

1 Machine learning to increase the efficiency
2 of a literature surveillance system: a
3 performance evaluation
4 .

5 Cynthia Lokker^{1*}, Wael Abdelkader¹, Elham Bagheri¹, Rick Parrish¹, Chris Cotoi¹, Tamara
6 Navarro¹, Federico Germini^{1,2}, Lori-Ann Linkins², R. Brian Haynes^{1,2}, Lingyang Chu³, Muhammad
7 Afzal⁴, Alfonso Iorio^{1,2},

8

9 ¹Health Information Research Unit, Department of Health Research Methods, Evidence, and
10 Impact, McMaster University, Hamilton, Ontario, Canada

11 ²Department of Medicine, McMaster University, Hamilton, Ontario, Canada

12 ³Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada

13 ⁴Department of Computing and Data Science, Birmingham City University, Birmingham, UK

14 *Corresponding author:

15 lokkerc@mcmaster.ca

16

17

18 ABSTRACT

19 Background: Given suboptimal performance of Boolean searching to identify methodologically
20 sound and clinically relevant studies in large bibliographic databases such as MEDLINE,
21 exploring the performance of machine learning (ML) tools is warranted.

22 Objective: Using a large internationally recognized dataset of articles tagged for methodological
23 rigor, we trained and tested binary classification models to predict the probability of clinical
24 research articles being of high methodologic quality to support a literature surveillance
25 program.

26 Materials and Methods: Using an automated machine learning approach, over 12,000 models
27 were trained on a dataset of 97,805 articles indexed in PubMed from 2012-2018 which were
28 manually appraised for rigor by highly trained research associates with expertise in research
29 methods and critical appraisal. As the dataset is unbalanced, with more articles that do not
30 meet criteria for rigor, we used the unbalanced dataset and over- and under-sampled datasets.
31 Models that maintained sensitivity for high rigor at 99% and maximized specificity were
32 selected and tested in a retrospective set of 30,424 articles from 2020 and validated
33 prospectively in a blinded study of 5253 articles.

34 Results: The final selected algorithm, combining a model trained in each dataset, maintained
35 high sensitivity and achieved 57% specificity in the retrospective validation test and 53% in the
36 prospective study. The number of articles needed to read to find one that met appraisal criteria
37 was 3.68 (95% CI 3.52 to 3.85) in the prospective study, compared with 4.63 (95% CI 4.50 to 4.77)
38 when relying only on Boolean searching.

39 Conclusions: ML models improved by approximately 25% the efficiency of detecting high quality
40 clinical research publications for literature surveillance and subsequent dissemination to
41 clinicians and other evidence users.

42

43 Keywords: bioinformatics; machine learning; evidence-based medicine; literature retrieval; medical
44 informatics; Natural Language Processing; biomedical informatics.

45 INTRODUCTION

46 The increasing pace with which medical literature is produced is well established. So is the challenge in
47 filtering the high-quality, clinically relevant articles from those not ready for clinical practice. Validated
48 search strategies that filter articles by research methods, such as systematic reviews (1) and randomized
49 controlled trials (2) have been integrated into biomedical databases to improve the efficiency of finding
50 evidence. Though these strategies perform well, maximizing sensitivity or recall (i.e., the proportion of
51 all on-target articles that are retrieved) comes at the cost of lower specificity (the proportion of off-
52 target articles that are excluded from the result set) and precision (positive predictive value, i.e., the
53 proportion of retrieved articles that are on target). Low specificity leads to significant time and
54 resources needed to manually review and appraise the quality of the studies reported in the articles.

55 More recently, machine learning (ML) approaches have been applied to retrieve high quality evidence
56 from the biomedical literature (3). There are several types of ML approaches which are determined by
57 the mathematical method used (4,5). The most common approaches are supervised ML, unsupervised
58 ML, ensemble learning, and neural networks. Supervised ML relies on a prelabelled training dataset to
59 provide the machine with the necessary input to make accurate predictions (6). There are several
60 supervised ML algorithms used to train models. For example, authors have used Artificial Neural
61 Networks, Decision Tree, K-Nearest Neighbour, Naïve Bayes, Random Forest, and Support Vector

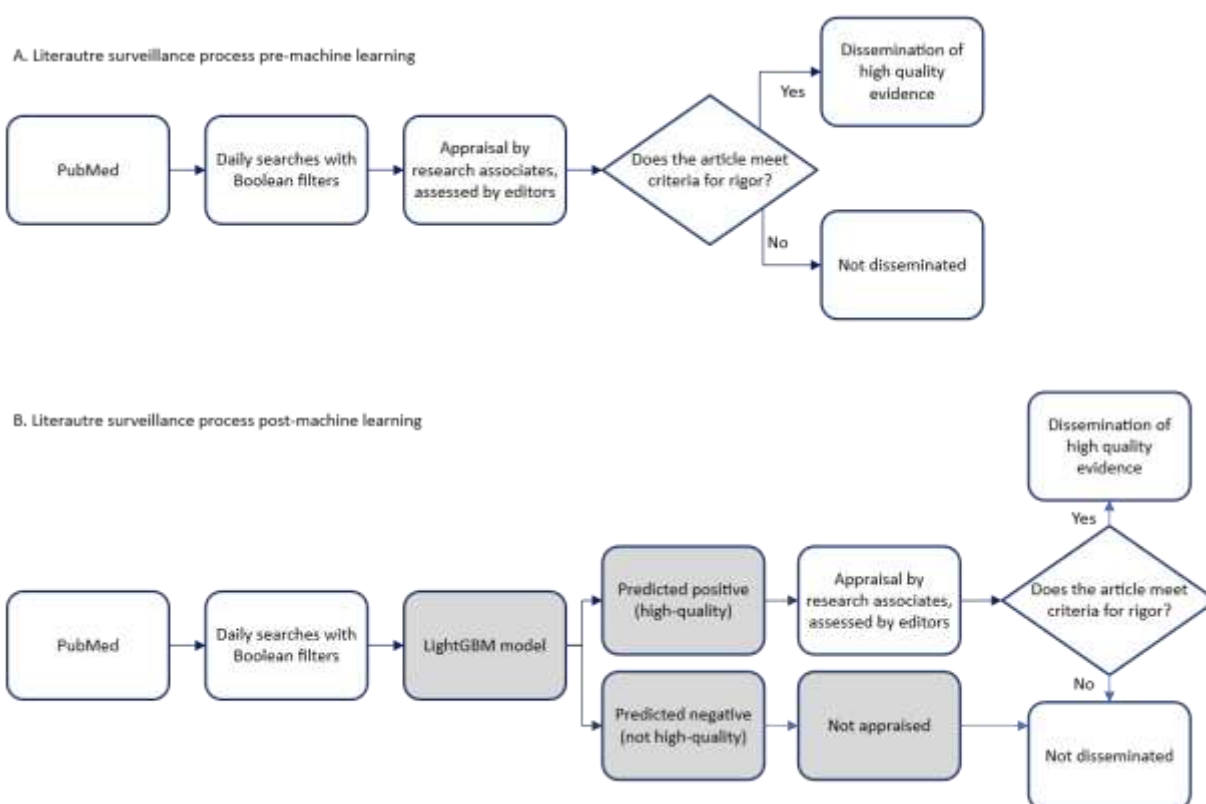
62 Machine algorithms for predicting diseases (7). Automated machine learning (AutoML) iterates, selects,
63 and optimizes ML models at multiple steps of the process (8) by automating the selection of promising
64 algorithms, hyperparameter tuning, pre-processing, and features selection (8,9). The system searches
65 through possible model and hyperparameter configurations and selects those that perform best on the
66 given task. This reduces the time needed to train and test models and inaccuracies in the model that
67 may arise from human errors and bias.

