RESEARCH ARTICLE

Can the quality of published academic journal articles be assessed with machine learning?

Mike Thelwall

University of Wolverhampton, UK

Keywords: citation analysis, machine learning, research evaluation, text mining

ABSTRACT

Formal assessments of the quality of the research produced by departments and universities are now conducted by many countries to monitor achievements and allocate performance-related funding. These evaluations are hugely time consuming if conducted by postpublication peer review and are simplistic if based on citations or journal impact factors. I investigate whether machine learning could help reduce the burden of peer review by using citations and metadata to learn how to score articles from a sample assessed by peer review. An experiment is used to underpin the discussion, attempting to predict journal citation thirds, as a proxy for article quality scores, for all Scopus narrow fields from 2014 to 2020. The results show that these proxy quality thirds can be predicted with above baseline accuracy in all 326 narrow fields, with Gradient Boosting Classifier, Random Forest Classifier, or Multinomial Naïve Bayes being the most accurate in nearly all cases. Nevertheless, the results partly leverage journal writing styles and topics, which are unwanted for some practical applications and cause substantial shifts in average scores between countries and between institutions within a country. There may be scope for predicting articles' scores when the predictions have the highest probability.

1. INTRODUCTION

Higher education has become increasingly monitored and managed by national states over the past half century (Amaral, Meek et al., 2003). As part of this, countries typically have competitive systems in place to fund academic research. In addition to project-based funding, some nations also now directly reward research quality through systematic procedures to assess this at the departmental level (e.g., Poland: Kulczycki, Korzeń, and Korytkowski (2017); but not Germany: Hinze, Butler et al. (2019)). The UK's Research Excellence Framework (REF) uses expert postpublication peer review to evaluate the outputs of academic researchers at approximately the departmental level every 6 or 7 years, combining the results with evaluations of case studies and institutional environments (Wilsdon, Allen et al., 2015a). New Zealand's Performance Based Research Fund (Buckle & Creedy, 2019) and Excellence in Research Australia (Hinze et al. 2019) are similar peer review schemes. Italy also has a performance-based funding scheme that includes evaluating the quality of researchers' outputs, albeit exempting journal articles meeting bibliometric thresholds (Franceschini & Maisano, 2017).

Large-scale post publication peer review is hugely time consuming, employing many researchers (over 1,000 in the United Kingdom) to make postpublication quality assessments on the outputs. It is therefore logical to investigate whether the peer review part of this process could be streamlined in any way, such as by automation. In the UK, academics in 2019 believed

an open access journal

Check for updates

Citation: Thelwall, M. (2022). Can the quality of published academic journal articles be assessed with machine learning? *Quantitative Science Studies*, *3*(1), 208–226. https://doi.org/10.1162/qss_a_00185

DOI: https://doi.org/10.1162/qss_a_00185

Peer Review: https://publons.com/publon/10.1162/qss_a_00185

Received: 5 January 2022
Accepted: 8 February 2022

Corresponding Author:
Mike Thelwall
m.thelwall@wlv.ac.uk

Handling Editor:
Ludo Waltman

The MIT Press

that technology would be used to enhance research assessment in the future (Parks, Rodriguez-Rincon et al., 2019), perhaps thinking of this. In November 2021, the four UK higher education funding bodies published a call for a systematic review of the potential for technology to support research assessment, and particularly the labor-intensive REF (Gov.uk, 2021). Motivated by this, the current article uses a dummy automated peer review exercise to underpin a discussion of the potential for artificial intelligence to replace peer review.

Few previous studies have attempted to automatically score the quality of academic research. This is presumably because only aggregate scores are made public by national evaluation exercises, so there is no output-level data to leverage to build effective algorithms. Nevertheless, some have attempted to find indicators that correlate with overall university quality profiles, such as hyperlinks to university websites (Thelwall, 2002) or citation-based indicators (Traag & Waltman, 2019). One report by the team organizing the UK REF has analyzed the raw output scores, however, finding moderate correlations with citation-based indicators and altmetrics (Wilsdon, Allen et al., 2015b). The task of predicting long-term citations for articles is related because citation counts in some fields are approximate indicators of scientific impact. Investigations of this possibility, often using regression rather than machine learning, have found a range of article metadata factors to associate with higher citation counts. These include the number of authors, the number of countries in the author team, the readability of the abstract, and keyword repetition (Hall, Vogel et al., 2018; Lei & Yan, 2016; Li, Zhao et al., 2019; McCannon, 2019; Sohrabi & Iraj, 2017; Stegehuis, Litvak, & Waltman, 2015). Text mining has rarely been used to predict citations but has been used to detect plagiarism (Foltýnek, Meuschke, & Gipp, 2019) or statistical errors (Nuijten & Polanin, 2020) and to investigate topics in fields (Heo, Kang et al., 2017), to identify research trends (Kim & Delen, 2018; Nie & Sun, 2017), to map science (Chen, 2017), and to predict journal or conference reviewing decisions (Checco, Bracciale et al., 2021; Thelwall, Papas et al., 2020).

A few papers have used machine learning approaches to predict article citations, using methods including Support Vector Machines (Fu & Aliferis, 2010), *k*-Nearest Neighbors, and Bagging (Wang, Jiao et al., 2020), Stochastic Gradient Descent, Random Forest, XGBoost classifier, LGBoost classifier (Klemiński, Kazienko, & Kajdanowicz, 2021), Decision Trees (Su, 2020), and CART (Yuan, Tang et al., 2018). Different deep learning architectures have been proposed and tested on small full text article sets, apparently convenience samples of collections of papers with full text online. These include long-term physics article sets (Zhao & Feng, 2022), library, information, and documentation articles (Ruan, Zhu et al., 2020), historical Markov chain articles (Xu, Li et al., 2019), two computational linguistics conferences (Li et al., 2019), and five prestigious journals (using annual citations rather than full text: Abrishami & Aliakbary, 2019). A comparison of deep learning with other approaches found Support Vector Machines to be the most accurate for computer science publications (Zhu & Ban, 2018). Almost all previous machine learning studies have focused on a single topic or a small set of fields. The only exception, which took a science-wide approach, sampled 12,374 random Web of Science articles from 2015, ignoring field classifications rather than comparing between fields (Akella, Alhoori et al., 2021). Also, no previous study seems to have used a development set that is separate from the training/testing sets, risking overfitting (the current article also does not use separate training/test sets but reports separate scores for 326 fields and does not customize the methods for any field). The machine learning studies so far have not given comparative information about the relative attractiveness of machine learning for different fields of science, and none have addressed the issue of scientific quality estimation.

The research goal for this article is to investigate whether it is possible to assess the quality of published academic journal articles with machine learning, but the research questions are

only indirectly related to this and hence need justification. Because there are no large-scale sources of postpublication quality control scores for academic articles (with the partial exception of biomedical science: Mohammadi & Thelwall, 2013), a proxy source of quality is used. For this article, the citation rate (defined in detail below) of the publishing journal is used as the quality proxy. This is a poor proxy but is used in the absence of a better one. Journal impact factors of various types are widely recognized and are considered to be indicators of scientific quality to some extent by some researchers, varying between countries and fields. This is because in some fields, citations are accepted as (very approximate) indicators of scientific worth and so journals attracting better articles tend to be more cited. UK evidence suggests that journal citation rates correlate moderately positively with the quality of the articles that they publish in the medical and physical sciences and economics, weakly in engineering and social sciences, and negatively or not at all in the arts and humanities (Wilsdon et al., 2015b, Table A18). Moreover, in some fields, Journal Impact Factors correlate with researchers' opinions of journals (Haddawy, Hassan et al., 2016; Serenko & Bontis, 2021; Serenko & Dohan, 2011), although not in others (Maier, 2006). In fields where this logic is accepted, there is a tendency for it to become truer over time because there is more competition to be published in journals with higher impact factors. On the other hand, citations are irrelevant in some fields and impact factors reflect journal specialisms to some extent. Thus, in this article, average journal citation rate is used as an approximate indicator of article quality, accepting that in some fields it is irrelevant to quality. Articles are split into three groups by journal citation rate, mirroring the UK REF, where articles are allocated a weighting of 0, 0.25 or 1 for quality-related funding (https://re.ukri.org/funding/quality-related-research-funding/). The specific research questions are therefore as follows.

