Khabsa**: Khabsa *et al.* [27] uses textual features, co-citations, and brown clustering as features to train a random forest model.
- **Norman**: Norman *et al.* [40] uses the same method as explained in Section 5.4.1 but using the intra-topic setting with 5 × 2 cross-validation.
- **Matwin**: Matwin *et al.* [33] uses similar features to Cohen *et al.* [15] but trained Complement Naive Bayes instead.

5.4.3 Fine-tuning setting. As we target the inter-topic setting that learns a model to classify articles as a function of article-article relations and article-topic relations, we propose *fine-tuning* our pre-trained model to evaluate our model in the intra-topic setting. Under the *fine-tuning* setting, we follow the inter-topic setting to pre-train our model, then use an intra-topic setting (5 × 2 cross-validation) to fine-tune the weights of the pre-trained model. For example, if we want to predict which articles are relevant for the "ACEInhibitors" SR, then we use the other 14 SR topics to pre-train MMiDaS-AE first, and then use one-half of articles in "ACEInhibitors" to fine-tune the weights of the pre-trained MMiDaS-AE, reserving the other half for testing. Then the roles of each half are switched. In other words for each of the 10 intra-topic experiments, we use the same pre-trained MMiDaS-AE that was trained with 14 SR topics but is then fine-tuned on 50% of the topic-specific data. We repeat this procedure 5 times, as same as 5 × 2 cross-validation. We report the average estimated score across the 10 experiments.

5.4.4 Hyperparameter tuning. We found empirically that using a unified length of the independent hidden layer performs better than other settings. The unified length of the independent hidden layer means learning all three representations into the same length. For example, we use a 256-dimensional vector representation for network representation and a 200-dimensional vector representation for document and topic representation. We add an independent hidden layer connected with the input representation with a 100-dimensional vector representation, thus after these layers, all three inputs will have equal dimensions. Also for the length of the shared representation (encoding dimensions), we empirically discovered that 50 works the best in our setting that balances the predictive power and the error in the reconstructed representation. For the activation functions in the multi-modal stacked autoencoder, we use Rectified Linear Units (ReLUs) for all encoders and sigmoid activation function for all decoders.

Between 15 topics from Cohen dataset, we left one topic out as the test set and used the other 14 topics as the training set. The other 4 datasets, COPD, proton beam, anemia, and clopidogrel, are used as a validation set to tune the hyperparameters and performed the testing on the topic that was being held-out. For a fair comparison,

[5] We attempted to construct higher-order citations but found that not only crawling the network took time, but LINE did not converge within 2 days on a machine with 16 CPU cores and 100GB RAM.

Table 2: Comparison between MMiDaS-AE and other approaches in the inter-topic setting. Cohen *et al.* [13] only reported AUC, thus we only compared WSS@95% score with Norman *et al.* [40]. Bold scores are the top scores while underlined scores are the second best scores.

SR	WSS@95%		AUC		
	MMiDaS-AE	Norman	MMiDaS-AE	Norman	Cohen (2008)
ACEInhibitors	**0.602**	0.566	**0.872**	0.817	0.806
ADHD	**0.661**	0.128	**0.727**	0.591	0.469
Antihistamines	**0.273**	0.073	**0.667**	0.652	0.62
AtypicalAntipsychotics	**0.244**	0.162	0.758	**0.759**	0.653
BetaBlockers	**0.445**	0.400	**0.850**	0.837	0.801
CalciumChannelBlockers	**0.381**	0.129	**0.894**	0.759	0.712
Estrogens	**0.256**	0.176	**0.705**	0.693	0.588
NSAIDS	0.654	**0.671**	0.901	**0.912**	0.899
Opiods	**0.678**	0.301	**0.885**	**0.885**	0.856
OralHypoglycemics	**0.115**	0.072	0.654	**0.657**	0.573
ProtonPumpInhibitors	**0.398**	0.377	**0.857**	0.823	0.793
SkeletalMuscleRelaxants	**0.502**	0.241	**0.848**	0.828	0.836
Statins	**0.341**	0.266	0.819	**0.826**	0.773
Triptans	**0.469**	0.464	**0.825**	0.819	0.823
UrinaryIncontinence	**0.451**	0.374	**0.895**	0.887	0.851

we fixed the validation set to be these 4 datasets, so that SRs from Cohen dataset are used only as training and test set. For the citation network, we used all articles in the partial citation networks (from all train, validation, and test sets), as LINE requires the entire graph as the input. For articles that are present in multiple topics, we remove the sample from the training to prevent data leakage and only use it for testing.

