Computers in Human Behavior 156 (2024) 108215

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Using machine learning for continuous updating of meta-analysis in educational context

Olga Chernikova [a,*], Matthias Stadler [b], Ivan Melev [c], Frank Fischer [a]

[a] Ludwig-Maximilians-Universität in Munich, Germany
[b] Institute of Medical Education, University Hospital, Ludwig-Maximilians-Universität in Munich, Germany
[c] Technical University of Munich, Germany

ARTICLE INFO

Handling editor: Nicolae Nistor

Keywords:
Machine learning
Abstract screening
Systematic literature review
Meta-analysis

ABSTRACT

Machine learning and learning analytics are powerful tools that not only support researchers in the detailed measurement and enhancement of learning processes in various learning environments, but also enable the aggregation and synthesis of evidence regarding effective educational practices. This paper describes the development and application of machine learning algorithms aimed at semi-automatic selection of abstracts for a meta-analysis on the effects of simulation-based learning in higher education. The goal was to reduce the workload while also maintaining the transparency and objectivity of the selection process. The algorithms were trained, validated, and tested on a set of 3187 studies on simulation-based learning found in medical and educational databases collected before April 2018. Subsequently, they were utilized to classify abstracts for a follow-up meta-analysis consisting of 2373 studies (published between 2018 and 2020). The aim of training the algorithms was to predict studies' abstract eligibility based on words and combinations of words used in these abstracts. The application of the algorithms reduced the number of studies that had to be manually screened from 2373 to 711. A total of 458 studies from automatically selected abstracts were included in the full-text screening, indicating the high precision of the algorithms (also compared to the performance of human raters). We conclude that machine learning algorithms can be trained and used to classify abstracts for their eligibility, significantly reducing the workload for the researchers without diminishing objectivity and quality when updating systematic literature reviews with or without a meta-analysis.

Funding

This research was funded by a grant from the German Research Community (Deutsche Forschungsgemeinschaft (DFG FOR2385; FI792/1)).

1. Problem statement

Systematic literature reviews with and without a meta-analysis that aggregate and systematize evidence from empirical research are essential for advancing research as well as for theory development and informing practical decisions. This is crucial in the field of education and educational psychology to keep up with rapid developments in educational technologies and to facilitate fair, inclusive, and high-quality education worldwide.

One of the biggest challenges in performing systematic literature reviews, with or without a meta-analysis across different contexts, is the time and resources needed to ensure quality standards. A group of researchers (Borah et al., 2017) estimated that the average amount of time needed to complete a meta-analysis is 67.3 weeks, and a large part of that time is spent on the manual selection of eligible studies, which in turn might be associated with errors, biases, and increased costs. Furthermore, the increased quantity of publications in recent years (e.g., Ware & Mabe, 2015) creates a range of complications for systematizing research, including research on education. Among these issues are appropriate location and selection of studies addressing the specific research question.

There is a range of different tools available to support researchers with proper documentation, screening, and evaluation of studies for meta-analyses (e.g., ASReview LAB developers, 2023; Kebede et al., 2023). Some researchers have successfully applied and reported on machine learning tools and algorithms (e.g., using R (R Core Team,

* Corresponding author. Chair of Education and Educational Psychology, Department Psychology, Ludwig-Maximilians-Universität in Munich, Germany.
E-mail address: o.chernikova@psy.lmu.de (O. Chernikova).

https://doi.org/10.1016/j.chb.2024.108215
Received 18 December 2023; Received in revised form 7 March 2024; Accepted 19 March 2024
Available online 26 March 2024

2020) or Python (Python Software Foundation, 2024)) to support systematic literature reviews with and without a meta-analysis in different fields (e.g., van de Schoot et al., 2021; Xiong et al., 2018 in medicine; or Banach-Brown et al., 2019 in animal studies). Yet, although many tools exist, including those enhanced with machine learning, generative artificial intelligence and natural language processing algorithms, there is lack of information on their performance and use. This in turn might lead to a perceived difficulty (e.g., Chai et al., 2021) and unsystematic use of such tools in the fields of education and educational psychology. Moreover, these tools are still evolving, and their performance may differ across domains and topics (e.g., Burgard & Bittermann, 2023; Chai et al., 2021; van de Schoot et al., 2021).

2. Theoretical background

2.1. Synthetizing research: systematic literature reviews with or without a meta-analysis

The domain of education relies on a range of different theories, and different approaches and strategies are utilized to promote effective learning across contexts. Maintaining a broad perspective, synthesizing research findings across different contexts, and mapping the effectiveness of teaching and learning technologies and strategies are all essential to informing practitioners and policymakers about possible implementation (e.g., Taylor & Hedges, 2023).

At its core, a systematic literature review, with or without a meta-analysis, is a process of aggregating results from a collection of studies. The main difference between reviews and primary empirical studies is the different unit of analysis: from the individual subject to the studies themselves (Borenstein et al., 2009). In other words, instead of performing statistical analysis on a sample of subjects, the analysis is done on a sample of studies. Therefore, the implications of the findings of a systematic literature review with or without a meta-analysis are considered more reliable, as they summarize a collection of studies. This makes it possible to draw more objective and robust conclusions and contribute to the further development, criticism, and improvement of existing theories as well as computational and conceptual models.

The main motivation for conducting meta-analytical studies in addition to a systematic literature review is a variance in the primary studies that needs to be quantified and explained. Meta-analysis is a powerful tool that helps to reduce potential possible methodological noise (e.g., Xiong et al., 2018). For example, research in educational contexts often uses convenience samples, which are relatively small, to statistically estimate the effect of the parameters in primary studies and thus might yield inconsistent results. In a meta-analysis, the effect size for each study is computed and combined with other effect sizes (Borenstein & Higgins, 2013). This makes it possible to determine whether the effects are consistent across studies and ensures that each study might be seen as control for other studies to minimize the potential impact of confounding variables.