1. How accurately can machine learning identify the journal impact third of published journal articles from other metadata in different fields and years?
2. Which textual features are most powerful at detecting the journal impact third of published journal articles in different fields and years?
3. Do the machine learning results have systematic biases against any genders, countries or institutions?

2. METHODS

The research design was to gather a reasonably comprehensive sample of academic journal articles, allocate journal impact-based thirds to the articles, and apply machine learning to detect the probable journal third of each article. The same parameters (sample size, machine learning method, feature set size) were applied to each field so that the results could be compared.

2.1. Data

Scopus was chosen as the source of journal articles for its wide coverage of academic literature and fine-grained field classification scheme. The Web of Science could have been used but Scopus is slightly larger, giving more data. Only documents of type journal article were included to give consistency. The UK REF excludes review articles, so these were not included.

All Scopus documents of type journal article with publication years between 2014 and 2020 were downloaded from Scopus in January 2021 using its API. The years 2014 to 2020 were chosen to mimic REF2021, and January 2021 citation data is appropriate because it would be available at the start of the original assessment period (although the start was delayed due to the COVID-19 pandemic). The quality of the citation data thus varies between years,

with 2014 citation data being relatively mature and 2020 citation data being poor quality due to December 2020 articles having almost no time to attract citations, whereas January 2020 article had about a year. This is taken into account in the discussion but not in the methods.

Journal thirds were identified for each Scopus narrow field with a multistage approach. First, the Normalized Log-transformed Citation Score (NLCS) (Thelwall, 2017) was calculated for each article. This uses log transformation to reduce the skewing of citation count data so that the result is not dominated by a small number of highly cited articles. The score for an article is 1 if its citation rate is the world average, with scores above 1 indicating a greater citation rate than the world average and scores below 1 indicating fewer citations than the world average. These scores are normalized within the Scopus narrow field in which an article is classified (or the average of all fields for articles in multiple fields). Second, the arithmetic mean of the NLCS for all articles in each journal was calculated as its average citation score, called here JMNLCS (Journal Mean NLCS). This differs from the Journal Impact Factor (JIF) in that the average is calculated from articles in a single year, includes all citations to date, and uses NLCS instead of raw citation counts. This should be less affected by skewing than JIFs and should be more relevant to articles from the year with the data used for the calculation. Finally, JMNLCS thresholds were calculated to split the articles into approximately equal thirds. In some cases, this was not possible due to single very large journals dominating categories and the split generated approximate halves instead of thirds.

2.2. Features Analyzed

The machine learning stage requires a set of data about the articles to predict from. Because journal citation rates are used as a quality proxy for articles, they cannot also be used as inputs for the machine learning process. Also, because the purpose of quality control is to assess individuals or institutions, it is inappropriate to include these as inputs. The following features were included.

- NLCS for each article: This is a citation-based indicator, normalized for fields and year to be comparable between articles. The log transformation reduces skewing, which may make the feature more powerful for learning with linear-based algorithms.
- Number of authors: Articles with more authors are likely to be more cited in many fields.
- Number of country affiliations: Articles with authors from more countries are likely to be more cited in many fields.
- Word unigrams, bigrams, and trigrams: The quality of an article is presumably encoded in its text and figures. While full-text analysis is impractical and likely to confuse an algorithm with many irrelevant details from a paper, the title, abstract, and keywords may be helpful as a succinct summary. These were therefore extracted and added as features. Individual words and short phrases of two or three words were extracted, with those occurring only once in a field being discarded.

Abstracts were preprocessed to remove standard texts, such as publisher copyright statements and structured abstract headings. A large set of heuristics had been developed to remove these for previous automated text analyses of abstracts (Fairclough & Thelwall, 2022; Thelwall & Nevill, 2021) and these were reused for the current paper. These heuristics vary from generic (e.g., remove the first or last sentence if the first character is a copyright symbol or the first word is *Copyright*; remove the phrase, *All rights reserved.*) to publisher-specific (e.g., remove the first abstract sentence if it contains Elsevier and a copyright symbol; remove the first sentence if it contains *Maney & Son Ltd* starting in the first 20 characters). The

heuristics included a list of common structured abstract headings, such as "Results:" and "PARTICIPANTS". Articles with abstracts with fewer than 500 characters after this stage were removed. This standardizes the machine learning task by ignoring articles without abstracts or with relatively trivial abstracts.

Each field and year combination formed a separate data set for training and evaluation. There were 330 nonempty fields per year, on average, over 7 years, so this gave 2,310 data sets to analyze. The smaller sets had too few articles to analyze, however, so the final number of field/year combinations analyzed was slightly less. The total number of articles analyzed was 31,273,062, varying between 3,846,106 in 2014 and 5,694,904 in 2020. This counts articles multiple times when they occur in multiple narrow fields but excludes articles with short or no abstracts. Exact numbers for each field and year are in the online supplement (column B of worksheet "Acc aboveAll Fig 1" in spreadsheet "All files and methods - accuracy fig 1,6 chi square 2014.xlsx": https://doi.org/10.6084/m9.figshare.17912009).

2.3. Machine Learning

There are many different machine learning algorithms and all have advantages and disadvantages, so there is not an obvious candidate for the machine learning task. Twenty classification or regression algorithms were compared (Table 1), as implemented in the standard scientific machine learning system scikit-learn on Python with their default settings. These include three that are general-purpose and accurate on a wide range of tasks: Support Vector Machines (Linear Support Vector Classification here), Gradient Boosting Classifier, and Random Forest Classifier. Two were discarded for the full testing (see table footnotes). All the classifiers were run a second time as ordinal classifiers by classifying two separate two-class problems: 1 vs (2 and 3) and (1 and 2) vs. 3, giving the result 1 from the first problem, 3 from the second task, and otherwise 2. Any cases classified as both 1 and 3 were instead classed as 2. This procedure takes into account the ordering of the classifications, so should, in theory, be superior to both classification (unordered) and sometimes regression (when it assumes a linear relationship).

Feature reduction (i.e., selecting a subset of the inputs to feed into each algorithm) was performed using the chi-square method, except forcing the citation, author, and country information to be kept. Tests with a range of training set sizes and feature set sizes suggested that the performance of the algorithms increases as either or both increases, so there was not an optimal choice for either one. As a compromise, 1,000 features and 1,000 articles for training were selected as large enough to be close to the optimal accuracy without slowing the algorithms too much. Fields were trained on 90% of the articles or 1,000 articles (whichever was the smaller) and evaluated on the remainder. The algorithms were trained and evaluated on 30 separate random test/train splits (rather than, for example, 30-fold cross-validation, because the training set size needs to be fixed) and the average accuracy reported. The predictions from the first iteration on each field data set were saved for further analysis of the individual predictions.

2.4. Analysis

Overall accuracy statistics (precision) were calculated separately for each field/year combination as the average accuracy of the 30 algorithm iterations on the evaluation sets. Recall and F1 measure were not calculated because the small number of classes (three) means that they give little extra information, and they are in any case subsumed in the score-based tests of the influence of the results, discussed below. For ease of comparison between fields, the main statistic reported is the level of accuracy above the baseline (the percentage of articles in

Table 1. Machine learning methods initially tested for regression and classification. Those marked with "/o" have an ordinal version of the classification.

