Chapter 8
Using Machine Learning to Locate Evidence More Efficiently:
New Roles for Academic Research Librarians

Michelle A. Cawley
University of North Carolina at Chapel Hill, USA

ABSTRACT

Evidence that machine learning can assist article selection and minimize manual screening burden for scholarly research has been documented in the peer-reviewed literature for more than 20 years. Despite the robust evidence and continual technological advances, uptake has been slow among research teams. This chapter discusses the benefits of using machine learning (ML) and other automation tools on bibliographic data and argues that academic librarians are well-positioned to partner with research teams around this application of ML. An overview of the automation approaches used at UNC's Health Sciences Library (HSL) is discussed along with detailed accounts of multiple success stories of when HSL librarians partnered with research teams to locate evidence more efficiently. Finally, a discussion of likely barriers and possible solutions to increase uptake of this technology among academic librarians is provided.

INTRODUCTION

At UNC Chapel Hill's Health Sciences Library (HSL), health sciences librarians have been using multiple approaches, including machine learning (ML), to locate evidence more efficiently for researchers since 2018. Specifically, HSL librarians use ML and other automation techniques on literature search results with the primary objective of reducing the volume of literature that must be screened manually by research teams. This use of ML-assisted screening can save significant time and resources for research

DOI: 10.4018/978-1-7998-9702-6.ch008

teams, shortening the time it takes to conduct and publish scholarly research. ML-based automation methods also represent a paradigm shift and offer research teams an opportunity to broaden the scope of research questions without wading through vast amounts of literature.

In their white paper discussing the future of data science in libraries, Burton et al. (2018) recommend that academic libraries highlight success stories around data science. At HSL, cases where ML-based approaches have been effective at minimizing screening burden include large, comprehensive searches and search updates, including those for systematic reviews. HSL librarians have also used ML-based methods to get quick answers from relatively small datasets and identify evidence from very low precision searches (i.e., the needle in the haystack problem).

There is no single best approach for using ML on bibliographic data, which means experienced partners are needed to consult on projects to apply the technology appropriately. Academic librarians sit at a critical intersection to support research teams in applying this technology given their knowledge of bibliographic data and databases, term indexing, and keyword searching techniques.

This chapter provides an overview of several applications and case studies of ML-based approaches to reduce manual screening burden for article selection. A discussion of the barriers, including a skills gap among librarians, access to freely available or low-cost software, and hesitancy among research teams along with potential solutions is also provided.

BACKGROUND

AI-enabled approaches, including ML, have been developed, tested, and validated to minimize manual screening of search results for large, complex research questions. This application of ML has been documented in the peer-reviewed literature across multiple domains for several decades (Aphinyanaphongs et al., 2005; Bannach-Brown et al., 2019; Bekhuis & Demner-Fushman, 2012; Cohen et al., 2006; Mostafa & Lam, 2000; O'Mara-Eves et al., 2015; Shemilt et al., 2014; Thomas et al., 2021; Varghese et al., 2018; Wallace et al., 2010), yet application of this technology by research libraries has been nearly non-existent. U.S. Federal agencies, including the U.S. Environmental Protection Agency (U.S. EPA), have successfully applied ML-enabled approaches to their large-scale risk assessments for several years (Cawley et al., 2020), which affords the opportunity to locate relevant evidence in a large set of search results without relying entirely on keywords.
Applications of AI-enabled technology to bibliographic data may include:

- **Clustering or unsupervised learning** to assist with search strategy development and to identify a pocket of search results within a large result set to then review manually.
- **Supervised clustering, a form of semi-supervised learning**, to prioritize literature to screen manually with the ability to predict recall.
- **Machine learning or supervised learning** to predict the probability of an individual article being relevant to a particular research question.

The automation approaches used by HSL librarians, including ML, are typically employed to save time for research teams. Saving time translates to publishing scholarly research faster and often results in cost savings for research teams. Saving time can also mean researchers find answers and evidence more quickly, which can be valuable in any number of circumstances. It certainly has applications to

the current "infodemic," defined by the World Health Organization (WHO) as too much information that causes confusion or risk-taking behaviors that harm human health (WHO, n.d.). A simple search of "COVID-19" in PubMed returns over 200,000 articles published since 2020—an overwhelming number of results. Using ML-assisted methods on large search result sets such as this can be an effective and time-saving alternative to traditional methods.

In addition to ML-driven approaches to bring efficiencies to research teams, HSL librarians also use low-technology solutions to save time and resources and provide data-driven insights to optimize comprehensive search strategies.

DEFINITIONS

SAS Institute defines machine learning as "a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention" (SAS Institute, n.d.)

For clarity, this chapter will use the term "machine learning" or ML to refer broadly to the methodology that includes unsupervised learning, semi-supervised learning, and supervised learning (see Table 1 for definitions of all ML-related terms used in this chapter). "Supervised learning" as used in this chapter is often generally referred to as machine learning; however, supervised learning is used throughout to contrast with semi-supervised learning and unsupervised learning and to distinguish from the broader concept of ML. ML-based approaches used by HSL librarians include unsupervised learning, semi-supervised learning with an ensemble approach, and supervised learning (Table 1). The technical details of the ML algorithms and how they are applied to bibliographic data and other text are reviewed and explained in numerous other sources and will not be discussed here (Aphinyanaphongs et al., 2005; Bekhuis & Demner-Fushman, 2012; O'Mara-Eves et al., 2015; Shemilt et al., 2014; Varghese et al., 2018; Wallace et al., 2010).

This chapter will not discuss active machine learning or deep learning (i.e., neural networks). Active machine learning requires a robust user interface as the model is trained in real time, making this method likely to be available only through commercially available software. Recent advances to neural network-based ML models are making these approaches more accessible. For instance, Bidirectional Encoder Representations from Transformers (BERT) is open source and freely available to use in Google (Devlin et al., 2019). This computationally heavy approach can run in a few hours and is pretrained using plain text (e.g., Wikipedia pages), which saves resources needed to develop training data similar to other unsupervised methods such as clustering.

Table 1. Machine learning (ML) terms and descriptions

Term	Description
Artificial intelligence (AI)	A computer system able to perform tasks that normally require human intelligence.
Machine learning (ML)	A branch of AI; generally defined as using machines to complete human tasks. Often used interchangeably with the term AI.
Labeled data	Data that have been classified manually by one or more people. For example, an article citation categorized as relevant or not relevant by a researcher. Labeled data are used to make predictions on unlabeled data with ML models.
Training data	Refers to labeled data that are used to build a ML model for making predictions on unlabeled data.
Seeds*	Refers to training data used in semi-supervised learning. Seeds are known relevant studies identified through manual screening. Ideally seeds should be identified by reviewing batches of studies identified at random from the unclassified corpus.
Unsupervised learning	Type of ML that does not require training data; *k*-means, nonnegative matrix factorization (NMF), and latent Dirichlet allocation (LDA) are examples of clustering algorithms. ***Synonyms**: clustering; unsupervised machine learning, topic extraction**
Semi-supervised learning	Type of ML that uses a small amount of labeled data. ***Synonyms**: supervised clustering*; semi-supervised machine learning*
Ensemble approach*	Ensemble Approach refers to using multiple clustering models to increase the accuracy of predictions compared to using a single clustering model. It is essentially a voting method such that each model gets a "vote" as to whether a given study is likely to be relevant. The votes are tallied to get an ensemble score (ES) for each study. The ES is used to prioritize a set of studies run through semi-supervised learning with an ensemble approach. For example, using a six-model ensemble approach would result in ensemble scores ranging from 0 to 6. Studies with an ES = 0 were not "voted" as likely to be relevant by any of the 6 models. In contract, studies with an ES = 6 were "voted" as likely to be relevant by all 6 models (Figure 3).
Ensemble score (ES)*	
Supervised learning	Type of ML that uses a large set of training data (i.e., labeled data) to build an ML model. Naïve Bayes, k-nearest neighbors, and support vector machines (SVM) are examples of supervised learning algorithms. ***Synonyms**: machine learning; supervised machine learning*
*These terms refer to specific approach primarily used by author. For additional details see Cawley et al. (2020) and Varghese et al. (2018).	

