

1 **Machine learning algorithms for systematic review:**
2 **reducing workload in a preclinical review of animal**
3 **studies and reducing human screening error**

4 Bannach-Brown, Alexandra ^{1,5} (<https://orcid.org/0000-0002-3161-1395>) a.bannach-
5 brown@ed.ac.uk*,

6 Przybyła, Piotr ² piotr.przybyla@manchester.ac.uk,

7 Thomas, James ³ (<https://orcid.org/0000-0003-4805-4190>) james.thomas@ucl.ac.uk,

8 Rice, Andrew S.C. ⁴ (<https://orcid.org/0000-0002-1237-4769>) a.rice@imperial.ac.uk,

9 Ananiadou, Sophia ² Sophia.Ananiadou@manchester.ac.uk,

10 Liao, Jing ¹ jing.liao@ed.ac.uk,

11 Macleod, Malcolm Robert ¹ (<https://orcid.org/0000-0001-9187-9839>)

12 Malcolm.Macleod@ed.ac.uk,

13

14 1 Centre for Clinical Brain Sciences, University of Edinburgh.

15 2 National Centre for Text Mining, School of Computer Science, University of Manchester.

16 3 EPPI-Centre, Department of Social Science, University College London.

17 4 Pain Research, Department of Surgery and Cancer, Imperial College.

18 5 Translational Neuropsychiatry Unit, Aarhus University.

19 * Corresponding Author

20

21

22

23 Abstract:

24 Background: Here we outline a method of applying existing machine learning (ML) approaches to aid
25 citation screening in an on-going broad and shallow systematic review of preclinical animal studies,
26 with the aim of achieving a high performing algorithm comparable to human screening.

27 Methods: We applied ML approaches to a broad systematic review of animal models of depression
28 at the citation screening stage. We tested two independently developed ML approaches which used
29 different classification models and feature sets. We recorded the performance of the ML approaches
30 on an unseen validation set of papers using sensitivity, specificity and accuracy. We aimed to achieve
31 95% sensitivity and to maximise specificity. The classification model providing the most accurate
32 predictions was applied to the remaining unseen records in the dataset and will be used in the next
33 stage of the preclinical biomedical sciences systematic review. We used a cross validation technique
34 to assign ML inclusion likelihood scores to the human screened records, to identify potential errors
35 made during the human screening process (error analysis).

36 Results: ML approaches reached 98.7% sensitivity based on learning from a training set of 5749
37 records, with an inclusion prevalence of 13.2%. The highest level of specificity reached was 86%.
38 Performance was assessed on an independent validation dataset. Human errors in the training and
39 validation sets were successfully identified using assigned the inclusion likelihood from the ML
40 model to highlight discrepancies. Training the ML algorithm on the corrected dataset improved the
41 specificity of the algorithm without compromising sensitivity. Error analysis correction leads to a 3%
42 improvement in sensitivity and specificity, which increases precision and accuracy of the ML
43 algorithm.

44 Conclusions: This work has confirmed the performance and application of ML algorithms for
45 screening in systematic reviews of preclinical animal studies. It has highlighted the novel use of ML
46 algorithms to identify human error. This needs to be confirmed in other reviews, , but represents a
47 promising approach to integrating human decisions and automation in systematic review
48 methodology.

49

50 Key-words: machine learning, systematic review, analysis of human error, citation screening,
51 automation tools

52 Background:

53 The rate of publication of primary research is increasing exponentially within biomedicine [1].
54 Researchers find it increasingly difficult to keep up with new findings and discoveries even within a
55 single biomedical domain, an issue that has been emerging for a number of years [2]. Synthesising
56 research – either informally or through systematic reviews - becomes increasingly resource
57 intensive as searches retrieve larger and larger corpuses of potentially relevant papers for reviewers
58 to screen for relevance to the research question at hand.

59 This increase in rate of publication is seen in the animal literature. In an update to a systematic
60 review of animal models of neuropathic pain, 11,880 further unique records were retrieved in 2015,
61 to add to 33,184 unique records identified in a search conducted in 2012. In the field of animal
62 models of depression, the number of unique records retrieved from a systematic search increased
63 from 70,365 in May 2016 to 76,679 in August 2017.

64 The use of text-mining tools and machine learning (ML) algorithms to aid systematic review is
65 becoming an increasingly popular approach to reduce human burden and monetary resources
66 required and to reduce the time taken to complete such reviews [3; 4; 5]. ML algorithms are
67 primarily employed at the screening stage in the systematic review process. This screening stage
68 involves categorising records identified from the search into 'Relevant' or 'Not-Relevant' to the
69 research question, typically performed by two independent human reviewers with discrepancies
70 reconciled by a third. This decision is typically made on the basis of the title and abstract of an article
71 in the first instance. In previous experience at CAMARADES (Collaborative Approach to Meta-
72 Analysis and Review of Animal Data from Experimental Studies), screening a preclinical systematic
73 review with 33,184 unique search results took 9 months, representing (because of dual screening)
74 around 18 person months in total. Based partly on this, we estimate that a systematic review with
75 roughly 10,000 publications retrieved takes a minimum of 40 weeks. In clinical systematic reviews,
76 Borah and colleagues [6] showed the average clinical systematic review registered on PROSPERO

77 (International Prospective Register of Systematic Reviews) takes an average 67.3 weeks to complete.
78 ML algorithms can be employed to learn this categorisation ability, based on training instances that
79 have been screened by human reviewers [7].

80 Several applications of ML are possible. The least burdensome is when a review is being updated,
81 where categorisations from the original review are used to train a classifier, which is then applied to
82 new documents identified in the updated search [7; 8; 9]. When a screening is performed de novo,
83 without such previous collection, humans first categorise an initial set of search returns, which are
84 used to train an ML model. The performance of the model is then tested (either in a validation set or
85 with k fold cross validation); if performance does not meet a required threshold then more records
86 are screened, chosen either through random sampling or, using active learning [10], on the basis
87 either of those with highest uncertainty of predictions [11; 12] or alternatively from those most
88 likely to be included[13; 14; 15]. Here we use a de novo search with subsequent training sets
89 identified by random sampling, and we introduce a novel use of machine prediction, in identifying
90 human error in screening decisions.

91 Machine learning approaches have been evaluated in context of systematic reviews of several
92 medical problems including drug class efficacy assessment [7; 8; 12], genetic associations [9], public
93 health [16; 13], cost-effectiveness analyses [9], toxicology [3], treatment effectiveness [17; 18] and
94 nutrition [17]. To the best of our knowledge there have been only two attempts to apply such
95 techniques to reviews of preclinical animal studies [3; 19]. These can be broad and shallow reviews
96 or focussed and detailed reviews, and can have varying prevalence of inclusion.

97 Here we outline the ML approach taken to assist in screening a corpus for a broad and shallow
98 systematic review seeking to summarise studies using non-human animal models of depression,
99 based on a corpus of 70,365 records retrieved from two online biomedical databases. *In this paper,*
100 *our aim was to identify the amount of training data required for an algorithm to achieve the level of*

101 *performance of two independent human screeners, so that we might reduce the human resource*
102 *required.*

103 Sena and colleagues developed guidelines for the appraisal of systematic reviews of animal studies
104 [20]. These guidelines consider dual extraction by two independent human reviewers as a feature of
105 a high quality review. From a large corpus of reviews conducted by CAMARADES we estimate the
106 inter-screener agreement to be between 95% and 99%. Errors may occur at random (due to fatigue
107 or distraction) or, more consequentially, systematic error, which, if included in a training set, might
108 be propagated into a ML algorithm. Sources of systematic errors with certain types of records being
109 at greater risk of misclassification. To our knowledge the nature of this 5% residual human error in
110 systematic review methodology has not been formally investigated. The training data used for ML
111 categorisation is based on training instances that has been screened by two independent human
112 screeners.

