








# Few-shot Approach for Systematic Literature Review Classifications

Maísa Kely de Melo<sup>1,2</sup><sup>a</sup>, Allan Victor Almeida Faria<sup>1,7</sup><sup>b</sup>, Li Weigang<sup>1,3</sup><sup>c</sup>,  
Arthur Gomes Nery<sup>1,5</sup><sup>d</sup>, Flávio Augusto R. de Oliveira<sup>1,8</sup><sup>e</sup>, Ian Teixeira Barreiro<sup>1,6</sup><sup>f</sup>  
and Victor Rafael Rezende Celestino<sup>1,4</sup><sup>g</sup>

<sup>1</sup>LAMFO - Lab. of ML in Finance and Organizations, University of Brasilia, Campus Darcy Ribeiro, Brasilia, Brazil

<sup>2</sup>Department of Mathematics, Instituto Federal de Minas Gerais Campus Formiga, Formiga, Brazil

<sup>3</sup>Department of Computer Science, University of Brasilia, Campus Darcy Ribeiro, Brasilia, Brazil

<sup>4</sup>Department of Business Administration, University of Brasilia, Federal District, Brazil

<sup>5</sup>Department of Economics, University of Brasilia, Federal District, Brazil

<sup>6</sup>Department of Economics, University of São Paulo, Ribeirão Preto, Brazil

<sup>7</sup>Department of Statistics, University of Brasília, Federal District, Brazil

<sup>8</sup>Ministry of Science, Technology and Innovation of Brazil, Federal District, Brazil

**Keywords:** Automation of Systematic Literature Review, Few-shot Learning, Meta-Learning, Transformers.


**Abstract:** Systematic Literature Review (SLR) studies aim to leverage relevant insights from scientific publications to achieve a comprehensive overview of the academic progress of a specific field. In recent years, a major effort has been expended in automating the SLR process by extracting, processing, and presenting the synthesized findings. However, implementations capable of few-shot classification for fields of study with a smaller amount of material available seem to be lacking. This study aims to present a system capable of conducting automated systematic literature reviews on classification constraint by a few-shot learning. We propose an open-source, domain-agnostic meta-learning SLR framework for few-shot classification, which has been validated using 64 SLR datasets. We also define an Adjusted Work Saved over Sampling (AWSS) metric to take into account the class imbalance during validation. The initial results show that AWSS@95% scored as high as 0.9 when validating our learner with data from 32 domains (just 16 examples were used for training in each domain), and only four of them resulted in scores lower than 0.1. These findings indicate significant savings in screening time for literature reviewers.


## 1 INTRODUCTION


The Systematic Literature Review (SLR) is a key tool for comprehending the status quo of a research area. It is a type of study that summarizes all available data fitting pre-specified criteria to answer precise research questions providing evidence of directions taken in recent years and the next steps indicated by the scientific community (Kusa et al., 2022). SLR is a means of identifying, evaluating, and synthesizing available research relevant to a particular research question (van Dinter et al., 2021b). Citation screening


is the stage where reviewers need to read and comprehend hundreds (or thousands) of documents and decide whether or not they should be included in the systematic review (Kusa et al., 2022). The collection, extraction, and synthesizing of the required data for systematic reviews are known to be highly manual, error-prone, and labor-intensive tasks (van Dinter et al., 2021c; Kusa et al., 2022). The workload involved in the SLR process is enormous and the process is slow, which motivates the effort to automate this process.


Despite the benefits obtained by the automation of SLR, its exploration is still in its infancy, and the theme has been scarcely discussed (Borges et al., 2021). A systematic overview of the current state-of-the-art in SLR automation seems to be lacking (van Dinter et al., 2021c). The primary purpose of SLR automation is to cut down the cost of systematic reviews and to reduce human error (van Dinter et al., 2021c).


<sup>a</sup> <https://orcid.org/0000-0001-8120-9778>


<sup>b</sup> <https://orcid.org/0000-0002-4300-9334>

<sup>c</sup> <https://orcid.org/0000-0003-1826-1850>

<sup>d</sup> <https://orcid.org/0000-0001-6844-807X>

<sup>e</sup> <https://orcid.org/0000-0001-5801-7620>

<sup>f</sup> <https://orcid.org/0000-0002-8498-9500>

<sup>g</sup> <https://orcid.org/0000-0001-5913-2997>

Traditional Machine Learning algorithms have been efficient in reducing researcher time, but they have labeling accuracy limitations (Howard et al., 2016). These methods may not always converge to a high recall performance of at least 95%, which is a key requirement of the citation screening task (Bekhuis and Demner-Fushman, 2012).

The basic methods of Natural Language Processing (NLP) are used for automated SLR processes. NLP refers to a branch of computer science and more specifically, a branch of artificial intelligence concerned with giving computers the ability to understand texts and spoken words in much the same way human beings can (IBM Cloud Education, 2021). Researchers should take advantage of the advances in NLP capability and the usage of its technologies to do further research (Collins et al., 2021).

Concerning the text classification, Support Vector Machines (SVM) have historically always performed well on this task because of their ability to generalize well on a large number of features (van den Bulk et al., 2022). However, they have been surpassed by Pre-trained Language Models (PLMs) based on Artificial Neural Networks (ANN) such as Embeddings from Language Model (ELMo) (Peters et al., 2018), or in transformers such as Bidirectional Encoder Representations from Transformers (BERT), (van den Bulk et al., 2022). ELMo and BERT have been effective in text classification tasks as they gain a deep understanding of the behavior of the textual language. Nonetheless, neural networks often perform optimally when there is an extensive dataset (Kontonatsios et al., 2020; van Dinter et al., 2021a).

For an automated system to be beneficial to systematic reviewers, it should save time and miss as few relevant papers as possible (Kusa et al., 2022). Besides, the automated SLR process displays a trade off issue: the state-of-the-art (SOTA) provides neural network models that perform well in this problem. However, these models require a large labeled domain dataset. Labeling on a large domain takes a significant amount of time, which wastes a researcher's time and budget constraints (van Dinter et al., 2021c). Therefore, the SOTA of automated SLR process requires a skillful algorithm that performs well on this problem using as little labeled domain data as possible. On the other hand, it is well-known that pre-training and fine-tuning even when evaluating small-scale data can suffer from instability, and results may change dramatically given a new split of data (Wang et al., 2021).

Regarding the challenge indicated above, Meta-learning has emerged as an approach for learning from small amounts of data (Nichol et al., 2018). To achieve the best performance of the machine learning

algorithms, the mechanism of learning to learn (meta-learning) should be broad in order to adapt to different tasks and computations required to complete these tasks. The model (or learner) is trained during a meta-learning phase on a set of tasks, such that the trained model can quickly be adapted to new tasks using only a small number of examples or trials (Finn et al., 2017). Few-shot learning (Weigang and da Silva, 1999; Fei-Fei et al., 2003) is well-studied in the domain of supervised tasks, where the goal is to learn a new function from only a few input/output pairs for that task, using primary data from similar tasks for meta-learning (Finn et al., 2017).