ROLE OF ACADEMIC LIBRARIANS

Although a skills gap exists among academic librarians in terms of the technical skills needed to offer value in data-rich research and academic settings, Burton et al. (2018) maintain that librarians have "a crucial role in the future of the data science ecosystem." Similarly, the Obama Administration's *Federal Big Data Research and Development Strategic Plan* (NITRD/NCO, 2016) notes that to meet demand around data services, domain experts, including librarians, trained in data science, are needed.

Barriers aside, academic research librarians are well-positioned to apply automation tools to bibliographic data. Their expertise with scholarly databases, term indexing, and keyword search strategies can be leveraged when they partner and consult with research teams to improve the efficiency with which relevant evidence is identified. ML-based approaches are widely used across multiple domains on a wide variety of data, so research partners are familiar with the terminology and approaches. At HSL, research teams view librarians as valued partners in applying this technology to bibliographic data.

Health sciences librarians in particular should seize the opportunity to leverage their deep understanding of bibliographic data and databases to improve the efficiency of locating evidence for large, comprehensive research questions, including systematic reviews. Systematic review methodology is being adopted by other disciplines, such as public health, where the volume of literature returned is often untenable. ML-based approaches are not one-size-fits-all; research teams need an experienced domain expert to apply this technology effectively. Similar to developing a search strategy, an iterative approach is necessary when applying ML to bibliographic data. Ultimately, academic librarians have an opening to serve on research teams as subject matter experts around automation techniques for bibliographic data.

TRADITIONAL APPROACH

Typically, academic librarians partner with subject matter experts to design a keyword search strategy with multiple sets of terms. This process is often iterative as it requires several rounds of back and forth with the research team to balance specificity and sensitivity. In a comprehensive literature search, as is needed for a systematic review, sensitivity or maximizing recall is paramount and often requires that specificity or precision is sacrificed. This often results in a low precision search where most studies are not relevant (Cohen et al., 2006; Golder et al., 2006). For example, Belter (2016) reported search precision of less than 5% for 14 systematic reviews.

While balancing search precision and search recall, the research team and academic librarian also must consider available resources. A research question might need to be narrowed or otherwise modified if resources (e.g., time, staffing, funding) are insufficient to allow the total number of search results returned by the search strategy to be manually screened. In practice, assuming the search is returning relevant results, this often necessitates modifying the search strategy until an acceptable number of results is returned.

For example, in a search for health effects of disinfectants used in hospital settings it may be untenable to use the term *disinfectant* due to the high number of records returned. In this example, the research team may be forced to identify all the specific disinfectants by name and synonyms that could be used in hospital settings while omitting the term *disinfectant* to avoid a substantial increase in search results (or vice versa). An actual search conducted on this topic included over 400 terms for specific disinfectants; when adding the general terms *disinfectant*, *disinfectants*, and *disinfection* to the keywords, the already large set of results increased by more than 20%.

PARADIGM SHIFT

Using ML to reduce the volume of literature that must be manually screened is a paradigm shift from the traditional approach. With ML, librarians and research teams have another tool at their disposal beyond keywords to locate relevant articles. Although the initial corpus of literature will still be defined using a keyword search strategy, the strategy need not be developed with such a close eye on total number of results returned. From the example above of disinfectants in hospital settings, the research team would be free to use the general terms (i.e., *disinfectant*, *disinfectants*, *disinfection*) without being overly concerned about the increase in search results.

Further, using the ML paradigm, the research team may have the option of using a simpler search strategy and expanding the scope of their question beyond what is already known. For example, for a research question designed to investigate the human health outcomes associated with exposure to arsenic—a question that U.S. EPA and other public health agencies consider for various toxicants—the traditional approach might include two or three sets of keywords, as follows:

- **Set 1**: Terms related to arsenic, including related compounds and synonyms
  AND
- **Set 2:** Terms related to possible health outcomes
  AND
- **Set 3**: Terms related to exposure settings

The traditional paradigm would require that the research team determine all possible health outcomes, including those that are known (e.g., bladder cancer) and those that are unknown but still plausible—a much more difficult exercise. Under the traditional paradigm, the onus to identify all possible keywords is on the research team and librarian. Using the ML paradigm, the research team has the option to use keywords only for the toxicant (i.e., Set 1 in this example) to develop the initial corpus of results. This relieves the librarian of the burden of defining all possible health outcomes and exposure settings. The librarian would then apply ML to the corpus to identify a subset of studies to screen manually, i.e., studies most likely to be related to health outcomes resulting from arsenic exposure.

OVERVIEW OF AUTOMATION APPROACHES

The application of ML discussed here is designed to reduce the volume of literature that must be screened manually and includes multiple approaches such as unsupervised learning, semi-supervised learning with an ensemble approach, and supervised learning.

Technical details around the approaches along with relevant success stories are laid out below. In each success story the projects used biomedical literature. Citations are provided when applicable for projects that resulted in peer-reviewed publications with a discussion of the ML approach used. For all success stories detailed below, HSL used DoCTER software (Document Classification and Topic Extraction Resource; ICF, n.d.). DoCTER prioritizes search results using the text of titles and abstracts and has functions for unsupervised learning, semi-supervised learning, and supervised learning.

Each success story provided below includes an estimate of time saved. Systematic review methodology stipulates that each article returned from a search strategy is reviewed by two independent screeners for title-abstract screening and again by two independent screeners for studies that move forward to full-text screening. At the title-abstract screening phase, 1 min/study per screener (i.e., 2 min per study) is a general rule of thumb (Cawley et al., 2020). This estimate was used to derive the estimated time saved screening for systematic reviews; other projects that do not require two screeners used an estimate of 1 min/study. There is some overhead cost in terms of time for using these approaches; however, it is typically not significant (i.e., less than 4 hr) and may be offset by a lower logistical cost of using a smaller team to screen studies.

Unsupervised Learning

At HSL unsupervised learning or clustering is used by librarians to get quick, data-driven insights into a set of search results. Unsupervised learning is used to "refine or shine"—that is *refine* a search strategy or *shine* a light on a subset of the search results that is likely to contain results of interest. Unsupervised learning can also provide insight into unexpected results. For example, a search was designed to locate research related to homelessness among youth. Using unsupervised learning on the preliminary search results revealed that the search was returning studies on children orphaned due to HIV/AIDS. This aspect of youth homelessness was out of scope for the project and the research team could choose to refine the search strategy or omit this cluster to exclude these results from consideration.

In practice, HSL librarians use unsupervised learning by clustering title and abstract text from PubMed search results. The *k*-means algorithm is typically used on PubMed results, and these results are placed into 10 clusters. PubMed results are used because it is fast to download a large result set and unsupervised learning can easily be repeated for each iteration of the search strategy. Running 20,000 or more results through unsupervised learning only takes a few minutes and does not require training data or seeds. Each result (i.e., title and abstract text for a single study) is placed into a single cluster and the algorithm generates a set of keywords for each cluster (Figure 1).