113 *We therefore aimed to explore the use of established ML algorithms as part of a preclinical*
114 *systematic review framework at the classification stage, to investigate if the ML algorithms could be*
115 *used to improve the human gold standard by identifying human screening errors and thus improve*
116 *the overall performance of ML.*

117

118 **Methods:**

119 We applied two independent machine learning approaches to the screening of a large (70,365
120 records) systematic review. Because we did could not predict how many training instances would be
121 required we first selected 2000 records at random to provide the first training set. Of these, only
122 1993 were suitable due to data deposition errors. These were then screened by 2 human reviewers
123 with previous experience with reviews of animal studies, with a third expert reviewer reconciling any
124 differences. The resulting ML algorithms gave a score between 0 and 1. To ensure that the true

125 sensitivity was likely to be 95% or higher we chose as our cut-point the value for which the lower
126 bound of the 95% confidence interval of the observed sensitivity exceeded 95% when applied to the
127 unseen validation dataset. We then repeated this process adding a further 1000 randomly selected
128 (996 useable) citations to the training set; and then again adding a further 3000 randomly selected
129 (2760 useable) citations to the training set. At each stage, performance of the approaches was
130 assessed on a validation set of unseen documents, using a number of different metrics. Next, the
131 best performing algorithm was used to identify human errors in the training and validation sets by
132 selecting those with the largest discrepancy between the human decision (characterised as 0 for
133 exclude or 1 for include) and the machine prediction (a continuous variable between 0 and 1).
134 Performance of the approaches trained on the full 5749 records is reported here, and of each of the
135 iterations is available in Supplementary Materials 1. The error analysis was assessed on the net
136 reclassification index, and the performance of the ML approach is compared before and after
137 correcting the errors in human screening using AUC.

138

139 **Step 1: Application of ML tools to screening of a large preclinical systematic review.**

140

141 **Training Sets:**

142 70,365 potentially relevant records were identified from Pubmed and EMBASE. The search strings
143 were composed of the animal filters devised by the Systematic Review Center for Laboratory animal
144 Experimentation (SYRCLE) [21; 22], NOT reviews, comments, or letters AND a depression disorder
145 string (for full search strings see [23]). The training set and the validation set were chosen at random
146 from the 70,365 by assigning each record a random number using the RAND function in excel and
147 ranking them from smallest to largest. The training set consisted of 5749 records. The validation set
148 consisted of 1251 records. The training set and validation set were screened by two independent

149 human screeners with any discrepancies reconciled by a third independent human screener. The
150 human screening process involved an online tool (app.syrf.org), which randomly presents a reviewer
151 with a record, with the title and abstract displayed. The reviewer makes a decision about the record,
152 included (1) or excluded (0). A second reviewer is also randomly presented with records. If a record
153 receives two 'included' decisions, the screening for this record is considered complete. If reviewer 1
154 and reviewer 2 disagree, the record gets presented to a third reviewer who makes a decision. The
155 record then has an average inclusion score of 0.666 or 0.333. Any record that has an inclusion score
156 above 0.6 is included, those scoring less than 0.6 are excluded, and screening is considered
157 complete. Datasets are available on Zenodo, as described in "Availability of Data & Materials" below,
158 Performance was assessed at each level on a validation set of unseen records. The training and
159 validation set were selected consecutively from the initial random ordering. For the training set of
160 5749 records, the validation set was the subsequent 1251 records. This validation set had more than
161 150 "included" records, which can give reasonably precise 95% confidence intervals for sensitivity
162 and specificity.

163 << Insert Experimental Setup Diagram here >>

164 Figure 1. Diagram of the Layout of the Study.

165

166 **Feature Generation:**

167 First, documents in the training set were transformed into a representation appropriate for the
168 machine learning algorithms. Documents were created by concatenating the title and the abstract.
169 Every case (document) is represented by a fixed number of features, numerical quantities describing
170 certain properties that might be used by the classifier to extract rules and make predictions about
171 inclusion. The classifiers described below used generally similar approaches

172 We used “bag-of-words” (BoW) to characterise document titles and abstracts in both classifiers. To
173 account for the relative importance of words within a given document, and difference in words used
174 between documents we used ‘Term Frequency – Inverse Document Frequency’ (TD-IDF). This is
175 defined as:

176

$$tfidf(w_i, d_j) = tf(w_i, d_j) * \frac{|D|}{|\{d: w_i \in d\}|}$$

177 The score for the i-th word in context of the j-th document takes into account not only how many
178 times the word occurred there (tf), but also how many other documents (d) from the whole corpus
179 (D) contain it as well. This helps to reduce the score for words that are common for all documents
180 and therefore have little predictive power. This helps the classifier to focus on terms which help to
181 distinguish between documents, rather than on terms which occur frequently [24]. We allowed n-
182 grams; did not use stemming; and used the MySQL text indexing functionality “stopword” list to
183 remove frequently occurring words which provide little relevant information for classification
184 purposes. [25]

185 Because bag-of-words representation generates as many unigram features as there are words in the
186 collection (typically at least several thousand); and many more when using higher-order n-grams, we
187 used additional approaches. Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA)
188 represent textual data in a more efficient way. In LSI [26], the training set is represented as a matrix,
189 where rows correspond to documents, columns to terms (words or n-grams), while cells contain
190 frequency or TF/IDF score of a given term in a given document. The matrix is then decomposed
191 using a general matrix factorisation technique known as Singular Value Decomposition (SVD) and
192 truncated to the first n dimensions. Because of the properties of SVD the new features will be such
193 linear combinations of features of the old space that minimise the differences between the original
194 and the transformed space. In case of textual data it means that those words that frequently occur

195 in the same documents (probably because of the similar meaning) will be treated in the same way.
196 The n is set a-priori to a reasonably low value – usually a few hundred. LDA exploits distributional
197 similarities between words, but based on explaining document contents using a Bayesian network
198 [27]. This method is based on the premise that every document is a mixture of topics, which in turn
199 consist of related words. The correspondence between documents and topics and between topics
200 and words can be inferred via Gibbs sampling process. As a result, similarly to LSI, every document is
201 represented by a sequence of n numbers, indicating how related it is to every topic [28]. Unlike in
202 SVD, the model fitness to the data cannot be expressed through the amount of variance of the
203 original matrix it explains and the optimal number of topics may be different for every collection and
204 classification task. Following previous work in the domain [13] and the user guide for MALLET (the
205 tool we use for LDA, which recommends values between 200 and 400) we elected to generate 300
206 topics Here we use three feature sets, BoW, LDA and SVD (LSI) individually, in pairs and finally all
207 together; preliminary evaluation through the cross-validation on the training set suggests that
208 LDA+SVD and bag-of-words with a simple linear classifier deliver the most robust performance.

209

210

211 **Classifiers:**

212 Following the transformations made in feature selection, the documents are then used to train the
213 machine learning classifier. The classifier most commonly used for document classification in context
214 of systematic reviews [11; 13; 8; 9; 12; 14; 15; 17] is the Support Vector Machine (SVM) as it has
215 frequently been used for tasks involving text.). SVM is a supervised learning algorithm, learning to
216 classify new documents based on a training set of labelled documents [31]. This algorithm
217 represents training documents as points in a multi-dimensional space defined by all available
218 features. To be able to classify cases into positive and negative category, it seeks a hyperplane
219 dividing the space into one side corresponding to included documents and the other to excluded

220 ones. Based on the training data, the optimal hyperplane is constructed so that it maximises both
221 the number of training cases located on the “correct” side of decision boundary and their distance
222 from the plane (margin). The new, unseen, documents are then ranked according to their location
223 with respect to the boundary. Those far from it are confidently predicted as included or excluded,
224 according to which side of the plane they lie. The cases which the model has less confidence about
225 will be located close to the hyperplane. Logistic regression is a similar linear classifier, which instead
226 of hyperplane, seeks such coefficients of a linear combination of feature values that will give high
227 values for positive cases (included documents) and low for negative (excluded documents). Both of
228 these approaches could be enriched with feature selection elements to mitigate the problems with
229 multitude of features.