Code	Method	Type
bnb/o	Bernoulli Naive Bayes	Class
cnb/o	Complement Naive Bayes	Class
gbc/o	Gradient Boosting Classifier	Class
knn/o	*k* Nearest Neighbors	Class
lsvc/o	Linear Support Vector Classification	Class
log/o	Logistic Regression	Class
mnb/o	Multinomial Naive Bayes	Class
pac/o	Passive Aggressive Classifier	Class
per/o	Perceptron	Class
rfc/o	Random Forest Classifier	Class
rid/o	Ridge classifier	Class
sgd/o	Stochastic Gradient Descent	Class
svc/o	Support Vector Classification	Class*
elnr	Elastic-net regression	Reg
krr	Kernel Ridge Regression	Reg
lasr	Lasso Regression	Reg
lr	Linear Regression	Reg
ridr	Ridge Regression	Reg
sgdr	Stochastic Gradient Descent Regressor	Reg
svr	Support Vector Regression	Reg**

* Almost the same results as lsvc and so was not used for the full testing.

** Inaccurate and slow in all tests with 1,000 features and so was not use for the full testing.

the most common class). This is fairer than comparing accuracy between fields, because some fields have substantially higher baseline accuracies than others.

To assess the influence of the machine learning on the overall scores of countries, institutions, and two genders, for each field, the weighted average true score (JMNLCS thirds) and machine learning predicted scores were compared. Within each country, the results were compared between institutions and male/female first author genders. Any differences suggest a machine learning bias (accidental or systematic) towards or away from the group in question. First author genders were assigned by checking their first name against a list of country-based gendered first names from Gender-API.com, allocating a gender only when the probability of a correct assignment was above 95%. These tests were reported for the main three machine learning methods only, to avoid reporting low-value information.

To identify the types of term with the greatest discriminatory power in the machine learning, chi-square tests were conducted on all terms used for the machine learning stage in each field

and the top term selected for all 314 Scopus narrow fields in 2014 with three categories. These terms were informally investigated using the Key Word In Context (KWIC) approach by identifying their most common single context in the field that caused their high chi-square values.

3. RESULTS

3.1. Comparison of Methods and Years

The single most accurate method was gbc 46% of the time, followed by rfc (45%) and mnb (3%). The mnb method was rarely accurate for the early years but was relatively more accurate on the 2020 data. Overall, gbc and rfc had similar levels of accuracy, but all were substantially more accurate than all 30 other methods, on average. The regression classifiers had relatively poor accuracy, and the ordinal versions of classifiers surprisingly tended to be less accurate overall than the standard versions (Figure 1).

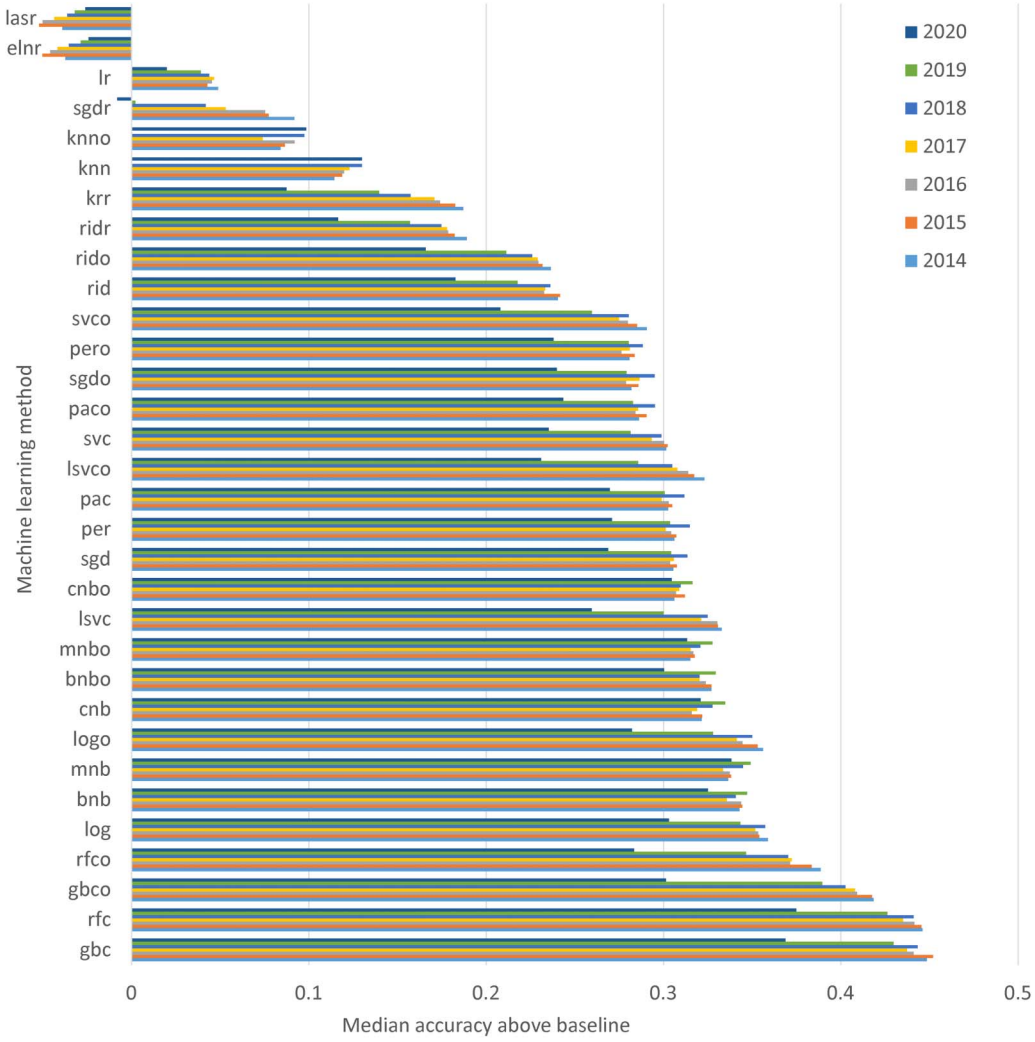

Figure 1. Average (from 30 tries) accuracy above baseline (on a scale of 0 = baseline to 1 = 100% accurate) across the 326 Scopus narrow fields for 32 different machine learning methods. Each has a training set of 1,000 articles, using 1,000 features selected by chi-square, and evaluated on the remaining articles.

Accuracy is generally highest in 2014 and substantially lower in 2020 than in other years, with 2019 also being lower than 2014-2018. For 2014, this confirms that citations are useful in helping to predict journal thirds. Accuracy is presumably lower in 2019 and 2020 because early published articles have a substantial citation advantage over late published articles due to longer citation windows. This tends to confirm that citations are less valuable for newer articles. Another possible explanation is that the journal thirds are less coherent for 2020 because the citation data has had less time to mature.

3.2. Fields with Highest and Lowest Relative Accuracy

The Scopus narrow fields in 2014 with the highest accuracy for any machine learning method tended to be small (fewer than 1,000 articles) for "miscellaneous" (mixed) fields, or both (Table 2). Small fields may be easier to predict because they contain fewer journals, so prediction is a simpler problem. For example, Review and Exam Preparation had three journals: one in the top third (*Clinical Teacher*), two in the mid third (*Journal for Nurses in Professional Development* and *Journal of Continuing Education in Nursing*), and none in the bottom third.