Interpreting results from unsupervised learning may require subject matter knowledge to make sense of keywords for each cluster generated by the algorithm. There are also no quantifiable predictions of the effect on search recall if a research team chooses to exclude one or more clusters from manual review.

Success Story 1 below demonstrates how HSL librarians have used unsupervised learning to identify a pocket of relevant literature from a relatively small dataset without the need for training data (Table 2). For an additional demonstration of unsupervised learning, Success Story 3 includes unsupervised learning as part of a two-phase approach with supervised learning.

Figure 1. Sample output for unsupervised learning using k-means algorithm and 10 clusters. Search results are related to effects of feral cat colonies on humans and the environment in this example. Results can be used to refine search strategy or identify one or more clusters to look at more closely. For example, reviewing the keywords for Clusters 3 and 10 indicates that studies tagged to Cluster 3 are less likely to be relevant compared to Cluster 10. Cluster 3 appears to contain studies related to indoor allergies resulting from domestic pets, not feral cat colonies.
Source: Author, 2022.

Cluster	# of Studies	Keywords
1	3082	cats | dogs | infection | animals | disease | domestic | virus | species | feline | rabies | cases | animal | infected | wild | human | prevalence | infections | humans | samples | transmission
2	389	users | download | email | permission | print | holder | listserv | copied | emailed | posted | abridged | articles | warranty | written | express | property | copy | refer | accuracy | version
3	357	allergen | dust | allergens | indoor | fel | mite | homes | levels | asthma | exposure | der | air | airborne | children | concentrations | allergic | home | samples | house | dog
4	1349	enzyme | activity | substrate | catalytic | ph | degrees | purified | active | kinetic | site | acid | protein | mutant | binding | amino | enzymes | fold | residues | temperature | values
5	436	asthma | children | allergic | allergens | ige | sensitization | atopic | allergen | allergy | skin | dust | risk | exposure | ci | symptoms | 95 | atopy | age | patients | prick
6	1079	gene | expression | cells | promoter | cell | protein | genes | type | transcription | virus | dna | binding | chloramphenicol | activity | wild | reporter | sequence | mrna | mutant | induced
7	1570	neurons | visual | responses | cortex | response | stimulation | stimuli | nerve | auditory | stimulus | cortical | activity | cats | cells | frequency | spinal | nucleus | environment | noise | neural
8	531	gondii | toxoplasma | infection | oocysts | toxoplasmosis | antibodies | cats | prevalence | seroprevalence | infected | parasite | test | mice | samples | transmission | agglutination | risk | domestic | animals | positive
9	6096	cats | species | environment | study | results | effects | different | stress | used | time | effect | animals | increased | using | activity | water | treatment | high | temperature | higher
10	2905	feral | cats | species | island | population | conservation | felis | predation | prey | habitat | catus | predators | populations | islands | wildlife | mammals | australia | areas | birds

Success Story 1: Unsupervised Learning for Quick Answers

For this project, the researcher asked the health sciences librarian to carry out a quick turnaround search. Unlike a systematic review or other comprehensive search, the requestor did not need an exhaustive list of literature, but rather wanted a targeted list of citations as fast as possible. Summary information is presented in Table 2. This project had the following objective with respect to ML:

- Identify a priority set of studies most likely to be relevant within a set of search results.

Table 2. Summary information for unsupervised learning (Success Story 1)

Component	Summary Information
Training data	N/A
Summary of approach	Unsupervised learning to identify a subset of studies to screen
Total unique citations	3,559 studies
Total studies screened manually	931 studies (26% of total search results)
Total studies excluded using ML	2,628 studies (74% of total search results)
Estimated time saved*	44 hr (1.1 work weeks)
*Given that this project was not a systematic review and therefore did not require two independent screeners, the estimate of time saved is based on 1 min x 1 screener/study.	

Semi-Supervised Learning

Semi-supervised learning uses the same underlying process as unsupervised learning. The difference is that a set of known relevant studies, or seeds, are clustered along with the unclassified corpus. The seeds are tracked so that clusters containing a high percentage of seeds can be prioritized for manual review (Figure 2). The logic is that similar studies cluster together, so unlabeled studies appearing in the same cluster as labeled studies (or seeds) are more likely to be relevant.

The seeds used in semi-supervised learning provide a quantitative signal beyond the algorithm-generated keywords (Figure 2; note that Figure 2 is identical to Figure 1 except for a column indicating how many seeds were found in each cluster). Semi-supervised learning requires relatively few seeds (i.e., training data) and can take the guesswork out of which clusters to review. At HSL, semi-supervised learning with an ensemble approach (see below) is typically used, so that recall can be predicted.

To identify seeds, a random sample of 250 studies from the unclassified corpus is identified. These studies are reviewed by a subject matter expert to identify relevant articles with the goal of retrieving 25–50 seeds. This process may continue by pulling additional random samples if necessary. However, if the team has not identified 25 positive studies after reviewing 500 studies selected at random, another approach may be necessary (see Success Story 3).

Another option if seed studies are overly burdensome to locate would be to use an outside source (i.e., studies not in the search results). Cawley et al. (2020) demonstrated that externally derived training data can be effective at locating relevant studies in a related unclassified corpus. Similarly, Success Story 5 details a project where seed studies from an original search were used to identify studies most likely to be relevant in a search update.

Figure 2. Sample output from semi-supervised learning using k-means algorithm and 10 clusters. Search results are related to effects of feral cat colonies on humans and the environment in this example. The percentage of seeds provides a quantitative signal (i.e., clusters with a high percentage of seeds are more likely to contain relevant results) to use in addition to the algorithm-generated keywords.
Source: Author, 2022.

Cluster	# of Studies	% of Seeds	Keywords
1	3082	16	cats | dogs | infection | animals | disease | domestic | virus | species | feline | rabies | cases | animal | infected | wild | human | prevalence | infections | humans | samples | transmission
2	389	0	users | download | email | permission | print | holder | listserv | copied | emailed | posted | abridged | articles | warranty | written | express | property | copy | refer | accuracy | version
3	357	1	allergen | dust | allergens | indoor | fel | mite | homes | levels | asthma | exposure | der | air | airborne | children | concentrations | allergic | home | samples | house | dog
4	1349	2	enzyme | activity | substrate | catalytic | ph | degrees | purified | active | kinetic | site | acid | protein | mutant | binding | amino | enzymes | fold | residues | temperature | values
5	436	0	asthma | children | allergic | allergens | ige | sensitization | atopic | allergen | allergy | skin | dust | risk | exposure | ci | symptoms | 95 | atopy | age | patients | prick
6	1079	5	gene | expression | cells | promoter | cell | protein | genes | type | transcription | virus | dna | binding | chloramphenicol | activity | wild | reporter | sequence | mrna | mutant | induced
7	1570	4	neurons | visual | responses | cortex | response | stimulation | stimuli | nerve | auditory | stimulus | cortical | activity | cats | cells | frequency | spinal | nucleus | environment | noise | neural
8	531	11	gondii | toxoplasma | infection | oocysts | toxoplasmosis | antibodies | cats | prevalence | seroprevalence | infected | parasite | test | mice | samples | transmission | agglutination | risk | domestic | animals | positive
9	6096	3	cats | species | environment | study | results | effects | different | stress | used | time | effect | animals | increased | using | activity | water | treatment | high | temperature | higher
10	2905	58	feral | cats | species | island | population | conservation | felis | predation | prey | habitat | catus | predators | populations | islands | wildlife | mammals | australia | areas | birds

Semi-Supervised Learning with an Ensemble Approach

An ensemble approach refers to using multiple clustering models to increase the accuracy of predictions compared to using a single clustering model. The underlying method is the same as semi-supervised learning. At HSL, semi-supervised learning with an ensemble approach is used to prioritize studies into batches for manual review. This method is preferred to prioritize studies because unlike unsupervised learning, recall can be predicted, and unlike supervised learning, fewer training data are needed.