6 EMPIRICAL RESULTS

In this section, we discuss the results from two different settings, inter-topic and fine-tuning. Then we evaluate variants of MMiDaS-AE using just one of the three features, different autoencoders (shallow versus stacked), and our proposed imputation method in the ablation study section.

6.1 Inter-topic Results

As MMiDaS-AE targets a general SR process where we do not assume that we have any labels of the topic, we first use the inter-topic setting. For the setting, we compare the obtained WSS@95% with the values of WSS@95% reported in existing approaches discussed in Section 5.4.1. Table 2 summarizes the results of our model in the inter-topic setting. While Norman reported the scores for both WSS@95% and AUC, Cohen (2008) only reported the AUC score. Thus, we also computed the AUC score of MMiDaS-AE to be comparable with Cohen (2008).

As shown in the table, MMiDaS-AE outperforms Norman in WSS@95% in a range from 1% (Triptans) to 416% (ADHD) except one SR (NSAIDS). Based on the WSS@95% scores, MMiDaS-AE reduces the reviewers' workload by 464 articles compared with Norman, which screens out 157 articles in CalciumChannelBlockers. For Opioids, MMiDaS-AE excludes 1,298 articles while Norman saves 576.

For AUC, MMiDaS-AE mostly outperforms other approaches. For the topics of AtypicalAntipsychotics, NSAIDS, OralHypoglycemics, and Statins, the AUC is lower than Norman but not by a substantial difference. This, coupled with the WSS@95% scores suggests that MMiDaS-AE may not perform as well on lower recall on these topics. Overall, the results show that with an inter-topic setting (non-topic specific setting) MMiDaS-AE performs well with a reasonable score. In other words, MMiDaS-AE works in a general case when we first start SR.

6.2 Fine-tuning and Intra-topic Results

While the inter-topic setting assumes that we do not have any labels for the SR topic, we can also assume that reviewers have labeled small batches of articles. To make this comparison, we use the fine-tuning setting, as discussed in Section 5.4.3, and compare the results against other intra-topic approaches introduced in Section 5.4.2. As they report the score only in WSS@95%, we only compare our results in WSS@95% for this setting. The results are shown in Table 3.

For Antihistamines and SkeletalMuscleRelaxants, according to Cohen *et al.* [15], the classification process did not provide any savings, thus are marked as 0.000 in "Cohen (2006)" column. Except for two SRs, ADHD and Estrogens, MMiDaS-AE outperforms other existing models. For ADHD, the size of the total articles as well as the list of articles that pass the full-text screening are small, thus, the fine-tuning process only marginally improves the results (0.661 in the inter-topic setting versus 0.674 in the fine-tuining setting). We also posit a similar issue with Estrogens, which is that the total number of articles is small and thus fine-tuning only marginally helps. More notably, for Statins, MMiDaS-AE saves reviewers' workload by 1,583 articles while Cohen (2006) saves 856, Khabsa saves 1,386, and Matwin saves 1,091 articles.

Table 3: Comparison between MMiDaS-AE with fine-tuning setting and other approaches in an intra-topic setting that uses 5×2 cross-validation. The scores are in WSS@95%. Bold scores are the top scores while underlined scores are the second-best scores.