230

231 Three feature sets (BoW, LDA and SVD (LSI)) were tested on SVMs, logistic regression and random
232 forests [32]. The two algorithms described below performed best for this dataset of 70,365 records,
233 on the broad topic of preclinical animal models of depression.

234

235 **Approaches:**

236 Here, two approaches were developed independently, using different classification models and
237 feature representations, but sharing the linear classification principles.

238 Approach 1:

239 Approach one used a tri-gram ‘bag-of-words’ model for feature selection and implemented a linear
240 support vector machine with Stochastic Gradient Descent (SGD) as supported by the SciKit-Learn
241 python library [33]. This classifier was chosen it is efficient, scales well to large numbers of records,
242 and provides an easily interpretable list of probability estimates when predicting class membership
243 (i.e. scores for each document lying between 0 and 1). Efficiency and interpretability are important,

244 as this classifier is already deployed in a large systematic review platform [34], and any deployed
245 algorithm therefore needs not to be too computationally demanding, and its results understood by
246 users who are not machine learning specialists. The tri-gram feature selection approach without any
247 additional feature engineering also reflects the generalist need of deployment on a platform used in
248 a wide range of reviews: the algorithm needs to be generalisable across disciplines and literatures,
249 and not ‘over-fitted’ to a specific area. For example, the tri-gram “randomised controlled trial” has
250 quite different implications for classification compared with “randomised controlled trials” (i.e.
251 ‘trials’ in plural). The former might be a report of a randomised controlled trial; while the latter is
252 often found in reports of systematic reviews of randomised trials. Stemming would remove the ‘s’ on
253 trials and thus lose this important information. Here, the algorithm needs to be generalisable across
254 disciplines and literatures, and not be ‘over-fitted’ to a specific area. This approach aims to give the
255 best compromise between reliable performance across a wide range of domains and that achievable
256 from a workflow that has been highly tuned to a specific context.

257

258 Approach 2:

259 Approach 2 used a regularised logistic regression model built on LDA and SVD features. Namely, the
260 document text (consisting of title and abstract) was first lemmatised with the tool GENIA tagger [35]
261 and then converted into bag of words representation of unigrams, which was then used to create
262 two types of features. First, the word frequencies were converted into a matrix TF/IDF scores, which
263 was then decomposed via SVD implemented in scikit-learn library and truncated to the first 300
264 dimensions. Second, an LDA model was built using MALLET library [36], setting 300 as a number of
265 topics. As a result each document was represented by 600 features, and an L1-regularised logistic
266 regression model was built using glmnet package [37] in R statistical framework [38].

267 In this procedure every document is represented with a constant, manageable number of features,
268 irrespective of corpus or vocabulary size. As a result, we can use a relatively simple classification

269 algorithm and expect good performance with short processing time even for very large collections.
270 This feature is particularly useful when running the procedure numerous times in cross-validation
271 mode for error analysis (see below).

272

273 For a given unseen test instance, the logistic regression returns a score corresponding to the
274 probability of it being relevant according to the current model. An optimal cut-off score that gives
275 the best performance is calculated as described above.

276

277 **Assessing Machine Learning Performance:**

278 The facets of a machine learning algorithm performance that would be most beneficial to this field
279 of research are high sensitivity (see table 1), at a level comparable to the 95% we estimate is
280 achieved by two independent human screeners. We therefore need to be confident that the
281 sensitivity is 95% or higher, which we do by setting our cut point such that the lower bound of the
282 95% confidence interval of the observed sensitivity is 95% or higher. Once the level of sensitivity has
283 been reached, the aim is to maximise specificity, to reduce the number of irrelevant records
284 included by an algorithm. Although specificity at 95% sensitivity is our goal, we provide values of
285 other measures for better illustration of the performance.

286 *Performance metrics:*

287 Performance was assessed using sensitivity (or recall), specificity, precision, accuracy, and Work
288 Saved over Sampling (WSS) (see table 1), carried out in R (R version 3.4.2; [38]) using the 'caret'
289 package [39]. 95% Confidence Intervals were calculated using the efficient-score method [40]. Cut-
290 offs for were determined manually for each approach by taking the score that achieved 95%
291 sensitivity (including the lower 95% confidence level), and the specificity at this score was calculated.

292

Table 1. Equations used to assess performance of machine learning algorithms

Sensitivity or Recall	$TP / (TP+FN)$
Specificity	$TN / (TN+FP)$
Precision	$TP / (TP+FP)$
Accuracy	$(TP+TN) / (TP+FP+FN+TN)$
WSS@95%	$((TN+FN) / N) - (1.0 - 0.95)$

All equations from [5].

293

294

295 **Step 2: Application of ML tools to training datasets to identify human error.**

296 **Error Analysis Methods:**

297 The methodology for the error analysis was outlined in an *a priori* protocol, published on the
298 CAMARADES website on 18th December 2016 [41]. To generate the machine learning scores for the
299 set of records that were originally used to train the machine (5749 records), the non-exhaustive
300 cross-validation method, 5-fold validation, was used. This method involved randomly partitioning
301 the set of records into 5 equal sized subsamples. One subsample was set aside, and the remaining 4
302 subsamples were used to train the algorithm [42]. Thanks to this process, every record has a score
303 computed by a machine learning model built without including it in the training portion. These
304 scores were used to highlight discrepancies or disagreements between machine decision and human
305 decision. The documents were ordered by the machine assigned labels in order of predictive
306 probability, from most likely to be relevant to least likely to be relevant. The original human assigned
307 scores were placed next to the machine-assigned scores, to highlight potential errors in the human
308 decision. A single human reviewer (experienced in animal systematic reviews) manually reassessed
309 the records where discrepancies were highlighted starting with the most discrepant. To avoid
310 reassessing the full 5749 record dataset, a stopping rule was established such that if the initial
311 human decision was correct for five consecutive records, further records were not reassessed.

312

313 << Insert Error Analysis Diagram here Figure 2 >>

314 Figure 2. Error Analysis.

315 *The methodology for using cross-validation to assign ML predicted probability scores. The ML*
316 *predicted probability scores for the records were checked against the original human inclusion*
317 *decision.*

318

319 After the errors in the training set were investigated and corrected as described above a new model
320 was built on the updated training data. The outcome of error analysis is presented as reclassification
321 tables, the area under the curve (AUC) being used to compare the performance of the ML algorithm
322 trained on the 'old' training set of records, and the net reclassification index (NRI) [43] used to
323 compare the performance of the classifier built on the updated training data with the performance
324 of the classifier built on the original training data. The following equation was used:

325
$$\text{NRI}_{\text{binary outcomes}} = (\text{Sensitivity} + \text{Specificity})_{\text{second test}} - (\text{Sensitivity} + \text{Specificity})_{\text{first test}}$$

326 [44]

327 The AUC was calculated using the DeLong method in the 'pROC' package in R [45].

328 Further, we applied the same technique as above to identify human screening errors in the
329 validation dataset. Due to the small number of records in the validation set (1251 records), it was
330 assumed that every error would be likely to impact measured performance, and so the manual
331 screening of the validation set involved revisiting every record where the human and machine
332 decision were incongruent. The number of reclassified records was noted. The inter-rater reliability
333 of all screening decisions on training set and validation set between Reviewer 1 and Reviewer 2 were
334 analysed using the 'Kappa.test' function in the 'fmsb' package in R [46].

335

336 Results:

337 In this section we first describe the performance from the ML algorithms. We then show the results
338 from the analysis of human error, and finally describe the performance of the ML algorithm after
339 human errors in the training and validation set have been corrected.