Hence, we propose a Model-Agnostic Meta-Learning (MAML) using Few-Shot Learning for Systematic Literature Review Classification in this work, applying a similar approach reported in the literature (Finn et al., 2017). The key idea is to train the model's initial parameters. The model has maximal performance on a new task after the parameters have been updated through one or more gradient steps computed with a small amount of data from that new task. Concerning the few-shot learning, the model requires the reviewer to label only a few papers previously. We also define an Adjusted Work Saved over Sampling (AWSS@R) to create a normalized metric to compare across different domains scores required when dealing to evaluate the same learner. Finally, we discuss the fine-tuning methodology, shedding light on the challenges of running an automated SLR, particularly the automated step of citation screening.

The contributions of this paper can be summarized as follows: we propose a model-agnostic meta-learning (MAML) with few-shot learning for automated SLR classification. This methodology can be used in SLR from different research fields. We conduct large-scale experiments across a total of 64 systematic review datasets to evaluate the effectiveness of the proposed method. From the Work Saved over Sampling (WSS@R) metric (Cohen et al., 2006), we define an Adapted WSS@R (AWSS@R) metric to make a fair comparison between the datasets using a model-agnostic meta-learner. Our method yields significant workload savings. AWSS@R metric scores of up to 0.9 were achieved when validating our learner using 32 data domains, only four of which resulted in scores below 0.1. This achievement is meaningful as these results were obtained using only 16 examples per domain to train the model. At the same time, the benchmark baseline used more than 60% of the examples in its domains as labeled data to train its models. Finally, our project is publicly available and open source<sup>1</sup>.

<sup>1</sup><https://github.com/BecomeAllan/ML-SLRC>

## 2 RELATED WORKS

The SLR process is separated into several steps to increase rigor and reproducibility (van Dinter et al., 2021b). The “Selection of primary studies (citation screening)” step is admittedly the most time-consuming (Bannach-Brown et al., 2019; Sellak et al., 2015; Tsafnat et al., 2018; van Dinter et al., 2021b). As time is crucial when dealing with research at the frontier of knowledge, finding a way to speed up the systematic review process is a priority.

Recently, deep learning algorithms have been useful in automating the citation screening process (Kontonatsios et al., 2020). Van Dinter and others (van Dinter et al., 2021b) presented the first end-to-end solution to citation screening with a deep neural network. Both models claim to yield the significant workload savings of at least 10% indicating notable savings in screening time in most domain data benchmark analyses (Kusa et al., 2022; Cohen et al., 2006).

The BERT model and its variants have pushed the state of the art for many NLP tasks (Devlin et al., 2019). BERT is based on a multi-layer bidirectional transformer. Its training is done by conditioning both left and right contexts, simultaneously optimizing for tasks of a masked word and next sentence prediction (Sun et al., 2019). SciBERT follows the same architecture as BERT but it is instead pre-trained on scientific text (Beltagy et al., 2019).

The use of BERT models for the specific purpose of document screening is very recent (Ioannidis, 2021; Qin et al., 2021). Kusa and others (Kusa et al., 2022) conducted a replicability study of the first two deep learning papers for citation screening (van Dinter et al., 2021b; Kontonatsios et al., 2020) and evaluated their performance on 23 publicly available domains data. Kontonatsios and others (Kontonatsios et al., 2020) presented an automatic text classification approach that aims to prioritize eligible citations earlier than ineligible ones. Van Dinter and others (van Dinter et al., 2021b) proposed a Multi-Channel Convolutional Neural Network approach to support the automated classification of primary studies.

The pre-trained algorithms used to automate the citation screening process demand that the reviewer labels a number of papers in order to train the model using the new dataset. This category of neural networks is prone to overfitting when trained on small datasets (Brownlee, 2018). Thus, a large amount of labeled domain data is necessary. To reduce the required effort to label a large dataset, it is necessary to explore alternative methods. The combination between Meta-learning and Few-Shot learners is a promising alternative.

Finn and others (Finn et al., 2017) proposed an algorithm for meta-learning that is model-agnostic. It is compatible with any model trained with gradient descent and applicable to various learning problems, including classification, regression, and reinforcement learning. An approximation to this model-agnostic meta-learning (MAML) can be obtained by ignoring second-order derivatives and using generalized first-order MAML (Nichol et al., 2018). Wang and others (Wang et al., 2021) reformulated traditional classification/regression tasks as textual entailment tasks.

This work uses the MAML strategy proposed by Finn and others (Finn et al., 2017) with the first-order approximation of data proposed by Nichols and others (Nichol et al., 2018), and the entailment Few-Shot learner strategy proposed by Wang and others (Wang et al., 2021). Here, these approaches are applied to the context of paper classification (include or do not include) used in the SLR process.

## 3 METHODS AND TRAINING FRAMEWORK

### 3.1 Data

In order to achieve a wide range of predictability, we propose a domain-agnostic dataset comprised of domains data from SLR on 64 topics, listed in Table 2. As pointed out by (van Dinter et al., 2021c), most studies use domain-dependent document meta-data from the medical research field. A strength of our work is the use of domains data from research fields other than medical, such as the ASReview Project (van de Schoot et al., 2021), Sciome Workbench for Interactive computer-Facilitated Text-mining (SWIFT-Review) (Howard et al., 2016), and Cereals and Leafy Greens (van den Bulk et al., 2022). The chosen data were reduced to titles, abstracts, and the labels as included or not included according to the revision criteria. We will consider included and not included to refer to those papers classified to be included or not included in the SLR, respectively.

### 3.2 SLR Classifier

In this work, we use the SLR Classifier model based on the SciBERT PLM, due to its previously presented advantages. The standard text classification training task passes the input text,  $P_1$ , to the learner and predicts a label based on this text. Here, this prediction is rethought as an entailment classification, where two bits of text,  $P_1$ ,  $P_2$ , are passed in as the input and the

output represents the judgment of whether the text  $P_1$  has entailment on  $P_2$  (Wang et al., 2021).  $P_1$  is the concatenation of the title and abstract, and  $P_2$  is the entailment text chosen as “It is a great text”.

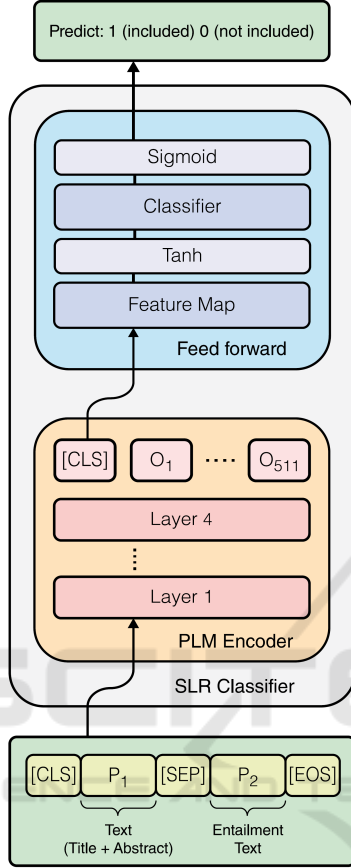


Figure 1: Architecture of SLR Classifier.