SR	MMiDaS-AE	Cohen (2006)	Khabsa	Norman	Matwin
ACEInhibitors	**0.693**	0.566	0.469	0.629	0.523
ADHD	0.674	**0.680**	0.447	0.616	0.622
Antihistamines	**0.287**	0.000	0.03	0.149	0.149
AtypicalAntipsychotics	**0.249**	0.141	0.199	0.21	0.206
BetaBlockers	**0.529**	0.284	0.361	0.511	0.367
CalciumChannelBlockers	**0.439**	0.122	0.287	0.398	0.234
Estrogens	0.262	0.183	0.18	0.292	**0.375**
NSAIDS	**0.671**	0.497	0.404	0.537	0.528
Opiods	**0.694**	0.133	0.455	0.590	0.554
OralHypoglycemics	**0.132**	0.090	0.074	0.111	0.085
ProtonPumpInhibitors	**0.431**	0.277	0.288	0.307	0.229
SkeletalMuscleRelaxants	**0.519**	0.000	0.371	0.429	0.265
Statins	**0.457**	0.247	0.400	0.436	0.315
Triptans	**0.485**	0.034	0.312	0.303	0.274
UrinaryIncontinence	**0.461**	0.261	0.411	0.422	0.296

Table 4: Ablation study on each component. The scores are in WSS@95% using the inter-topic setting. The results of the document, topic and citation network representation are using a basic autoencoder with a single input. A Shallow-AE is using the concatenation of three representations as an input of the autoencoder. MMS-AE uses the multi-modal stacked autoencoder implementation. And Imputation is the result when we apply the imputation technique to multi-modal stacked autoencoder. All results are using binary classification for simplicity. The left side of the table denotes the individual component features, and the right side of the table denotes different autoencoder settings with all three features. Bold scores are the best scores in the right table while underlined scores are the best scores in the left table. We also compare the results using two different pre-trained language models, PMCVec and SciBert.

SR	Document (SciBert)	Document (PMCVec)	Topic	Citation	Shallow-AE	MMS-AE (SciBert)	MMS-AE (PMCVec)	Imputation
ACEInhibitors	0.284	0.128	0.080	0.104	0.196	0.325	0.430	**0.488**
ADHD	0.122	0.179	0.124	0.283	0.133	0.210	0.212	**0.297**
Antihistamines	0.097	0.097	0.090	0.166	0.077	0.214	0.243	**0.246**
AtypicalAntipsychotics	0.064	0.057	0.061	0.119	0.053	0.094	0.156	**0.171**
BetaBlockers	0.127	0.279	0.088	0.141	0.235	0.291	0.319	**0.377**
CalciumChannelBlockers	0.063	0.028	0.107	0.137	0.060	0.117	0.123	**0.152**
Estrogens	0.054	0.055	0.086	0.158	0.078	0.165	0.194	**0.217**
NSAIDS	0.168	0.077	0.226	0.397	0.208	0.523	0.528	**0.597**
Opiods	0.182	0.180	0.124	0.232	0.020	0.220	0.268	**0.379**
OralHypoglycemics	0.034	0.029	0.082	0.028	0.019	0.082	0.080	**0.108**
ProtonPumpInhibitors	0.182	0.175	0.032	0.294	0.178	0.336	0.315	**0.381**
SkeletalMuscleRelaxants	0.238	0.225	0.167	0.364	0.154	0.406	0.489	**0.495**
Statins	0.139	0.174	0.125	0.066	0.157	0.234	0.237	**0.292**
Triptans	0.234	0.102	0.216	0.204	0.199	0.278	0.330	**0.410**
UrinaryIncontinence	0.040	0.124	0.215	0.273	0.314	0.316	0.317	**0.323**

6.3 Ablation Study

In addition to the results for the two settings discussed, we evaluate the results achieved when we ablate the different components of MMiDaS-AE, summarized in Table 4. First, we use a basic autoencoder to compress each of the three representations, ("Document", "Topic", and "Citation") and only train on the individual representation. "Shallow-AE" concatenates the features of all three representations and passes it to a single auto-encoder which is then passed to a softmax layer. The "MMS-AE" is the multi-modal stacked autoencoder implementation [9] without any imputation. And finally, we show the results of our proposed imputation process. All the results are shown in Table 4 are only using binary classification with full-text screening as a label (not multi-label classification task we proposed in Section 4.3) with an inter-topic setting. Therefore, the results are different from the results reported in Table 2 which also demonstrates the added benefit of using multi-label classification task. Also to evaluate the results with a larger pre-trained language model, we compare the PMCVec representation with SciBert representation using the "Document" and "MMS-AE" settings. We only evaluated "Topic" using PMCVec as we also evaluate the result on "MMS-AE".

