an open access journal

RESEARCH ARTICLE

Predicting article quality scores with machine learning: The U.K. Research Excellence Framework

Mike Thelwall[1], Kayvan Kousha[1], Paul Wilson[1], Meiko Makita[1], Mahshid Abdoli[1], Emma Stuart[1], Jonathan Levitt[1], Petr Knoth[2], and Matteo Cancellieri[2]

[1]Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, Wolverhampton, UK
[2]Knowledge Media Institute, Open University, Milton Keynes, UK

Keywords: artificial intelligence, bibliometrics, citation analysis, machine learning, scientometrics

Citation: Thelwall, M., Kousha, K., Wilson, P., Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., & Cancellieri, M. (2023). Predicting article quality scores with machine learning: The U.K. Research Excellence Framework. *Quantitative Science Studies*, *4*(2), 547–573. https://doi.org/10.1162/qss_a_00258

DOI:
https://doi.org/10.1162/qss_a_00258

Peer Review:
https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00258

Received: 12 December 2022
Accepted: 10 April 2023

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

Handling Editor:
Ludo Waltman

The MIT Press

ABSTRACT

National research evaluation initiatives and incentive schemes choose between simplistic quantitative indicators and time-consuming peer/expert review, sometimes supported by bibliometrics. Here we assess whether machine learning could provide a third alternative, estimating article quality using more multiple bibliometric and metadata inputs. We investigated this using provisional three-level REF2021 peer review scores for 84,966 articles submitted to the U.K. Research Excellence Framework 2021, matching a Scopus record 2014–18 and with a substantial abstract. We found that accuracy is highest in the medical and physical sciences Units of Assessment (UoAs) and economics, reaching 42% above the baseline (72% overall) in the best case. This is based on 1,000 bibliometric inputs and half of the articles used for training in each UoA. Prediction accuracies above the baseline for the social science, mathematics, engineering, arts, and humanities UoAs were much lower or close to zero. The Random Forest Classifier (standard or ordinal) and Extreme Gradient Boosting Classifier algorithms performed best from the 32 tested. Accuracy was lower if UoAs were merged or replaced by Scopus broad categories. We increased accuracy with an active learning strategy and by selecting articles with higher prediction probabilities, but this substantially reduced the number of scores predicted.

1. INTRODUCTION

Many countries systematically assess the outputs of their academic researchers to monitor progress or reward achievements. A simple mechanism for this is to set bibliometric criteria to gain rewards, such as awarding funding for articles with a given Journal Impact Factor (JIF). Several nations have periodic evaluations of research units instead. These might be simultaneous nationwide evaluations (e.g., Australia, New Zealand, United Kingdom: Buckle & Creedy, 2019; Hinze, Butler et al., 2019; Wilsdon, Allen et al., 2015), or rolling evaluations for departments, fields, or funding initiatives (e.g., The Netherlands' Standard Evaluation Protocol: Prins, Spaapen, & van Vree, 2016). Peer/expert review, although imperfect, seems to be the most desirable system because reliance on bibliometric indicators can disadvantage some research groups, such as those that focus on applications rather than theory or methods development, and bibliometrics are therefore recommended for a supporting role (CoARA, 2022; Hicks, Wouters et al., 2015; Wilsdon et al., 2015). Nevertheless, extensive peer review requires a substantial time investment from experts with the skill to assess academic research quality and there is a risk of human bias, which are major disadvantages. In response, some systems inform peer review with bibliometric indicators (United Kingdom: Wilsdon et al., 2015) or automatically score outputs that meet certain criteria, reserving human reviewing for the remainder (as Italy has done: Franceschini & Maisano, 2017). In this article we assess a third approach: machine learning to estimate the score of some or all outputs in a periodic research assessment, as previously proposed (Thelwall, 2022). It is evaluated for the first time here with postpublication peer review quality scores for a large set of journal articles (although prepublication quality scores mainly for conference papers in computational linguistics have been predicted: Kang, Ammar et al., 2018).

The background literature relevant to predicting article scores with machine learning has been introduced in an article (Thelwall, 2022) that also reported experiments with machine learning to predict the citation rate of an article's publishing journal as a proxy article quality measurement. This study found that the Gradience Boosting Classifier was the most accurate out of 32 classifiers tested. Its accuracy varied substantially between the 326 Scopus narrow fields that it was applied to, but it had above-chance accuracy in all fields. The text inputs into the algorithm seemed to leverage journal-related style and boilerplate text information rather than more direct indicators of article quality, however. The level of error in the results was large enough to generate substantial differences between institutions through changed average scores. Another previous study had used simple statistical regression to predict REF scores for individual articles in the 2014 REF, finding that the value of journal impact and article citation counts varied substantially between Units of Assessment (UoAs) (HEFCE, 2015).