Figure 1 shows the proposed SLR Classifier. The proposed model uses four hidden layers of transformer-based PLM encoders from SciBERT, followed by a feedforward network (FNN) with an output of size 200 called “Feature map”, connected to the special token [CLS] of this chosen final SciBERT output layer. Then, the 200 outputs are passed to a hyperbolic tangent (tanh) activation function followed by the FNN classification layer called “Classifier”, which has a single output  $\hat{y}$ . Between the chosen SciBERT layer output and the “Feature map”, a batch normalization was set (Ioffe and Szegedy, 2015), and before the tanh activation, a dropout of 0.2 was set.

The classifier layer uses a sigmoid activation to output a prediction between 0 and 1,  $\sigma(\hat{y}) \in (0, 1)$ . The use of this activation is beneficial for the SLR Classifier. It can be used as a confidence ranker for the predictions, by establishing a threshold and only including in the review papers that the learner predicts

with higher confidence than this value (van Dinter et al., 2021b).

The Binary cross-entropy was used to compute the loss of this entailment classification task, adding a weight ( $p > 1$ ) to include examples to retrieve more recall to the predictions. The loss function is given by  $Loss(\hat{y}, y) = -[p \cdot y \cdot \log(\sigma(\hat{y})) + (1 - y) \cdot \log(1 - \sigma(\hat{y}))]$  where  $\hat{y}$  is the logit of the classifier layer followed by a sigmoid function  $\sigma(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}}$ , and the  $y$  is the correct label.

The choice of four layers of SciBERT for the SLR Classifier was conducted empirically. The minimal number of layers that the model should have were experimented with to improve its results. We aimed to reduce the number of model parameters, seeking that the researcher could perform an automated SLR Classification task (SLRC) with minimum GPU resources.

### 3.3 Task Learner

Table 2 presents the classification frequencies of their respective domain data. These domains have a severe data imbalance. The challenge is to build a task training framework that can handle a range of imbalanced data, such as the SLR datasets and resulting Task Learner being domain-independent across various classification tasks. As we aim for this learner to perform classification with as few examples as possible in a domain, the use of few-shot learning is promising.

The proposed learner is designed to train the SLR Classifier model with a few balanced domain data examples to perform a classification task. We call this framework a meta-learner - SLR Classifier (ML-SLRC). For this proposal, we train the SLR Classifier in as meta-learning phase with batches of tasks  $S = \{S_1, \dots, S_k\}$  with a few balanced examples (F examples) for each  $S_k$  task. Then we apply this model as a starting point to learn a new specific task  $S_{k+1}$  and perform predictions on this  $k + 1$  task domain. Figure 2 illustrates this meta learner.

In practice, this framework follows a method of MAML, which trains the learner based on a few shots, named N-way-F-shots. The idea is to try to use efficient starting parameters to optimize training tasks of different domains (Finn et al., 2017). In this work, we perform this training as 2-way-8-shots, where 2 is the number of classes of the task (1: included, 0: not included) with eight shots for each class. Each batch of tasks  $S$ , is called a support set. Many batches of test tasks, named queries and denoted as  $Q$ , are used to test the learner’s performance in the meta-learning and domain learning phase.



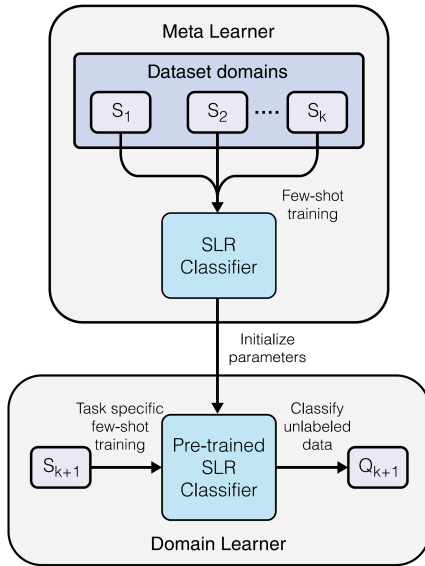


Figure 2: ML-SLRC framework.

Since the purpose is to use a PLM like SciBERT to have a good understanding of the text, the number of parameters can turn into a computational problem while calculating gradients used in MAML. The Reptile algorithm is applied to handle this issue, as it is a good approximation for this method (Nichol et al., 2018).

### 3.4 Metrics

When dealing with SLR automation problems, especially those that involve optimizing the selection of primary studies, the most used metrics to evaluate model performance are precision, recall, and the F1-score (van Dinter et al., 2021c). However, these traditional metrics are insufficient for learner evaluation as they do not indicate how much effort was spared for the researcher through the use of the learner (Kusa et al., 2022). Furthermore, when creating a learner to optimize the literature review process, the cost of failing to detect relevant new literature is high, and as such high recall is demanded (Cohen et al., 2006).

These demands led to the creation of the WSS@R metric (Cohen et al., 2006), which was defined as the fraction of work saved at a specific recall rate, as stated below:

$$WSS@R = \frac{TN + FN}{Sample\ Size} - (1 - R)$$

where TN is the number of true negatives, FN is the number of false negatives, and R is the recall. The WSS@R score measures the percentage of papers that meet the original search criteria that the reviewers do not have to read (because they have been screened out

by the classifier). Using automatic citation classification can effectively reduce the workload of preparation of systematic review.

The WSS@R score scale depends on the sample imbalance variation. This large variation is further illustrated in Figure 3, which presents heatmaps of WSS@R values considering sample sizes with different class imbalances that lead to different WSS@R’s perspectives. As defined in Section 3.3, the proposed ML-SLRC method trains the model by crossing different domains data. For the learner to have “impartiality” when evaluate with different domains, it is necessary to normalize them. Hence, an alternative metric to WSS@R is required. We propose a normalized metric in Section 4.

## 4 META-LEARNING EXPERIMENT

### 4.1 Adapted Work Saved over Sampling

The WSS@R score is well suited to the problem, although it has clear limitations. It is susceptible to high variation stemming from changes in class imbalance (van Dinter et al., 2021b). This paper uses many datasets to form the domain model-agnostic of SLRs. The same starting model (learner) is used to learn each SLR Classification task (SLRC), and the learner’s performance must be comparable between different datasets domains. This comparison could not be attainable with the WSS@R metric, justifying our proposition of a new adapted WSS@R metric. To the best of our knowledge, it is the first time such a metric has been suggested. We propose a case of WSS@R that is stable regardless of class imbalance in the sample used.