2.2. Data collection for systematic reviews: title and abstract screening

Just as adequate sampling is critical in primary studies, one of the most essential steps in systematic literature reviews with or without a meta-analysis is selecting the appropriate studies to synthetize evidence to answer the target research question. Considering the immense effort it would require to read all studies in their entirety, title and abstract screening is a critical component of the systematic literature review process (Chai et al., 2021). There are many different traditions and approaches in the educational field, and this is also reflected in titles and abstracts. This includes different reporting standards (e.g., mentioning the results of the study vs. only mentioning problem statement in the abstract) and different attitudes toward and conceptualizations of (as well as publisher recommendations) what information an abstract should represent and include (e.g., different content and structure of abstracts for teacher or medical education, business, psychology). This makes study selection a challenging task, which depends on a range of researcher decisions (e.g., selecting a broad vs. narrow research focus, justifying eligibility criteria). Thus, ensuring objectivity, transparency, and replicability of the selection process are important challenges to maintain the quality standards for a meta-analysis (Page et al., 2021).

2.3. Machine learning algorithms for abstract screening

The use of artificial intelligence, machine learning, and natural language processing methods for semi-automated systematic literature reviews with or without a meta-analysis constitutes an independent research field that spans across various domains (e.g., Cierco Jimenez et al., 2022; Marshall & Wallace, 2019). These methods might support researchers in text classification (e.g., title and abstract screening, coding moderators) and data extraction (e.g., identifying frequent topics, moderators or statistical data). Text classification enables sorting manuscripts or chunks of text (e.g., abstracts) into predefined categories of interest (e.g., empirical vs. not). Some examples of the tools that support researchers in text classifications include ASReview (ASReview LAB developers, 2023) and Abstrakr (Wallace et al., 2012), which use machine learning (ML) algorithms (see van de Schoot et al., 2021 for a more detailed review of these tools). Data-extraction models (e.g., RAKE package in Python; see Rose et al., 2010) can identify text elements (some words or their combinations; particular numbers) that correspond to a variable of interest, such as the number of people in experimental or control conditions or the use of a type of scaffolding or software.

Yet, although quite a few tools, packages, and algorithms exist, there is lack of information regarding their performance across different fields of study and different types of data (e.g., empirical vs. conceptual research articles). Some tools are evolving, and their performance may differ across and within domains (e.g., Burgard & Bittermann, 2023; Chai et al., 2021; van de Schoot et al., 2021). This study does not aim at performing a comprehensive review of all of the existing possibilities. Rather, it seeks to contribute to the existing tools and approaches by describing procedures we developed for updating a meta-analysis in educational context.

2.4. Screening abstracts as a ML classification task

ML algorithms can accelerate abstract screening for a meta-analysis by using statistical and data-driven approaches to learn patterns and structures from text data to classify text based on its relevance. They can then prioritize the relevant abstracts, reducing the number of studies that require human review (van de Schoot et al., 2021), or even classify the studies if trained on a sufficient amount of pre-screened studies. To enable this performance on classification tasks, ML can be trained in supervised or semi-supervised manner if the data is fully labeled, or some labeling exists, respectively. Alternatively, one can also use unsupervised learning in order to split the data in multiple clusters, which in this case will be eligible or non-eligible studies for a meta-analysis (e.g., Mankolli & Guliashki, 2020).

Below, we describe a few important steps that should be considered when planning the use of ML tools for an abstract classification task. We further describe the decisions made for this study in the method section.

Feature extraction. The first essential step in training ML algorithms to identify text patterns is transforming raw text data into meaningful and representative variables that can be used as input for ML algorithms (Sammons et al., 2016). Effective feature extraction is essential for classification to capture relevant information from text data and can be done with different methodologies, such as the bag-of-words technique (Kwartler, 2017). A bag-of-words model (BoW) is a way of extracting features from text. It describes the occurrence of words within a document, whereas information about the order or structure of words in the document is discarded. The model is only concerned with whether and how often particular words occur in the document, and each word count

is considered a feature (Kwartler, 2017).

Misclassification. Due to the nature of the literature search process, there might be many irrelevant studies in the search results, which were not possible to exclude using the search terms. For example, the bench meta-analysis (Chernikova et al., 2020) used to train the algorithms in this study indicated that only 7.3% of the initial hits were relevant for the analysis. This can lead to two highly imbalanced classes of eligible and non-eligible studies. The algorithms can misclassify the eligible studies, resulting in a failure to consider relevant studies for analysis and therefore having a biased picture of the empirical evidence. To address the problem of imbalanced classes, it is important to find ways to penalize the misclassification of the relevant class as opposed to penalizing the misclassification of the irrelevant class (Chawla, 2005). In other words, selecting some false positive studies is less problematic than failing to identify relevant studies (as irrelevant studies can be removed later in the analysis).

Determining performance metrics. Another important step is to define what would constitute an indicator of good performance for the ML algorithm. Different performance metrics exist and can be selected for different tasks (Foody, 2023). There are many metrics that can be used to evaluate a classification model, and different metrics can lead to different conclusions regarding the model's performance. However, despite the large number of performance measures that are available, there are four that are most commonly used (James et al., 2017): accuracy, recall, precision, and the F score. All of them are scored between 0 and 1 (Foody, 2023).

1. *Accuracy* is calculated as the ratio of the sum of true positives and true negatives to the sum of all classified cases: true positives, true negatives, false positives, and false negatives. However, this measure has a major disadvantage when used with highly imbalanced datasets: The accuracy of the model will be high, even though it might not be a particularly useful algorithm (Foody, 2023).

For example, our training dataset contained 7% of the data belonging to the eligible studies and 93% belonging to non-eligible studies. A model that classifies all of the data as non-eligible would achieve an accuracy of 93%, which is a very high accuracy by all standards. However, it would be a useless model, as we are interested in a model that identifies the eligible studies.