340

341 Performance of Machine Learning Algorithms

342 Table 2 shows the performance of the two machine learning approaches from the SLIM (Systematic
343 Living Information Machine) collaboration. The desired sensitivity of 95% (including lower bound
344 95% CI) has been reach by both approaches. Both approaches reached 98.7% sensitivity based on
345 learning from a training set of 5749 records, with an inclusion prevalence of 13.2% (see below).
346 Approach 1 reached a higher specificity level of 86%. This is visualised on an AUC curve (figure 1).

347

Table 2. Performance of machine learning approaches on depression training dataset.

	Approach 1	Approach 2
Training Set Size	5749	5749
Optimal Cut-Off Score	0.1	0.07
Sensitivity	98.7%	98.7%
Upper 95% CI	0.997	0.997
Lower 95% CI	0.949	0.949
Specificity	86.0%	84.7%
Precision	50%	47.66%
Accuracy	1096/1251 = 87.6%	1081/1251= 86.4%
WSS@95%	0.705	0.693

348

349

350 Figure 3. Performance of Machine Learning Approaches.

351 *For the interactive version of this plot with cut-off values, see code and data at*
352 *[https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/ML-fig3.html)*
353 *[systematic-reviews/blob/master/ML-fig3.html](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/ML-fig3.html)*

354

355 < Figure 3 here >

356

357

358 **Error Analysis & Reclassification**

359 Cohen's κ was run to determine the interrater agreement of screening decisions between Reviewer
360 1 and Reviewer 2. $\kappa = 0.791$ (95% CI, 0.769 to 0.811), $p < 0.0001$, with 281 records requiring a third
361 reviewer decision. To assess whether machine learning algorithms can identify human error and
362 therefore improve the training data, error analysis was conducted. Seventy-five papers out of 5749
363 papers had predictive scores very far from the human assigned labels, so were reassessed to see if
364 these were due to human errors. Out of 75 rescreened papers, the machine corrected the human
365 decision 47 times. The machine was wrong, (i.e. the initial human decision was correct) 28 times.
366 The validation set was also rescreened. Ten papers out of the 1251 records were identified as
367 potential human errors. Out of 10 errors, the machine corrected 8 human decisions. These 8 records
368 were all falsely excluded by the human and were now included. The initial human decision was
369 correct twice.

370 To calculate human error in the training set, the number of errors identified (47) out of the training
371 set (5749 records) was calculated to be at least 0.8%. Of the 47 records reclassified, 11 records were
372 falsely included in the original screening process and were now correctly excluded, and 36 records
373 were falsely excluded in the original screening process and were now correctly included. The
374 machine correctly identified human screening errors, which were calculated to be just under 1% of

375 the dual screened training set. Forty-seven papers out of 760 were ‘correctly’ reclassified, 6% of the
376 included papers.

377 Similarly, the human error rate in the validation set (1251 records) was 0.6%. Again looking at the
378 prevalence of inclusion in this dataset (155/1251), which is 12.4%, the 8 records of out the now 163
379 were correctly reclassified which is 4.9% reclassified. All 8 records we falsely excluded in the original
380 screening process and are now correctly included.

381

382 Test 1: $98.7\% + 86\% = 184.7\%$

383 Test 2: $98.2\% + 89.3\% = 187.5\%$

384 **NRI = 3.2%**

385

386 We consider the updated validation set to be the new gold standard as 8 records were now
387 included. The confusion matrix for the performance of the machine learning algorithm after the
388 error analysis update on the training records is displayed below in table 3.

389

Table 3. Reclassification of records in validation after error analysis.

		Test 1 – Original Machine Learning Algorithms results			
Test 2 – Post-error analysis ML results	In	In	Out	Total	
	Out	153	160	116	276
		2	943	945	975
	Total	155	1096	1251	
	163	1088			

390

391 Analysing the human errors identified by the machine learning algorithm and correcting for these
392 errors and re-teaching the algorithm leads to improved performance of the algorithm, particularly its

393 sensitivity. This can save considerable human time in the screening stage of a systematic review.
394 Consider the remaining approximately 64,000 papers, if the ML algorithm results are 3% more
395 accurate, that is approximately 2000 papers that are correctly 'excluded' that would not be
396 forwarded for data extraction.

397

398 **After Error Analysis: Improving Machine Learning**

399 Using the error analysis technique above, of the 47 errors identified in the full training dataset of
400 5749 records, 0.8% were corrected. We retrained approach 1 on the corrected training set and
401 measured performance on the corrected validation set of 1251 records as we consider this to be the
402 'new' gold standard. The performance of the original approach 1 and updated approach 1 was
403 assessed on the corrected validation set of 1251 records. The performance of this retrained
404 algorithm in comparison to the performance of the original classifier 5 on the updated validation set
405 is shown in table 4.

406

Table 4. Performance of machine learning approach after error analysis.

	Updated Approach 1	Original Approach 1
Cut-Off	0.09	0.10
Sensitivity	98.7%	98.7%
Upper 95% CI of Sensitivity	0.997	0.997
Lower 95% CI of Sensitivity	0.949	0.949
Specificity	88.3%	86.7%
Precision	55.9%	52.61%
Accuracy	89.7%	88.2%
WSS@95%	961/ 1251 – (0.05) = 0.718	945/1251 – (0.05) = 0.705

407

408

409 Figure 4. Performance of Approach 1 after error analysis.

410 *The updated approach is retrained on the corrected training set after error analysis correction.*
411 *Performance on both the original and the updated approach is measured on the corrected validation*
412 *set (with error analysis correction). For the interactive version of this plot with exact cut-off values,*
413 *see code and data at [https://github.com/abannachbrown/The-use-of-text-mining-and-machine-](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html)*
414 *[learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html)*

415 < **Figure 4 here** >

416

417 We compared the area under the ROC curve for the original approach 1 and the updated approach
418 1. The AUC for the original approach 1 was 0.9272 (95% CI calculated using DeLong method; 0.914-
419 0.9404). The AUC for the updated approach 1 was 0.9355 (95% CI calculated using DeLong method;
420 0.9227-0.9483). DeLong's test to compare the AUC between the ROC of the two approaches was
421 applied, $Z = -2.3685$, $p = 0.0178$.

422 Discussion:

423 Document Classification:

424 We have shown machine learning algorithms to have high levels of performance, with 98.7%
425 sensitivity and 88.3% specificity; this sensitivity is comparable to two independent human screeners.
426 The objectives for selecting ML approaches in this project was to achieve a minimum 95% sensitivity
427 (including lower bound confidence intervals), to minimise the number of potentially relevant papers
428 which are wrongly excluded. Thereafter, algorithms were then chosen on the basis of their
429 specificity, to reduce the subsequent human time required to sort through and assess papers.

430 The two approaches have similar performance. The slight differences may reflect the method of
431 feature generation. These algorithms have high performance on this specific topic of animal models
432 of depression. As demonstrated previously, the performance of various classifiers can alter
433 depending on the topic and specificity of the research question [3].

434 In this study, the cut-off points were selected using the decisions on the validation set to achieve the
435 desired performance. Although this allows the measurement of the maximum possible gain using a
436 given approach in an evaluation setting, in practice (e.g. when updating a review), the true scores
437 would not be available. The problem of choosing a cut-off threshold, equivalent to deciding when to
438 stop when using a model for prioritising relevant documents, remains an open research question in
439 information retrieval. Based on their experience with a given tool, a reviewer may come up with a
440 heuristic fitting their workflow, e.g. if no new includes are seen in the 100 highest-ranked
441 documents, then everything else could be discarded as well. More sophisticated approaches have
442 also been tested [47], but they do not guarantee achieving a desired sensitivity level. It has to be
443 noted that ML-based prioritisation could be useful even if no cut-off is used and all documents are
444 screened manually, since seeing the relevant documents first can help to organise the process and
445 thus reduce the workload [5]. In a similar broad preclinical research project in neuropathic pain it
446 took 18 person months to screen 33,814 unique records – based on these numbers it would take an
447 estimated 40 person months to screen 70,365 unique records. Performance of machine learning
448 tools demonstrated in this paper can greatly reduce the amount of human resource needed for
449 initial title and abstract screening of a large corpus of records retrieved from a broad search.