Miscellaneous fields may be easier to predict because they contain journals from relatively different topics, so making journal-related predictions from text may be easier (because abstracts would contain more distinctive terms for each journal). For example, Veterinary (misc.) includes relatively different titles, such as *Brazilian Journal of Veterinary Pathology*, *Journal of Fish Diseases*, and *Parasite*. The top chi-square words/phrases for this narrow field suggested another cause, however. They were "opinion," "opinion on," "opinion on the," "scientific," "scientific opinion," and "scientific opinion on," which all originated primarily from the titles of articles in the *EFSA Journal* (e.g., https://doi.org/10.2903/j.efsa.2014.3573). This is a publication of the European Food Standards Agency that published its outputs, which seem to be conclusions of expert scientific committees after deliberation. It is not a peer reviewed academic journal, even though it publishes expert scientific outputs. The standardization of title phrases has made its articles easily identified by machine learning. No similar standard phrases were discovered for the other three miscellaneous categories, although some near-

Table 2. The 10 out of 324 Scopus narrow fields in 2014 with the *highest* accuracy above baseline for any machine learning algorithm (training set 1,000, or 90% if under 1,000; 1,000 features)

Narrow field (2014)	Articles	Baseline	Top	AOB
2923 Review & Exam Preparation	150	57%	mnb	99.0%
3401 Veterinary (misc.)	957	49%	gbc	86.8%
3604 Emergency Medical Services	137	69%	mnb	86.8%
3603 Complementary & Manual Therapy	405	45%	cnb	83.4%
1504 Chemical Health and Safety	365	55%	cnb	82.9%
2920 Pharmacology (nursing)	114	81%	mnb	82.7%
1301 Biochemistry, Genetics & Molecular Biology (misc.)	702	40%	cnb	80.2%
1501 Chemical Engineering (misc.)	1,313	41%	gbc	80.1%
3601 Health Professions (misc.)	508	45%	cnb	78.7%
2917 Oncology (nursing)	592	35%	gbc	78.3%

Table 3. The 10 out of 324 Scopus narrow fields in 2014 with the *lowest* accuracy above baseline for any machine learning algorithm (training set 1,000, or 90% if under 1,000; 1,000 features)

Narrow field (2014)	Articles	Baseline	Top	AOB
2602 Algebra & Number Theory	2,841	42%	rfc	27.2%
2000 Economics, Econometrics and Finance (all)	8,178	49%	gbc	26.6%
1205 Classics	627	41%	mnb	25.0%
3206 Neuropsychology & Physiological Psychology	3,507	42%	rfc	25.0%
2607 Discrete Mathematics & Combinatorics	2,546	44%	gbc	24.3%
2608 Geometry & Topology	2,378	44%	gbc	24.3%
1208 Literature & Literary Theory	5,496	44%	rfco	24.1%
2600 Mathematics (all)	12,392	48%	gbc	23.2%
1212 Religious Studies	4,785	42%	gbc	23.0%
1200 Arts & Humanities (all)	4,820	84%	mnb	19.3%

top terms (sixth to ninth highest) from Health Professions (misc.) were from an unusual structured abstract phrase, "Conclusions and implications for practice," in the *Psychiatric Rehabilitation Journal* that had not been filtered out.

The Scopus narrow fields with the lowest accuracy attainable with the 32 methods include three general fields with "all" in their title, four humanities fields, and two mathematical fields. General "all" fields may contain journals with relatively similar or general scopes that are difficult to detect through word frequency analyses because their abstract texts tend to contain similar words. Humanities fields may contain many small specialist journals with highly diverse topics, complicating the machine learning problem (Table 3). Mathematics journals,

Table 4. The 10 out of 326 Scopus narrow fields in 2020 with the *highest* accuracy above baseline for any machine learning algorithm (training set 1,000, or 90% if under 1,000; 1,000 features)

Narrow field (2020)	Articles	Baseline	Top	AOB
3503 Dental Hygiene	120	82%	gbc, lsvc	100.0%
2923 Review & Exam Preparation	170	46%	logo	87.7%
3613 Podiatry	316	51%	rfc	85.9%
2920 Pharmacology (nursing)	254	70%	mnbo	84.6%
3404 Small Animals	1,331	40%	gbc	84.3%
3402 Equine	1,243	43%	gbc	84.3%
1801 Decision Sciences (misc.)	211	38%	mnb, cnb	83.7%
3401 Veterinary (misc.)	1,222	46%	gbc	83.0%
3001 Pharmacology, Toxicology & Pharmaceutics (misc.)	2,042	59%	rfc	82.3%
3601 Health Professions (misc.)	2,510	56%	mnb	81.9%

Table 5. The 10 out of 326 Scopus narrow fields in 2020 with the *lowest* accuracy above baseline for any machine learning algorithm (training set 1,000, or 90% if under 1,000; 1,000 features)

Narrow field (2020)	Articles	Baseline	Top	AOB
3304 Education	55,756	35%	rfc	22.4%
1205 Classics	671	70%	mnb	22.1%
2603 Analysis	7,297	36%	mnb	21.5%
3002 Drug Discovery	22,715	48%	gbc,rfc	21.0%
3200 Psychology (all)	15,162	42%	gbc	20.9%
1202 History	15,986	36%	mnb	20.9%
2602 Algebra & Number Theory	4,585	43%	rfc	17.7%
2600 Mathematics (all)	22,277	41%	rfc	16.1%
1507 Fluid Flow & Transfer Processes	17,493	79%	rfc	5.2%
2101 Energy (misc.)	6,818	88%	mnb	4.9%

in contrast, may be jargon-dense, with little overlap between articles in the terminologies used for the relatively specialist topic addressed in each one. As citations have little relevance to mathematics and the humanities, the citation data may also not be useful in these fields.

The high accuracy set for 2020 has four overlaps with the 2014 set and a similar pattern of mainly small or miscellaneous narrow fields (Table 4). The top chi-square terms for the highest scoring narrow field with at least 1,000 articles, Small Animals, suggest a combination of animal-specificity and incompletely cleaned structured abstracts. The top and third terms are "and relevance" and "relevance", from the nonstandard structured abstract phrase "Conclusions and relevance" in the *Journal of Feline Medicine and Surgery*. This journal also had other nonstandard abstract headings, such as "Case summary" and "Relevance and novel information." The second and fifth terms were animal-specific, "cats" and "in cats," associating with the two feline journals. The high accuracy was also helped by the presence of a journal specializing in reproduction, *Theriogenology*, which is associated with a set of relatively unique terminology with high chi-square scores, including "embryo," "sperm," "pregnancy," and "oocytes."

The low accuracy set for 2020 has three overlaps with the 2014 set and a similar pattern of two general "all" narrow fields, three mathematics fields, and two humanities fields (Table 5). Nevertheless, some other fields do not fit this pattern. In particular, Energy (misc.) is an anomaly. This Scopus field had its top third dominated by a single general journal, *Energies* (5,219 articles), and the generalities of the topics in this journal make the task of machine learning difficult from text. A similar issue occurred for Fluid Flow and Transfer Processes, with a different single large general top third journal: *Applied Sciences* (8,396 articles).

3.3. Terms with the Highest Chi-Square Value in Each Field

A manual analysis of the terms with the highest chi-square value for each of the 314 Scopus narrow fields from 2014 with three categories (the remainder had two) revealed three main contexts (Table 6). In 13% of cases, the term most discriminating between journal thirds, at least in terms of the highest chi-square value, originated from journal mandatory text, such

Table 6. Contexts found for the top chi-square term for the 314 Scopus narrow fields from 2014 with three categories

Context	Fields	Description	Examples
Topic	139 (44%)	The term associates with a topic or method.	"for nursing management," "early childhood," "brain injury," "fixed point," "librarians," "Romania," "setting participants," "urban," "consumers," "electrochemical," "education," "energy," "p," "painter," "vaccine," "wastewater," "wound"
Style	134 (43%)	The term is a stylistic device, whether optional or journal mandated.	"we," "our," "was proposed," "this paper," "this article," "this letter," "find that," "here we," "study on," "the author," "results show," "the present study"
Boilerplate	41 (13%)	The term occurs within journal boilerplate text, such as structured abstract headings.	*Structured abstract headings*: "Key points," "Research purpose," "Statement of problem"; *Journal added keyword*: "issue" (e.g., "issue 91"); *Other*: "available online".

as structured abstract headings or (presumably) mandated keywords. While the initial data cleaning was designed to remove all structured abstract headings, many rare structured headings had not been removed. In theory, these could be removed with additional data filtering steps, although this is time-consuming.