In practice at HSL, a six-model ensemble approach with semi-supervised learning is used, meaning the unclassified corpus and seed studies are run through six models. Models are based on using different algorithms (e.g., *k*-means, LDA, NMF) or different number of clusters (10, 20, or 30 clusters) (Cawley et al., 2020).

For each model, a study is given a model score of 0 or 1. A study receives a model score of 1 if it is "voted" relevant, indicating it was part of a cluster with a high proportion of seed studies. Otherwise, it receives a model score of 0. Model scores are tallied to calculate the ensemble score (ES; Table 3). With a six-model ensemble approach the ES will range from 0 to 6. For studies with ES = 0, none of the six

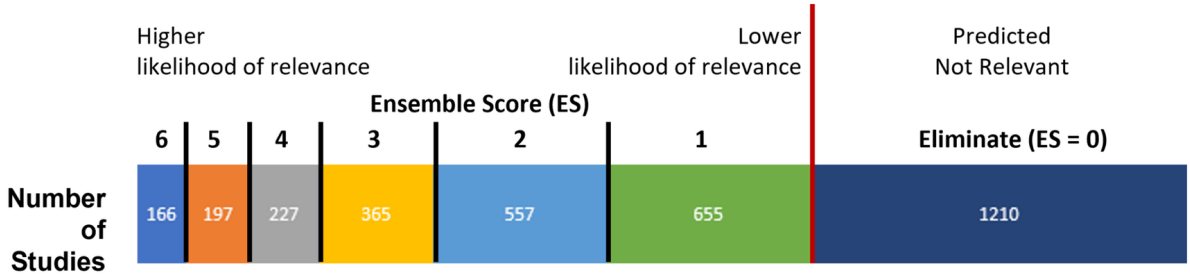
models "voted" the study relevant and ES = 6 indicates the study was "voted" relevant by all six models (i.e., it was part of a relevant cluster in all six models). The research team can begin by reviewing studies with an ES = 6, followed by studies with an ES = 5, and so on (Figure 3). Studies with an ES = 0 were not found in a relevant cluster by any of the six models and can be excluded without manual review.

Recall can be predicted when using semi-supervised learning with an ensemble approach. For example, 90% recall of relevant studies in the unclassified corpus can be expected if clusters are selected with up to or greater than 90% of the seed studies.

Table 3. Sample results from semi-supervised learning with three-model ensemble approach. A model score of 1 indicates the study was found in a relevant cluster (i.e., a cluster with a high percentage of seed studies). The ensemble score is the sum of model scores with ES = 3, indicating the study was found in a relevant cluster for all three models.

Unique Study ID	Model Score 1	Model Score 2	Model Score 3	Ensemble Score
5202	1	1	1	3
16055	0	1	1	2
16097	0	0	0	0
17115	1	1	0	2
22056	1	1	1	3
26039	1	0	0	1

Figure 3. Example results from semi-supervised learning with six-model ensemble approach. Ensemble score (ES) ranges from 6 to 0 indicating how many of the six models in which the study was found in a relevant cluster. For example, 166 studies have an ES = 6, indicating they were found in a relevant cluster by all six models and are more likely to be relevant.
Source: Author, 2022.



Success Story 2 describes how semi-supervised learning was used with a stratified approach, which can be helpful when a particular topic in a set of search results is likely to have relatively few results.

At HSL, semi-supervised learning is typically used in a two-phase approach with supervised learning (see Success Story 5).

Success Story 2: Stratified Approach with Semi-supervised Learning and Ensemble Approach

For this project, the research team conducted a comprehensive search that returned approximately 7,000 studies after duplicates were removed. The search strategy was designed to capture results from low- and middle-income countries (LMICs) and high-income countries (HICs). Search results were divided into two groups, including those set in LMICs (813 citations) and all others (6,075 citations). Given the size imbalance between these two groups, it would be inappropriate to use a single ML-based approach on all 7,000 citations. The concern was that the non-LMIC studies would drown out the signal from the smaller group of studies set in LMICs. To overcome this barrier, a stratified approach was used to prioritize studies for manual screening (Table 4). This project had the following objectives with respect to ML:

- Reduce the volume of literature requiring manual screening from the literature search.
- Design a stratified approach suitable for this dataset to ensure relevant studies set in LMICs were identified.

For this project all studies set in LMICs (813 studies) were manually screened and semi-supervised learning with an ensemble approach was used to prioritize the remaining results for manual screening (6,075 studies). The purpose of this project was to conduct a bibliometric analysis on a set of studies addressing the research question. Bibliometric analysis has a different purpose than a systematic review and generally there is less burden to be comprehensive if results are representative.

Table 4. Summary information for stratified approach (Success Story 2)

	Group 1: Studies Set in LMICs	Group 2: Remaining Studies
Training data	N/A	45 seed studies identified at random for semi-supervised learning
Summary of approach	All studies screened manually	Prioritization using semi-supervised learning with an ensemble approach (2,016 studies screened manually)
Total unique citations	813 studies	6,075 studies
Total studies screened manually	813 studies	2,016 studies (33% of total search results)
Total studies excluded using ML	N/A	4,059 studies (67% of total search results)
Estimated time saved*	68 hr (1.7 work weeks)	
*Given that this project was not a systematic review and therefore did not require two independent screeners, the estimate of time saved is based on 1 min x 1 screener/study.		

Supervised Learning

Supervised learning uses different algorithms (e.g., naïve Bayes, support vector machines) than unsupervised and semi-supervised learning and requires a relatively large training dataset. Frunza et al. (2011) noted that the cost of developing a training dataset is not trivial and may outweigh the benefit of using a ML-based approach to save time.

In practice, HSL librarians use supervised learning on search results as part of a two-phase process. The output of supervised learning is the probability of relevance for each study in the unclassified corpus. The probability can be used to set the order of screening (see Figures 4 and 5). The research team can set the recall threshold (typically 90%–95%) and review all studies with a greater than 0.5 probability of being relevant. At HSL, as part of an "insurance step," it is typically recommended that research teams review another batch of studies beyond the articles predicted most likely to be relevant by supervised learning. The intention for this step is to locate "the cliff" when precision drops significantly and approaches zero, meaning that no (or very few) relevant studies are identified beyond this point (Panel 2 of Figures 4 and 5).

While semi-supervised learning is the preferred approach by HSL librarians, it is sometimes necessary to modify the approach, such as with a very low precision dataset. In this case, supervised learning can be used in conjunction with unsupervised learning. Success Story 3 demonstrates how unsupervised learning was used with supervised learning for a low precision dataset (Table 5).