68 At the McMaster Health Research Information Unit (HiRU), we evaluate, at the time of publication,
69 original studies, systematic reviews, and evidence-based guidelines in ~120 top health care journals and
70 research synthesis services (10) through the Premium Literature Service (PLUS). Candidate studies are
71 retrieved from PubMed daily using validated, highly sensitive Boolean search strategies to maximize
72 recall of high-quality studies. Articles are then manually appraised by highly trained research associates
73 with expertise in health research methods and critical appraisal to determine if they meet explicit
74 criteria for scientific merit. Those that meet the criteria are reviewed by a clinical editor and rated for
75 clinical relevancy and newsworthiness by a cadre of >6000 clinicians worldwide (11). The identified
76 research is packaged into several evidence information services, tailored to the needs of knowledge
77 users (e.g., publishers, authors, guideline developers, policy makers) and end users (e.g., clinicians)
78 (Figure 1A). This process and application of critical appraisal criteria is consistent with the methods the
79 HiRU team used in creating the HEDGES dataset of articles published in 2006 which has been used in the
80 development of numerous Boolean search strategies (2,12–14) and ML models (15–17). Through PLUS,
81 we have curated a database of articles manually classified according to methodological rigor and clinical
82 relevance since 2012. For example, in 2019, 59 052 items indexed in PubMed in the journal set were
83 reduced to 17 349 (29.3%) by the sensitive Boolean search filters, all of which were manually appraised
84 by research associates. Of these, 3749 articles met critical appraisal criteria (18), giving a number of
85 articles needed to read to identify one that met criteria (number needed to read; NNR), measured as the

86 inverse of precision, of 4.63 (95% CI 4.50 to 4.77). The NNR provides a measure of human effort required
87 during the critical appraisal step and a proxy for efficiency; a lower NNR reflects fewer off target articles
88 and reduced time and effort for research associates to screen them out.

89 Maintaining PLUS is a resource-intensive activity, and currently limited to a subset of ~120 journal titles
90 (11). Reducing the NNR, by having staff focus on fewer articles that are more likely to meet criteria for
91 rigor, while maintaining high recall (sensitivity >99%), can improve the efficiency of the process. This is
92 particularly important as PLUS has expanded to include appraisal of all COVID-19 publications since
93 March 2020 across all of PubMed.

94 Objective: To improve the efficiency of identifying high-quality clinical research to support a literature
95 surveillance service while maintaining sensitivity at 99% (to ensure high quality articles are not missed)
96 and reducing the NNR.



97
98

99 Figure 1. Illustration of the literature surveillance process A. before and B. after addition of a machine
100 learning algorithm to predict quality of the article.

101 MATERIALS AND METHODS

102

103 We performed a retrospective study using a labelled dataset of articles that were critically appraised for
104 methodologic rigor and clinical relevance to train, validate, and test algorithms that predict the
105 likelihood of a clinical article meeting appraisal criteria for rigor. We used automated ML as an efficient
106 approach to training multiple models. Selected models were prospectively evaluated by having trained
107 research associates, blinded to model predictions, appraise incoming articles in the literature
108 surveillance program, as a test of the external validity of model predictions.

109 *Quality standard database*

110 We define high-quality or rigor as meeting at least all critical appraisal criteria for a particular article
111 type (review, guideline, original study) or purpose category (treatment, diagnosis, prognosis, etiology for
112 harm primary prevention, quality improvement, economics, or clinical prediction guides) based on
113 established evidence assessment criteria (18). The critical appraisal step, conducted manually by
114 research associates, has previously documented high inter-rater agreement ($\kappa > 0.80$ for all
115 categories)(10). Articles that meet methodological criteria are then reviewed by a clinical editor with
116 advanced research methods training and at least three members of an online community of >4000
117 clinicians who rate the methodologically rigorous articles for clinical relevance and newsworthiness (11).
118 Over the course of two decades, we have reviewed more than 500,000 articles and have curated an
119 internal database that also includes articles that did not meet methodological rigor criteria or clinical
120 relevance or newsworthiness. Notably, the database is unbalanced, with about 4.5 times the number of
121 articles that fail to meet methodologic rigor or clinical relevance than those that pass. The growing
122 database now includes articles on COVID-19 indexed in PubMed not limited to the core journal set.