Unsurprisingly, in almost half (44%) of all narrow fields checked from 2014, topic-related terms were the most discriminatory between journal thirds. This category includes some methods-related terms (e.g., "p" [-value], "setting participants") that may primarily differentiate between empirical and conceptual papers, but this dichotomy was not explored due to the difficulty in making this distinction. Some of the topic words also specified a geographic location that may be secondary to the main topic of a paper (e.g., "Romania") for categories with nationally focused journals. The commonness of topic terms is unsurprising because almost all journals have topic specializations, although generalist journals might span the entire scope of a Scopus narrow field. Of course, a topic term can be discriminatory if only one journal in a Scopus narrow field has a narrow scope, as its topic terms will associate with its journal third. Thus, it seems likely that some topic terms are discriminatory in all Scopus narrow fields, even though they are the top terms in under a half.

Perhaps more surprisingly, stylistic devices are the top discriminatory terms in 43% of all Scopus narrow fields. These terms may be optional custom and practice followed by authors in some journals. Conversely, some stylistic terms might be mandated by journals, such as the use of the active voice or first-person plural "we" rather that the passive voice when describing methods. Journals might also suggest or give examples of phrases that might be useful for authors to include in their abstract (e.g., "in this study we show") to ensure that key points are not omitted.

3.4. Prediction Accuracy by Gender, Country, and Institution

If the predicted scores are compared to the actual scores for each article separately for male and female first authors, it is possible to detect whether the prediction algorithms indirectly favor one of these two genders compared to the other (Figure 2). The results do not show universal patterns. The predictions favor females in the United States, Japan, and Brazil but males in Germany and France. The gender advantage varies between method for the other five of the

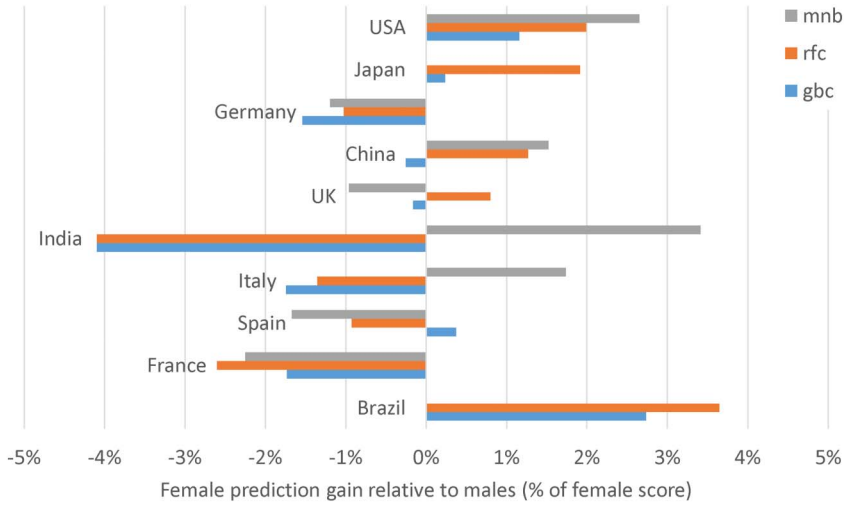

Figure 2. The 2014 relative score increases for women compared to men for the 10 countries with the most articles with a gendered first author. The scores are expressed as a percentage of the original female score.

10 countries with the most gendered first authors. No method seems to systematically favor one of the two genders. The effects are relatively large, however, accounting for a gender shift of up to 4%.

Changing the proxy scores with predictions would have an even more substantial impact on the overall scores of individual institutions (Figure 3). Taking the 10 largest institutional affiliation addresses in the United Kingdom as an example (not merging different affiliations for the same overall university), machine learning could introduce a 7% shift in relative score

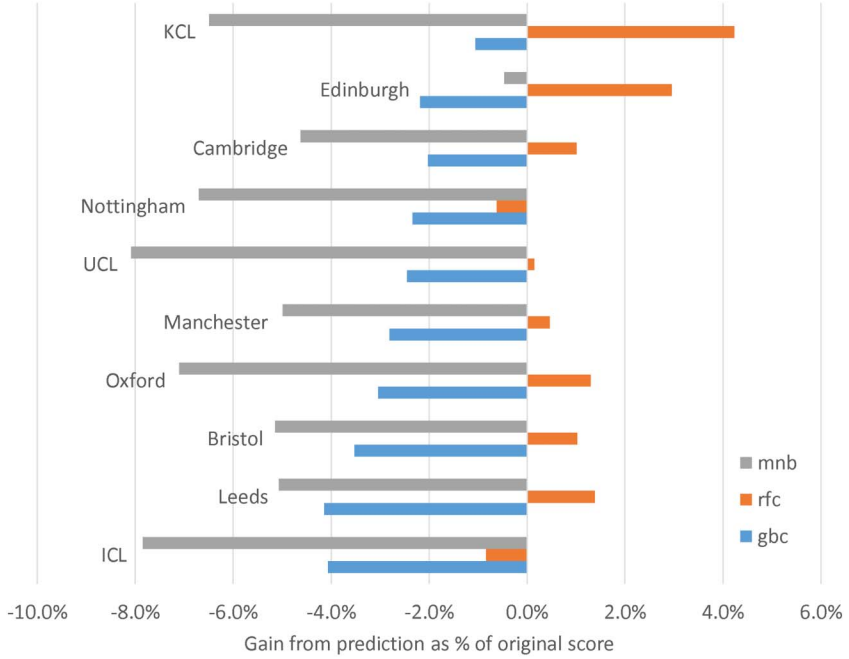

Figure 3. The 2014 relative score increases for the 10 UK institutions with the most articles. The scores are expressed as a percentage of the original institution's score.

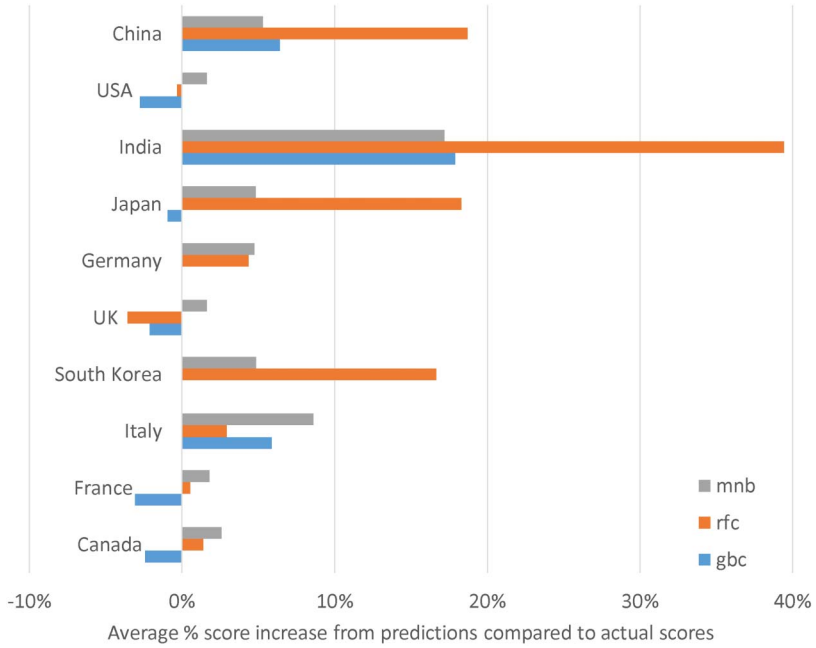

Figure 4. The 2014 relative score increases for the 10 countries with the most articles. The scores are expressed as a percentage of the original institution's score.

between institutions. Because the scores translate into money, this would mean a relative 7% shift in research funding. There is a tendency for the Random Forest Classifier to make prediction gains and the others to make prediction losses for these institutions, but this is not relevant because the funding is shared from a fixed amount for the United Kingdom. The swing is largest for mnb (7%), followed by rfc (5%) and gbc (3%). As the ultimate goal of the REF is to allocate funding to institutions on the basis of the scores, gbc has a substantial advantage over the other two algorithms.