Success Story 3: Modified Approach for Low Precision Dataset

For this project, the research team conducted a comprehensive search that returned 9,460 studies after duplicates were removed. Semi-supervised learning was attempted; however, after reviewing 1,000 studies at random the team had only identified nine seed studies, indicating very low search precision (i.e., less than 1%). The process was modified by using a two-phase approach with unsupervised learning, followed by supervised learning (Table 5). This project had the following objectives with respect to ML:

- Reduce the volume of literature requiring manual screening from the literature search.
- Develop a ML-based approach for a low precision dataset.

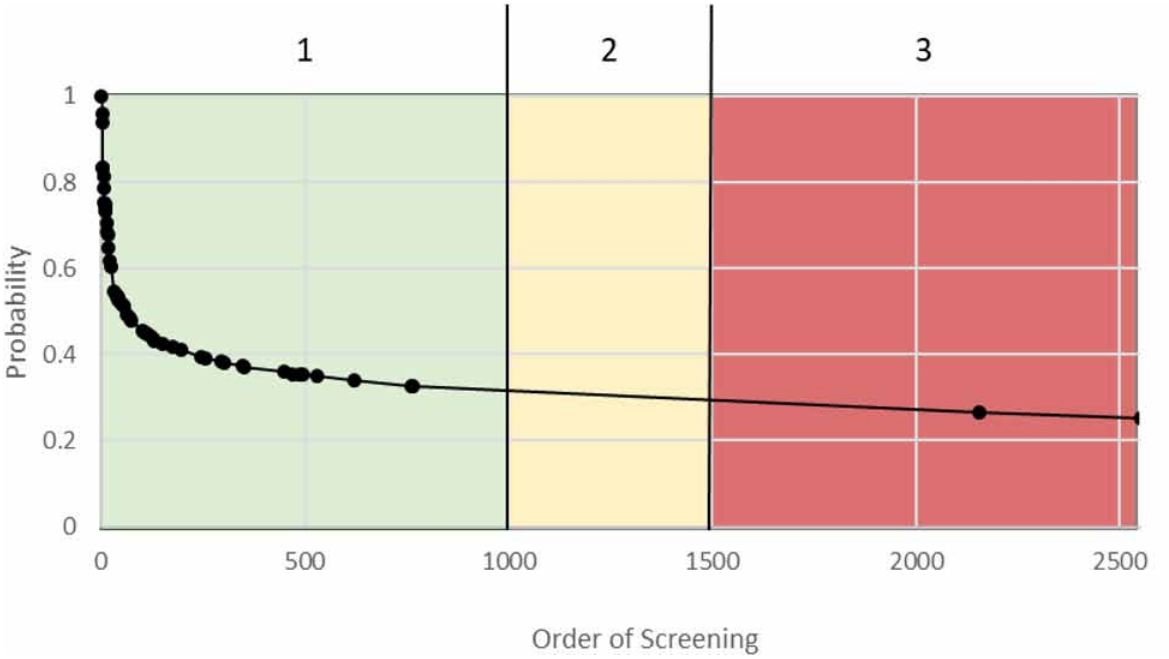
For this project *k*-means clustering was used (Phase 1: Unsupervised Learning) to identify a group of studies for manual review. Because the approach was unsupervised, no training data were needed. A cluster of studies was identified to screen based on subject matter expert review of keywords generated by the *k*-means algorithm. (Notably, all nine seed studies identified in the failed attempt to build training data for semi-supervised learning were found in this cluster.) The cluster identified as likely to be relevant contained 2,057 studies that were screened manually using title/abstract text; 183 studies were labeled relevant.

For Phase 2, supervised learning was run with training data labeled in Phase 1, including 183 relevant studies and 579 studies randomly selected from studies manually labeled nonrelevant using title/abstract text. In total, the unclassified corpus contained 7,248 studies that were prioritized using supervised learning. From the unclassified corpus, approximately 2,500 studies were screened manually for relevance based on their probability scores; 54 studies from this step were labeled relevant (Figure 4).

Table 5. Summary information for low precision data set (Success Story 3)

	Phase 1: Unsupervised Learning	Phase 2: Supervised Learning
Training data	N/A	183 studies labeled relevant from title/abstract screening of priority cluster; 579 studies randomly selected from all studies excluded during title-abstract screening
Summary of approach	Prioritization using unsupervised learning (2,057 studies from single priority cluster screened manually)	Prioritization using supervised learning (2,551 studies screened manually)
Total unique citations	9,460 studies	
Total studies screened manually	2,057 studies	2,551 studies
	Total: 4,608 studies (49% of search results)	
Total studies excluded using ML	4,852 studies (51% of search results)	
Estimated time saved*	162 hr (4 work weeks)	
*Estimate of time saved based on 1 min x 2 screeners/study (Cawley et al., 2020)		

Figure 4. Order of screening based on probability score derived from supervised learning. Black dots indicate studies identified as relevant in manual screening. Panel 1 indicates studies most likely to be relevant based on probability score and recommended for manual review. Panel 2 is the "insurance step"—the next 500 studies most likely to be relevant after the studies recommended for review by the model. Panel 3 indicates studies with low probability of being relevant that are not recommended for manual screening; however, the research team elected to screen them. The black dot in panel 3 was a study labeled "interesting" by the research team but not necessarily relevant to the project.
Source: Author, 2022.



Two-Phased Approaches

Success Story 3 demonstrated how unsupervised learning and supervised learning can be used in a two-step approach. At HSL, semi-supervised learning is often used as part of a two-phase approach in conjunction with supervised learning.

Supervised learning is typically used on results less likely to be relevant based on semi-supervised learning as follows:

- **Phase 1**: Semi-supervised learning, training data are seed studies and identified at random from unclassified corpus.
- **Phase 2**: Supervised learning, training data derived from studies reviewed manually in Phase 1.

For example, in Phase 1 the research team may manually review studies with an ES of 3 or higher. For Phase 2, the librarian will further prioritize studies with an ES = 2 or 1 using supervised learning. The training data would come from the studies manually reviewed in Phase 1. Studies with an ES = 0 can also be prioritized in Phase 2 if a more conservative approach is requested by the research team. Success Story 4 provides simulation data on the efficacy of this approach.

Success Story 4: Evidence for Two-Phase Approach

For this project, the health sciences librarian conducted a comprehensive search resulting in over 17,000 studies. Semi-supervised learning with an ensemble approach was used to prioritize studies. Of these results, the team manually screened 15,003 studies and excluded 2,061 studies automatically without manual screening. (See Christenson et al., 2021 for additional details.)

This project had the following objectives with respect to ML:

- Reduce the volume of literature requiring manual screening.
- Simulate a two-phase approach of supervised clustering + machine learning to test performance.

To simulate a two-phase approach, the 15,003 studies manually screened (i.e., labeled data) for the actual project were split into two groups as follows:

- **Group 1 (source of training data)**: 7,326 studies with higher likelihood of relevance (ensemble score = 6, 5, or 4).
- **Group 2 (unlabeled data for simulation)**: 7,677 studies with lower likelihood of relevance (ensemble score = 3, 2, or 1).

The purpose of the simulation was to test whether supervised learning could be used to eliminate additional studies in Group 2 from manual screening without missing any studies labeled relevant during full-text screening.

Training data derived from Group 1 included 387 studies labeled relevant during title-abstract screening and 1,061 studies randomly selected from all studies excluded during title-abstract screening.