123 *Model training and performance*

124 Our approach to model training was to use automated machine learning (AutoML), a process that allows
125 for running multiple sequential experiments with varying settings. The process, depicted in Figure 2,
126 automatically iterates model training using the combinations of pre-processing options, weighting
127 methods, feature selection, and hyper-parameters listed in Table 1, and optimizes selections to identify
128 the best performing combinations—essentially the approach optimizes performance and abandons
129 steps that do not lead to better performing models. The performance of an AutoML system depends on
130 the quality of the data and the specific task at hand. We chose AutoML for this study since our dataset
131 was of high quality as it was reviewed and appraised by human experts, and we wanted to remove our
132 biases and grow our understanding of the best approaches for our dataset. AutoML allowed for
133 experimentation while developing expertise. We used Microsoft’s ML.NET AutoML (19) to train and test
134 binary classification models that predicted if an article was of high-quality or not to help us get a highly
135 optimized model, driven by a set goal of improving specificity while maintaining sensitivity above 99%.
136 We tested weighting by term frequency (TF), inverse document frequency (IDF), and TF-IDF to account
137 for frequency of words within titles and abstracts of articles and their frequency across a dataset. A
138 convenience sample of algorithms available in the public domain and in ML.NET that provided a
139 probability score as an output measure was selected for training. This allowed us to set a threshold of
140 99% sensitivity rather than the default 50%. The available algorithms at the time of training were
141 FastTree, Limited-memory Broyden-Fletcher-Goldfarb-Shanno Logistic Regression, Stochastic Dual
142 Coordinate Ascent Logistic Regression, Stochastic Gradient Descent Calibrated Logistic Regression,
143 Symbolic SGD Logistic Regression, and Light Gradient Boosting Machine (LightGBM) .

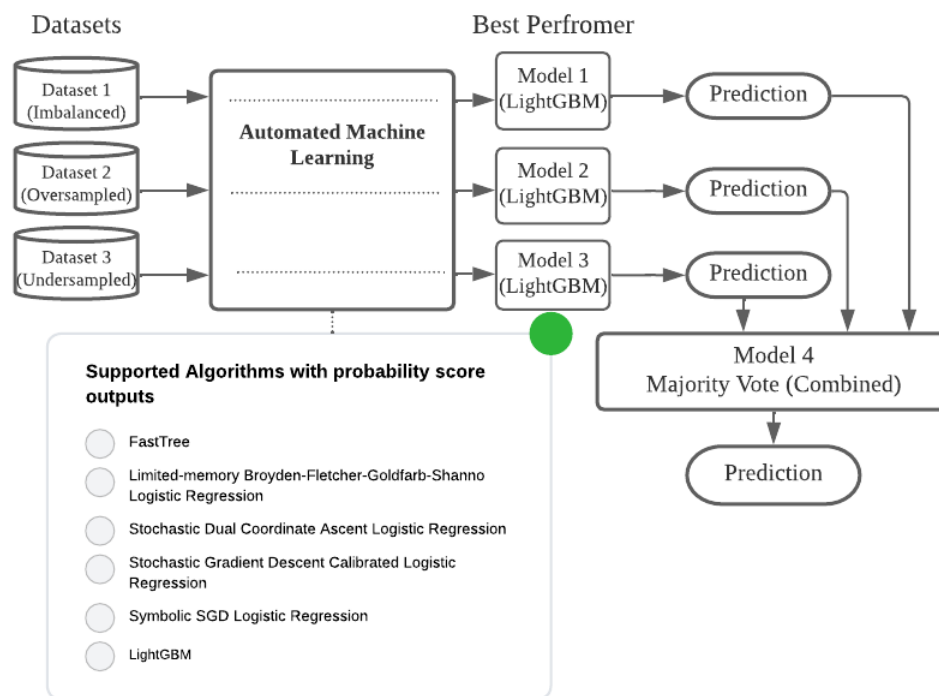
144

145 Table 1. Parameters and features used in the training of models using automated ML.

Preprocessing/featurization	Options	Datasets applied to
Case	Lowercase or unchanged	All
Numbers	Removed or left as is	All
Punctuation	Removed or left as is	All
Stop words	Removed or left as is	All
Normalization	L1, L2, infinity, or none	All
Ngram length	1 or 2	All
	3	Undersampled dataset only
All lengths*	Yes or no	All
Weighting	TF, IDF, or TF-IDF	All

146 IDF = inverse document frequency; TF = term frequency.

147 *All lengths applies when ngram length is >1 and indicates whether it only uses ngrams of the specified
 148 length (use all lengths = false) or uses ngrams of all lengths up to and including the specified length (use
 149 all lengths = true).



150
 151 Figure 2. Example depiction of the autoML process.
 152 Models with >99% sensitivity were ranked by maximal specificity. The classification models were trained
 153 using titles and abstracts of a random 80% of articles from 2012-2018 (n = 97,805). Of these, 17,824 met

154 criteria for rigor for one or more article categories; 79,981 did not. To address the imbalance in articles,
155 we created 3 training datasets: 80% of the full dataset (unbalanced; n = 97,805), and two additional
156 datasets to achieve balance through oversampling (articles meeting criteria were included multiple
157 times to equal the number of articles that did not; n = 159,962) and undersampling (random subset of
158 articles not meeting criteria were matched to the number that did; n = 35,648).

159 Trained models were tested on the remaining hold-out set of 20% (n = 24,678) of articles from 2012-
160 2018. Models with $\geq 99\%$ sensitivity with the best specificity for each of the full, over-, and under-
161 sampled datasets were retained, and one model per dataset was selected from the leaderboard. Models
162 return a probability score ranging from 0 (does not meet criteria) to 1 (meets criteria) for each article.
163 The probability threshold was determined as the point where sensitivity was 99%. To determine if
164 ensembling the three models improved performance compared with the individual models, we tested
165 their performance individually and combined—using a majority vote such that articles predicted to pass
166 in ≥ 2 of the 3 models were classified as ‘pass’ (or classified as ‘fail’ if 0 or 1 model predicted a pass)—in a
167 retrospective sample of 30,424 articles in our dataset that were published in 2020.