2. *Recall* is the ratio between the true positives (TP) and the sum of the true positives (TP) and false negatives (FN): TP/TP + FN. Recall can be seen as a measure of how many of the cases from a set of eligible studies have been identified compared to how many have been missed. Thus, recall is a measure of identifying relevant information (Foody, 2023). Continuing the classification example, if the algorithm classifies all of the studies as irrelevant, because it has not identified any of the studies that are of interest, its recall is zero. This is a good measure for imbalanced data because it allows gaining information about the relevant class as opposed to a measure that provides information about both classes, one of which might not be of interest at all.
3. *Precision* is formally defined as the ratio between the true positives (TP) and the sum of the true positives (TP) and false positives (FP): TP/TP + FP (Foody, 2023). This measure can be interpreted as one that gives information about how many of the studies that have been identified as eligible are in fact eligible and how many of them are non-eligible studies that the algorithm has incorrectly classified. As such, this is also a useful measure for imbalanced data.
4. *F score* is a function of the precision and recall measures that were previously discussed (Foody, 2023). It is calculated as the doubled product of precision and recall divided by their sum. This measure is useful as a global estimate measure of the algorithm's performance, as it considers not just the amount of relevant studies successfully identified but also the identification process.

If we reverse the scenario, and all studies are identified as eligible, even though that amounts to only 7%, it will lead to a recall of 100%. However, the precision will be close to zero, as all of the non-eligible studies have not been classified as such by the algorithm. Thus, it is a particularly inefficient algorithm, even though it successfully identifies all of the eligible studies.

One of the disadvantages of all of these measures is that there is no objective comparative reference; rather, they are measures that compare models against each other. As such, it is necessary to set an arbitrary threshold that is considered to represent acceptable performance for the future application of the algorithm. The performance of human raters can be set as such a threshold to ensure the performance of ML algorithms is not worse than that of human raters.

Overfitting vs. Generalizability. Another issue to consider when training and using ML algorithms is overfitting. Overfitting refers to the tendency of algorithms to learn from and adapt to sample-specific error variance, which results in reduced generalizability (e.g., Yarkoni & Westfall, 2017). The reasons why overfitting happens are technical, but it can be interpreted as the model learning the particularities of the training dataset "by heart" instead of extracting the relevant features of the data and generalizing it to yet unseen data (James et al., 2017).

2.5. Updating existing systematic literature reviews: keeping up with increasing publication rates

Systematic literature reviews with or without a meta-analysis usually take a long time to perform (e.g., Borah, et al., 2017; Smith et al., 2011) – and even longer for the results to get published. Some new studies might appear as the meta-analysis is finished and offer new evidence, which had been missing. This is particularly true in rapidly developing fields in education (e.g., research on simulation-based learning, artificial intelligence), as new technologies and approaches emerge. The aggregated evidence becomes outdated relatively quickly, and (continuous) updates are needed to inform practical decisions as well as to further develop theory. Although the screening process for a completely new meta-analysis or a follow-up have much in common, screening and classifying abstracts to update systematic literature reviews with or without a meta-analysis have some specific features, which offer an opportunity for more automatization and speeding up the screening process.

First, most conceptual preliminary steps, such as formulating search strings or eligibility criteria, are already performed. The search can be limited to studies that appeared in a limited period of time (e.g., studies published after the date of last search). Second, a great deal of labeled or coded data already exists from previously conducting the meta-analysis (e.g., manually coded abstracts, documentation on decisions made about eligibility). Although documentation and protocols of the procedures remain available, some additional time might still be needed to train new research assistants to classify the abstracts, as the trained raters may no longer be available. Training human raters can be resource demanding and subject to availability constraints. Meanwhile, due to the large volume of available training data, ML algorithms can be trained and reused multiple times, allowing the same solutions to the same tasks and facilitating objective, reproducible, and transparent classification. Furthermore, for ML algorithms, there is little difference between classifying 100 or 1000 abstracts in terms of time resources, and the algorithms are not affected by fatigue and a related increase in error rates (e.g., Xiong et al., 2018).

3. Goals of this study

The aim of this study was to train and test a few commonly used supervised ML algorithms to support updating meta-analytic findings in an educational context by supporting abstract screening and classifying abstracts as eligible vs. not eligible, based on existing manual ratings. Specifically, the ML algorithms were used to methodologically enhance

a follow-up meta-analysis on the effects of simulation-based learning in higher education and decrease the time needed in the screening phase while ensuring objectivity and transparency. Through this study, we aim to provide evidence on the performance of ML tools, encouraging researchers to invest time and resources in ML algorithms to support updating meta-analytic findings in an educational context and further enrich the evidence base. We also believe that using ML algorithms for systematic research in educational contexts contributes to theory development and might inform practical decisions (see Sailer et al, 2024; this issue). We used four different ML algorithms, which represent different ways of dealing with the classification task and imply different levels of intrinsic bias, different levels of dependence on the data (e.g., including all features or examples vs. only some of them), as well as different model interpretability (e.g., James et al., 2017). We included *regularized logistic regression*, which is one of the most common models for classification tasks. It represents a parametric model, which formulates the decision as probabilities belonging to one class or another (eligible vs. non-eligible) and is relatively easy to interpret. We also included the *support vector machine* classifier, which works similarly to logistic regression, but unlike the former has a better ability to generalize unseen data. The *random forest* model does not require knowledge of the shape of data distribution, which is required by parametric models. It uses only some features and examples for each of the decision trees, which means that influential cases and features do not determine its performance. Further, the *feedforward neural network* model was chosen due to its ability to fit complex models (with different types of relationship between the features, including non-linear) and achieve high accuracy.

We discuss the details of each of the algorithms, including their strengths, challenges, and disadvantages, in the following section. Our main goal was to find the most reliable algorithm for our follow-up meta-analysis, which can effectively generalize from the training dataset to unseen data and thus reduce the workload for human raters. This in turn could contribute to enabling the continuous updating of systematic literature reviews with or without a meta-analysis.