It is relevant to examine international differences in the effect of replacing scores with predictions, in case international organizations, such as the European Union, adopt this approach. The results show that the predictions have an enormous impact on the relative scores of countries, with some increasing substantially and others decreasing (Figure 4). Although based on an artificial experiment, these figures suggest that international comparisons with machine learning would be highly problematic.

4. DISCUSSION

The results are limited by the coverage of Scopus and its categorization of articles into fields primarily at the journal level, which is not optimal (Klavans & Boyack, 2017). Although a large set of machine learning algorithms have been tested, different results may have been gained from others, including an appropriate deep learning framework. Different training set sizes and feature set sizes may also change the results. Similarly, more accurate predictions could be expected if additional features had been included, such as author-level career achievements and citing-cited document information. The results are also limited by the incomplete removal of journal boilerplate text, although this seemed to influence a minority of fields. The JMNLCS is equal to the NLCS for journals with a single article in a year, giving an unfair advantage, although this did not seem to be common. The lack of a development set to select the model to

use for each field is also likely to have resulted in slight overestimation of the accuracy achievable with machine learning models in Tables 2 and 4 reporting the highest accuracy of any method, although it should not affect the scores for the individual methods (e.g., Figure 1). Finally, from an interpretation perspective, recall that associating average citation levels with the quality of the articles in them is inaccurate in all fields. While there may be a moderate statistical association between article quality and journal citation rates in some fields (e.g., Biological Sciences, Clinical Medicine, Economics) there is a weak or even a negative association in others (e.g., the arts and humanities, some social sciences) (Wilsdon et al., 2015b, Table A18).

Compared to previous investigations of machine learning for citation prediction, this study evaluates the most different algorithms and analyzes the most different separate fields. It also has a different target to all previous studies (predicting journal thirds rather than citation counts or citation percentiles), so the results are not directly comparable. Nevertheless, the results confirm the relative accuracy of Support Vector Machines (Zhu & Ban, 2018) and Random Forest and Gradient Boosting Classifiers (Klemiński et al., 2021) for citation-related tasks. In contrast, Multinomial Naïve Bayes is suggested here for the first time as the most accurate algorithm for a minority of narrow fields.

The results show that machine learning can predict the citation-based journal third of articles in all Scopus narrow fields based on its citations and article metadata (excluding journal-related information). While the prediction accuracy tends to be higher for older papers (2014), the same is also true in the worst case for 2020 articles, with citation information collected at the end of the publication year (i.e., January 2021 for articles published in 2020). One implicit factor that text mining machine learning studies can exploit is the topic of papers (Chen & Zhang, 2015): By learning highly cited topics, they can predict how often a paper is likely to be cited from its topic. This factor may help to explain the above-chance predictions for all Scopus narrow fields in 2020, but the predictions can also leverage natural variations between journals in topics (including methods and contexts) and writing styles. Thus, the predictions may be based on topic rather than citations. The substantially greater accuracy for older years suggests that citation factors are important, however, so the predictions tend to be more successful when they can leverage citation-related factors.

4.1. Prediction Accuracy for Individual Documents

The accuracy of the predictions for each narrow field can be increased by focusing on a subset of articles for which the algorithm reports a higher probability of a correct prediction, as follows. Some of the algorithms (including the top three) report a probability that each document falls within each class. The documents for which the probability for one class is much higher than for the other two classes tend to have a higher probability of the prediction being correct than average. If the documents predicted are arranged in descending order of this difference (highest class probability minus second-highest class probability) then a subset of documents can be identified with relatively high prediction accuracy. This would allow an accuracy threshold to be set, accepting the machine learning results for documents falling above the threshold and using an additional round of human reviewer evaluation for the remaining documents. The proportion of articles that can be predicted with a high level of accuracy varies between fields, however. Taking the materials science narrow fields as an example, 40% of the articles in both Materials Science (all) and Ceramics and Composites can have their classes predicted with above 90% accuracy, compared to 5% for Materials Science (misc.) and 2% for Electrical, Optical and Magnetic Materials (Figure 5).

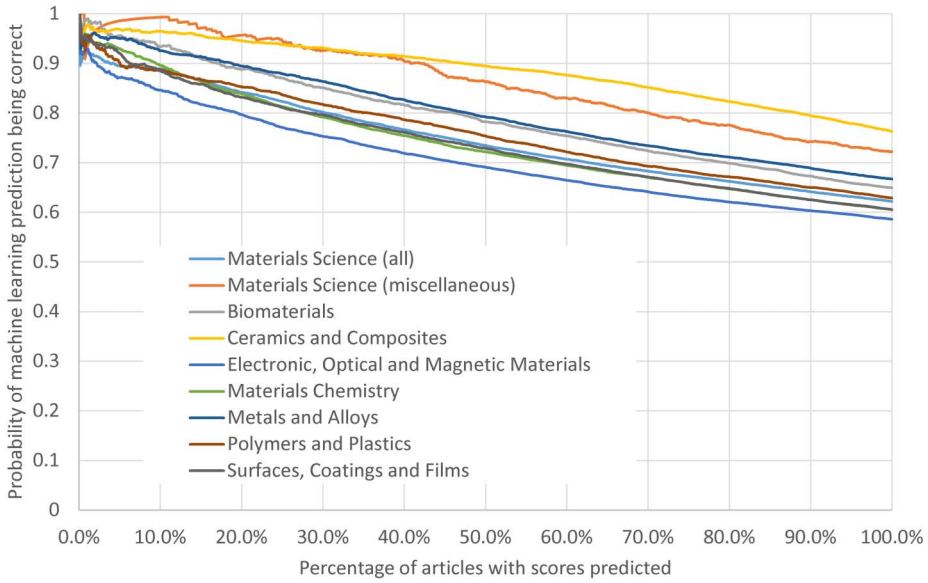

Figure 5. The probability of the machine learning results for the Gradient Boosting Classifier being correct against the prediction likelihood percentile, as reported by the algorithm. The results are for Scopus narrow fields within the Materials Science Scopus broad field in 2014.

4.2. Prediction Accuracy Without Text

Text factors can be excluded from the machine learning inputs to reduce dependency on article topics, leaving three factors: NLCS, number of authors, and number of countries. This reduces the median accuracy for most algorithms (Figure 6, compared to Figure 1) and years, especially for 2020. This reduced number of inputs especially reduces the accuracy of Multinomial Naïve Bayes and the Random Forest Classifier. The leading partial exception is the Gradient Boosting Classifier in 2019 and 2018, although it is not clear why it performs relatively well in these two years. The Gradient Boosting Classifier and its ordinal variant are still considerably more accurate than the remaining algorithms on this reduced set of inputs.

4.3. Implications for Post Peer Review Score Prediction

Returning to the motivating goal of this paper, the results give some insights into the potentials and limitations of estimating the quality of an academic article from citation counts and metadata available at publication time, including title, abstract and keyword text and the number of authors and countries. Author-related factors (e.g., *h*-index) and journal-related factors (e.g., JIF) were not included because they were potentially inappropriate for this type of exercise, where the focus is on evaluating the quality of individual outputs, irrespective of publishing platform or context. Using publishing journal thirds as approximate proxies for article quality (e.g., an article is assumed to be more likely to be high quality if it is in a journal in the top citation third than if it is in a journal in the other two citation thirds), the results suggest that article quality prediction with machine learning is possible to some extent for all fields. Nevertheless, the results are not convincing because the algorithms partly leverage topic and style. More importantly, the fact that the algorithms leverage both topic and style suggests that both have strong associations with journals. Algorithms directly learning article quality from human reviewer scores (rather than using journal impact as a proxy for article quality, as used here) are therefore likely to indirectly learn which journals predominantly publish from one quality category (high, medium, low). As