168 The performance of the models in the hold-out test set is akin to internal validation. Since our goal is to
169 implement an algorithm into a literature surveillance program, we assessed its performance in real-time
170 in an external test on unseen data. We prospectively evaluated the accuracy of the majority vote
171 algorithm by applying it after Boolean searches of PubMed and before critical appraisal by our research
172 associates, who were blinded to the predictions of 5253 articles published between March 9 to May 11,
173 2021. Staff appraised all articles predicted to pass and a random subset of those predicted to fail. False
174 negative articles were assessed by a senior clinical researcher (BH) to determine clinical relevance and
175 newsworthiness .

176 *Evaluation metrics*

177 For all trained models, during the testing phase we calculated sensitivity (recall), specificity, accuracy,
178 precision, NNR (1/precision), and F-score (harmonic mean of recall and precision metrics) in the 20%
179 hold-out set of articles from 2012-2018. We also calculated the area-under-the-curve (AUC) of the
180 receiver operating characteristic (ROC) curve. The ROC curve is created by plotting the true positive rate
181 (sensitivity) against the false positive rate (1-specificity) by varying the threshold applied to the
182 probability outputs of a classifier. AUC is thus a threshold-independent parameter which demonstrates
183 the overall performance of the classifier. The statistical probability was calculated for the three selected
184 models and majority vote algorithm in the 2020 data and the prospective evaluation. For the
185 prospective evaluation, we estimated the bias-corrected sensitivity and specificity with corresponding
186 95% confidence intervals (CIs) using the Begg and Greenes (20) formula that corrects for any bias when
187 a only subsample is verified to account for the articles that were predicted to fail and that were not
188 verified by design. The bias correction models the diagnostic distribution of the articles that were
189 verified (20).

190

191 RESULTS

192 *Selected models and their performance*

193 We trained 3456 models using the unbalanced and oversampled datasets and 5760 models using the
194 undersampled dataset. The preprocessing steps and parameters used in the selected top performing
195 models are shown in Table 2; each of the three selected models used the LightGBM binary classification
196 algorithm (21,22). LightGBM (light gradient-boosting machine)(21) is a gradient boosting framework that
197 uses decision tree algorithms. It is a more efficient implementation of gradient boosting decision tree
198 (23) which is an ensemble model of decision trees trained in sequence and a widely-used machine
199 learning algorithm, thanks to its efficiency, accuracy, and interpretability. The performance

200 characteristics of each of the three models in the test datasets from 2012-18 and 2020 are listed in Table
 201 3. The oversampled dataset shows more variation in the ROC curves of all trained classifiers, which could
 202 be due to having higher number of training examples resulting in underfitting of some classifiers; the
 203 ROC curves are available in Appendix A. The classifiers trained on undersampled data also have slightly
 204 more variation in performance compared to unbalanced data, which may be because some information
 205 was lost compared to using all available data. Nevertheless, the AUC values for the three top performing
 206 models are very close to each other indicating a high performance for the selected LightGBM model in
 207 all three cases.

208
 209 Table 2. Characteristics of the dataset, preprocessing, and feature extraction steps employed by
 210 AutoML in the training of the model selected from each dataset experiment*

	Model 1 (Unbalanced dataset)	Model 2 (Balanced by over-sampling)	Model 3 (Balanced by under- sampling)
Number of articles in training datasets	97 805	159 962	35 648
Ratio of negative:positive articles (or %positive)	4.5:1	1:1	1:1
Number of models trained	3456	3456	5760
Features employed in the selected best model:			
Text converted to lowercase	Yes	Yes	No
Removal of punctuation	Yes	Yes	Yes
Removal of stop words	Yes	No	No
Removal of Diacritics	Yes	Yes	Yes
Removal of numbers	Yes	Yes	Yes
Weighting method	TF-IDF	TF-IDF	TF-IDF
Normalization technique	None	None	L1
N-grams	Uni-grams	Bi-grams	Tri-grams

211 L1 = Manhattan Distance or Taxicab norm. TF-IDF = term frequency - inverse document frequency.
 212 *All selected models used LightGBM binary classification model.

213

214 Table 3. Performance characteristics for the three models in the testing datasets (20% from 2012-2018,
215 and 2020).