In comparing individual components in Table 4, if the test set has a large number of articles in total, it leads to a high WSS@95% when using the document representation only. For example, ACEInhibitors has 2,544 articles in total, and Statins has 3,465 articles in total, and both SRs have a relatively higher WSS@95% than other individual components. There are cases when using only the citation representation is better. This also depends on the number of articles that lack citation information. For example, ADHD has only 6% of articles missing citation information and Opioids has 8% of articles missing citation information, and both have higher WSS@95% for the citation representation than other individual components. However, for ACEInhibitors 17% of articles are missing citation information and Statins has 15% of articles missing citation information, thus both have a lower WSS@95% than using only document representation.

Learning a classifier for SR is a difficult task as we only use partial information (title, abstract, and MeSH terms) to predict whether the article passed the full-text screening (where full-text is not included as a feature). Our intuition is that citation network representation can complement the lack of full-text information to improve the overall performance as citations are used in the full-text. In SR, although reviewers consider texts (title, abstract, and full-text), this implicitly considers the co-citation information. By comparing "Citation" and "MMS-AE" in Table 4, we can see cases when WSS@95% of using only citation representation outperforms multi-modal settings such as ADHD and CalciumChannelBlockers. This demonstrates the usefulness of the citation information.

In most cases, Shallow-AE performs worse than individual components which implies that the simple concatenation of representations does not learn a robust shared representation that encodes all three representations. However, if we use MMS-AE, it performs better in all topics compared to Shallow-AE. This suggests that MMS-AE is learning a more robust shared representation than Shallow-AE. Finally, if we apply the imputation technique proposed in Section 4.2, it performs the best and can reduce the workload by up to 59.7% compared to the MMS-AE. In addition, comparison between the WSS@95% scores in the "Imputation" column in Table 4 and the MMiDaS-AEcolumn in Table 2, shows a significant improvement through the introduction of the multi-label formulation discussed in Section 4.3.

All the results shown in Table 2 and Table 3 are using PMCVec for the document and topic features. However, we wanted to evaluate the difference in using approaches that exploit pre-trained representations induced by large transformers such as SciBert [7]. We compared the results using SciBert and PMCVec on "Document" and "MMS-AE" in Table 4. As shown, for most of the cases when using single-component (only document as a feature), SciBert performs better than PMCVec. But when using the MMS-AE setting, PMCVec outperforms SciBert in most of the cases. This illustrates the importance of the number of dimensions in MMS-AE. We note that the dimension of SciBert is 768 while the dimension of PMCVec is 200. When using a single-component, we can select the size of the hidden layer based on the input dimension. However, for MMS-AE, the three features are encoded into a shared representation, and it becomes difficult when the dimension of one input differs greatly from the other input. In other words, there will be information lost from the input feature with a larger dimension when learning the shared representation. Thus, in MMS-AE, information from Document and Topic is lost when learning the shared representation and consequently performs worse than the single-component.

7 CONCLUSIONS AND LIMITATIONS

In this paper we proposed using the Multi-modal Missing Data aware Stacked Autoencoder (MMiDaS-AE) — inspired by [9] — for biomedical citation screening. The aim is to reduce the workload involved in the systematic review process via semi-automation. We showed that this multi-modal approach, which treats title/abstract texts, citation networks, and topics as separate modalities and explicitly models these, outperforms prior models in inter-topic settings. Further, in the topic-specific (intra-topic) setting, our fine-tuned MMiDaS-AE outperforms alternative approaches.