450 We have applied the algorithm to the full dataset (remaining 63,365 records) and are in the process
451 of full-text screening. Following this process, it will allow a more in depth learning on the part of the
452 machine that it can apply to any updates to the search.

453

454 **Error Analysis:**

455 By using the ML algorithm to classify the likelihood of inclusion for each record in the training set,
456 we highlighted discrepancies between the human inclusion or exclusion decision and the machine
457 decision. Using this technique, we identified human errors, which were then corrected to update the
458 training set.

459 Human screening of the training set was conducted using the “majority vote” system; it is interesting
460 to consider the potential reasons for errors or ‘misclassifications’ arising in this process. Reviewers’
461 interpretation of the “breadth” of this wide review might be one contributing factor to
462 discrepancies. With a less clear cut-off, reviewers are unsure of where some articles should be
463 included. Discrepancies arising where Reviewer 1 was more inclusive and where Reviewer 2 was less
464 inclusive, thereby Reviewer 3 will be the deciding factor. A different approach whereby Reviewer 1
465 and 2 discuss discrepancies might be a pinpoint the exact reasons for misunderstandings or different
466 interpretations of the inclusion criteria. However, for larger projects when using a crowd-sourcing
467 approach with many individual people contributing to each Reviewer, this may not be a practical
468 solution.

469 We have successfully identified human screening errors which were calculated to be just under 1%
470 of the training set which was dual screened by two independent human reviewers. The prevalence
471 of inclusion in this training set is 13.2% (760 out of the 5749), so an error of 0.8% is likely to be
472 important Therefore errors of false inclusion or exclusion in the training sets may have a substantial
473 impact on the learning of the ML algorithm. This error analysis results in a 3% increase or change in
474 sensitivity and specificity, with increased precision, accuracy, and work saved over sampling of the
475 algorithm. We observed an increase in specificity of 1.6% without compromise to sensitivity. In a
476 systematic review with this number of records this saves considerable human resources, as the
477 number of records required to screen reduces by at least 1125.

478 This error analysis was an initial pilot with stopping criteria where if the initial human decision was
479 correct five consecutive times, further records were not reassessed. It is possible and likely that
480 there are further errors in the human screened training set. A more in-depth analysis of the training
481 dataset, investigating every instance where the human and machine decision were incongruent,
482 might identify more errors and further increase the precision and accuracy of machine learning
483 approaches, further reducing human resources required for this stage of systematic review. We have

484 shown here that even with minimal intervention (only assessing incongruent records until the
485 original human decision was correct 5 consecutive times), the performance of ML approaches can be
486 improved.

487

488 **Limitations & Future Directions:**

489 Here we show the best performing algorithms for this dataset with a broad research question. Other
490 dissimilar research questions or topics may require different levels of training data to achieve the
491 same levels of performance, or may require different topic modelling approaches or classifiers. The
492 best performing algorithm, outlined in this paper, is being applied in an ongoing research project,
493 therefore the 'true' inclusion and exclusion results for the remaining 63365 records is not yet known.
494 The 'true' results will unfold with the fullness of time.

495 These machine learning algorithms are deployed in an existing systematic review online platform,
496 EPPI-Reviewer [34], and are in the process of being integrated into the Systematic Review Facility
497 (SyRF) tool, which is focused on the preclinical domain (www.app.syrf.org). This will improve the
498 ease of use of machine learning functions for systematic reviewers, increase the usage of machine
499 learning algorithms for systematic review and significantly reduce the amount of human resources
500 required to conduct systematic review across a range of topics. By allowing a degree of user control
501 over which classifiers and the levels of performance are required for each specific research project.
502 With a broad collaboration such as SLIM we aim to test many ML algorithms across a range of
503 research topics to identify which classifiers perform best under which circumstances, to be able to
504 provide recommendations to users of SyRF.

505

506 This paper outlines a pilot approach to using machine learning algorithms to identify human errors in
507 current systematic review methodology. Future research can investigate this concept more

508 thoroughly by setting up a more comprehensive experimental design. After further investigation into
509 the extent of human error in dual reviewing, the picture will be clearer as to the scale of human
510 error and to what extent a machine learning algorithm can identify and aid in rectifying this. These
511 tools can could be integrated into systematic review platforms, such as SyRF (www.app.syrf.org),
512 and may provide feedback to the systematic reviewer during screening, and could ultimately flag
513 incorrectly screened records as the human screens them for inclusion in a dataset for machine
514 training.

515

516 **Conclusions:**

517 We have demonstrated that machine learning techniques can be successfully applied to an ongoing,
518 broad pre-clinical systematic review. We have demonstrated that machine learning techniques can
519 be used to identify human errors in the training and validation datasets. We have demonstrated that
520 updating the learning of the algorithm after error analysis improves performance. This error analysis
521 technique requires further detailed elucidation and validation. These machine learning techniques
522 are in the process of being integrated into existing systematic review applications to enable more
523 wide-spread use. In future, machine learning and error analysis techniques that are optimised for
524 different types of review topics and research questions can be applied seamlessly within the existing
525 methodological framework.

526

527

528 **List of Abbreviations:**

529

530 1. Area Under the Curve (AUC)

531 2. Bag-of-Words (BoW)

- 532 3. Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental
533 Studies (CAMARADES)
- 534 4. Latent Dirichlet Allocation (LDA)
- 535 5. Latent Semantic Indexing (LSI)
- 536 6. Machine learning (ML)
- 537 7. Net Reclassification Index (NRI)
- 538 8. PROSPERO (International Prospective Register of Systematic Reviews)
- 539 9. Singular Value Decomposition (SVD)
- 540 10. SLIM (Systematic Living Information Machine) collaboration
- 541 11. Stochastic Gradient Descent (SGD)
- 542 12. Support Vector Machine (SVM)
- 543 13. Systematic Review Center for Laboratory animal Experimentation (SYRCLE)
- 544 14. Systematic Review Facility (SyRF)
- 545 15. Term Frequency – Inverse Document Frequency (TD-IDF)
- 546 16. Work Saved over Sampling (WSS)

547

548 **Declarations:**

549 **Ethical Approval:**

550 Not applicable.

551

552 **Availability of Data & Materials:**

553 The training and validation datasets, error analysis datasheets, as well as all the records in the
554 depression systematic review are available on Zenodo: DOI [10.5281/zenodo.60269](https://doi.org/10.5281/zenodo.60269)

555 The protocol for the systematic review of animal models of depression is available from:

556 <http://onlinelibrary.wiley.com/doi/10.1002/ebm2.24/pdf>

557 The protocol for the Error Analysis is available via the CAMARADES website and can be accessed
558 directly from this link: <https://drive.google.com/file/d/0BxckMffc78BYTm0tUzJJZkc1alk/view>

559 The results of the classification algorithms and the R code used to generate the results is available on
560 GitHub: [https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews)
561 [algorithms-in-systematic-reviews](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews).

562

563 **Competing Interests:**

564 The authors declare that they have no competing interests.

565

566 **Funding:**

567 This work is supported by a grant from the Wellcome Trust & Medical Research Council (Grant
568 Number: MR/N015665/1). ABB is supported by a scholarship from the Aarhus-Edinburgh Excellence
569 in European Doctoral Education Project.

570

571 **Authors' Contributions:**

572 ABB screened and analysed the datasets. JT & PB conducted feature selection and built the
573 classifiers. ABB, JT & PB wrote the manuscript. ABB, JT, PB, MRM, JL, AR & SA devised the study. JL,
574 MRM & SA supervised the study. All authors edited and approved the final manuscript.