4. Method

The training, validation, and initial testing of the algorithms was based on a total of 3187 studies that had been manually labeled as eligible or non-eligible for the purpose of conducting a meta-analysis on the effects of simulation-based learning in higher education (Chernikova et al., 2020). These studies were labeled by trained raters. The final agreement reached 100%, and the interrater agreement was estimated as *kappa* = 0.90. Out of the 3187 studies, only 235 (7.3%) were classified as eligible, indicating a highly imbalanced data set. The 3187 studies were randomly split into training (80%), validation (10%), and test (10%) datasets for the analysis. The test dataset consisted of 319 abstracts, of which 25 were previously coded as eligible. The codes used for the analysis can be found online (https://osf.io/5z6n2/?view_only=71372cf2ac2f40fea75f89511fdb39cc).

A follow-up meta-analysis (Chernikova et al., 2023) was conducted to include new published studies. The follow-up included 2373 studies (published between 2018 and 2020) identified for title and abstract screening. The search string, included databases, and eligibility criteria were identical to those used for the initial meta-analysis (see Chernikova et al., 2020). The growing numbers support the claim about the increased amount of publications and the need to use ML algorithms to manage the challenge of keeping the summary of research evidence up to date.

4.1. Pre-processing and feature extraction

Abstracts of empirical studies in educational research journals are very diverse, ranging from 75 to 350 words and utilizing different structures and standards of reporting. The goal of the feature-extraction process was to find a way to mine for the valuable, quantitative aspects of the abstracts of the texts to enable working with the data using statistical techniques or ML algorithms to make predictions about the eligibility of these studies. The extraction of quantitative aspects of the raw data was performed in several steps. First, the punctuation and numerals in the text were removed, and all words were standardized to a lowercase format. Second, words that contained the same word stem were reduced to their word stem and considered the same. Third, words that only contributed to the sentence structure (e.g., articles, connector words, prepositions) were removed. In this way, the fictional example sentence from the abstract, "This study used simulated scenarios to facilitate development of diagnostic skills in pre-service teachers," was transformed into "study use simulate scenario facilitate development diagnose skill pre-service teacher."

The extraction of the features was achieved using a bag-of-words extraction approach. A matrix containing each of the primary studies as rows was created, with the columns representing the extracted variables (individual words in this study). For instance, if "simulation" appeared 10 times in the abstract of a particular study, the matrix cell of the intersection between the row that represents that study and the column that represents the word "simulation" would contain the value 10. The training data resulted in the identification of 17,060 features. To reduce the amount of features of the training data to meaningful amount, a lower (10) and upper frequency boundary (1000) was set to focus on the words that distinguished abstracts from one another. In other words, if a word was frequently used across all abstracts in the training data (e.g., "result") or only used in very few abstracts (e.g., name of particular simulation or a tool), the word (feature) was disregarded in the further analysis. This restriction made it possible to reduce the initial feature count to 2519 features. The amount of features was further reduced using linear association measures with the class-belonging variable (i.e., eligible vs. non-eligible). We used the Pearson correlation as a measure of linear association, setting the lower and upper boundaries of the strength to be one standard deviation above or below the mean correlation coefficient, respectively. This led to another reduction in the number of features from 2519 to 528. This step was only performed for the training dataset to prevent possible data leakage. Stricter rules could be set to further reduce the amount of features, but we were satisfied with the level of reduction, and so stricter rules were not applied.

4.2. Algorithms used in the study

To achieve the goals of this study and address the features of the data (e.g., highly imbalanced dataset), we trained and applied four different classification algorithms to select eligible abstracts: 1) logistic regression with least absolute shrinkage and selection operator (LASSO) regularization, 2) support vector machine classifier, 3) random forest classifier, and 4) feedforward neural network. The choice of algorithms was based on the attempt to train classifiers with different levels of intrinsic bias, different levels of dependence on the data, as well as different level of model interpretability (see James et al., 2017). All these classification algorithms have their pros and cons. For example, the logistic regression classifier can be easier to interpret (e.g., James et al., 2017), but it might perform worse than the other algorithms, while the opposite is true for feedforward neural networks (Goodfellow et al., 2017).

4.2.1. Logistic regression with LASSO regularization

Logistic regression (James et al., 2017, Wright, 1995) is one of the oldest and best-understood statistical modeling approaches for classification. The technique itself is an extension of the linear regression technique and belongs to the family of generalized linear models. Logistic regression has a linear component, which is the decision boundary between two classes, and produces outputs as probabilities belonging to one class or the other (James et al., 2017). One of the advantages of this method is getting the same solution when the same data are used.

Furthermore, logistic regression can predict probabilities instead of an exact class value (e.g., a logistic regression predicts values like 0.7 belonging to class 1 instead of just giving class 1 as the output). Disadvantages of the model include an inclination to overfitting, which means finding the optimal solution for the training dataset, which is not generalizable to unseen data (e.g., James et al., 2017). To overcome overfitting, regularization techniques are recommended (Goodfellow et al., 2017; Tibshirani, 1996). Logistic regression with LASSO regularization is considered beneficial in situations where a sparse solution is desired, as it automatically selects important features and ignores less relevant ones, thus potentially improving the model's generalization to new data and preventing overfitting (e.g., James et al., 2017).

4.2.2. Support vector machine classifier

A support vector machine (SVM) classifier is a powerful supervised ML algorithm used for classification (Bishop, 2006). It aims to find a solution that best separates different classes of the given dataset. SVM can be seen as an extension of logistic regression that is used in cases of perfectly separable classes (James et al., 2017). Although such classes are rarely found in reality, the logic behind perfectly separable classes still holds with a small adjustment, which allows for some level of misclassification. In other words, by specifying the level of misclassification, the SVM optimization is solved for the decision boundary, with the largest distance between the two classes taking the misclassification into account. Additionally, unlike logistic regression, SVM is not a probabilistic algorithm, and thus it is cannot assign levels of belonging to a particular class. However, it has the ability to generalize well on unseen data, which is an important feature for abstract classification (James et al., 2017).