The simulation of supervised learning on Group 2 studies prioritized 1,196 studies for manual review to reach 90% recall (Figure 5, Panel 1). As an insurance step, screening an additional 500 studies most

likely to be relevant by probability score (Figure 5, Panel 2) is typically recommended. In total, the simulation included 1,696 studies from Group 2 for screening. This would have automatically excluded an additional 5,949 studies beyond the 2,061 studies excluded by semi-supervised learning in Phase 1 (Table 6).

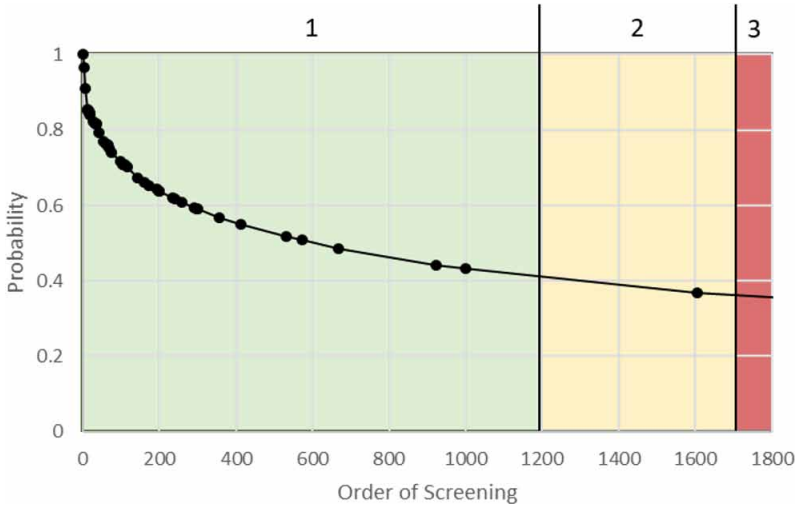
The simulation of supervised learning on Group 2 captured known relevant studies as follows:

- 94% of studies (n = 234) labeled relevant during title/abstract screening.
- 100% of studies (n = 44) labeled relevant during full-text screening.

Table 6. Summary information for simulation of two-phase approach (Success Story 4)

Component	Summary Information
Training data	387 studies labeled relevant from title/abstract screening of studies with ensemble score of 6, 5, or 4; 1,061 studies randomly selected from all studies excluded during title-abstract screening of Group 1
Summary of approach	Simulation of two-phase approach: semi-supervised learning + supervised learning
Total unique citations	17,064 studies
Total studies screened manually	9,054 studies (53% of total search results)
Total studies excluded using ML	8,010 studies (47% of total search results)
Estimated time saved*	267 hr (6.7 work weeks)
*Estimate of time saved based on 1 min x 2 screeners/study (Cawley et al., 2020)	

Figure 5. Order of screening based on probability score derived from supervised learning. Black dots indicate studies identified as relevant in manual screening. Panel 1 indicates studies most likely to be relevant based on probability score and recommended for manual review. Panel 2 is the "insurance step"—the next 500 studies most likely to be relevant after the studies recommended for review by the model. Panel 3 indicates studies with low probability of being relevant that are not screened manually. (Note that not all studies shown for Panel 3).
Source: Author, 2022.



Alternate Sources for Training Data

The time needed to locate training data can be significant and costly for a project. While semi-supervised learning requires less training data than supervised learning, seed studies are still required. Cawley et al. (2020) demonstrated that externally derived training data can be effective in locating relevant articles in a separate unclassified corpus. Success Story 5 describes how training data were derived from an initial search and applied to the literature search update.

Success Story 5: Externally Derived Training Data for Comprehensive Search Update

For this project, the research team conducted an initial comprehensive search in 2017 for a systematic review and manually screened all search results. Prior to publishing the manuscript in 2020, the team needed to conduct a search update, which returned more than 9,000 unique results after duplicates were removed. This project had the following objectives with respect to ML:

- Reduce the volume of literature requiring manual screening from the literature search update.
- Use externally derived training data from initial search results that were manually screened.

Externally derived training data were used to prioritize studies in a two-phase approach with semi-supervised learning with ensemble approach plus supervised learning (Table 7). Of the 2,241 studies screened manually from the search update, only four were included in the final analysis. (See Anderson et al., 2021 for additional details.)

Table 7. Summary information for externally derived training data (Success Story 5)

	Phase 1: Semi-supervised Learning With Ensemble Approach	**Phase 2: Supervised Learning**
Training data	154 studies labeled relevant from initial search	57 studies labeled relevant from title/abstract screening of Phase 1; 171 studies randomly selected from all studies excluded during title-abstract screening in Phase 1
Summary of approach	Prioritization using semi-supervised learning with an ensemble approach (880 studies screened manually)	Prioritization using supervised learning (1,541 studies screened manually)
Total unique citations	9,667 studies	
Total studies screened manually	2,241 studies (23% of search update)	
Total studies excluded using ML	7,426 (77% of search update)	
Estimated time saved*	247 hr (6.2 work weeks)	
*Estimate of time saved based on 1 min x 2 screeners/study (Cawley et al., 2020)		

Other Automation Applications for Bibliographic Data

At HSL, librarians also use a simple, non-ML process to analyze keywords in draft search strategies and initial search results. This approach matches a list of keywords against titles and abstracts in a set of search results. The tool is not commercially or publicly available at this time; however, the code was developed using Python and would be relatively simple for individuals with some programming experience to replicate.

The output indicates the percent and number of records that contain each keyword (Table 8, columns 2 and 3) and can be used as a diagnostic tool to provide quick information on which keywords are driving the search results. The tool also gives a count of the number of records where only the given keyword is contained in the record (Table 8, column 4), meaning the record was only returned due to a single keyword and no other keywords (i.e., synonyms connected with ORs) co-occur in that record.

The latter is especially useful when making decisions around removing particular keywords in an effort to reduce the number of search results. Research librarians developing complicated search strategies with many keywords often do this analysis based on experience. For example, librarians may run a search with a set of terms and then may advise removing a particular keyword because it *might* be causing a high rate of false positives. While the tool does not confirm if a record is truly a false positive, it does provide some indication to that effect.

This simple tool provides quick, data-driven insights into the search strategy and results to support decision making by the librarian and research team. It is also a good example of a simple, low-cost solution that improves how librarians do their work.

Table 8. Sample output from keyword analysis. In this example, 168 keywords were run against 9,686 records. (Top 10 keywords are shown.) Columns 2 and 3 indicate percent and number of documents, respectively. Column 4 indicates number of records where only that keyword appears. For example, the term "intensive care" appears in 2,598 documents and of these 2,174 records only contain the term "intensive care" and none of the other 167 keywords from the set analyzed.

1. Keyword	2. Percentage Occurrence	3. Number of Documents With Keyword	4. Number of Unique Documents With Keyword
intensive care	26.82	2598	2174
clinical setting	9.45	915	867
medical center	6.1	591	509
teaching hospital	5.44	527	394
operating room	5.09	493	440
general hospital	3.24	314	267
hospital environment	2.86	277	199
critical care	2.4	232	154
outpatient clinic	2.22	215	197
health facilities	2.1	203	146

OVERCOMING BARRIERS

Significant barriers around using ML-based approaches on bibliographic data include limited skills among academic librarians, lack of access to freely available or affordable resources, and hesitancy among research teams.