To derive this new metric, we suppose the worst scenario for a given SLRC where the number of not included (N) papers is much higher than the included (P) papers. Statistically, this can be represented by  $N \gg P$  for the population distribution. Given a sample of this SLRC scenario and a learner to predict the classifications, the WSS@R measures the quality of the learner’s exclusion criteria on the sample by the examples classified as not included ( $tn + fn$ ,  $tn$  and  $fn$  are true and false negative on sample’s predictions, respectively), given how many were correctly classified as included ( $tp$ , true positive on sample’s prediction). The total sample’s classification examples, not included and included, are given by  $n = tn + fp$  and  $p = tp + fn$ , respectively, and the sample size is

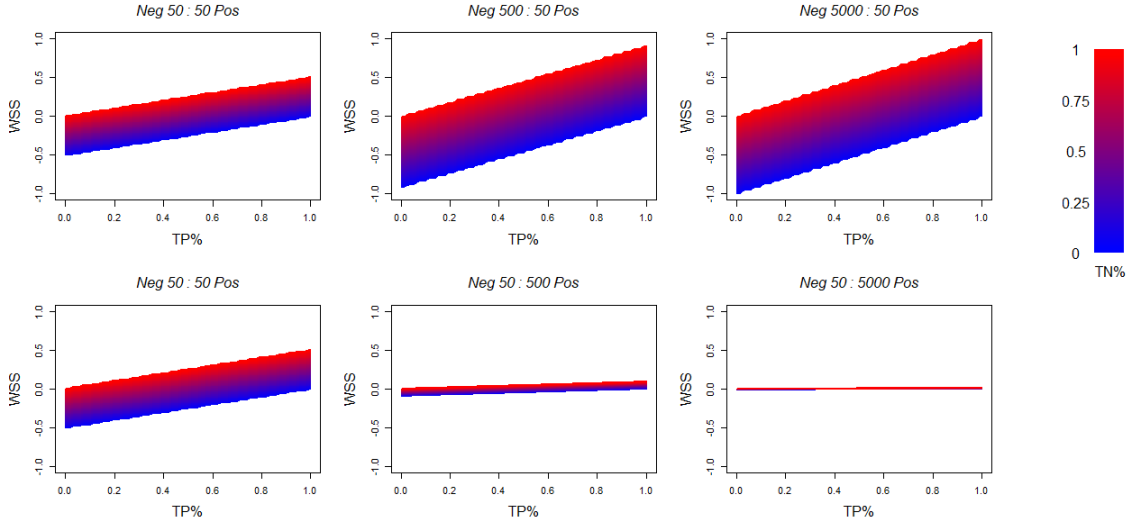


Figure 3: WSS@R by the ratio of the true labels, not included ( $TN\%$ ) and included ( $TP\%$ ) simulated.

$M = n + p$ . The WSS@R can be derived as

$$WSS@R = \frac{tn + fn}{M} - (1 - R)$$

$$WSS@R = \frac{tn}{M} + \frac{fn}{M} - (1 - tp\%).$$

Thus, with the restriction  $N \gg P$  for the SLRC scenario, when the sample grows to the corresponding population size,  $(n, p) \rightarrow (N, P) = (TN + FP, TP + FN)$ , the learner's exclusion criteria also will converge the predictions to some distribution of TN, FN (true and false negative on population's predictions, respectively) and TP, FP (true and false positive on population's predictions, respectively). It is reasonable to approximate  $N \approx N + P$  and  $fn/M \rightarrow 0$ . Therefore, the WSS@R can be approximated by<sup>2</sup>

$$WSS@R \xrightarrow{M \rightarrow \infty} TN\% - (1 - TP\%) = AWSS@R.$$

We named this convergence approximation as Adapted WSS (AWSS@R). This new formula is now used with the assumption that a sample was given by the worst-case scenario of the SLRC, where  $TN\%$  and  $TP\%$  are the proportion's predictions of the true classifications of not included and included, respectively.

In the case of a relatively large sample, this metric is a good approximation of the WSS@R in the sample case of  $n \gg p$ . In other cases like  $n = p$  or  $n \ll p$ , the WSS@R is susceptible to changes in sample distribution to evaluate the respective work saved. In contrast, the AWSS@R is not susceptible because it considers the sample's work saved rate with the SLRC scenario assumption. The evaluation is a relative perspective of WSS@R and can be compared

<sup>2</sup> $tp\% = \frac{tp}{p}$ ,  $TN\% = \frac{TN}{N}$ , and  $TP\% = \frac{TP}{P}$ .

between the sample cases to evaluate the learner's exclusion criteria. Table 1 shows this sample's proportions cases with WSS@R and AWSS@R. If 100% of included papers in the SLR were included by the classifier, the WSS@100% measures the percentage of the sample's examples that are not included in the SLR and were not selected to be read, therefore reducing this percentage of the sample to be read. The AWSS@100% measures the same percentage if the sample was given by the worst SLRC perspective. The AWSS@R is numerically comparable and can infer the proposed scenario, as can be seen in Figure 4. Different ratios are simulated to map the AWSS@R plane, where in all cases of the sample's ratios, the metric is stabilized in contrast to WSS@R in Figure 3. Using both formulas is advisable for the sample and worst-case perspective evaluation.

Table 1: AWSS@R and WSS@R values considering different class imbalanced examples with  $TN\% = 100\%$ .

Sample		AWSS		WSS	
Not included	Included	@0%	@100%	@0%	@100%
550	550	0	1	0	0.50
1000	100	0	1	0	0.90
100	1000	0	1	0	0.09

## 4.2 Split Training 50-50

When training and validating our learner, it is important to note the impact of task imbalance during the meta-learning phase. The proportion of included and not included labels in some domain tasks considered varies vastly (Table 2). Therefore, it is necessary to split the tasks into train, test, and validation task sets that consider the need for the meta-learner to perform

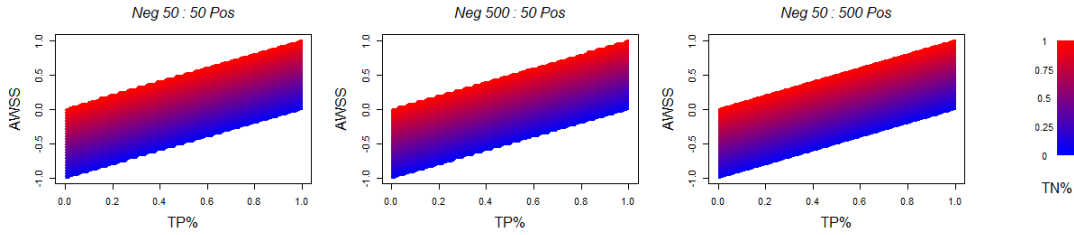


Figure 4: AWSS@R by the ratio of the true labels, not included ( $TN\%$ ) and included ( $TP\%$ ) simulated.

generically. Given that the training task set is a randomly selected batch of tasks, if a given domain task was already included in this batch, different examples of this task must be chosen to train the learner. Thus, the learner learns a task independently of the task example, avoiding task’s example overfit.

Due to the issue raised above, we created a training task set selecting the domains from Table 2 with over 50 entries for both not included and included classifications. In the meta-learning phase, the learner was set to train during four inner epochs (the number of epochs to learn the task  $S_k$ ) and 20 outer epochs (the number of epochs to pass over all the batch of tasks  $S$ ). For each outer epoch, a new batch of tasks  $S$  containing 20 tasks with 16 examples for each task (eight included and eight not included examples) is randomly initialized. To update the parameters with the training task set, the inner phase (“learn the task”) was set to create batches of four examples, and to update the parameters, the outer phase (“learn to learn tasks”) was set to have five tasks. All examples that pass through the SLR Classifier have 512 tokens, and to compute the loss, the weight of included examples was set to 1.5 to retrieve more recall and the learning rate of the inner and outer update step were set to  $5 \times 10^{-5}$ .