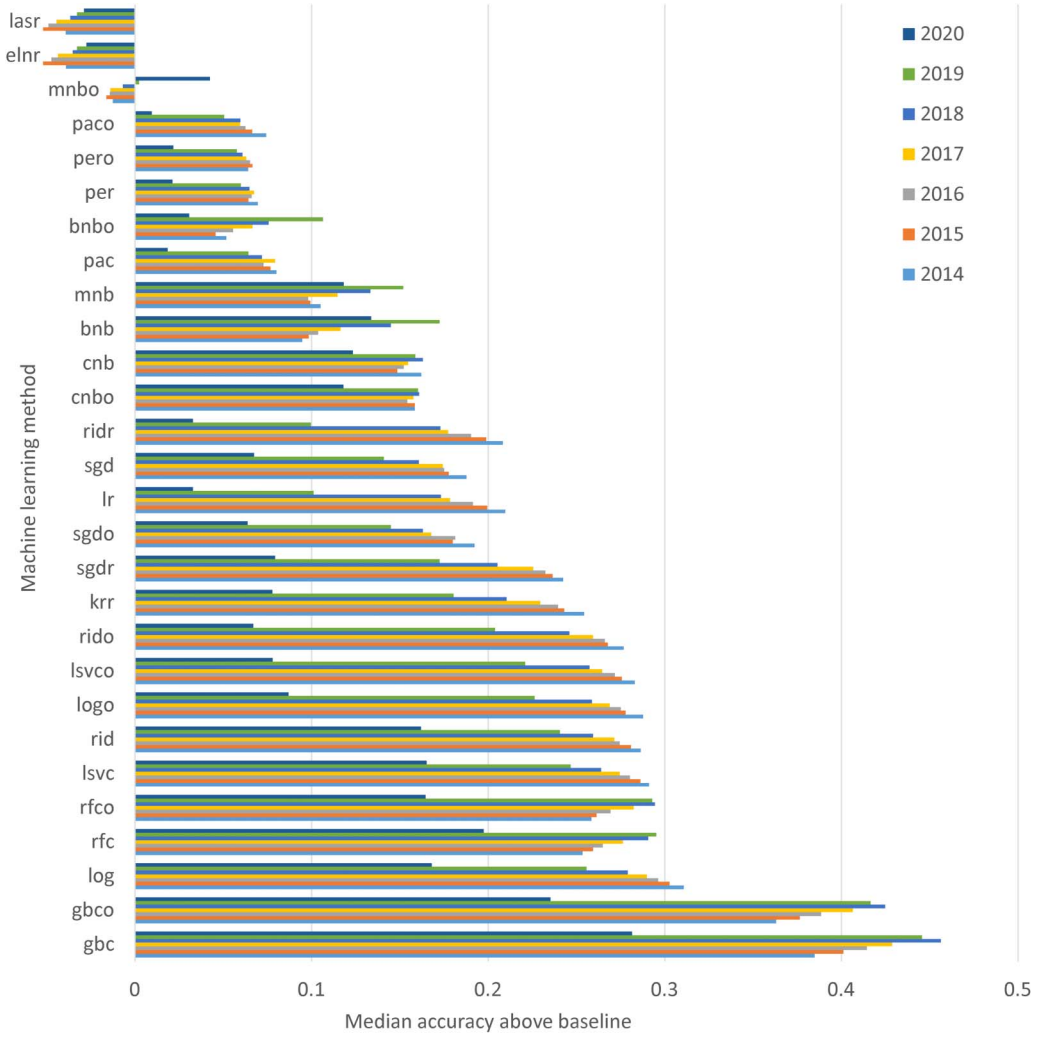

Figure 6. Average (from 30 tries) accuracy above baseline (on a scale of 0 = baseline to 1 = 100% accurate) across the 326 Scopus narrow fields for 29 different machine learning methods (excluding three slow and inaccurate methods compared to Figure 1). Each has a training set of 1,000 articles, using three features (NLCS, authors, countries), and evaluated on the remaining articles.

journals can be learned from article styles or topics, algorithms can predict article quality thirds at least partly from the publishing journal. In the UK REF this is a potential cause for concern because of the explicit instructions to reviewers to ignore publication venues when allocating quality scores. Of course, the human reviewers may consciously or subconsciously leverage similar factors to the machine learning algorithms when making their decisions.

In addition to the journal-related factors discussed above, leveraging topic when predicting article quality (rather than journal third) may be an unwanted characteristic for machine learning for another reason. It would reward weak articles on topics with generally strong results and penalize strong articles in weak research areas (e.g., debunking the weak research or surpassing it by a quantum increase in quality). If topic-related information is excluded then machine learning accuracy is reduced (Figure 6), so this issue needs careful consideration.

In terms of the potential for machine learning to introduce biases, the country-level results strongly suggest that international comparisons based on machine learning are inappropriate due to the potential for substantial shifts in scores caused by predictions. There is also a small

gender effect of prediction, but varying in size and direction between countries. This may be a second-order effect of gender differences in research topics and methods (Thelwall, Bailey et al., 2019). Most worryingly for the United Kingdom, the results suggest that replacing some or all human peer review scores with machine learning predictions could result in substantial shifts between institutions in the allocation of the block funding grant based on output scores (Figure 3). If the accurate classifier giving the least variation, gbc, is used, then this amounts to 3% in the worst case for the 10 largest institutions, which represents a substantial amount of money. For institutions with fewer articles, the shift can be larger, such as the 18.5% shift from Kingston University (277 articles in 2014; loss of 9.7% from predictions) to Edge Hill University (113 articles in 2014; gain of 8.8% from predictions). Although these figures are based on an artificial task, they illustrate the potential for machine learning to systematically skew results for or against institutions even in the absence of institutional and author career information.

5. CONCLUSIONS

The results suggest that journal citation-based thirds can be predicted with above baseline accuracy in all Scopus narrow fields, even at the end of the year of publication. They also show that the Gradient Boosting Classifier or Random Forest Classifier are the most accurate from the set tested in almost all fields, with Multinomial Naïve Bayes being the most accurate in a minority. Deep learning methods were not tested. The results also show that machine learning can leverage topics and writing styles, which can associate with journals. Thus, even if all journal-level information is excluded from article quality prediction and all journal boilerplate text is removed, algorithms can still leverage indirect indicators of the publishing journal. This undermines the goal of generating algorithms that predict the quality of an article on its own merits rather than indirectly through its publishing journal. Organizations wishing to evaluate article quality without journal-level influences must therefore investigate and discuss further to consider whether the influences found are substantial enough to rule out machine learning altogether. Organizations should also carefully consider the influence of topic on the predictions, irrespective of publishing journals. As topic seems much easier to detect than quality, machine learning algorithms may tend to predict quality primarily based on topic, which may generate a perverse incentive to focus on high citation topics.

In terms of practical applications of machine learning for article quality prediction for comparisons between countries or institutions, the results seem to rule out its use for international comparisons due to variability between countries in results. Differences in average scores between institutions are also substantially affected by the machine learning methods for the task here, so this aspect needs to be seriously considered with testing on post peer review scores before any such approach can be implemented. The gender differences in results are perhaps less of a concern because they are smaller than institutional differences but should also be considered as a potential unwanted indirect influence on the scores.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

This research was not funded, but was preparatory work for a project (Gov.uk, 2021) that was funded after the final version of this article was written.

DATA AVAILABILITY

Code and processed data, including the data behind all figures is on Figshare: https://doi.org/10.6084/m9.figshare.17912009.