This provides evidence that capitalizing on three (potentially complementary) representations is a promising approach. The main strength of MMiDaS-AE is that the model is able to handle missing data via imputation. Imputation while training makes the multi-modal stacked autoencoder learn a more robust shared representation. Even if some of the test data is missing a particular modality, MMiDaS-AE can find a robust shared representation. Also as shown in the ablation study, co-citation information and the citation network representation plays an important role in the performance. Thus, citation representation can support the SR process.

There are important limitations to this study. First, we have only experimented on a small collection of 15 systematic reviews, which are very related in topic. This is an ideal scenario for transfer learning, and it is not clear if fine-tuning pre-trained models would be as efficient when considering a more diverse set of topics. We aim to next evaluate the model on a completely held out set of systematic reviews to assess generalizability.

8 ACKNOWLEDGEMENTS

Eric W. Lee and Joyce C. Ho were supported by the National Science Foundation award IIS-#1838200 and the National Institute of Health award 1K01LM012924-01. Karla I. Galaviz was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (P30DK111024). Byron C. Wallace was supported by the National Library of Medicine of the National Institutes of Health under award number 2R01LM012086. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] JJ García Adeva, JM Pikatza Atxa, M Ubeda Carrillo, and E Ansuategi Zengotitabengoa. 2014. Automatic text classification to support systematic reviews

in medicine. *Expert Systems with Applications* 41, 4 (2014), 1498–1508.

[2] I Elaine Allen and Ingram Olkin. 1999. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA* 282, 7 (Aug. 1999), 634–635.

[3] Ethan Balk, Gowri Raman, Mei Chung, Stanley Ip, Athina Tatsioni, Alvaro Alonso, Priscilla Chew, Scott J Gilbert, and Joseph Lau. 2006. Effectiveness of management strategies for renal artery stenosis: a systematic review. *Annals of internal medicine* 145, 12 (2006), 901–912.

[4] Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew SC Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews* 8, 1 (2019), 23.

[5] Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLOS Medicine* 7, 9 (Sept. 2010), e1000326.

[6] Tanja Bekhuis and Dina Demner-Fushman. 2010. Towards automating the initial screening phase of a systematic review.. In *MedInfo*. 146–150.

[7] Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[9] Cesar Cadena, Anthony R Dick, and Ian D Reid. 2016. Multi-modal Auto-Encoders as Joint Estimators for Robotics Scene Understanding.. In *Robotics: Science and Systems*, Vol. 5. 1.

[10] Iain Chalmers, Larry V Hedges, and Harris Cooper. 2016. A brief history of research synthesis. *Evaluation & the Health Professions* 25, 1 (June 2016), 12–37.

[11] Jackie Chandler, Rachel Churchill, Julian Higgins, Toby Lasserson, David Tovey, et al. 2013. Methodological standards for the conduct of new Cochrane Intervention Reviews. *Sl: Cochrane Collaboration* (2013).

[12] Mei Chung, Ethan M Balk, Stanley Ip, Gowri Raman, Winifred W Yu, Thomas A Trikalinos, Alice H Lichtenstein, Elizabeth A Yetley, and Joseph Lau. 2009. Reporting of systematic reviews of micronutrients and health: a critical appraisal. *The American journal of clinical nutrition* 89, 4 (2009), 1099–1113.

[13] Aaron M Cohen. 2008. Optimizing feature representation for automated systematic review work prioritization. In *AMIA annual symposium proceedings*, Vol. 2008. American Medical Informatics Association, 121.

[14] Aaron M Cohen, Kyle Ambert, and Marian McDonagh. 2009. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association* 16, 5 (2009), 690–704.

[15] Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13, 2 (2006), 206–219.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Karla Ivette Galaviz, Mary Beth Weber, Audrey Straus, Jeehea Sonya Haw, KM Venkat Narayan, and Mohammed Kumail Ali. 2018. Global diabetes prevention interventions: a systematic review and network meta-analysis of the real-world impact on incidence, weight, and glucose. *Diabetes Care* 41, 7 (2018), 1526–1534.

[18] Zelalem Gero and Joyce Ho. 2019. PMCVec: Distributed phrase representation for biomedical text processing. *Journal of Biomedical Informatics: X* 3 (2019), 100047.