Some previous studies have attempted to estimate the quality of computational linguistics conference submissions (e.g., Kang et al., 2018; Li, Sato et al., 2020), with some even writing reviews (Yuan, Liu, & Neubig, 2022). This is a very different task to postpublication peer review for journal articles across fields because of the narrow and unusual topic combined with a proportion of easy predictions not relevant to journal articles, such as out-of-scope, poorly structured, or short submissions. It also cannot take advantage of two powerful postpublication features: citation counts and publication venue.

Although comparisons between human scores and computer predictions for journal article quality tend to assume that the human scores are correct, even experts are likely to disagree and can be biased. An underlying reason for disagreements is that many aspects of peer review are not well understood, including the criteria that articles should be assessed against (Tennant & Ross-Hellauer, 2020). For REF2014 and REF2021, articles were assessed for originality, significance, and rigor (REF2021, 2019), which is the same as the Italian Valutazione della Qualità della Ricerca requirement for originality, impact, and rigor (Bonaccorsi, 2020). These criteria are probably the same as for most journal article peer reviewing, except that some journals mainly require rigorous methods (e.g., *PLOS*, 2022). Bias in peer review can be thought of as any judgment that deviates from the true quality of the article assessed, although this is impossible to measure directly (Lee, Sugimoto et al., 2013). Nonsystematic judgement differences are also common (Jackson, Srinivasan et al., 2011; Kravitz, Franks et al., 2010). Nonsystematic differences may be due to unskilled reviewers, differing levels of leniency or experience (Haffar, Bazerbachi, & Murad, 2019; Jukola, 2017), weak disciplinary norms (Hemlin, 2009), and perhaps also to teams of reviewers focusing on different aspects of a paper (e.g., methods, contribution, originality). Weak disciplinary norms can occur because a field's research objects/subjects and methods are fragmented or because there are different schools of thought about which theories, methods, or paradigms are most suitable (Whitley, 2000).

Sources of systematic bias that have been suggested for nonblinded peer review include malicious bias or favoritism towards individuals (Medoff, 2003), gender (Morgan, Hawkins,

& Lundine, 2018), nationality (Thelwall, Allen et al., 2021), and individual or institutional prestige (Bol, de Vaan, & van de Rijt, 2018). Systematic peer review bias may also be based on language (Herrera, 1999; Ross, Gross et al., 2006), and study topic or approach (Lee et al., 2013). There can also be systematic bias against challenging findings (Wessely, 1998), complex methods (Kitayama, 2017), or negative results (Gershoni, Ishai et al., 2018). Studies that find review outcomes differing between groups may be unable to demonstrate bias rather than other factors (e.g., Fox & Paine, 2019). For example, worse peer review outcomes for some groups might be due to lower quality publications because of limited access to resources or unpopular topic choices. A study finding some evidence of same country reviewer systematic bias that accounted for this difference could not rule out the possibility that it was a second-order effect due to differing country specialisms and same-specialism systematic reviewer bias rather than national bias (Thelwall et al., 2021).

In this article we assess whether it is reasonable to use machine learning to estimate any U.K. REF output scores for journal articles. It is a condensed and partly rephrased version of a longer report (Thelwall, Kousha et al., 2022), with some additional analyses. The main report found that it was not possible to replace all human scores with predictions, and this could also be inferred from the findings reported here. The purpose of the analysis is not to check whether machines can understand research contributions but only to see if they can use available inputs to guess research quality scores accurately enough to be useful in any contexts. The social desirability of this type of solution is out of scope for this article, although it informs the main report. We detail three approaches:

- human scoring for a fraction of the outputs, then machine learning predictions for the remainder;
- human scoring for a fraction of the outputs, then machine learning predictions for a subset of the rest where the predictions have a high probability of being correct, with human scoring for the remaining articles; and
- the active learning strategy to identify sets of articles that meet a given probability threshold.

These are assessed with expert peer review quality scores for most of the journal articles submitted to REF2021. The research questions are as follows, with the final research question introduced to test if the results change with a different standard classification schema. While the research questions mention "accuracy," the experiments instead measure agreement with expert scores from the REF, which are therefore treated as a gold standard. As mentioned above and discussed at the end, the expert scores are in fact also only estimates. Although the results focus on article-level accuracy to start with, the most important type of accuracy is institutional level (Traag & Waltman, 2019), as reported towards the end.