4.2.3. Random forest classifier

The previously two models belong to the class of parametric models. They assume a particular shape of the distribution and a particular shape of the curve that is being fit to the data, namely, a linear decision boundary. However, often it is not possible to assume the shape of the curve that is being fit to the data. One possible approach in this case is a decision tree, which uses simple Boolean (i.e., binary logic) statements to determine the classification output. One particular case, random forest (RF), allows building a model that has a high number of trees. This leads to different final answers but, on average, improves the model's performance compared to other models that produce just one answer for each input (Breiman, 2001). The RF model does not use all of the examples and features for each of the trees, which means that influential cases and features do not determine the performance of the ensemble, although they do determine the behavior of individual decision trees (Goodfellow et al., 2017; James et al., 2017). This means that an RF model is not strongly dependent on any specific examples or features. Thus, reproducibility and interpretability might be problematic (James et al., 2017). Nevertheless, the model can generalize well on unseen data.

In this study, in order to deal with the problem of imbalanced classes, the trained RF model was prompted to use all of the examples from the eligible class—and twice that number from the non-eligible studies class. This resulted in a model that placed much more weight on the underrepresented class and thus imposed a greater penalty for misclassification.

4.2.4. Feedforward neural network

A feedforward neural network (FNN) might be seen as an extension of generalized linear models, using interconnected layers of neurons (extracted features) to process data through linear and non-linear transformations (James et al., 2017). As a statistical learning model, neural networks have several advantages, such as their ability to fit complex models and to achieve high accuracy. These two properties make neural network models a popular choice for tasks that are complex and require extremely high performance. However, neural network models also have certain disadvantages. The solution is not uniquely defined or guaranteed, different training processes can lead to different solutions of the parameters, the models are too complex to be interpretable (e.g., James et al., 2017), and the network architecture can sometimes be arbitrary (i.e., there are no guidelines on what type of an architecture can be used for what type of a problem).

In this study, the neural network consisted of an input layer with the same number of units as the number of features extracted using the bag-of-words technique, along with two hidden layers with a Leaky ReLU activation function (PyTorch, 2023a, 2023b), which is characterized by a small linear slope in the negative part of the number line and linear function in the positive part of the number line (Goodfellow et al., 2017). The output layer consisted of a single node with a binary logistic activation function, also known as a logistic function, which is a representation of the probability that a set of inputs belongs to one of the two classes (eligible vs. non-eligible abstracts).

Given that the data consist of two highly imbalanced classes, for the training process we employed stochastic gradient descent optimization, which uses each of the examples from the eligible class and an equal number of examples from the non-eligible class for each of the training steps. Since neural networks have a large capacity, regularization techniques were employed to avoid overfitting. In particular, weight decay regularization was utilized as well as the so-called drop-out technique, which randomly turns off neurons in each training step. With the weight decay approach, the neural network is biased to decrease the values of the weights, resulting in a smaller loss value (hence the name weight decay). Meanwhile, the dropout regularization does not allow any neuron to become too specialized for a particular feature, which in turn leads to better generalization to unseen data.

4.2.5. Human raters

A 10% test set (N = 319) was also screened by a new cohort of trained raters to compare the ML models with the performance of human raters. The human classifiers demonstrated a recall of 100%, but their precision was only 15%. This means that even though humans are able to recall all of the studies, there are many false positives, which is undesirable when seeking to saves time and effort in abstract screening. The false-positive rate suggests that humans are using a strategy whereby they allow too many studies to be considered eligible in order to prevent missing some relevant studies. Most ML models can be trained to do this by adjusting (twitching) the threshold that is used for belonging to one class or another. However, this twitching will lead to a decrease in precision, as seen with the human raters.

5. Analysis

5.1. Evaluation and performance metrics

In the present study, we selected recall, precision, and F score as performance metrics. As we were working with highly imbalanced classes, accuracy was not a useful measure because it does not indicate actual performance in terms of classifying eligible studies (which we are interested in). In order to evaluate the models, we first tested them on unseen data (10% of the dataset) and selected the best-performing algorithms to complete abstract screening for a subsequent literature search. It was also important to set realistic expectations regarding how well the algorithm could balance between precision and recall. We set the expectation for recall to 0.8 and the minimum precision level to 0.15. Our goal was to ensure that the ML algorithms did not perform worse than human raters in terms of precision. With this level of precision and recall, it is possible to extract most of eligible studies while simultaneously reducing the workload for human raters in full-text screening and subsequent coding.

5.2. Used tools and packages

To develop and apply the classification algorithms, we used two programming platforms: R (R Core Team, 2020) and Python (Python Software Foundation, 2024). The extraction of the features was done using the R package "tm" (Feinerer et al., 2008) together with base R functionalities. The tm package was also used in the preprocessing. The model building was done using different packages. Logistic regression was performed using the "glmnet" package (Simon et al., 2011), and the SVM model was trained using the "e1071" package in R (Meyer et al., 2023). Additionally, this package offers tools for cross-validation, non-linear transformations of the decision boundary, and manually setting the weighting of each of the classes (Meyer et al., 2023). The RF was set up and trained with the "randomForest" package in R (Breiman, 2001)). It offers the ability to manually set up the number of features that are going to be randomly chosen and to set restrictions for the examples to be chosen from each class. This is important because this is one particular solution for the problem of imbalanced data (Liaw & Wiener, 2002). FNN was set up using the "PyTorch" library in Python (PyTorch, 2023a, 2023b). PyTorch offers tools for setting up highly customizable FNNs, along with tools for their training (Paszke et al., 2019). The codes used for the analysis can be found online (https://osf.io/5z6n2/?view_only=71372cf2ac2f40fea75f89511fdb39cc).

6. Results

6.1. Algorithm training and testing results

The regularized logistic regression model used LASSO regularization and was trained on 80% and validated on 10% of the available data. The choice of the regularization coefficient (0.94) was based on 10-fold cross-validation. The model used different weighting of the two classes, setting the importance of the eligible class to be 13 times higher than that of the non-eligible class (due to the distribution of the two classes). The model used only linear combinations of the coefficients. For the rest 10% of data (test dataset) the model achieved a recall of 0.71, with a precision of 0.17 (F score = 0.27).