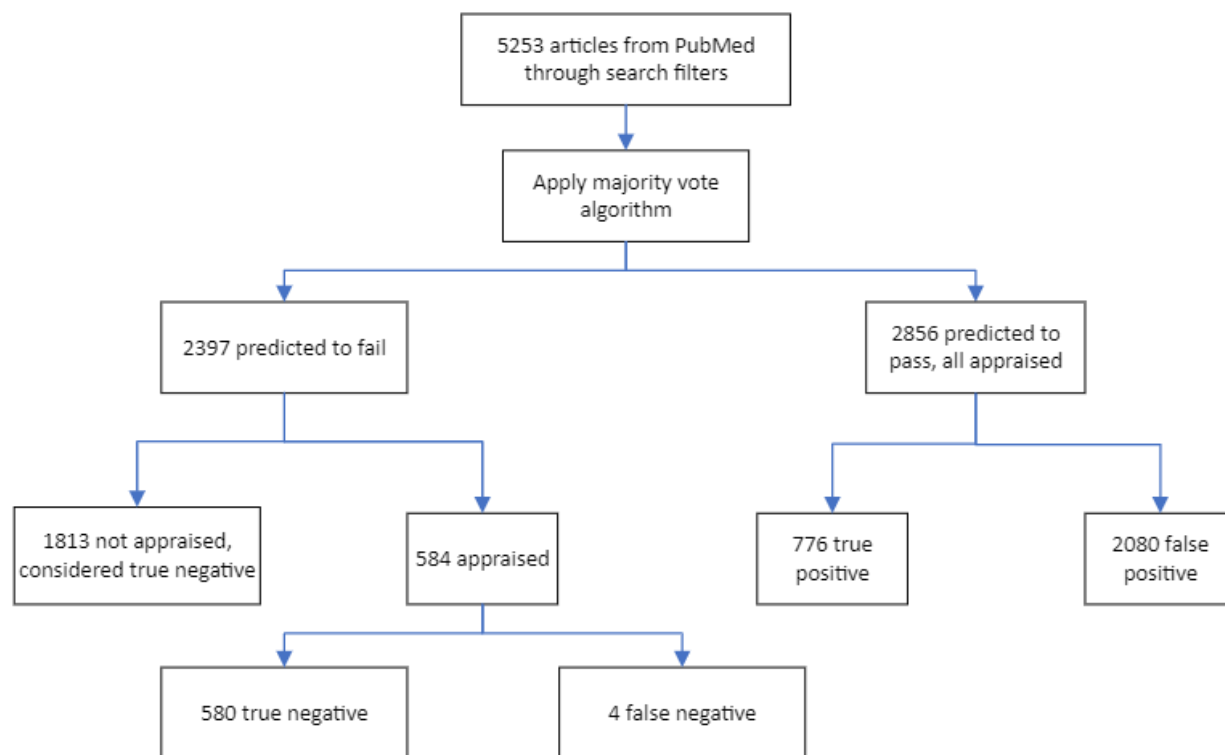
Testing dataset	Model	Sensitivity (95% CI)	Specificity (CI)	Precision	F-score	Accuracy	NNR (CI)	AUC (CI)
2012-2018*	2	99.0% (98.7 to 99.3)	53.7% (53.0 to 54.4)	32.6%	0.490	62.1%	3.07 (3.02 to 3.13)	0.952 (0.949 to 0.956)
	1	99.0% (98.7 to 99.3)	51.0% (50.3 to 51.6)	31.3%	0.475	59.8%	3.20 (3.14 to 3.26)	0.952 (0.949 to 0.955)
	3	99.0% (98.7 to 99.3)	51.8% (51.1 to 52.5)	31.6%	0.480	60.5%	3.16 (3.10 to 3.22)	0.948 (0.944 to 0.951)
2020†	Combined‡	99.2% (98.8 to 99.4)	57.5% (56.9 to 58.1)	26.2%	0.415	63.0%	3.86 (3.79 to 3.93)	NA
	2	99.1% (98.7 to 99.4)	57.3% (56.7 to 57.9)	26.1%	0.413	62.8%	3.87 (3.80 to 3.95)	0.962 (0.959 to 0.964)
	1	99.0% (98.7 to 99.3)	56.4% (55.8 to 57.0)	25.7%	0.408	62.0%	3.94 (3.86 to 4.01)	0.959 (0.956 to 0.962)
	3	99.0% (98.7 to 99.3)	56.0% (55.4 to 56.6)	25.5%	0.406	61.7%	3.96 (3.88 to 4.04)	0.956 (0.953 to 0.959)

216 NA = not applicable; NNR = number needed to read. *20% of the articles from 2012-2018 for internal
217 testing, n=24,677. †n=30,424. ‡Predictions determined by majority vote of articles meeting 2 of 3
218 probability thresholds from the unbalanced, over- and under-sampled models to pass.

219

220 *Prospective evaluation*

221 For the prospective evaluation, we opted to use the majority vote algorithm to classify 5253 consecutive
222 articles entering the surveillance system; 2856 (54%) were predicted to be high quality and 2397 (46%)
223 were not (Figure 3). All the 2856 predicted to be high quality and a random sample of 584 of the 2397
224 predicted to not be high quality were assessed by human appraisers. The remaining 1813 (90%) were
225 not assessed and considered true negatives. Of the random sample predicted to not be high quality and
226 appraised by staff, four were adjudicated to be high quality (false negatives), all of which required using
227 information from the full text of the manuscript to confirm they met the appraisal criteria for their
228 article categories. Sensitivity was 99.5% (CI, 98.7 to 99.9), specificity was 53.5% (CI, 52.0 to 55.0), and
229 the F-score was 0.427 (Table 4). The results of the corrected analysis that adjusts for 1813 articles that
230 were not assessed (bias corrected calculation) overlapped with the uncorrected values (Table 4).



231
232

233 Figure 3. Prospective evaluation of model performance in >5000 articles retrieved from PubMed.

234 Table 4. Prospective performance of the majority vote ML algorithm.

Article subset	N	Sensitivity (95% CI)	Specificity (CI)	Precision	F-score	Accuracy	NNR (CI)
All	5253	99.5% (98.7 to 99.9)	53.5% (52.0 to 55.0)	27.2%	0.427	60.3%	3.68 (3.52 to 3.85)
All-corrected*	5253	97.9% (95.9 to 99.9)	53.4% (51.3 to 55.4)	NA	NA	NA	NA
COVID	3317	100% (97.4 to 100)†	59.3% (57.6 to 61.1)	9.7%	0.177	61.0%	10.29 (9.33 to 11.49)

235 NA = not applicable; NNR = number needed to read.

236 *Bias correction to account for the 1813 articles that were predicted to fail but not verified (20).

237 †97.5% one-sided CI.

238

239 DISCUSSION

240 *Model training and performance*

241 The initial approach of using AutoML and supervised machine learning led to efficient development of
242 models for identifying articles pre-filtered by highly sensitive Boolean searches likely to be found
243 rigorous and clinically relevant at critical appraisal. Adopting AutoML was time efficient and allowed for
244 the system to test various permutations of preprocessing steps and algorithms with minimal
245 programmer time. Each of the selected highest performing models used the LightGBM binary
246 classification algorithm (21). Our selected top performing models used TF-IDF, which accounts for both
247 the number of times a word appears in a document and the inverse of the number of documents in the
248 dataset that includes the word; this essentially eliminates naturally occurring English terms and gives
249 higher values to words that are less common across the documents, or articles, in this case.

250 Training the models with datasets of varying size and balanced/unbalanced allowed us to assess the
251 value in data augmentation. We also explored the effect of combining models to determine if such an
252 approach would improve performance. Though the improvement was very small, our decision to test
253 the ensemble and implement it was based solely on our efforts to maximize specificity to reduce the
254 NNR. Keeping sensitivity high at 99%, the specificity of the trained models was >50% in the random test

255 set from 2012-2018, with slightly better performance with the model trained using the larger
256 oversampled dataset compared with the unbalanced and undersampled datasets. Though this offered a
257 larger sample, it came at the cost of time required for model training. Despite having more models
258 trained using the undersampled dataset, the performance of the top models was consistent with the
259 unbalanced dataset model. All models had similar specificity in the 2020 dataset and performed
260 marginally better than in the 2012-18 set. This could be the result of a larger sample and a broader
261 range of journal titles and article types with the inclusion of COVID-19 publications.