- RQ1: How accurately can machine learning estimate article quality from article metadata and bibliometric information in each scientific field?
- RQ2: Which machine learning methods are the most accurate for predicting article quality in each scientific field?
- RQ3: Can higher accuracy be achieved on subsets of articles using machine learning prediction probabilities or active learning?
- RQ4: How accurate are machine learning article quality estimates when aggregated over institutions?
- RQ5: Is the machine learning accuracy similar for articles organized into Scopus broad fields?

2. METHODS

The research design was to assess a range of machine learning algorithms in a traditional training/testing validation format: training each algorithm on a subset of the data and evaluating it on the remaining data. Additional details and experiments are available in the report that this article was partly derived from (Thelwall et al., 2022).

2.1. Data: Articles and Scores

We used data from two data sources. First, we downloaded records for all Scopus-indexed journal articles published 2014–2020 in January–February 2021 using the Scopus Application Programming Interface (API). This matches the date when the human REF2021 assessments were originally scheduled to begin, so is from the time frame when a machine learning stage could be activated. We excluded reviews and other nonarticle records in Scopus for consistency. The second source was a set of 148,977 provisional article quality scores assigned by the expert REF subpanel members to the articles in 34 UoAs, excluding all data from the University of Wolverhampton. This was confidential data that could not be shared and had to be deleted before May 10, 2022. The distribution of the scores for these articles is online (Figure 3.2.2 of Thelwall et al., 2022). Many articles had been submitted by multiple authors from different institutions and sometimes to different UoAs. These duplicates were eliminated, and the median score retained, or a random median when there were two (for more details, see Section 3.2.1 of Thelwall et al., 2022).

The REF data included article DOIs (used for matching with Scopus, and validated by the REF team), evaluating UoA (one of 34), and provisional score (0, 1*, 2*, 3*, or 4*). We merged the REF scores into three groups for analysis: 1 (0, 1* and 2*), 2 (3*), and 3 (4*). The grouping was necessary because there were few articles with scores of 0 or 1*, which gives a class imbalance that can be problematic for machine learning. This is a reasonable adjustment because 0, 1*, and 2* all have the same level of REF funding (zero), so they are financially equivalent.

We matched the REF outputs with journal articles in Scopus with a registered publication date from 2014 to 2020 (Table 1). Matching was primarily achieved through DOIs. We checked papers without a matching DOI in Scopus against Scopus by title, after removing nonalphabetic characters (including spaces) and converting to lowercase. We manually checked title matches for publication year, journal name, and author affiliations. When there was a disagreement between the REF registered publication year and the Scopus publication year, we always used the Scopus publication year. The few articles scoring 0 appeared to be mainly anomalies, seeming to have been judged unsuitable for review due to lack of evidence of substantial author contributions or being an inappropriate type of output. We excluded these because no scope-related information was available to predict score 0s from.

Finally, we also examined a sample of articles without an abstract. For the five fields with the highest machine learning accuracy in preliminary tests, these were mainly short format (e.g., letters, communications) or nonstandard articles (e.g., guidelines), although data processing errors accounted for a minority of articles with short or missing abstracts. Short format and unusual articles seem likely to be difficult to predict with machine learning, so we excluded articles with abstracts shorter than 500 characters. Unusual articles are likely to be difficult to predict because machine learning relies on detecting patterns, and short format articles could be letter-like (sometimes highly cited) or article-like, with the former being the difficult to predict type. Thus, the final set was cleansed of articles that could be identified in advance as unsuitable for machine learning predictions. The most accurate predictions were found for the years 2014–18, with at least 2 full years of citation data, so we reported these for the main analysis as the highest accuracy subset.

Table 1. Descriptive statistics for creation of the experimental data set

Set of articles	Journal articles
REF2021 journal articles supplied	148,977
With DOI	147,164 (98.8%)
With DOI and matching Scopus 2014–20 by DOI	133,218 (89.4%)
Not matching Scopus by DOI but matching with Scopus 2014–20 by title	997 (0.7%)
Not matched in Scopus and excluded from analysis	14,762 (9.9%)
All REF2021 journal articles matched in Scopus 2014–20	134,215 (90.1%)
All REF2021 journal articles matched in Scopus 2014–20 except score 0	134,031 (90.0%)
All nonduplicate REF2021 journal articles matched in Scopus 2014–20 except score 0	122,331 [90.0% effective]
All nonduplicate REF2021 journal articles matched in Scopus **2014–18** except score 0. These are the most accurate prediction years	87,739
All nonduplicate REF2021 journal articles matched in Scopus **2014–18** except score 0 and except articles with less than 500 character cleaned abstracts	84,966