575

576 **Acknowledgements:**

577 Thank you to Kaitlyn Hair, Paula Grill, Monica Dingwall & Zsanett Bahor for their assistance in
578 second screening the training and validation datasets.

579 **References:**

580

581 [1] Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based
582 on the number of publications and cited references. *Journal of the Association for Information
583 Science and Technology*, 66(11), 2215-2222.

584 [2] Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., ... & Smalheiser, N. R. (2010,
585 November). Evidence-based medicine, the essential role of systematic reviews, and the need for
586 automated text mining tools. In *Proceedings of the 1st ACM international Health Informatics
587 Symposium* (pp. 376-380). ACM.

588 [3] Howard, B.E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M.R., Holmgren, S., Pelch, K.E.,
589 Walker, V., Rooney, A.A. and Macleod, M., 2016. SWIFT-Review: a text-mining workbench for
590 systematic review. *Systematic reviews*, 5(1), p.87.

591 [4] Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F. and Coiera, E., 2014. Systematic
592 review automation technologies. *Systematic reviews*, 3(1), p.74.

593 [5] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining
594 for study identification in systematic reviews: a systematic review of current approaches. *Systematic
595 reviews*, 4(1), 5.

596 [6] Borah, R., Brown, A.W., Capers, P.L., *et al.* (2017). Analysis of the time and workers needed to
597 conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ
598 Open*;7:e012545. doi: 10.1136/bmjopen-2016-012545

599 [7] Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P. Y. (2006). Reducing workload in systematic
600 review preparation using automated citation classification. *Journal of the American Medical
601 Informatics Association*, 13(2), 206–219. <http://doi.org/10.1197/jamia.M1929>

- 602 [8] Cohen, A. M., Ambert, K., & McDonagh, M. (2012). Studying the potential impact of automated
603 document classification on scheduling a systematic review update. *BMC Medical Informatics and*
604 *Decision Making*, 12(1), 33.
- 605 [9] Wallace, B. C., Small, K., Brodley, C. E., Lau, J., Schmid, C. H., Bertram, L., ... Trikalinos, T. A. (2012).
606 Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining.
607 *Genetics in Medicine*, 14(7), 663–669.
- 608 [10] Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In
609 *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development*
610 *in Information Retrieval* (pp. 3–12).
- 611 [11] Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). Active learning for biomedical
612 citation screening. In *Proceedings of the 16th ACM SIGKDD International conference on Knowledge*
613 *Discovery and Data mining* (pp. 173-182). ACM.
- 614 [12] Liu, J., Timsina, P., & El-Gayar, O. (2016). A comparative analysis of semi-supervised learning:
615 The case of article selection for medical systematic reviews. *Information Systems Frontiers*, 1–13.
616 <http://doi.org/10.1007/s10796-016-9724-0>
- 617 [13] Miwa, M., Thomas, J., O'Mara-Eves, A. and Ananiadou, S., 2014. Reducing systematic review
618 workload through certainty-based screening. *Journal of biomedical informatics*, 51, pp.242-253.
- 619 [14] Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. a. (2012). Deploying an interactive
620 machine learning system in an evidence-based practice center: abstractkr. *Proceedings of the 2nd*
621 *ACM SIGHIT Symposium on International Health Informatics - IHI '12*, 819.
622 <http://doi.org/10.1145/2110363.2110464>
- 623 [15] Kontonatsios, G., Brockmeier, A. J., Przybyła, P., McNaught, J., Mu, T., Goulermas, J. Y., &
624 Ananiadou, S. (2017). A semi-supervised approach using label propagation to support citation
625 screening. *Journal of Biomedical Informatics*, 72, 67–76. <http://doi.org/10.1016/j.jbi.2017.06.018>

- 626 [16] Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., ... Thomas, J.
627 (2014). Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening
628 workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1), 31–49.
- 629 [17] Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated
630 screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11, 1.
- 631 [18] Rathbone, J., Hoffmann, T., & Glasziou, P. (2015). Faster title and abstract screening? Evaluating
632 Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic*
633 *Reviews*, 4(1), 80. <http://doi.org/10.1186/s13643-015-0067-6>
- 634 [19] Liao, J., Ananiadou, S., Currie, G.L., Howard, B.E., Rice, A., Sena, E.S., Thomas, J., Varghese, A.,
635 Macleod, M.R. (2018) Automation of citation screening in pre-clinical systematic reviews
636 bioRxiv 280131; doi: <https://doi.org/10.1101/280131>
- 637 [20] Sena, E. S., Currie, G. L., McCann, S. K., Macleod, M. R., & Howells, D. W. (2014). Systematic
638 reviews and meta-analysis of preclinical studies: why perform them and how to appraise them
639 critically. *Journal of Cerebral Blood Flow & Metabolism*, 34(5), 737-742.
- 640 [21] de Vries, R. B., Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2014). Letter
641 to the Editor. *Laboratory Animals*, 48(1), 88-88. <https://doi.org/10.1177/0023677213494374>. ;
- 642 [22] Hooijmans, C. R., Tillema, A., Leenaars, M., & Ritskes-Hoitinga, M. (2010). Enhancing search
643 efficiency by means of a search filter for finding all studies on animal experimentation in
644 PubMed. *Laboratory Animals*, 44(3), 170–175. <http://doi.org/10.1258/la.2010.009117>
- 645 [23] Bannach-Brown, A., Liao, J., Wegener, G., & Macleod, M.R. (2016). Understanding in vivo
646 modelling of depression in non-human animals: a systematic review protocol. *Evidence-based*
647 *Preclinical Medicine*, 3(2), 20-27.

- 648 [24] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval,
649 Cambridge University Press: USA.
- 650 [25] Oracle (2018). MySQL 8.0 Reference Manual: Full-Text Stopwords. Accessed from:
651 <https://dev.mysql.com/doc/refman/8.0/en/fulltext-stopwords.html> on: 14/05/2018
- 652 [26] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by
653 latent semantic analysis. *Journal of the American society for information science*, 41(6), 391.
- 654 [27] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine*
655 *Learning research*, 3(Jan), 993-1022.
- 656 [28] Kontonatsios, G., Brockmeier, A. J., Przybyla, P., McNaught, J., Mu, J., Goulermas, J.Y., &
657 Ananiadou, S. (2017), A semi-supervised approach using label propagation to support citation
658 screening, *Journal of Biomedical Informatics*, 72, 67-76.
- 659 [29] Hashimoto, K., Kontonatsios, G., Miwa, M., & Ananiadou, S. (2016). Topic detection using
660 paragraph vectors to support active learning in systematic reviews. *Journal of biomedical*
661 *informatics*, 62, 59-65.
- 662 [30] Mu, T., Goulermas, J. Y., Korkontzelos, I., & Ananiadou, S. (2016). Descriptive document
663 clustering via discriminant learning in a co-embedded space of multilevel similarities. *Journal of the*
664 *Association for Information Science and Technology*, 67(1), 106-133.
- 665 [31] Mertsalov, K., & McCreary, M. (2009). Document classification with support vector machines.
666 Rational Enterprise: White Paper. Accessed from:
667 [http://www.rationalenterprise.com/assets/content/files/Classification_with_Support_Vector_Machi](http://www.rationalenterprise.com/assets/content/files/Classification_with_Support_Vector_Machines.pdf)
668 [nes.pdf](http://www.rationalenterprise.com/assets/content/files/Classification_with_Support_Vector_Machines.pdf), on: 05/09/2016.
- 669 [32] Breiman L (2001). "Random Forests". *Machine Learning*. 45 (1): 5–32.
670 doi:10.1023/A:1010933404324.