[19] David Gough and Diana Elbourne. 2002. Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society* 1, 3 (July 2002), 225–236.

[20] David Gough, Sandy Oliver, and James Thomas. 2017. *An introduction to systematic reviews* (second ed.). Sage.

[21] Neal R Haddaway and Martin J Westgate. 2018. Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology* 6 (Oct. 2018), 136.

[22] Kazuma Hashimoto, Georgios Kontonatsios, Makoto Miwa, and Sophia Ananiadou. 2016. Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of biomedical informatics* 62 (2016), 59–65.

[23] Brian E Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R Shah, Stephanie Holmgren, Katherine E Pelch, Vickie Walker, Andrew A Rooney, Malcolm Macleod, Ruchir R Shah, and Kristina Thayer. 2016. SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews* 5, 1 (May 2016), 87.

[24] Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J Marshall, and Byron C Wallace. 2018. Learning disentangled representations of texts with application to biomedical abstracts. *arXiv preprint arXiv:1804.07212* (2018).

[25] Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis* 6, 5 (2002), 429–449.

[26] Tom Kenter, Alexey Borisov, and Maarten De Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).

[27] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning* 102, 3 (2016), 465–482.

[28] Georgios Kontonatsios, Austin J Brockmeier, Piotr Przybyła, John McNaught, Tingting Mu, John Y Goulermas, and Sophia Ananiadou. 2017. A semi-supervised approach using label propagation to support citation screening. *Journal of biomedical informatics* 72 (2017), 67–76.

[29] Ioannis Koulouridis, Mansour Alfayez, Thomas A Trikalinos, Ethan M Balk, and Bertrand L Jaber. 2013. Dose of erythropoiesis-stimulating agents and adverse outcomes in CKD: a metaregression analysis. *American Journal of Kidney Diseases* 61, 1 (2013), 44–56.

[30] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[31] Ivan Lerner, Perrine Créquit, Philippe Ravaud, and Ignacio Atal. 2019. Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *Journal of clinical epidemiology* 108 (2019), 86–94.

[32] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*. 5040–5048.

[33] Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. 2010. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association* 17, 4 (2010), 446–453.

[34] Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics* 51 (2014), 242–253.

[35] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and The PRISMA Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine* 6, 7 (July 2009), e1000097.

[36] Zoë Slote Morris, Steven Wooding, and Jonathan Grant. 2011. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine* 104, 12 (Dec. 2011), 510–520.

[37] Sally Morton, Alfred Berg, Laura Levit, Jill Eden, et al. 2011. *Finding what works in health care: standards for systematic reviews.* National Academies Press.

[38] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[39] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4920–4928.

[40] Christopher Norman, Mariska Leeflang, Pierre Zweigenbaum, and Aurélie Névéol. 2018. Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat.

[41] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 4, 1 (2015), 5.

[42] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan-a web and mobile app for systematic reviews. *Systematic reviews* 5, 1 (2016), 210.

[43] Pascal Vincent PASCALVINCENT and Hugo Larochelle LAROCHEH. 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion pierre-antoine manzagol. *Journal of Machine Learning Research* 11 (2010), 3371–3408.

[44] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[45] Ian Shemilt, Antonia Simon, Gareth J Hollands, Theresa M Marteau, David Ogilvie, Alison O'Mara-Eves, Michael P Kelly, and James Thomas. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* 5, 1 (2014), 31–49.

[46] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[47] Teruhiko Terasawa, Tomas Dvorak, Stanley Ip, Gowri Raman, Joseph Lau, and Thomas A Trikalinos. 2009. Systematic review: charged-particle radiation therapy for cancer. *Annals of internal medicine* 151, 8 (2009), 556–565.

[48] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics*. Springer, 382–392.

[49] Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, and Thomas A Trikalinos. 2012. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *proceedings of the 2nd ACM SIGHIT International*

Health Informatics Symposium. ACM, 819–824.

[50] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. 2010. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 173–182.

[51] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. 2011. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*. IEEE, 754–763.

[52] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 11, 1 (2010), 55.