Table 3 summarizes the results of five attempts for each task on the test task set<sup>3</sup>. Once the SLR Classifier model is trained in the meta-learning phase, a new domain task is passed to be learned with 16 examples (eight included and eight not included) in the domain learning phase to validate the learner. Table 3 lists the mean of these five attempts. In some cases, at 95% recall rate, the learner performs well once the threshold (the confidence level to predict if the example is included in the SLR task domain) surpasses 80%. Overall, with as few as 16 examples, the learner achieves an AWSS@95% metric close to one, representing a satisfactory performance and saving considerable time when labeling these different SLR domains.

To demonstrate the learner predictions after train-

<sup>3</sup>When using @R means, the threshold was defined as 0.9 to evaluate the learner confidence level.

ing with 16 examples in one domain, the 200 outputs of the SLR Classifier model from the feature map layer were mapped to 2 dimensions with t-distributed Stochastic Neighbor Embedding t-SNE (van der Maaten and Hinton, 2008) and plotted as dispersion points. To map the confidence of each point, the sigmoid-activated classifier layer output was used. The colors blue (0, not included) and orange (1, included) are the true labels of these examples.

Figure 5 describes the t-SNE technique for one attempt in the “Fluoride” test task domain. In particular, Figure 5 (b) contains all included classified papers with a threshold@95 greater than 0.798. For this case, the accuracy is 0.9, showing a high hit rate of the true label. Figure 5 (c) represents the predicted included papers for a 90% threshold. Figures 5 (a), (b), and (c) show that the learner is sorting the predicted papers, and the learner aims to prioritize those classified as included. Figure 5 (d) represents the ROC curve for this attempt. The red line is the recall at 95%, confirming the learner’s reasonable confidence in classifying the papers using different thresholds.

### 4.3 Benchmark Datasets Comparison

To validate our proposed training framework, we performed a comparison with a literature baseline (Kusa et al., 2022). The comparison considered these baseline domains for the test tasks phase and the other domains defined in Table 2 for the train and validation tasks phases.

Table 5 hands out the WSS@95% values presented in the baseline domain (Kusa et al., 2022) concatenated with WSS@95% and AWSS@95% of the proposed Task Learner. The AWSS@95% shows the normalization effect. As we can see, in general, the WSS@95% is smaller than AWSS@95% due to the class imbalance defined in Table 2.

Concerning these metrics, even when the learner does not exceed the values of the benchmark domains, there is time saved in the SLR. The average of the AWSS@95%, considering all the domains, is 24.5%, more than double the minimum (10%) considered acceptable (Cohen et al., 2006). The contribution of

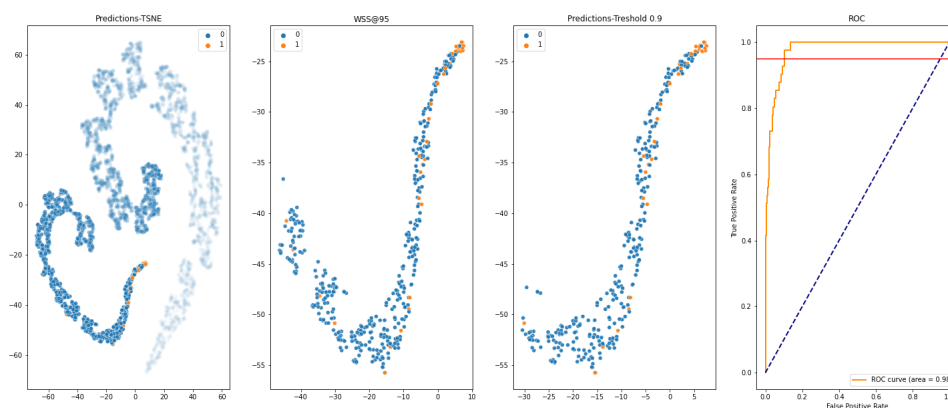


Figure 5: t-SNE predictions and ROC curve of one attempt (Fluoride).

the proposed Task Learner is significant given the 16 labeled papers, a tiny amount compared to the number of articles used to train and test the benchmarks baselines. Several studies (Cohen et al., 2006; van Dinter et al., 2021a; Kontonatsios et al., 2020; Kusa et al., 2022) used more than 60% of the articles present in the Atypical Antipsychotics domain to train the learner.

Figure 6 (a), (b) and (c) describes the t-SNE technique for one attempt in the test domain, Atypical Antipsychotics. This figure shows that the learner can sort the predictions. According to the opacity of the dots, this sorting firstly classifies the papers predicted as true included. Figure 6 (d) shows the ROC curve for this case. This curve identifies that the learner’s confidence is low when approximating the recall at 95% (red line). This fact also is exhibited in Table 4, where the threshold at recall 95% (threshold@95%) for the Atypical Antipsychotics domain is 29,5%. The same is observed in all other domains included in the test task set.

## 5 DISCUSSION

In this paper, we use neural network-based pre-trained language models with MAML components to train a model to perform primary citation screening on various subject domains. To the best of our knowledge, the proposed learner expands the research in the field of automated systematic literature reviews, innovating with the unprecedented use of meta-learning coupled with neural network-based methods in the area.

Out of the 12 domains, the SLR Classifier learner trained with only 16 examples in the domain learning phase, all of them had scores superior to 10%, as shown in Table 3. In the case of the learner trained on the benchmark domains, shown in Table 4, out of 20 tasks, the learner had scores superior to 10% in

15 of them. This indicates that the learner can contribute to reducing citation screening workload. Once again, when using the AWSS@95% as a reference, no domains had scores below 10% in the results shown in Table 3, and only four tasks had scores below this value in the results seen in Table 4. The minimum WSS@95% score for the learner was 0.05, and the maximum score was 0.9.

In general, when comparing the results of our learner with those of the literature (Cohen et al., 2006; Matwin et al., 2010; Cohen, 2008; Cohen, 2011; Howard et al., 2016), utilizing the AWSS@95% metric, it can be inferred that our learner had inferior results, with some exceptions, as can be seen in Table 5. Nevertheless, the few-shot learner herein proposed is domain-agnostic, while all the comparing models were specifically trained for each domain with a relatively large dataset and, as such, should be much less flexible and adaptable.

In Figure 5, we can see an example where the performance of the algorithm was suitable. As can be seen, there is a large AUC for the ROC curve, resulting in a low false-positive rate at a recall of 95%. Furthermore, separability between included and not included values can emerge in the first three images. On the other hand, Figure 6 shows an example that stresses room for improvement where the learner showed poor performance in disentangling the classification confidence of the examples, necessary in order to improve its ability to create separability between classes.