- 671 [33] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P,
672 Weiss R, Dubourg V, Vanderplas J. (2011). Scikit-learn: Machine learning in Python. *Journal of*
673 *machine learning research*.12(Oct):2825-30.
- 674 [34] Thomas, J., Brunton, J., Graziosi, S., (2010). EPPI-Reviewer 4.0: software for research synthesis.
675 EPPI-Centre Software. London: Social Science Research Unit, Institute of Education.
- 676 [35] Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. I. (2005).
677 Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on*
678 *Informatics* (pp. 382-392). Springer, Berlin, Heidelberg.
- 679 [36] McCallum, Andrew Kachites. (2002). "MALLET: A Machine Learning for Language Toolkit."
680 <http://mallet.cs.umass.edu>.
- 681 [37] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear
682 models via coordinate descent. *Journal of statistical software*, 33(1), 1.
- 683 [38] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2013.
- 684 [39] Kuhn, M., (2017) "The caret package". <https://topepo.github.io/caret/>
- 685 [40] Newcombe, R.G. (1998). "Two-Sided Confidence Intervals for the Single Proportion: Comparison
686 of Seven Methods," *Statistics in Medicine*, 17, 857-872
- 687 [41] Bannach-Brown, A., Thomas, J., Przybyła, P., Liao, J., (2016). "Protocol for Error Analysis:
688 Machine learning and text mining solutions for systematic reviews of animal models of depression".
689 Published on CAMARADES Website. www.CAMARADES.info. Direct Access:
690 <https://drive.google.com/file/d/0BxckMffc78BYTm0tUzJJZkc1alk/view>
- 691 [42] Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in
692 prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3),
693 569-575.

- 694 [43] Kerr, K. F., Wang, Z., Janes, H., McClelland, R. L., Psaty, B. M., & Pepe, M. S. (2014). Net
695 Reclassification Indices for Evaluating Risk-Prediction Instruments: A Critical Review. *Epidemiology*
696 (*Cambridge, Mass.*), 25(1), 114–121. <http://doi.org/10.1097/EDE.000000000000018>
- 697 [44] Pencina, M.J., D'Agostino, R.B. and Vasan, R.S., (2008). Evaluating the added predictive ability of
698 a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in*
699 *medicine*, 27(2), 157-172.
- 700 [45] Robin, X. (2017). “pROC” Package. <https://cran.r-project.org/web/packages/pROC/pROC.pdf>
- 701 [46] Nakazawa, M., (2018). “fmsb” Package. <https://cran.r-project.org/web/packages/fmsb/fmsb.pdf>
- 702 [47] Cormack, G. V., & Grossman, M. R. (2016). Engineering Quality and Reliability in Technology-
703 Assisted Review. In Proceedings of the 39th International ACM SIGIR conference on Research and
704 Development in Information Retrieval - SIGIR '16 (pp. 75–84). New York, New York, USA: ACM Press.
705 <http://doi.org/10.1145/2911451.2911510>

706

707

708

709 Figure Titles & Legends:

710

711 Figure 1. Diagram of the Layout of the Study.

712

713

714 Figure 2. Error Analysis.

715 *The methodology for using cross-validation to assign ML predicted probability scores. The ML*
716 *predicted probability scores for the records were checked against the original human inclusion*
717 *decision.*

718

719 Figure 3. Performance of Machine Learning Approaches.

720 *For the interactive version of this plot with cut-off values, see code and data at*
721 *[https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/ML-fig3.html)*
722 *[systematic-reviews/blob/master/ML-fig3.html](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/ML-fig3.html)*

723

724 Figure 4. Performance of Approach 1 after Error Analysis.

725 *The updated approach is retrained on the corrected training set after error analysis correction.*
726 *Performance on both the original and the updated approach is measured on the corrected validation*
727 *set (with error analysis correction). For the interactive version of this plot with exact cut-off values,*
728 *see code and data at [https://github.com/abannachbrown/The-use-of-text-mining-and-machine-](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html)*
729 *[learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html](https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html)*

730

Fig 1. Diagram of Experimental Setup.

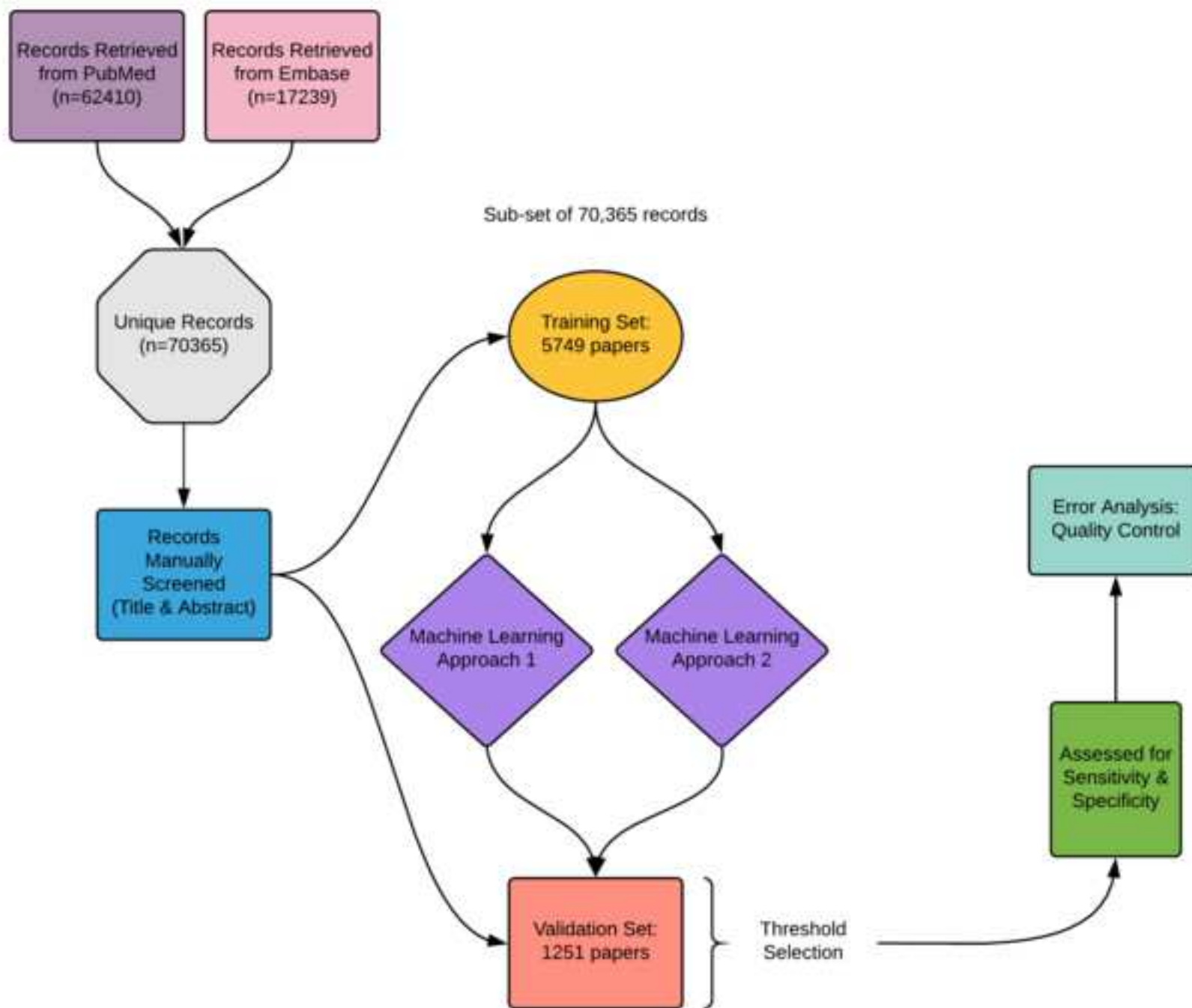


Fig 2. Diagram displaying the methodology for using cross-validation to assign ML predicted probability scores. The ML predicted probability scores for the records were