SVM was also trained on 80%, validated on 10% of the data. Since there were more than 500 features (high-dimensional space), we opted for a model with a linear decision boundary, as any of the more complex models were more likely to overfit. The value of the cost was chosen using 10-fold cross validation. The eligible class was weighted 13 times more than the non-eligible class. Since the linear SVM classifier can be seen as a type of logistic regression with better generalizability, it was expected that the model would perform better than the logistic regression. This was indeed the case. For the rest 10% of data (test dataset) the SVM classifier correctly classified 21 of the 25 relevant examples, representing a recall of 84%. At the same time, its precision was 20%. This result is an improvement compared to the logistic regression model, which had a slightly worse overall recall and precision. The F score was 0.32.

The RF model was trained on 80% of the available data, and 10% was used for the validation. The model consisted of 4500 classification trees. Each tree was built using all of the available examples from the eligible class and twice that many examples from the non-eligible class. Additionally, the relevant class was weighted 13 times more than the irrelevant class. For the rest 10% of data (test dataset) the recall of the model on the test dataset was 0.60 (15 out of 25 relevant studies), while the precision was 0.13. This model, even though it was more complex than the logistic regression and SVM models, exhibited poor performance (F score = 0.21).

FNN is a generalization of the generalized linear model with a number of linear and non-linear transformations of the original data. For the purposes of this study, we once again used only 80% of the available data, with 10% dedicated to a validation set and 10% to the test set. This model demonstrated the best performance (F score = 0.76). The recall was 0.8, while the precision was 0.72. This precision was significantly better than that of any of the other models, while the recall was slightly worse than the recall of the SVM. A frequency interpretation of these results shows that the neural network model is able to extract most of the studies (i.e., 80% of them), and of those that are extracted, 72% are likely to be truly relevant. (see Table 1 to The summary of the results for all models is presented in Table 1.

6.2. Performance on a follow-up meta-analysis

A follow-up meta-analysis was conducted to include new published studies. The follow-up included 2373 studies (published between 2018 and 2020) found in databases and identified for title and abstract screening (after removing duplicates). The search string, included databases, and eligibility criteria were identical to those used for the initial meta-analysis (see Chernikova et al., 2020).Based on the results from model training and testing, we selected SVM and RLR (which performed similarly to human raters) to classify the set of 2373 abstracts from studies published between 2018 and 2020. The algorithms agreed on 2062 (87%) abstracts (400 were labeled as eligible and 1662 as non-eligible). The other 311 (13%) abstracts were labeled as eligible by one of the two algorithms. These abstracts were reviewed by an experienced human rater, and 58 abstracts were included in full-text review. We also conducted an adequacy check to see if any eligible studies might have been labeled as "not eligible": 10% of randomly selected abstracts (N = 165) labeled as "not eligible" by both algorithms were rated by an experienced rater blind to the algorithm's decisions. No abstracts from this set were included in the full-text review, which supported the algorithm decisions. In total, 458 abstracts (19% from the initial abstracts) were subjected to full-text review. The use of algorithms saved approximately two-thirds of the time required for human raters to perform the abstract screening (based on records from the initial meta-analysis).

7. Discussion

7.1. Summary of the results

Four different models were trained, and the performance of each was evaluated, using recall, precision, and F score as performance measures. Of all the models that were trained, the FNN showed the best results, with a recall of 80% and a precision of 72%. Notably, in real-life applications involving systematic literature reviews with or without a meta-analysis, recall is (and should be) getting higher value than precision, as missing a relevant study is considered more problematic than including a study that can later be excluded by the researcher during the full-text review. However, precision results in the most time-efficient work. For the follow-up meta-analysis, we implemented two different algorithms: SVM had the highest recall (outperforming FNN) and performing similarly to the human raters in precision. RLR was selected as the one with the best interpretability and still relatively high recall and precision comparable with human raters. Furthermore, we simulated double coding for the abstract screening (i.e., letting different algorithms code the same data and assess their agreement) to estimate the practical value of the trained algorithms. The algorithms were successful

Table 1
Summary of the findings on a 10% test data set (N = 319; eligible studies = 25).

Algorithm/Metrics	RLR	SVM	RF	FNN	HR
Recall	0.71	0.84	0.6	0.8	1
Precision	0.17	0.2	0.13	0.72	0.15
F score	0.27	0.32	0.21	0.76	0.26

Note. Abbreviations stand as follows: RLR = Regularized Logistic Regression, SVM = Support Vector Machine, RF = Random Forest, FNN = Feedforward Neural Network, HR = Human Raters.

in reducing the amount of time spent on abstract screening, and they were also were accurate in assessing the eligibility of the abstracts. We believe that using multiple different algorithms might offer further benefits for ensuring quality and efficiency of the text classification.

7.2. Scope and limitations

This study aimed at tackling the problem of hyper-publication in the scientific literature by developing ML algorithmic tools that assist, and ideally automate, the process of classifying the eligibility of studies for a meta-analysis in an educational context based on their abstracts. The development, validation, and testing of the algorithms was exemplary performed on studies that had been manually coded as eligible or non-eligible for a meta-analysis on the effects of simulation-based learning in higher education and a follow up meta-analysis, aiming to update state-of-the-art research evidence by aggregating empirical primary studies.

We argue that approaches and procedures presented in this paper can be applied to different subjects within the educational context, as they share similar features, including the highly heterogeneous structure and reporting standards for abstracts, as well as the imbalanced distribution of eligible and non-eligible studies. These tools are particularly effective for updating the results of systematic literature reviews when a significant amount of primary data already exists and can be used for training.

For new systematic literature reviews with or without a meta-analysis synthesizing empirical or conceptual papers, other tools and software implementing different ML algorithms and natural language processing tools should also be considered (for a review, see Campos et al., 2023).

One of important limitations of this study is that the algorithms were trained without using any prompts or prior information (e.g., eligibility criteria). The use of prior information is a standard practice for human raters and statistical modeling using the Bayesian approach. New training approaches, such as for Bayesian neural networks, allow the use of prior information in the model training and can improve the model performance, especially if full automatization is desired.