## 6 CONCLUSION AND FUTURE WORK

We developed a domain-agnostic ML-SLRC framework using few-shot classification in this work. With



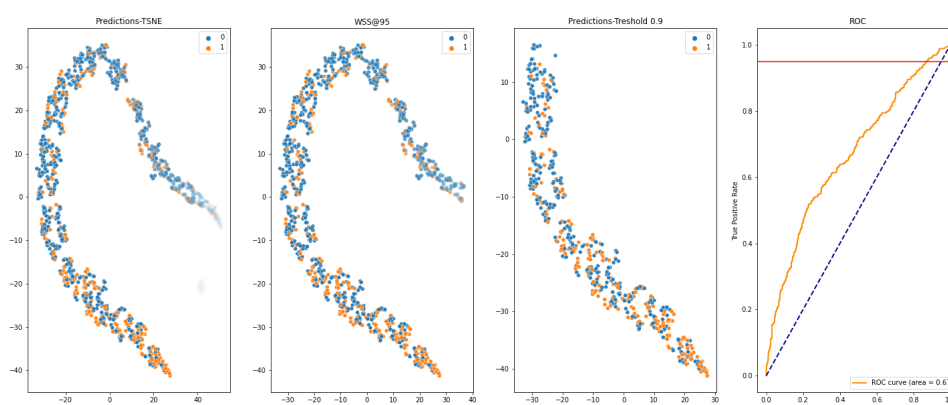


Figure 6: T-SNE predictions and ROC curve of one attempt (Atypical Antipsychotics).

a pre-trained SLR Classifier as a learner, it is possible to conduct the citation screening of a SLR Classification task utilizing a small number of previously labeled papers. We also proposed a new metric called Adjusted Work Saved on Sampling at recall R%, capable of representing the proportion of examples not included in the SLR separated to be not read if it is the worst perspective scenario of an SLR, resulting in comparable scores in different SLR domains with variable classes imbalances. The results showed that the proposed learner can classify the articles within the domain as included or not included, presenting reasonable WSS@R and AWSS@R values. Our proposed learner saves time, preventing the reviewer from reading papers that are not to be included in the SLR, using just as few papers as possible to train and perform a citation screening. The domains utilized in our experiments considered a wide range of different fields of research.

This paper is a part of a systematic study to develop an SLR solution for the end-user. All code and databases are publicly available and open-source, so this proposed model has a practical application to society. Future work is required to improve the classification accuracy. Therefore, we suggest using Active Learning, where the learning method includes the reviewer's interference in the training loop (Yi et al., 2022). The learner is expected to select key papers for the reviewer to label online among those with the lowest probability of hitting. When the reviewer specifically labels the papers that the model has the most difficulty classifying, it is also expected to increase the learner accuracy.

## REFERENCES

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., and 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):1–12.

Bekhuis, T. and Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial intelligence in medicine*, 55(3):197–207.

Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Borges, A. F., Laurindo, F. J., Spínola, M. M., Gonçalves, R. F., and Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, 57:102225.

Brownlee, J. (2018). *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery.

Cohen, A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, page 121–125.

Cohen, A. M. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the wss@95 measure: Table 1. *Journal of the American Medical Informatics Association*, 18(1).

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

Collins, C., Dennehy, D., Conboy, K., and Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, 60(June):102383.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

- Fei-Fei, L., Fergus, R., and Perona, P. (2003). A bayesian approach to unsupervised one-shot learning of object categories. In *proceedings ninth IEEE international conference on computer vision*, pages 1134–1141. IEEE.
- Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400.
- Harmsen, W. W., Groot, J. d., and Dusseldorp, I. v. (2021). Medical guidelines dutch association medical specialists. Datastorage of published guidelins on the Dutch Medical Guideline Database.
- Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., Holmgren, S., Pelch, K. E., Walker, V., Rooney, A. A., et al. (2016). Swift-review: a text-mining workbench for systematic review. *Systematic reviews*, 5(1):1–16.
- IBM Cloud Education (2021). Natural language processing (NLP). <https://www.ibm.com/cloud/learn/natural-language-processing>. [Online; accessed 08-March-2022].
- Ioannidis, A. (2021). An analysis of a bert deep learning strategy on a technology assisted review task. *arXiv preprint arXiv:2104.08340*.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.
- Kontonatsios, G., Spencer, S., Matthew, P., and Korkontzelos, I. (2020). Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, 6:100030.
- Kusa, W., Hanbury, A., and Knoth, P. (2022). Automation of citation screening for systematic literature reviews using neural networks: A replicability study. *arXiv preprint arXiv:2201.07534*.
- Lanera, C., Minto, C., Sharma, A., Gregori, D., Berchiolla, P., and Baldi, I. (2018). Extending pubmed searches to clinicaltrials.gov through a machine learning approach for systematic reviews. *Journal of Clinical Epidemiology*, 103:22–30.
- Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., and O’Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453.
- Nichol, A., Achiam, J., and Schulman, J. (2018). On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.
- Qin, X., Liu, J., Wang, Y., Liu, Y., Deng, K., Ma, Y., Zou, K., Li, L., and Sun, X. (2021). Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *Journal of Clinical Epidemiology*, 133:121–129.
- Sellak, H., Ouhbi, B., and Frikh, B. (2015). Using rule-based classifiers in systematic reviews: a semantic class association rules approach. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, pages 1–5.
- Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics*, pages 194–206. Cham. Springer International Publishing.
- Tsafnat, G., Glasziou, P., Karystianis, G., and Coiera, E. (2018). Automated screening of research studies for systematic reviews using study characteristics. *Systematic reviews*, 7(1):1–9.
- van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdemans, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133.
- van den Bulk, L. M., Bouzembrak, Y., Gavaï, A., Liu, N., van den Heuvel, L. J., and Marvin, H. J. (2022). Automatic classification of literature in systematic reviews on food safety using machine learning. *Current Research in Food Science*, 5:84–95.
- van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.
- van Dinter, R., Catal, C., and Tekinerdogan, B. (2021a). A decision support system for automating document retrieval and citation screening. *Expert Systems with Applications*, 182:115261.
- van Dinter, R., Catal, C., and Tekinerdogan, B. (2021b). A multi-channel convolutional neural network approach to automate the citation screening process. *Applied Soft Computing*, 112:107765.
- van Dinter, R., Tekinerdogan, B., and Catal, C. (2021c). Automation of systematic literature reviews: A systematic literature review. *Inf. Softw. Technol.*, 136:106589.
- Wang, S., Fang, H., Khabsa, M., Mao, H., and Ma, H. (2021). Entailment as few-shot learner. *CoRR*, abs/2104.14690.
- Weigang, L. and da Silva, N. C. (1999). A study of parallel neural networks. In *IJCNN’99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 2, pages 1113–1116. IEEE.
- Yi, J. S. K., Seo, M., Park, J., and Choi, D.-G. (2022). Using self-supervised pretext tasks for active learning. *arXiv preprint arXiv:2201.07459*.

## APPENDIX

Tables containing the datasets description and experiments results.

Table 2: Citations for the 64 publicly available datasets used in the experiments on automated citation screening for SLRs.