REFERENCES

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*(2), 485–499. https://doi.org/10.1016/j.joi.2019.02.011

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, *15*(2), 101128. https://doi.org/10.1016/j.joi.2020.101128

Amaral, A., Meek, V. L., Larsen, I. M., Larsen, I. M., & Lars, W. (Eds.). (2003). *The higher education managerial revolution?* Springer Science & Business Media. https://doi.org/10.1007/978-94-010-0072-7

Buckle, R. A., & Creedy, J. (2019). The evolution of research quality in New Zealand universities as measured by the performance-based research fund process. *New Zealand Economic Papers*, *53*(2), 144–165. https://doi.org/10.1080/00779954.2018.1429486

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1), 1–11. https://doi.org/10.1057/s41599-020-00703-8

Chen, C. (2017). Science mapping: a systematic review of the literature. *Journal of Data and Information Science*, *2*(2), 1–40. https://doi.org/10.1515/jdis-2017-0006

Chen, J., & Zhang, C. (2015). Predicting citation counts of papers. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI&CC)* (pp. 434–440). Los Alamitos: IEEE Press. https://doi.org/10.1109/ICCI-CC.2015.7259421

Fairclough, R., & Thelwall, M. (2022). Questionnaires mentioned in academic research 1996–2019: Rapid increase but declining citation impact. *Learned Publishing*. https://doi.org/10.1002/leap.1417

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys (CSUR)*, *52*(6), 1–42. https://doi.org/10.1145/3345317

Franceschini, F., & Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, *11*(2), 337–357. https://doi.org/10.1016/j.joi.2017.02.005

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270. https://doi.org/10.1007/s11192-010-0160-5

Gov.uk. (2021). PS21219 *The responsible use of technology-assisted research assessment*. https://www.contractsfinder.service.gov.uk/Notice/a24d1724-c7cd-48eb-8baf-6ed34c8af428

Haddawy, P., Hassan, S. U., Asghar, A., & Amin, S. (2016). A comprehensive examination of the relation of three citation-based journal metrics to expert judgment of journal quality. *Journal of Informetrics*, *10*(1), 162–173. https://doi.org/10.1016/j.joi.2015.12.005

Hall, K. L., Vogel, A. L., Huang, G. C., Serrano, K. J., Rice, E. L., ... Fiore, S. M. (2018). The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, *73*(4), 532. https://doi.org/10.1037/amp0000319, PubMed: 29792466

Heo, G. E., Kang, K. Y., Song, M., & Lee, J. H. (2017). Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. *BMC Bioinformatics*, *18*(7), 45–57. https://doi.org/10.1186/s12859-017-1640-x, PubMed: 28617229

Hinze, S., Butler, L., Donner, P., & McAllister, I. (2019). Different processes, similar results? A comparison of performance assessment in three countries. In *Springer handbook of science and technology indicators* (pp. 465–484). Berlin: Springer. https://doi.org/10.1007/978-3-030-02511-3_18

Kim, Y. M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. *Health Informatics Journal*, *24*(4), 432–452. https://doi.org/10.1177/1460458216678443, PubMed: 30376768

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Klemiński, R., Kazienko, P., & Kajdanowicz, T. (2021). Where should I publish? Heterogeneous, networks-based prediction of paper's citation success. *Journal of Informetrics*, *15*(3), 101200. https://doi.org/10.1016/j.joi.2021.101200

Kulczycki, E., Korzeń, M., & Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics*, *11*(1), 282–298. https://doi.org/10.1016/j.joi.2017.01.001

Lei, L., & Yan, S. (2016). Readability and citations in information science: Evidence from abstracts and articles of four journals (2003–2012). *Scientometrics*, *108*(3), 1155–1169. https://doi.org/10.1007/s11192-016-2036-9

Li, S., Zhao, W. X., Yin, E. J., & Wen, J. R. (2019). A neural citation count prediction model based on peer review text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4914–4924). https://doi.org/10.18653/v1/D19-1497

Maier, G. (2006). Impact factors and peer judgment: The case of regional science journals. *Scientometrics*, *69*(3), 651–667. https://doi.org/10.1007/s11192-006-0175-0

McCannon, B. C. (2019). Readability and research impact. *Economics Letters*, *180*, 76–79. https://doi.org/10.1016/j.econlet.2019.02.017

Mohammadi, E., & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. *Scientometrics*, *97*(2), 383–395. https://doi.org/10.1007/s11192-013-0993-9

Nie, B., & Sun, S. (2017). Using text mining techniques to identify research trends: A case study of design research. *Applied Sciences*, *7*(4), 401. https://doi.org/10.3390/app7040401

Nuijten, M. B., & Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, *11*(5), 574–579. https://doi.org/10.1002/jrsm.1408, PubMed: 32275351

Parks, S., Rodriguez-Rincon, D., Parkinson, S., & Manville, C. (2019). *The changing research landscape and reflections on national research assessment in the future*. https://dera.ioe.ac.uk

/34336/2/RAND%20summary.pdf. https://doi.org/10.7249/RR3200
Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, *14*(3), 101039. https://doi.org/10.1016/j.joi.2020.101039
Serenko, A., & Bontis, N. (2021). Global ranking of knowledge management and intellectual capital academic journals: A 2021 update. *Journal of Knowledge Management*, *26*(1), 126–145. https://doi.org/10.1108/JKM-11-2020-0814
Serenko, A., & Dohan, M. (2011). Comparing the expert survey and citation impact journal ranking methods: Example from the field of artificial intelligence. *Journal of Informetrics*, *5*(4), 629–648. https://doi.org/10.1016/j.joi.2011.06.002
Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics*, *110*(1), 243–251. https://doi.org/10.1007/s11192-016-2161-5
Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics*, *9*(3), 642–657. https://doi.org/10.1016/j.joi.2015.06.005
Su, Z. (2020). Prediction of future citation count with machine learning and neural network. In *2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)* (pp. 101–104). Los Alamitos, CA: IEEE Press. https://doi.org/10.1109/IPEC49694.2020.9114959
Thelwall, M. (2002). Conceptualizing documentation on the web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society for Information Science and Technology*, *53*(12), 995–1005. https://doi.org/10.1002/asi.10135
Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, *11*(1), 128–151. https://doi.org/10.1016/j.joi.2016.12.002
Thelwall, M., Bailey, C., Tobin, C., & Bradshaw, N. A. (2019). Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? *Journal of Informetrics*, *13*(1), 149–169. https://doi.org/10.1016/j.joi.2018.12.002
Thelwall, M., Papas, E. R., Nyakoojo, Z., Allen, L., & Weigert, V. (2020). Automatically detecting open academic review praise and criticism. *Online Information Review*, *44*(5), 1057–1076. https://doi.org/10.1108/OIR-11-2019-0347
Thelwall, M. & Nevill, T. (2021). Is research with qualitative data more prevalent and impactful now? Interviews, case studies, focus groups and ethnographies. *Library & Information Science Research*, *43*(2), 101094. https://doi.org/10.1016/j.lisr.2021.101094
Traag, V. A., & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, *5*(1), 1–12. https://doi.org/10.1057/s41599-019-0233-x
Wang, M., Jiao, S., Zhang, J., Zhang, X., & Zhu, N. (2020). Identification high influential articles by considering the topic characteristics of articles. *IEEE Access*, *8*, 107887–107899. https://doi.org/10.1109/ACCESS.2020.3001190
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., ... Johnson, B. (2015a). *The metric tide. Report of the independent review of the role of metrics in research assessment and management*. https://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html. https://doi.org/10.4135/9781473978782
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., ... Johnson, B. (2015b). *The metric tide. Report of the independent review of the role of metrics in research assessment and management. Correlation analysis supplement*. https://doi.org/10.6084/m9.figshare.17912009
Xu, J., Li, M., Jiang, J., Ge, B., & Cai, M. (2019). Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network. *IEEE Access*, *7*, 92248–92258. https://doi.org/10.1109/ACCESS.2019.2927011
Yuan, S., Tang, J., Zhang, Y., Wang, Y., & Xiao, T. (2018). Modeling and predicting citation count via recurrent neural network with long short-term memory. *arXiv*, arXiv:1811.02129. https://doi.org/10.48550/arXiv.1811.02129
Zhao, Q., & Feng, X. (2022). Utilizing citation network structure to predict paper citation counts: A deep learning approach. *Journal of Informetrics*, *16*(1), 101235. https://doi.org/10.1016/j.joi.2021.101235
Zhu, X. P., & Ban, Z. (2018). Citation count prediction based on academic network features. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)* (pp. 534–541). Los Alamitos, CA: IEEE Press. https://doi.org/10.1109/AINA.2018.00084