262 The results for the majority vote combined models, where articles predicted to pass for at least two of
263 the three models, did not factually improve the performance in the three testing datasets across years.
264 Such ensemble approaches of combining models have been used by Aphinyanaphongs et al., (24) and
265 Kilicoglu et al. (17) and showed improved F-scores. Ensemble techniques are used to reduce variability
266 across models by averaging out the errors made by each, assuming they are making different errors
267 (25). Ensemble models generally perform better when the base models they combine are as diverse as
268 possible (26). Our three models were built to represent the full unbalanced dataset, a balanced
269 undersampled dataset, and a larger oversampled dataset, but they include the same positive class of
270 articles and employed the same type of ML model and are likely not diverse enough to boost
271 performance when combined.

272 Testing and application of the ML models improved specificity compared with our traditional approach
273 of Boolean filters alone. Our goal was to maximize recall/sensitivity and specificity and reduce the NNR.
274 Prior to applying the ML models to the PLUS process (and before COVID-19), our NNR in 2019 was 4.63
275 (95% CI 4.50 to 4.77). With the addition of COVID-19 articles, in 2020 our overall NNR was 7.11 (CI 6.92
276 to 7.31). In the 2021 prospective evaluation with the addition of the ML models, the NNR was reduced
277 to 3.68 (CI 3.52 to 3.85) for all article categories. For the four false negative articles, the main apparent
278 reason for being missed was insufficient information in the title or abstract to be judged as valid.

279 *Machine learning for biomedical evidence*

280 Our approach is consistent with reported methods in our recent systematic review of ML applied to
281 improve the identification of high-quality articles (3). We used an established gold standard for high
282 quality articles produced through our PLUS process. Seven studies included in the review trained their
283 models using the Hedges dataset or articles included in ACP Journal Club, both of which are produced by
284 the same process in HiRU (3). Like other studies, we used title and abstracts as training features. Of the
285 10 studies included in our earlier systematic review, seven used datasets of articles that had been
286 critically appraised by our process (3).

287 Our models optimized recall to reduce the loss of relevant articles but that came at the cost of reducing
288 specificity and precision. The precision of our models, which ranged from 26% to 33%, was surpassed by
289 Kilicoglu et al. (17) who used ensemble models (74%), and Del Fiol et al.(16) (34%) and Afzal et al. (27)
290 (86%) who used neural network models. The high precision achieved is likely attributed to the targeting
291 particular categories of articles. Kilicoglu et al. (17) used an ensemble model which achieved a precision
292 of 37% and recall of 63% when applied to articles in general, and precision of 74% and recall of 86%
293 when used to identify rigorous treatment articles—all of which are randomized controlled trials (RCTs)—
294 a category with established terminology and structure for reporting. Afzal et al. (27) used the Cochrane
295 library as training dataset for their neural network, which includes systematic reviews and RCTs, which
296 again use explicit study design terminologies in the title, abstract, or commonly both. This facilitates the
297 retrieval function for the model and improves the overall model performance. The use of additional
298 features, like MeSH terms and MEDLINE metadata could also explain the improved performance of their
299 model, though these elements are not readily available for an article when it is first posted in PubMed;
300 there is a delay from PubMed creation date and indexing being applied, a delay that varies by journal
301 title (28). Aphinyanaphongs et al. trained models using treatment, diagnosis, prognosis, and etiology

302 articles from ACP Journal Club (29,30) (24,31) which reflects the range of article types included in our
303 dataset.

304 Deep learning is also being applied to address information retrieval and evidence classification.

305 Ambalavanan and Devarakonda (32) trained sciBERT, a pretrained deep learning algorithm, and looked
306 at both class ratios and size of the training sets for classifiers of treatment articles using the Clinical
307 Hedges dataset. They found that recall was maximized when there were more positive to negative
308 articles, precision was improved in larger training sets though there appeared to be a point at which
309 bigger did not mean better, and the F-score was optimal using a reasonably large set of balanced articles
310 (15,000:15,000). They modeled a number of steps in the article classification process (e.g., of interest to
311 humans, original study, treatment article, rigorous), and found that the F-score was lowest for predicting
312 rigor, which is a more difficult task. Notably, their study focused on articles in the treatment category
313 while our model covers articles from the full range of categories covered in the surveillance process.

314 We recently published results on models for classifying articles for rigor that we developed by finetuning
315 BioBERT, another pretrained language model (33). Our selected model outperformed the model
316 reported here, saving 60% of the manual assessments required by research associates. That model has
317 been integrated into our process but does require more computational power. Future work will
318 continue with both deep and shallow learning as each has optimal uses depending on resources
319 required for implementation.

320 F-score is the balance between recall (not missing a significant number of instances) and precision (how
321 many instances it classifies correctly) and it provides an intuitive value of the robustness of the
322 developed models. The article classification tasks assigned to the model were binary, with recall
323 optimized to increase the model robustness over its precision. This intentional optimization towards
324 higher recall was guided by our motive to minimize the chance of losing relevant articles. This limited

325 our flexibility in maximizing precision and resulted in a lower overall F-score. The wide range of article
326 categories in both training dataset and the stream of articles screened by the model would also have
327 reduced the F-scores. Had we sought to classify articles from a particular purpose category, such as
328 treatment studies using RCT designs, we expect the F-score would be higher.

329 *Implications for evidence surveillance*

330 Retrieving the best quality evidence to clinicians has driven research into the creation of initial Boolean
331 search strategies and now the advancements made applying ML models. We implemented the majority
332 vote ML algorithm into our process in May 2021 (see Figure 1B). Between May 11, 2021 to Mar 11,
333 2022, 25 867 articles were retrieved from PubMed with the Boolean searches; 11 776 (45.5%) were
334 predicted not to meet criteria and were removed from the critical appraisal queue. With a conservative
335 estimated time of 5 minutes of human resources to appraise each article, this saved >981 hours of
336 research associate time during that period while maintaining the integrity of the evidence processed.
337 This has been particularly important as we added COVID-19 related articles from all indexed journals to
338 our surveillance program in 2020 to support quick access for practitioners, policy makers, and lay
339 persons to appraised emerging research through the COVID-19 Evidence Alerts website (34). The ML
340 model has offset some of the additional burden of this growing body of COVID-19 literature.