Source	Size	Incl.	Excl.	Source	Size	Incl.	Excl.
ACE Inhibitors	2214	167	2047	Animal Depression	1599	251	1348
ADHD	781	80	701	Anxiety-Related Disorders	10515	770	9745
Antihistamines	277	87	190	Dementia	5609	11	5598
Atypical Antipsychotics	999	329	670	Heart Disease	4212	19	4193
Beta Blockers	1819	266	1553	Nudging	1850	383	1467
Calcium Channel Blockers	1069	246	823	PTSD Trajectories	5425	359	5066
Estrogens	337	77	260	Software Defect Detection	7002	62	6940
NSAIDs	348	81	267	Software Engineering	1700	45	1655
Opioids	1717	41	1676	Software Fault Metrics	6000	48	5952
Oral Hypoglycemics	462	134	328	Software Fault Prediction	8911	104	8807
Proton Pump Inhibitors	1171	220	951	Virus Metagenomics	2305	114	2191
Skeletal Muscle Relaxants	1318	26	1292	Wilson Disease	2358	161	2197
Statins	2659	150	2509	ASReview (van de Schoot et al., 2021)	57486	2327	55159
Triptans	573	200	373	Cereals	674	292	382
Urinary Incontinence	271	65	206	Cereals Future set	147	71	76
Drug Reviews (Cohen et al., 2006)	16015	2169	13846	Leafy Greens	224	66	158
Distal Radius Fractures Approach	182	10	172	Leafy Greens Future set	95	62	33
Distal Radius Fractures Closed Reduction	180	4	176	Food Safety (van den Bulk et al., 2022)	1140	491	649
Hallux Valgus Prognostic	59	2	57	Alzheimers	832	32	800
Head and Neck Cancer Bone	228	32	196	Angiotensin	209	9	200
Head and Neck Cancer Imaging	6	2	4	Anticoagulation	418	18	400
Obstetric Emergency Training	150	17	133	Atorvastatin	416	16	400
Post Intensive Care Treatment	291	77	214	Bivalirudin	414	14	400
Pregnancy Medication	321	39	282	Cetuximab	412	12	400
Shoulder Replacement Diagnostic	329	3	326	Colorectal Cancer	413	13	400
Shoulder Replacement Surgery	376	6	370	Dabigatran	413	13	400
Shoulder dystocia Positioning	146	6	140	Gastric Cancer	206	6	200
Shoulder dystocia Recurrence	281	5	276	Metformin	623	23	600
Total Knee Replacement	311	25	286	Parkinsons	209	9	200
Vascular Access	728	19	709	Rheumatoid	1675	75	1600
Medical Guidelines (Harmsen et al., 2021)	3588	247	3341	Tyrosine Kinase	1043	43	1000
Bisphenol A (BPA)	7093	102	6991	Ustekinumab	209	9	200
Fluoride and Neurotoxicity	3870	49	3821	PubMed Abstracts (Lanera et al., 2018)	7492	292	7200
Neuropathic pain	29202	5009	24193				
PFOA/PFOS	5950	95	5855				
Transgenerational	46147	606	45541				
SWIFT (Howard et al., 2016)	92262	5861	86401				

Table 3: Summary of the mean (and the standard deviation) of five validation considering 16 examples (eight positive and eight negative samples) in the domain learner phase, after training the SLR classifier in the meta learner phase. defined in Section 4.2.

Source	WSS		AWSS		Recall WSS		Accuracy		F1 Score		Threshold
	@95%	@R	@95%	@R	@95%	@R	@95%	@R	@95%	@R	@95%
Dementia	.90 (.04)	.73 (.11)	.90 (.04)	.73 (.11)	1 (0)	.87 (.16)	.90 (.04)	.87 (.06)	.01 (.01)	.01 (.01)	.89 (.06)
Fluoride	.69 (.09)	.50 (.27)	.70 (.09)	.50 (.27)	1 (0)	.56 (.32)	.70 (.09)	.94 (.05)	.02 (.01)	.09 (.05)	.80 (.07)
Head Cancer	.84 (.02)	.81 (.02)	.85 (.02)	.82 (.02)	.95 (0)	.89 (.03)	.90 (.02)	.93 (.02)	.17 (.02)	.21 (.04)	.80 (.12)
Heart Disease	.32 (.09)	.11 (.12)	.34 (.10)	.12 (.13)	.96 (0)	.20 (.22)	.43 (.09)	.86 (.06)	.23 (.03)	.14 (.12)	.53 (.11)
Leafy Greens F.	.15 (.04)	.01 (.01)	.47 (.13)	.03 (.03)	.97 (.01)	.04 (.05)	.83 (.04)	.33 (.03)	.89 (.02)	.07 (.09)	.43 (.17)
Opioids	.39 (.07)	.35 (.07)	.40 (.07)	.36 (.08)	.97 (0)	.53 (.13)	.44 (.07)	.83 (.06)	.06 (.01)	.11 (.01)	.39 (.16)
PFOS-PFOA	.72 (.03)	.43 (.21)	.73 (.03)	.44 (.21)	.95 (0)	.49 (.24)	.78 (.03)	.95 (.03)	.11 (.01)	.21 (.04)	.65 (.14)
Skeletal Muscle	.39 (.07)	.14 (.08)	.40 (.07)	.14 (.08)	1 (0)	.16 (.10)	.40 (.07)	.97 (.01)	.05 (.01)	.12 (.06)	.37 (.10)
Software Eng.	.34 (.06)	.05 (.09)	.34 (.06)	.05 (.09)	.97 (0)	.06 (.11)	.38 (.06)	.97 (.01)	.07 (.01)	.04 (.07)	.30 (.16)
Software Fault	.78 (.02)	.59 (.24)	.79 (.02)	.59 (.24)	.97 (0)	.64 (.28)	.82 (.02)	.95 (.04)	.07 (.01)	.19 (.07)	.54 (.07)
Knee R.	.76 (.08)	.61 (.14)	.80 (.09)	.64 (.15)	.96 (0)	.68 (.18)	.85 (.08)	.94 (.03)	.46 (.15)	.60 (.09)	.81 (.06)
Vascular A.	.51 (.10)	.27 (.13)	.52 (.10)	.27 (.13)	1 (0)	.33 (.16)	.53 (.10)	.93 (.04)	.07 (.01)	.16 (.09)	.56 (.14)

Table 4: Summary of the mean of five (and the standard deviation) validation considering 16 examples (eight positive and eight negative samples) in the domain learner phase, after training the SLR classifier in the meta learner phase defined in Section 4.3.