7.3. Outlook

An alternative approach to the supervised ML algorithms described here could be active ML, unsupervised ML methods, generative artificial intelligence (AI), etc., which might offer further advancements to the performance of text classification tasks. Multiple studies using and evaluating the use of AI and prioritized screening report encouraging results (e.g., Hamel et al., 2021; O'Mara-Eves et al., 2015).

Unsupervised methods can be used for clustering and classifying data points based on similarities between them. Ideally, the method should be able to identify an optimal way to group studies on a specific topic that is aligned with the process of screening for a meta-analysis. Potentially, an unsupervised algorithm could be used on a large set of publications in a certain domain and generate multiple classes based on similarities in these studies, providing a basis for conducting multiple meta-analyses on these different classes and leading to a more efficient meta-analysis research output. However, unsupervised ML tools are not guaranteed to result in the same clusters that we aim to identify qualitatively, as extracted features can have some statistical commonalities (e.g., frequently reoccur together without being actually connected) but not linguistic ones, with the opposite scenario (having a conceptual or linguistic connection without being statistically related) also being possible (e.g., Alloghani et al., 2020). Furthermore, many studies also emphasize, that performance of unsupervised or fully automated text screening is rather poor, while semi-automated analysis is usually more reliable (e.g., Gartlehner et al., 2019; Gates et al., 2020).

Generative AI and Large Language Models (LLM). Further option to support text classification and abstract screening is offered by evolving field of LLM and generative AI tools. The recent preprint in medical education (Tran et al., 2023) discusses use of LLM and generative AI in context of medical research. These algorithms are capable of reducing the screening workload by 65%. However there are many challenges to ensure transparency and interpretability of such models. The conclusion drawn in this study is that LLM can provide highly sensitive and moderately specific recommendations for text classification during title and abstract screening in systematic reviews (Tran et al., 2023). However, AI and LLM models are capable and should rather complement, but not fully replace human assessment. No similar systematic investigations were performed in educational context so far. However, we assume that the domain-specific characteristics (e.g., reporting standards, specific terminology used, context differences across countries) might also be very important in training and evaluating performance of LLM algorithms.

Further aspects related to conducting systematic literature reviews with or without a meta-analysis could also be supported by supervised, semi-supervised, and unsupervised ML algorithms and LLM. For example, this might include the extraction of statistical data (see Ivimey-Cook et al., 2023) or moderator coding to further enhance efficiency and leverage the most up-to-date evidence to develop theory and inform practical decision-making. It is also worth noticing, that the performance of ML and AI tools can improve over time, not only within the systematic review due to training, but also as new methods and approaches might evolve. Therefore, further research can investigate the challenges with reproducibility and transparency of these models and tools and derive on recommendations on how to use them across different research fields.

7.4. Conclusion

In addressing the challenges of conducting systematic literature reviews with or without a meta-analysis in the educational context, this study demonstrates the benefits of ML algorithms in streamlining the title and abstract screening process, reducing the workload and maintaining transparency of the process. The method described in this study has the potential to speed up the process of updating systematic literature reviews with or without a meta-analysis in rapidly growing fields of research (e.g., technology-enhanced learning). By training and testing four different ML models, it was found that the FNN model outperformed other models and human raters in precision and recall, offering a promising approach to reduce the time and effort required for manual screening. The use of ML algorithms, particularly in updating existing systematic literature reviews with or without meta-analysis, not only saves considerable amount of time but also maintains consistency in the selection process. Despite these advancements, we acknowledge the necessity of human oversight in ensuring the accuracy and relevance of selected studies. Future research should explore the integration of active ML, generative AI and LLM to further enhance the efficiency and comprehensiveness of systematic literature reviews in educational context.

CRediT authorship contribution statement

Olga Chernikova: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Matthias Stadler:** Writing – original draft, Supervision, Methodology, Conceptualization. **Ivan Melev:** Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Frank Fischer:** Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In M. Berry, A. Mohamed, & B. Yap (Eds.), *Supervised and unsupervised learning for data science . Unsupervised and semi-supervised learning* (pp. 3–21). Cham: Springer. https://doi.org/10.1007/978-3-030-22475-2_1.

ASReview LAB developers. (2023). ASReview LAB – a tool for AI-assisted systematic reviews. *Zenodo* [Computer software] https://asreview.nl.

Banach-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*. https://doi.org/10.1186/s13643-019-0942-7. Article 23.

Bishop, C. M. (2006). *Pattern recognition and machine learning. New York, NY*.

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open, 7*(2), Article e012545. https://doi.org/10.1136/bmjopen-2016-012545.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science, 14*(2), 134–143. https://doi.org/10.1007/s11121-013-0377-7

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1/topics/randomForest.

Burgard, T., & Bittermann, A. (2023). Reducing literature screening workload with machine learning. *Zeitschrift für Psychologie, 231*(1), 3–15. https://doi.org/10.1027/2151-2604/a000509

Campos, D. G., Fütterer, T., Gfrörer, T., Lavelle-Hill, R., Murayama, K., König, L., Hecht, M., Zitzmann, S., & Scherer, R. (2023). Screening smarter, not harder: A comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. *Preprint*. https://doi.org/10.31234/osf.io/fpwc2

Chai, K. E. K., Lines, R. L. J., Gucciardi, D. F., & Ng, L. (2021). Research screener: A machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews, 10*(1), 93. https://doi.org/10.1186/s13643-021-01635-3

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook*. Springer. https://doi.org/10.1007/0-387-25465-X_40.