341
342

343 *Future model development*

344 Using Auto-ML, we were able to train and test models that improved the efficiency of our literature
345 surveillance process. There are several pretrained deep learning language models, such as BERT,
346 BioBERT, and PubMedBERT, available for application to clinical literature. We have begun preliminary

347 development of deep learning models using our dataset and the results are promising. Our future
348 research includes assessing model performance by category of articles and applying our models more
349 broadly beyond the titles monitored for PLUS. Given the richness of our dataset, including tagged
350 reasons for not meeting critical appraisal criteria and other article metadata captured at the time of
351 appraisal, we hope to enhance model performance by leveraging these data.

352 *Strengths and limitations*

353 Our models were trained using the largest tagged dataset of health care research articles across a range
354 of article categories to date and based on an established gold standard in the field. Although the critical
355 appraisal criteria are applied by a single reader, all included studies and those passing with questions are
356 assessed by a final editor. The dataset overcomes some of the challenges we identified in our review: 1)
357 the criteria applied to assess rigor is an established gold standard based on best evidence-based
358 medicine practices; 2) the dataset is the largest, yet, and the training dataset included 17 824 articles in
359 the high-quality class that allowed for creating of oversampled and undersampled datasets for training;
360 3) journals for a range of clinical domains are included in the dataset (current list of journals can be
361 found here: <https://hiru.mcmaster.ca/hiru/journalslist.asp>); and 4) the training dataset contemporary
362 and includes articles from 2012-2018 and was tested in 2020 dataset. The prospective, blinded
363 evaluation of the performance of the selected combined models highlights the value of real-world
364 application and impact.

365 The models, however, were derived using prefiltered articles from PubMed for a subset of ~120 journals
366 and generalizability to all the content in a literature database is uncertain. These concerns are allayed by
367 the performance of the models in the 2020 articles which are more numerous and cover a greater array
368 of journal titles as all pre-filtered COVID-19 articles were included. Though the number to read was
369 higher (not surprising given the amount of lower quality evidence in COVID-related studies), specificity

370 and accuracy were improved. We used logistic regression approaches, and more advanced deep
371 learning techniques expected to perform better, as seen in the results of Del Fiol et al (16) and Afzal et al
372 (27). We have started using deep learning approaches in furthering our work in the area, initially to
373 further increase the specificity of classifiers for all categories of articles. We plan to evaluate models for
374 applications other than literature surveillance and investigate questions about optimal class ratios and
375 training dataset size for model development. Further work will include training deep learning models for
376 specific article categories, using more of the features in our dataset that correspond to rigor, and
377 developing interpretable AI models.

378 *Conclusion*

379 Using ML-based probability ranking, we improved the specificity of identifying biomedical articles that
380 meet methodological rigor criteria while preserving a very high sensitivity. The selected models perform
381 well in an active surveillance program that supports knowledge translation to practicing clinicians.
382 Future work includes training deep learning models using the dataset to develop higher performing
383 models to facilitate identification of high-quality research soon after publication.

384

385 REFERENCES

- 386 1. Montori VM, Wilczynski NL, Morgan D, Haynes RB, Hedges Team. Optimal search strategies for
387 retrieving systematic reviews from Medline: analytical survey. *BMJ* [Internet]. 2005 Jan
388 8;330(7482):68. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15619601>
- 389 2. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, Hedges Team. Optimal search
390 strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey.
391 *BMJ* [Internet]. 2005 May 21;330(7501):1179. Available from:
392 <http://www.ncbi.nlm.nih.gov/pubmed/15894554>
- 393 3. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, et al. Machine Learning
394 Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical
395 Literature: Systematic Review. *JMIR Med Informatics* [Internet]. 2021 Sep 1 [cited 2021 Nov
396 21];9(9). Available from: [/pmc/articles/PMC8461527/](https://pmc/articles/PMC8461527/)
- 397 4. Alzubi J, Nayyar A, Kumar A. Machine Learning from Theory to Algorithms: An Overview. *J Phys*

- 398 Conf Ser [Internet]. 2018 Nov 1 [cited 2021 Nov 24];1142(1):012012. Available from:
399 <https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012>
- 400 5. Dey A. Machine Learning Algorithms: A Review. *Int J Comput Sci Inf Technol*. 2016;7(3):1174–9.
- 401 6. Burkov A. The Hundred-page Machine Learning Book [Internet]. Andriy Burkov; 2019. Available
402 from: <https://books.google.ca/books?id=ZF3KwQEACAAJ>
- 403 7. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning
404 algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019 Dec 21;19(1):281.
- 405 8. Wang Q, Ming Y, Jin Z, Shen Q, Liu D, Smith MJ, et al. ATMSeer. In: *Proceedings of the 2019 CHI*
406 *Conference on Human Factors in Computing Systems* [Internet]. New York, NY, USA: ACM; 2019.
407 p. 1–12. Available from: <https://dl.acm.org/doi/10.1145/3290605.3300911>
- 408 9. Drozdal J, Weisz J, Wang D, Dass G, Yao B, Zhao C, et al. Trust in AutoML: exploring information
409 needs for establishing trust in automated machine learning systems. In: *Proceedings of the 25th*
410 *International Conference on Intelligent User Interfaces* [Internet]. New York, NY, USA: ACM;
411 2020. p. 297–307. Available from: <https://dl.acm.org/doi/10.1145/3377325.3377501>
- 412 10. Holland J, Haynes RB, McMaster PLUS Team Health Information Research Unit. McMaster
413 Premium Literature Service (PLUS): an evidence-based medicine information service delivered on
414 the Web. *AMIA . Annu Symp proceedings AMIA Symp* [Internet]. 2005 [cited 2021 Dec
415 15];2005:340–4. Available from: </pmc/articles/PMC1560593/>
- 416 11. Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D, et al. Second-Order Peer
417 Review of the Medical Literature for Clinical Practitioners. *JAMA* [Internet]. 2006 Apr 19 [cited
418 2021 Dec 22];295(15):1801–8. Available from: [https://jamanetwork-](https://jamanetwork-com.libaccess.lib.mcmaster.ca/journals/jama/fullarticle/202708)
419 [com.libaccess.lib.mcmaster.ca/journals/jama/fullarticle/202708](https://jamanetwork-com.libaccess.lib.mcmaster.ca/journals/jama/fullarticle/202708)
- 420 12. Wong SSL, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically
421 relevant qualitative studies in MEDLINE. *Stud Health Technol Inform*. 2004;107:311–4.
- 422 13. Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound
423 causation studies in MEDLINE. *AMIA Annu Symp Proc*. 2003;719–23.
- 424 14. Lokker C, Brian Haynes R, Wilczynski NL, McKibbin K, Walter SD. Retrieval of diagnostic and
425 treatment studies for clinical use through PubMed and PubMed’s clinical queries filters. *J Am*
426 *Med Informatics Assoc*. 2011;18(5).
- 427 15. Afzal M, Hussain M, Haynes RB, Lee S. Context-aware grading of quality evidences for evidence-
428 based decision-making. *Health Informatics J*. 2019 Jun 1;25(2):429–45.
- 429 16. Del Fiol G, Michelson M, Iorio A, Cotoi C, Brian Haynes R, Haynes RB, et al. A Deep Learning
430 Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the
431 Biomedical Literature: Comparative Analytic Study. *J Med Internet Res* [Internet]. 2018 Jun 25
432 [cited 2021 Nov 21];20(6):e10281. Available from: </pmc/articles/PMC6037944/>
- 433 17. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic
434 recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc* [Internet].
435 2009 Jan [cited 2021 Nov 21];16(1):25–31. Available from:
436 <https://pubmed.ncbi.nlm.nih.gov/18952929/>