Skills Gap

Addressing the skills gap among librarians is a challenge for most institutions. This is partly due to time and finding training opportunities, but also due to what Burton et al. (2018) call the management gap. They argue that regressive organizational structures and other limitations (e.g., narrowly defined position descriptions) make technical professional development difficult. Further, a common refrain among librarians is that while they may learn new skills and take on new roles, they still need to perform their old job duties. Librarian-developed and -led programs like The Carpentries are a great option (The Carpentries, n.d.). At HSL, hiring a librarian with ML expertise who then led internal, informal training to build expertise among staff was a successful approach. With a recent grant from the Network of the National Library of Medicine (NNLM) grant, HSL intends to make this training more widely available to other institutions.

Ultimately, what is needed most is time on task to experiment, learn, and fail. Mani et al. (2020) proposed a tiered service framework for delivering data science-related services and partnering around instruction and research. The framework also makes specific recommendations around reskilling that may be broadly applicable including tying technical professional development to performance reviews, providing opportunities for peer-to-peer learning, creating and leveraging communities of practice, and establishing summer learning groups. Peer-to-peer learning, in particular, is greatly advanced by co-location of academic librarians with functional specialists such as staff with programming expertise.

Overcoming the skills gap barrier is complex and will take persistence at the individual, managerial, and institutional levels. However, as librarians acquire data skills there will be momentum as they learn what is possible for solving problems they have yet to encounter. This is where the magic really lies—when staff begin to understand what is technically feasible and a culture of innovation begins to flourish. An innovation mindset is a requirement for any paradigm shift including being open to ML-based approaches such as the ones described here. Librarians have had to continually reinvent their profession and attain new skills to stay relevant and defend their worth in the information age. Something new is always coming and historically librarians have been up to the challenge.

Access to Freely Available Tools

Access to resources is likely a perceived barrier to using ML-based approaches on bibliographic data. Commercial products exist but may require a costly enterprise-wide license. The Systematic Review Toolbox (SR Toolbox) is an excellent resource for identifying automation tools, including ML-based tools used to select articles (Marshall et al., 2021). Currently, the SR Toolbox references 14 software tools, many of which are free, that use ML for study selection in healthcare disciplines, including DoCTER. As the skills gap is addressed, it is likely that freely available resources such as these will become more widely used among academic librarians.

Further, as librarians acquire data skills and build partnerships with functional specialists, more low cost, locally designed solutions should be the result. The non-ML based approach to keyword analysis is a good example of how simple solutions can provide outsized impact.

Hesitancy among Research Teams

Addressing hesitancy to using ML-based approaches on bibliographic data can be challenging. There are some messages that are useful to convey when a team is considering using ML on search results to reduce the volume of literature to screen manually. For example, it has been helpful to communicate that automation-assisted literature screening using ML is not cutting-edge technology that can put them at risk of having a manuscript rejected due to untested methodology. At HSL, librarians have been able to adequately address inquiries around the methodology put forth during peer review. These methods have been documented, tested, and described in the peer-reviewed literature for several decades. The algorithms referenced in the success stories and in DoCTER are publicly available and well documented (e.g., *k*-means clustering algorithm has been in use for over 50 years).

Research teams with hesitancy most often need evidence of efficacy. This can be easily accomplished through simulations similar to what was described in Success Story 4: Evidence for Two-Phase Approach. For this project, the research team elected to screen most of the search results manually. This provided an opportunity to run a simulation and generate evidence that using a two-phase approach for this dataset would have been effective (i.e., resulted in high recall of relevant articles) while saving significantly more time. Highlighting success stories and conducting simulations to compare manual screening to automation approaches are compelling strategies for addressing hesitancy. They help to quell discomfort with non-traditional approaches and are useful to cite in methods sections to assure peer reviewers the methodology is sound.

The International Collaboration for the Automation of Systematic Reviews (ICASR) is a group of practitioners, methodologists, programmers, information specialists, and other experts with interest in improving the efficiency of systematic reviews through automation. The group holds annual meetings to share information, network, and actively engage in solutions (ICASR, n.d.). At the third annual meeting in 2017, there was discussion around how to overcome the barrier around acceptance for tools. O'Connor et al. (2019) recommend that other terms such as "machine-assisted," "computer-assisted," and "computer-supported" be used in place of machine learning to make clear that humans control and provide input into the process even when automation is used to increase efficiency.

It may also help to encourage acceptance of machine-assisted screening by challenging the deeply held belief that the "gold standard" is manual review and anything else is inferior. For example, Devlin and Chang (2018) report that BERT outperformed humans and previous state-of-the-art models. O'Connor et al. (2019) note that it is important to document the relative advantage of an innovation (in this case ML to assist with screening search results) over the current method. If automation approaches are seen as inferior to manual screening, adoption will always be an uphill battle. However, considering that human screeners can be imperfect, inconsistent, and biased, and that perhaps automation does confer a relative advantage over manual screening, innovation might be seen differently. A case can certainly be made that machine-assisted screening would be advantageous over manual screening by the large team of researchers that would be needed for large results sets (i.e., 10,000 or more articles).

Finally, it is helpful to take a soft-sell approach by offering expertise and providing a clear explanation of what to expect, including the tradeoffs. HSL librarians note that all searches can miss relevant

evidence, as no search strategy can be designed with perfect sensitivity. As a first step to gain comfort with the approaches, one possibility would be for the team to prioritize all search results to see the process in action, but not necessarily to automatically exclude any studies.

CONCLUSION

ML-driven approaches bring efficiencies to research teams and provide data-driven insights to optimize comprehensive search strategies. These approaches are especially relevant in the context of an infodemic. Further, using ML-assisted screening tools allows researchers to change the scope of research questions without being overly concerned about the number of search results.

Ultimately, ML-based approaches are another tool in the toolbox for academic librarians when conducting large-scale literature searches, including systematic reviews. They also offer an opportunity to partner with research teams in a new way that can provide significant advantages in terms of the time and resources necessary to conduct and publish scholarly research. The approaches are not new, but uptake has been slow, possibly due to limited capacity of librarians, lack of access to freely available resources, and hesitancy among research teams. Thomas et al. (2021) note that very high recall can be achieved using ML-assisted screening techniques and the approach has been sufficiently validated for real-world use. For example, Cochrane uses ML-assisted technology trained with crowdsourced data to classify randomized control trials (RCTs) and controlled clinical trials (CCTs) in the *Cochrane Register of Studies* (Cochrane Training, 2022). As more resources become available and training opportunities arise, academic librarians will be well positioned to guide research teams in the use of these ML approaches.

REFERENCES

Anderson, D. M., Cronk, R., Fejfar, D., Pak, E., Cawley, M., & Bartram, J. (2021). Safe healthcare facilities: A systematic review on the costs of establishing and maintaining environmental health in facilities in low- and middle-income countries. *International Journal of Environmental Research and Public Health*, *18*(2), 817. doi:10.3390/ijerph18020817 PMID:33477905

Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D., & Aliferis, C. F. (2005). Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association: JAMIA*, *12*(2), 207–216. doi:10.1197/jamia.M1641 PMID:15561789

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., & Macleod, M. R. (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), 23. doi:10.118613643-019-0942-7 PMID:30646959

Bekhuis, T., & Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artificial Intelligence in Medicine*, *55*(3), 197–207. doi:10.1016/j.artmed.2012.05.002 PMID:22677493

Belter, C. W. (2016). Citation analysis as a literature search method for systematic reviews. *Journal of the Association for Information Science and Technology*, *67*(11), 2766–2777. doi:10.1002/asi.23605

Burton, M., Lyon, L., Erdmann, C., & Tijerina, B. (2018). *Shifting to data savvy: The future of data science in libraries. Project report*. University of Pittsburgh.