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Chernikova, O., Holzberger, D., Heitzmann, N., Stadler, M., Seidel, T., & Fischer, F. (2023). Where salience goes beyond authenticity: A meta-analysis on simulation-based learning in higher education. *Zeitschrift für Padagogische Psychologie, 38*(1–2), 15–25. https://doi.org/10.1024/1010-0652/a000357

Cierco Jimenez, R., Lee, T., Rosillo, N., Cordova, R., Cree, I. A., Gonzalez, A., & Indave Ruiz, B. I. (2022). Machine learning computational tools to assist the performance of systematic reviews: A mapping review. *BMC Medical Research Methodology, 22*, 322. https://doi.org/10.1186/s12874-022-01805-4

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1–54. https://doi.org/10.18637/jss.v025.i05

Foody, G. M. (2023). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLoS One, 18*. https://doi.org/10.1371/journal.pone.0291908

Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A., & Viswanathan, M. (2019). Assessing the accuracy of machine-assisted abstract screening with DistillerAI: A user study. *Systematic Reviews, 8*, 277. https://doi.org/10.1186/s13643-019-1221-3

Gates, A., Gates, M., Sebastianski, M., Guitard, S., Elliott, S. A., & Hartling, L. (2020). The semi-automation of title and abstract screening: A retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. *BMC Medical Research Methodology, 20*, 139. https://doi.org/10.1186/s12874-020-01031-w

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. MIT Press.

Hamel, C., Hersi, M., Kelly, S. E., Tricco, A. C., Straus, S., Wells, G., Pham, B., & Hutton, B. (2021). Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. *BMC Medical Research Methodology, 21*, 285. https://doi.org/10.1186/s12874-021-01451-2

Ivimey-Cook, E. R., Noble, D. W. A., Nakagawa, S., Lajeunesse, M. J., & Pick, J. L. (2023). Advice for improving the reproducibility of data extraction in meta-analysis. *Research Synthesis Methods, 14*(6), 911–915. https://doi.org/10.1002/jrsm.1663

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical knowledge discovery handbook*. https://doi.org/10.1007/978-0-387-09823-4_45

Kebede, M. M., Le Cornet, C., & Forter, R. T. (2023). In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. *Research Synthesis Methods, 14*(2), 156–172. https://doi.org/10.1002/jrsm.1589

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Liaw, A., & Wiener, M. (2002). *Classification and regression by random forest*, 18–22. RNews2/3 https://journal.r-project.org/articles/RN-2002-022.

Mankolli, E., & Guliashki, V. (2020). Machine Learning and Natural Language Processing: Review of Models and Optimization Problems. In V. Dimitrova, & I. Dimitrovski (Eds.), *Communications in Computer and Information Science: 1316. ICT Innovations 2020. Machine Learning and Applications. ICT Innovations 2020*. Cham: Springer. https://doi.org/10.1007/978-3-030-62098-1_7.

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*, 163. https://doi.org/10.1186/s13643-019-1074-9

Meyer, D., Dimitriadou, E., Weingessel, A., Leisch, F., Chang, C., & Lin, C.-C. (2023). Misc functions of the department of statistics, probability theory group (Formerly:E1071), TUWien. *R Documentation*. https://CRAN.R-project.org/package=e1071.

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*, 5. https://doi.org/10.1186/2046-4053-4-5

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology, 134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*. https://arxiv.org/abs/1912.01703.

Python Software Foundation. (2024). *Python for beginners*. https://www.python.org/about/gettingstarted.

PyTorch. (2023a). *PyTorch documentation*. https://pytorch.org/docs/stable/index.html.

PyTorch. (2023b). *LeakyReLU*. https://pytorch.org/docs/stable/generated/torch.nn.LeakyReLU.html#.

R Core Team. (2020). *Radiokhimiya: A language and environment for statistical computing*. https://www.R-project.org/.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Auto-matic keyword extraction from individual documents. In I. M. W. Berry, & J. Kogan (Eds.), *Text mining: Applications and theory* (pp. 1–20). Wiley. https://onlinelibrary.wiley.com/doi/10.1002/9780470689646.ch1.

Sailer, M., Ninaus, M., Huber, S. E., Bauer, E., & Greiff, S. (2024). under revision). The End is the Beginning is the End: The closed-loop learning analytics framework. *Computers in Human Behavior*.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software, 39*(5), 1–13. https://doi.org/10.18637/jss.v039.i05. Articles.

Smith, V., Devane, D., Begley, C. M., & Clarke, M. (2011). Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology, 11*(1), 15. https://doi.org/10.1186/1471-2288-11-15

Taylor, J. A., & Hedges, L. V. (2023). Toward more rapid accumulation of knowledge about what works in physics education: The role of replication, reporting practices, and meta-analysis, pp. 23/1-23/34. In M. F. Taşar, & P. R. L. Heron (Eds.), *The international handbook of physics education research: Special topics*. AIP Publishing. https://doi.org/10.1063/9780735425514.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tran, V. T., Gartlehner, G., Yaacoub, S., Boutron, I., Schwingshackl, L., Stadelmaier, J., Sommer, I., Aboulayeh, F., Afach, S., Meerpohl, J., & Ravaud, P. (2023). *Sensitivity, specificity and avoidable workload of using a large language models for title and abstract screening in systematic reviews and meta-analyses*. https://doi.org/10.1101/2023.12.15.23300018. preprint.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence, 3*(2). https://doi.org/10.1038/s42256-020-00287-7. Article 2.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (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* (pp. 819–824). https://doi.org/10.1145/2110363.2110464

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*. https://www.stm-assoc.org/2015_12_11_STM_Report_2015.pdf.

Wright, R. E. (1995). Logistic Regression. In L. G. Grimm, & P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics* (pp. 217–244). Washington DC: American Psychological Association.

Xiong, Z., Liu, T., Tse, G., Gong, M., Gladding, P. A., Smaill, B. H., & Zhao, J. (2018). A machine learning aided systematic review and meta-analysis of the relative risk of atrial fibrillation in patients with diabetes mellitus. *Frontiers in Physiology, 9*.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Sammons, M., Christodoulopoulos, C., Kordjamshidi, P., Khashabi, D., Srikumar, V., Roth. D. (2016). EDISON: Feature Extraction for NLP, Simplified. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), pages 4085–4092, Portorož, Slovenia. European Language Resources Association (ELRA). Retrieved from: https://aclanthology.org/L16-1645/.