The 2014–18 articles were mainly from Main Panel A (33,256) overseeing UoAs 1–6, Main Panel B (30,354) overseeing UoAs 7–12, and Main Panel C (26,013) overseeing UoAs 13–24, with a much smaller number from Main Panel D (4,209) overseeing UoAs 25–34. The number per UoA 2014–18 varied by several orders of magnitude, from 56 (Classics) to 12,511 (Engineering), as shown below in a results table. The number of articles affects the accuracy of machine learning and there were too few in Classics to build machine learning models.

2.2. Machine Learning Inputs

We used textual and bibliometric data as inputs for all the machine learning algorithms. We used all inputs shown in previous research to be useful for predicting citations counts, as far as possible, as well as some new inputs that seemed likely to be useful. We also adapted inputs used in previous research to use bibliometric best practice, as described below. The starting point was the set of inputs used in a previous citation-based study (Thelwall, 2022) but this was extended. The nontext inputs were tested with ordinal regression before the machine learning to help select the final set.

The citation data for several inputs was based on the normalized log-transformed citation score (NLCS) or the mean normalized log-transformed citation score (MNLCS) (for detailed explanations and justification; see Thelwall, 2017). The NLCS of an article uses log-transformed citation counts, as follows. First, we transformed all citation counts with the natural log: ln(1 + x). This transformation was necessary because citation count data is highly skewed and taking the arithmetic mean of a skewed data set can give a poor central tendency estimate (e.g., in theory, one extremely highly cited article could raise the average above all the remaining citation counts). After this, we normalized the log-transformed citation count for each article by dividing by the average log-transformed citation count for its Scopus narrow field and year. We divided articles in multiple Scopus narrow fields instead by the average of the field averages for all these narrow fields. The result of this calculation is an NLCS for each article in the Scopus data set (including those not in the REF). The NLCS of an article is field and year normalized by design, so a score of 1 for any article in any field and year always means that the article has had average log-transformed citation impact for its field and year. We calculated the following from the NLCS values and used them as machine learning inputs.

- Author MNLCS: The average NLCS for all articles 2014–20 in the Scopus data set including the author (identified by Scopus ID).
- Journal MNLCS for a given year: The average NLCS for all articles in the Scopus data set in the specified year from the journal. Averaging log-transformed citation counts instead of raw citation counts gives a better estimate of central tendency for a journal (e.g., Thelwall & Fairclough, 2015).

2.2.1. Input set 1: bibliometrics

The following nine indicators have been shown in previous studies to associate with citation counts, including readability (e.g., Didegah & Thelwall, 2013), author affiliations (e.g., Fu & Aliferis, 2010; Li, Xu et al., 2019a; Qian, Rong et al., 2017; Zhu & Ban, 2018), and author career factors (e.g., Qian et al., 2017; Wen, Wu, & Chai, 2020; Xu, Li et al., 2019; Zhu & Ban, 2018). We selected the first author for some indicators because they are usually the most important (de Moya-Anegon, Guerrero-Bote et al., 2018; Mattsson, Sundberg, & Laget, 2011), although corresponding and last authors are sometimes more important in some fields. We used some indicators based on the maximum author in a team to catch important authors that might appear elsewhere in a list.

1. **Citation counts** (field and year normalized to allow parity between fields and years, log transformed to reduce skewing to support linear-based algorithms).
2. **Number of authors** (log transformed to reduce skewing). Articles with more authors tend to be more cited, so they are likely to also be more highly rated (Thelwall & Sud, 2016).
3. **Number of institutions** (log transformed to reduce skewing). Articles with more institutional affiliations tend to be more cited, so they are likely to also be more highly rated (Didegah & Thelwall, 2013).
4. **Number of countries** (log transformed to reduce skewing). Articles with more country affiliations tend to be more cited, so they are likely to also be more highly rated (Wagner, Whetsell, & Mukherjee, 2019).
5. **Number of Scopus-indexed journal articles of the first author** during the REF period (log transformed to reduce skewing). More productive authors tend to be more cited (Abramo, Cicero, & D'Angelo, 2014; Larivière & Costas, 2016), so this is a promising input.
6. **Average citation rate of Scopus-indexed journal articles by the first author** during the REF period (field and year normalized, log transformed: the MNLCS). Authors with a track record of highly cited articles seem likely to write higher quality articles. Note that the first author may not be the REF submitting author or from their institution because the goal is not to "reward" citations for the REF author but to predict the score of their article.
7. **Average citation rate of Scopus-indexed journal articles by any author** during the REF period (maximum) (field and year normalized, log transformed: the MNLCS). Again, authors with a track record of highly cited articles seem likely to write higher quality articles. The maximum bibliometric score in a team has been previously used in another context (van den Besselaar & Leydesdorff, 2009).
8. **Number of pages of article, as reported by Scopus, or the UoA/Main Panel median if missing from Scopus**. Longer papers may have more content but short papers may be required by more prestigious journals.