Source	WSS		AWSS		Recall WSS		Accuracy		F1 score		Threshold
	@95%	@R	@95%	@R	@95%	@R	@95%	@R	@95%	@R	@95%
ACE Inhibitors	.22 (.04)	.13 (.08)	.24 (.04)	.14 (.09)	.95 (0)	.22 (.14)	.33 (.04)	.87 (.05)	.17 (.01)	.17 (.08)	.23 (.14)
ADHD	.36 (.04)	.34 (.16)	.39 (.05)	.38 (.17)	.96 (0)	.47 (.20)	.48 (.04)	.87 (.02)	.25 (.02)	.38 (.10)	.10 (.05)
Antihistamines	.08 (.03)	.04 (.04)	.11 (.04)	.05 (.05)	.95 (0)	.08 (.08)	.38 (.03)	.72 (.01)	.47 (.01)	.12 (.12)	.26 (.09)
Antipsychotics	.05 (.01)	.14 (.02)	.08 (.02)	.21 (.03)	.95 (0)	.65 (.06)	.39 (.01)	.59 (.04)	.50 (0)	.50 (.01)	.29 (.17)
Beta Blockers	.14 (.01)	.19 (.07)	.17 (.02)	.23 (.08)	.95 (0)	.41 (.14)	.32 (.01)	.76 (.03)	.29 (0)	.32 (.05)	.22 (.10)
Bisphenol A	.70 (.03)	.53 (.14)	.71 (.03)	.54 (.14)	.95 (0)	.65 (.17)	.76 (.03)	.89 (.03)	.10 (.01)	.14 (.02)	.58 (.20)
Calcium C.	.07 (.03)	.07 (.07)	.09 (.03)	.09 (.10)	.96 (.01)	.15 (.18)	.32 (.03)	.76 (.03)	.39 (.01)	.17 (.16)	.14 (.11)
Estrogens	.03 (.03)	.08 (.09)	.04 (.03)	.09 (.11)	.96 (0)	.18 (.24)	.26 (.03)	.76 (.05)	.35 (.01)	.17 (.14)	.07 (.03)
Fluoride	.88 (.02)	.82 (.06)	.89 (.02)	.83 (.06)	.95 (0)	.90 (.09)	.93 (.02)	.93 (.02)	.23 (.05)	.22 (.08)	.88 (.08)
Hypoglycemics	.04 (.02)	0 (.01)	.05 (.02)	0 (.02)	.95 (0)	.03 (.01)	.33 (.02)	.72 (.02)	.43 (.01)	.05 (.03)	.14 (.11)
Incontinence	.19 (.04)	.31 (.02)	.25 (.05)	.40 (.02)	.96 (.01)	.57 (.08)	.44 (.04)	.77 (.04)	.44 (.02)	.53 (.02)	.21 (.11)
Neuro. Pain	.28 (.01)	.15 (.09)	.34 (.02)	.18 (.11)	.95 (0)	.25 (.16)	.49 (.01)	.81 (.02)	.39 (.01)	.28 (.15)	.29 (.15)
NSAIDS	.20 (.04)	.24 (.14)	.25 (.05)	.31 (.18)	.95 (0)	.43 (.29)	.44 (.04)	.78 (.03)	.42 (.02)	.39 (.18)	.36 (.15)
Opioids	.46 (.06)	.37 (.09)	.47 (.06)	.38 (.09)	.95 (0)	.49 (.16)	.53 (.06)	.88 (.07)	.08 (.01)	.16 (.03)	.26 (.17)
PFOS-PFOA	.67 (.06)	.29 (.24)	.67 (.07)	.30 (.25)	.95 (0)	.39 (.34)	.72 (.06)	.90 (.09)	.09 (.02)	.09 (.06)	.65 (.19)
Proton Pump	.11 (.02)	.06 (.07)	.13 (.02)	.08 (.08)	.95 (0)	.11 (.11)	.31 (.02)	.82 (.01)	.33 (.01)	.15 (.13)	.06 (.03)
Skeletal Muscle	.22 (.08)	.12 (.10)	.22 (.08)	.13 (.10)	.96 (0)	.17 (.15)	.27 (.08)	.94 (.05)	.04 (0)	.09 (.03)	.20 (.15)
Statins	.15 (.04)	.14 (.02)	.16 (.04)	.14 (.02)	.95 (0)	.18 (.03)	.24 (.04)	.92 (.02)	.11 (0)	.19 (.02)	.14 (.06)
Transgen.	.43 (.12)	.33 (.09)	.44 (.12)	.34 (.09)	.95 (0)	.36 (.09)	.50 (.12)	.97 (0)	.06 (.01)	.25 (.03)	.23 (.13)
Triptans	.19 (.04)	0 (0)	.28 (.06)	-.01 (0)	.95 (0)	.02 (.02)	.53 (.04)	.67 (.01)	.57 (.02)	.04 (.03)	.16 (.03)

Table 5: Means (and the standard deviation) of WSS@95% and AWSS@95% across five validation runs for each of the 20 review datasets presented in Table 4 compared with results presented in (Kusa et al., 2022).

Dataset	Results							[6] replicated by [8]	[7] replicated by [8]	WSS @95%	AWSS @95%
	[1]	[2]	[3,4]	[5]	[6]	[7]					
ACE Inhibitors	.566	.523	.733	.801	.787	.783	.785	.367	.224	.240	
ADHD	.680	.622	.526	.793	.665	.698	.639	.704	.356	.391	
Antihistamines	.000	.149	.236	.137	.310	.168	.275	.135	.075	.105	
Atypical Antipsychotics	.141	.206	.170	.251	.329	.212	.190	.081	.054	.800	
Beta Blockers	.284	.367	.465	.428	.587	.504	.462	.399	.142	.166	
Calcium Channel Blockers	.122	.234	.430	.448	.424	.159	.347	.069	.073	.950	
Estrogens	.183	.375	.414	.471	.397	.119	.369	.083	.034	.430	
Oral Hypoglycemics	.090	.085	.136	.117	.095	.065	.123	.013	.360	.490	
Urinary Incontinence	.261	.296	.432	.531	.531	.272	.483	.180	.195	.251	
NSAIDS	.497	.528	.672	.730	.723	.571	.735	.601	.199	.252	
Opioids	.133	.554	.364	.826	.533	.295	.580	.249	.457	.466	
Proton Pump Inhibitors	.277	.229	.328	.378	.400	.243	.299	.129	.106	.128	
Skeletal Muscle Relaxants	.000	.265	.374	.556	.286	.229	.286	.300	.217	.221	
Statins	.247	.315	.491	.435	.566	.443	.487	.283	.149	.157	
Triptans	.034	.274	.346	.412	.434	.266	.412	.440	.191	.281	
Average Drug Reviews	.234	.335	.408	.488	.471	.335	.431	.269	.167	.195	
Bisphenol A (BPA)					.752	.792	.780	.369	.704	.713	
Fluoride and Neurotoxicity					.870	.883	.806	.808	.879	.888	
Neurophatic Pain					.691	.620	.598	.091	.283	.341	
PFOA/PFOS					.805	.071	.838	.305	.665	.675	
Transgenerational					.714	.708	.718	.000	.434	.440	
Average SWIFT					.766	.615	.748	.315	.314	.294	
Average (all datasets)					.619	.475	.590	.292	.241	.245	

[1](Cohen et al., 2006) [2](Matwin et al., 2010) [3,4](Cohen, 2008) (Cohen, 2011) [5](Howard et al., 2016)[6](van Dinter et al., 2021a)[7](Kontonatsios et al., 2020)

[8](Kusa et al., 2022)