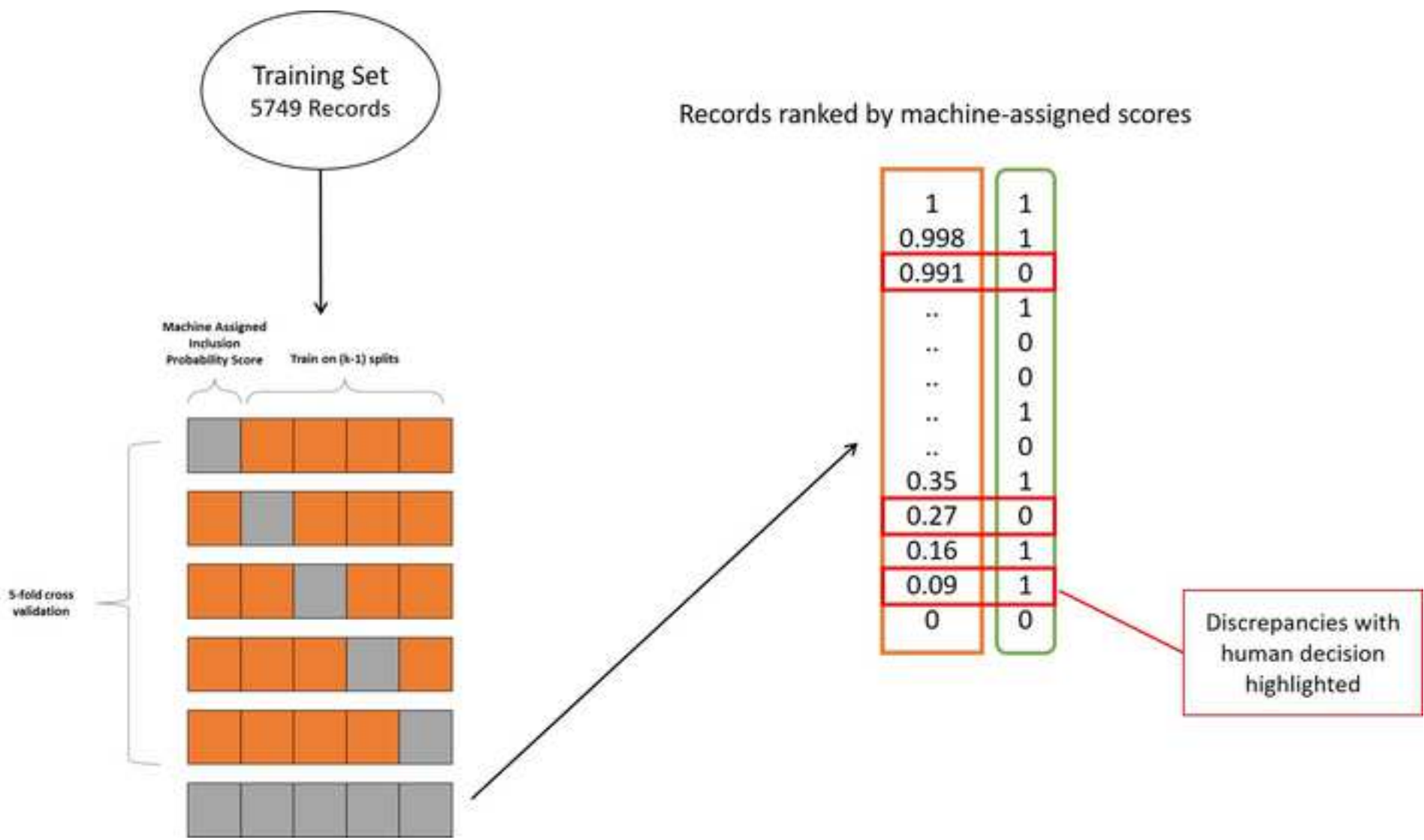


Fig 3. Performance of Machine Learning Approaches For the interactive version of this plot with cut-off values.

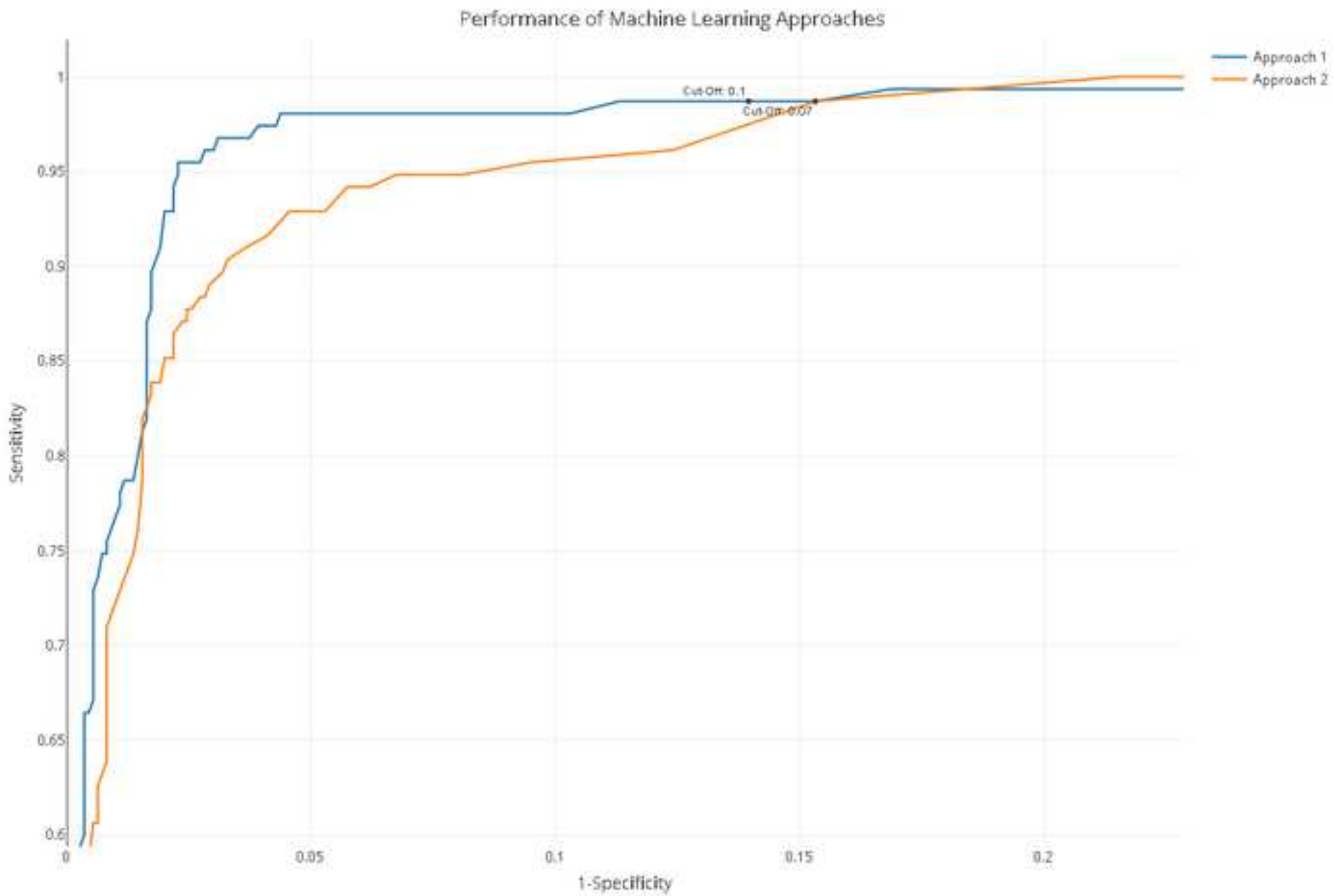


Fig 4. Performance of Approach 1 after error analysis. The updated approach is retained on the corrected training set after error analysis correction. Performance on