9. **Abstract readability**. Abstract readability was calculated using the Flesch-Kincaid grade level score and has shown to have a weak association with citation rates (Didegah & Thelwall, 2013).

2.2.2. Input set 2: bibliometrics + journal impact

Journal impact indicators are expected to be powerful in some fields, especially for newer articles (e.g., Levitt & Thelwall, 2011). The second input set adds a measure of journal impact to the first set. We used the journal MNLCS instead of JIFs as an indicator of average journal impact because field normalized values align better with human journal rankings (Haddawy, Hassan et al., 2016), probably due to comparability between disciplines. This is important because the 34 UoAs are relatively broad, all covering multiple Scopus narrow fields.

10. **Journal citation rate** (field normalized, log transformed [MNLCS], based on the current year for older years, based on 3 years for 1–2-year-old articles).

2.2.3. Input set 3: bibliometrics + journal impact + text

The final input set also includes text from article abstracts. Text mining may find words and phrases associated with good research (e.g., a simple formula has been identified for one psychology journal: Kitayama, 2017). Text mining for score prediction is likely to leverage hot topics in constituent fields (e.g., because popular topic keywords can associate with higher citation counts: Hu, Tai et al., 2020), as well as common methods (e.g., Fairclough & Thelwall, 2022; Thelwall & Nevill, 2021; Thelwall & Wilson, 2016), as these have been shown to associate with above average citation rates. Hot topics in some fields tend to be highly cited and probably have higher quality articles, as judged by peers. This would be more evident in the more stable arts and humanities related UoAs but these are mixed with social sciences and other fields (e.g., computing technology for music), so text mining may still pick out hot topics within these UoAs. While topics easily translate into obvious and common keywords, research quality has unknown and probably field-dependent translation into research quality (e.g., "improved accuracy" [computing] vs. "surprising connection" [humanities]). Thus, text-based predictions of quality are likely to leverage topic-relevant keywords and perhaps methods as indirect indicators of quality rather than more subtle textual expressions of quality. It is not clear whether input sets that include both citations and text information would leverage hot topics from the text, because the citations would point to the hot topics anyway. Similarly, machine learning applied to REF articles may identify the topics or methods of the best groups and learn to predict REF scores from them, which would be accurate but undesirable. Article abstracts were preprocessed with a large set of rules to remove publisher copyright messages, structured abstract headings, and other boilerplate texts (available: https://doi.org/10.6084/m9.figshare.22183441). See the Appendix for an explanation about why SciBERT (Beltagy, Lo, & Cohan, 2019) was not used.

We also included journal names on the basis that journals are key scientific gatekeepers and that a high average citation impact does not necessarily equate to publishing high-quality articles. Testing with and without journal names suggested that their inclusion tended to slightly improve accuracy.

11–1000. **Title and abstract word unigrams, bigrams, and trigrams** within sentences (i.e., words and phrases of two or three words). Feature selection was used (chi squared) to identify the best features, always keeping all Input Set 2 features. **Journal names** are also included, for a total of 990 text inputs, selected from the full set as described below.

2.3. Machine Learning Methods

We used machine learning stages that mirror those of a prior study predicting journal impact classes (Thelwall, 2022) with mostly the same settings. These represent a range of types of established regression and classification algorithms, including the generally best performing for tabular input data. As previously argued, predictions may leverage bibliometric data and text, the latter on the basis that the formula for good research may be identifiable from a text analysis of abstracts. We used 32 machine learning methods, including classification, regression, and ordinal algorithms (Table 2). Regression predictions are continuous and were converted to three class outputs by rounding to integers and rounding down (up) to the maximum (minimum) when out of scale. These include the methods of the prior study (Thelwall, 2022) and the Extreme Gradient Boosting Classifier, which has recently demonstrated good results (Klemiński, Kazienko, & Kajdanowicz, 2021). Accuracy was calculated after training on 10%, 25%, or 50% of the data and evaluated on the remaining articles. These percentages represent a range of realistic options for the REF. Although using 90% of the available data for training is

Table 2. Machine learning methods chosen for regression and classification. Those marked with /o have an ordinal version. Ordinal versions of classifiers conduct two binary classifications (1*–3* vs. 4* and 1*–2* vs 3*–4*) and then choose the trinary class by combining the probabilities from them

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

standard for machine learning, it is not realistic for the REF. Training and testing were repeated 10 times, reporting the average accuracy.