- 437 18. HiRU Inclusion Criteria [Internet]. [cited 2021 Aug 6]. Available from:
438 <https://hiru.mcmaster.ca/hiru/InclusionCriteria.html>
- 439 19. What is .NET? An open-source developer platform. [Internet]. [cited 2021 Dec 22]. Available
440 from: <https://dotnet.microsoft.com/en-us/learn/dotnet/what-is-dotnet>
- 441 20. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to
442 selection bias. *Biometrics* [Internet]. 1983 Mar;39(1):207–15. Available from:
443 <http://www.ncbi.nlm.nih.gov/pubmed/6871349>
- 444 21. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient
445 Boosting Decision Tree. In: NIPS. 2017.
- 446 22. Microsoft Corporation. Welcome to LightGBM's documentation! — LightGBM 3.3.1.99
447 documentation [Internet]. 2021 [cited 2021 Dec 14]. Available from:
448 <https://lightgbm.readthedocs.io/en/latest/>
- 449 23. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* [Internet].
450 2001 Oct 1;29(5). Available from: [https://projecteuclid.org/journals/annals-of-statistics/volume-](https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full)
451 [29/issue-5/Greedy-function-approximation-A-gradient-boosting-](https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full)
452 [machine/10.1214/aos/1013203451.full](https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full)
- 453 24. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models
454 for high-quality article retrieval in internal medicine. *J Am Med Informatics Assoc*.
455 2005;12(2):207–16.
- 456 25. Polikar R. Ensemble Learning. In: Zhang C, Ma Y, editors. *Ensemble Machine Learning: Methods*
457 *and Applications* [Internet]. Boston, MA: Springer US; 2012. p. 1–34. Available from:
458 https://doi.org/10.1007/978-1-4419-9326-7_1
- 459 26. Zhou Z-H. Ensemble Learning. In: *Encyclopedia of Biometrics* [Internet]. Boston, MA: Springer US;
460 2009. p. 270–3. Available from: http://link.springer.com/10.1007/978-0-387-73003-5_293
- 461 27. Afzal M, Park BJ, Hussain M, Lee S. Deep learning based biomedical literature classification using
462 criteria of scientific rigor. *Electron*. 2020 Aug 1;9(8):1–12.
- 463 28. Irwin AN, Rackham D. Comparison of the time-to-indexing in PubMed between biomedical
464 journals according to impact factor, discipline, and focus. *Res Soc Adm Pharm*. 2017 Mar
465 1;13(2):389–93.
- 466 29. Aphinyanaphongs Y, Aliferis C. Prospective validation of text categorization filters for identifying
467 high-quality, content-specific articles in MEDLINE. *AMIA . Annu Symp proceedings AMIA Symp*
468 [Internet]. 2006 Jan 1 [cited 2021 Nov 21];6–10. Available from:
469 <https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/17238292/?tool=EBI>
- 470 30. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models
471 for high-quality article retrieval in internal medicine. *J Am Med Informatics Assoc*.
472 2005;12(2):207–16.
- 473 31. Aphinyanaphongs Y, Aliferis CF. Text Categorization Models for Retrieval of High Quality Articles
474 in Internal Medicine. *AMIA Annu Symp Proc* [Internet]. 2003 [cited 2021 Nov 21];2003:31.
475 Available from: [/pmc/articles/PMC1480096/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480096/)

- 476 32. Ambalavanan AK, Devarakonda M V. Using the contextual language model BERT for multi-criteria
477 classification of scientific articles. J Biomed Inform [Internet]. 2020 Dec;112:103578. Available
478 from: <https://linkinghub.elsevier.com/retrieve/pii/S1532046420302069>
- 479 33. Lokker C, Bagheri E, Abdelkader W, Parrish R, Afzal M, Navarro T, et al. Deep learning to refine
480 the identification of high-quality clinical research articles from the biomedical literature:
481 Performance evaluation. J Biomed Inform [Internet]. 2023 Jun 1 [cited 2023 May 24];142:104384.
482 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1532046423001053>
- 483 34. McMaster HIRU. COVID-19 Evidence Alerts from McMaster PLUS | Home [Internet]. 2022 [cited
484 2022 Jun 29]. Available from: <https://plus.mcmaster.ca/Covid-19/>

485

486

487

488

489 Supporting Information:

490 Appendix A. Receiver operating characteristic (ROC) curves for the models trained on the 3 datasets