Cawley, M., Beardslee, R., Beverly, B., Hotchkiss, A., Kirrane, E., Sams, R. II, Varghese, A., Wignall, J., & Cowden, J. (2020). Novel text analytics approach to identify relevant literature for human health risk assessments: A pilot study with health effects of in utero exposures. *Environment International*, *134*, 105228. doi:10.1016/j.envint.2019.105228 PMID:31711016

Christenson, E. C., Cronk, R., Atkinson, H., Bhatt, A., Berdiel, E., Cawley, M., Cho, C., Coleman, C. K., Harrington, C., Heilferty, K., Fejfar, D., Grant, E. J., Grigg, K., Joshi, T., Mohan, S., Pelak, G., Shu, Y., & Bartram, J. (2021). Evidence map and systematic review of disinfection efficacy on environmental surfaces in healthcare facilities. *International Journal of Environmental Research and Public Health*, *18*(21), 11100. Advance online publication. doi:10.3390/ijerph182111100 PMID:34769620

Cochrane Training. (2022). *3.6 Features of the CRS*. https://training.cochrane.org/resource/tsc-induction-mentoring-training-guide/3-cochrane-register-studies/36-features-crs

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association: JAMIA*, *13*(2), 206–219. doi:10.1197/jamia.M1929 PMID:16357352

Devlin, J., & Chang, M.-W. (2018, November 2). Open sourcing BERT: State-of-the-art pre-training for natural language processing. *Google AI Blog*. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1. 10.18653/v1/N19-1423

Frunza, O., Inkpen, D., Matwin, S., Klement, W., & O'Blenis, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, *51*(1), 17–25. doi:10.1016/j.artmed.2010.10.005 PMID:21084178

Golder, S., McIntosh, H. M., Duffy, S., & Glanville, J. (2006). Developing efficient search strategies to identify reports of adverse effects in Medline and Embase. *Health Information and Libraries Journal*, *23*(1), 3–12. doi:10.1111/j.1471-1842.2006.00634.x PMID:16466494

ICF. (n.d.). *Document classification and topic extraction resource (DoCTER).* https://www.icf.com/technology/docter

International Collaboration for the Automation of Systematic Reviews (ICASR). (n.d.). https://icasr.github.io

Mani, N., Cawley, M., Henley, A., Triumph, T., Williams, J., Jansen, M., Dodd, A., Casden, J., Mc Keehan, M., Venlet, J., Bruckner, L., Morris, S., & Mc Garty, J. (2020). *University Libraries: Data Science Framework*. UNC., doi:10.17615/5c78-3t83

Marshall, C., Sutton, A., O'Keefe, H., & Johnson, E. (Eds.). (2021). *The systematic review toolbox*. http://www.systematicreviewtools.com/

Mostafa, J., & Lam, W. (2000). Automatic classification using supervised learning in a medical document filtering application. *Information Processing & Management*, *36*(3), 415–444. doi:10.1016/S0306-4573(99)00033-3

National Coordination Office for Networking and Information Technology Research and Development (NITRD/NCO). (2016). *The Federal big data research and development strategic plan*. Available from: https://www.nitrd.gov/pubs/bigdatardstrategicplan.pdf

O'Connor, A. M., Tsafnat, G., Gilbert, S. B., Thayer, K. A., Shemilt, I., Thomas, J., Glasziou, P., & Wolfe, M. S. (2019). Still moving toward automation of the systematic review process: A summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *8*(1), 57. doi:10.118613643-019-0975-y PMID:30786933

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, *4*(1), 5. doi:10.1186/2046-4053-4-5 PMID:25588314

SAS Institute. (n.d.) *Machine learning: What it is and why it matters*. https://www.sas.com/en_in/insights/analytics/machine-learning.html

Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., & Thomas, J. (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), 31–49. doi:10.1002/jrsm.1093 PMID:26054024

The Carpentries. (n.d.). https://carpentries.org/

Thomas, J., McDonald, S., Noel-Storr, A., Shemilt, I., Elliott, J., Mavergames, C., & Marshall, I. J. (2021). Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *Journal of Clinical Epidemiology*, *133*, 140–151. doi:10.1016/j.jclinepi.2020.11.003 PMID:33171275

Varghese, A., Cawley, M., & Hong, T. (2018). Supervised clustering for automated document classification and prioritization: A case study using toxicological abstracts. *Environment Systems & Decisions*, *38*(3), 398–414. doi:10.100710669-017-9670-5

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, *11*(1), 55. doi:10.1186/1471-2105-11-55 PMID:20102628

World Health Organization (WHO). (n.d.). *Infodemic*. World Health Organization. https://www.who.int/health-topics/infodemic#tab=tab_1

ADDITIONAL READING

Bishop, C. (2006). *Pattern recognition and machine learning* (Vol. 1). Springer.

Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT Press. doi:10.7551/mitpress/9780262033589.001.0001

Cohen, A. M., Ambert, K., & McDonagh, M. (2012). Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making*, *12*(1), 33. doi:10.1186/1472-6947-12-33 PMID:22515596

Marshall, I. J., Noel-Storr, A., Kuiper, J., Thomas, J., & Wallace, B. C. (2018). Machine learning for identifying randomized controlled trials: An evaluation and practitioner's guide. *Research Synthesis Methods*, *9*(4), 602–614. doi:10.1002/jrsm.1287 PMID:29314757

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8*(1), 163. doi:10.118613643-019-1074-9 PMID:31296265

Shekelle, P. G., Shetty, K., Newberry, S., Maglione, M., & Motala, A. (2017). Machine learning versus standard techniques for updating searches for systematic reviews: A diagnostic accuracy study. *Annals of Internal Medicine*, *167*(3), 213–215. doi:10.7326/L17-0124 PMID:28605762

Varghese, A., Agyeman-Badu, G., & Cawley, M. (2020). Deep learning in automated text classification: A case study using toxicological abstracts. *Environment Systems & Decisions*, *40*(4), 465–479. doi:10.100710669-020-09763-2

Varghese, A., Hong, T., Hunter, C., Agyeman-Badu, G., & Cawley, M. (2019). Active learning in automated text classification: A case study exploring bias in predicted model performance metrics. *Environment Systems & Decisions*, *39*(3), 269–280. doi:10.100710669-019-09717-3

KEY TERMS AND DEFINITIONS

Artificial Intelligence (AI): A computer system able to perform tasks that normally require human intelligence.

Labeled Data: Data that have been classified manually by one or more people. Labeled data are used to make predictions on unlabeled data with machine learning models.

Machine Learning: A branch of AI that is generally defined as using machines to complete human tasks; often used interchangeably with the term AI.

Precision: A measure of how many relevant records were found in a search relative to false positives. High precision searches have a higher percentage of relevant results.

Recall: A measure of how many relevant results were captured out of all possible relevant results. High recall is necessary in comprehensive searches such as those conducted for systematic reviews.

Semi-Supervised Learning: Type of machine learning that uses a small amount of labeled data to train the model.

Supervised Learning: Type of machine learning that uses a large set of training data (i.e., labeled data) to build the model. Naïve Bayes, *k*-nearest neighbors, and support vector machines are examples of supervised learning algorithms. Generally referred to as machine learning.

Training Data: Labeled data that are used to build a machine learning model for making predictions on unlabeled data.

Unsupervised Learning: Type of machine learning, such as clustering, that does not require training data; *k*-means, nonnegative matrix factorization, and latent Dirichlet allocation are examples of clustering algorithms.