The main differences between the current study and the prior paper (Thelwall, 2022) are as follows.

- Human REF scores instead of journal rankings converted to a three-point scale.
- An additional machine learning method, Extreme Gradient Boosting.
- Hyperparameter tuning with the most promising machine learning methods in an attempt to improve their accuracy.
- REF UoAs instead of Scopus narrow fields as the main analysis grouping, although we still used Scopus narrow fields for field normalization of the citation data (MNLCS and NLCS).
- Additional preprocessing rules to catch boilerplate text not caught by the rules used for the previous article but found during the analysis of the results for that article.
- Abstract readability, average journal impact, number of institutional affiliations, and first/maximum author productivity/impact inputs.
- In a switch from an experimental to a pragmatic perspective, we used percentages of the available data as the training set sizes for the algorithms rather than fixed numbers of articles.
- Merged years data sets: we combined the first 5 years (as these had similar results in the prior study) and all years as well as assessing years individually. The purpose of these is to assess whether grouping articles together can give additional efficiency in the sense of predicting article scores with less training data but similar accuracy.
- Merged subjects data sets: we combined all UoAs within each of the four Main Panel grouping of UoAs to produce four very broad disciplinary groupings. This assesses whether grouping articles together can give additional efficiency in the sense of predicting article scores with less training data but similar accuracy.
- Active learning (Settles, 2011). We used this in addition to standard machine learning. With this strategy, instead of a fixed percentage of the machine learning inputs having human scores, the algorithm selects the inputs for the humans to score. First, the system randomly selects a small proportion of the articles and the human scores for them (i.e., the provisional REF scores) are used to build a predictive model. Next, the system selects another set of articles having predicted scores with the lowest probability of being correct for human scoring (in our case supplying the provisional REF scores). This second set is then added to the machine learning model inputs to rebuild the machine learning model, repeating the process until a prespecified level of accuracy is achieved. For the current article, we used batches of 10% to mirror what might be practical for the REF. Thus, a random 10% of the articles were fed into the machine learning system, then up to eight further batches of 10% were added, selected to be the articles with the lowest AP prediction probability. Active learning has two theoretical advantages: Human coders score fewer of the easy cases that the machine learning system can reliably predict, and scores for the marginal decisions may help to train the system better.
- Correlations are reported.

The most accurate classifiers were based on the Gradient Boosting Classifier, the Extreme Gradient Boosting Classifier, and the Random Forest Classifier, so these are described here. All are based on large numbers of simple decision trees, which make classification suggestions based on a series of decisions about the inputs. For example, Traag and Waltman (2019) proposed citation thresholds for identifying likely 4* articles (top 10* cited in a field). A decision tree could mimic this by finding a threshold for the NLCS input, above which articles would be classified as 4*. It might then find a second, lower, threshold, below which articles would be classified as 1*/2*. A decision tree could also incorporate information from multiple inputs. For example, a previous VQR used dual thresholds for citations and journal impact factors and a decision tree could imitate this by classing an article as 4* if it exceeded an NLCS citation threshold (decision 1) and a MNLCS journal impact threshold (decision 2). Extra rules for falling below a lower citation threshold (decision 3) and lower journal impact threshold (decision 4) might then classify an article as 1*/2*. The remaining articles might be classified by further decisions involving other inputs or combinations of inputs. The three algorithms (gbc, rfc, xgb) all make at least 100 of these simple decision trees and then combine them using different algorithms to produce a powerful inference engine.

We did not use deep learning because there was too little data to exploit its power. For example, we had metadata for 84,966 articles analyzed in 34 nonoverlapping subsets, whereas one standard academic data set used in deep learning is a convenience sample of 124 million full text papers with narrower topic coverage (https://ogb.stanford.edu/docs/lsc/mag240m/). A literature review of technologies for research assessment found no deep learning architectures suitable for the available inputs and no evidence that deep learning would work on small input sets available (Kousha & Thelwall, 2022).

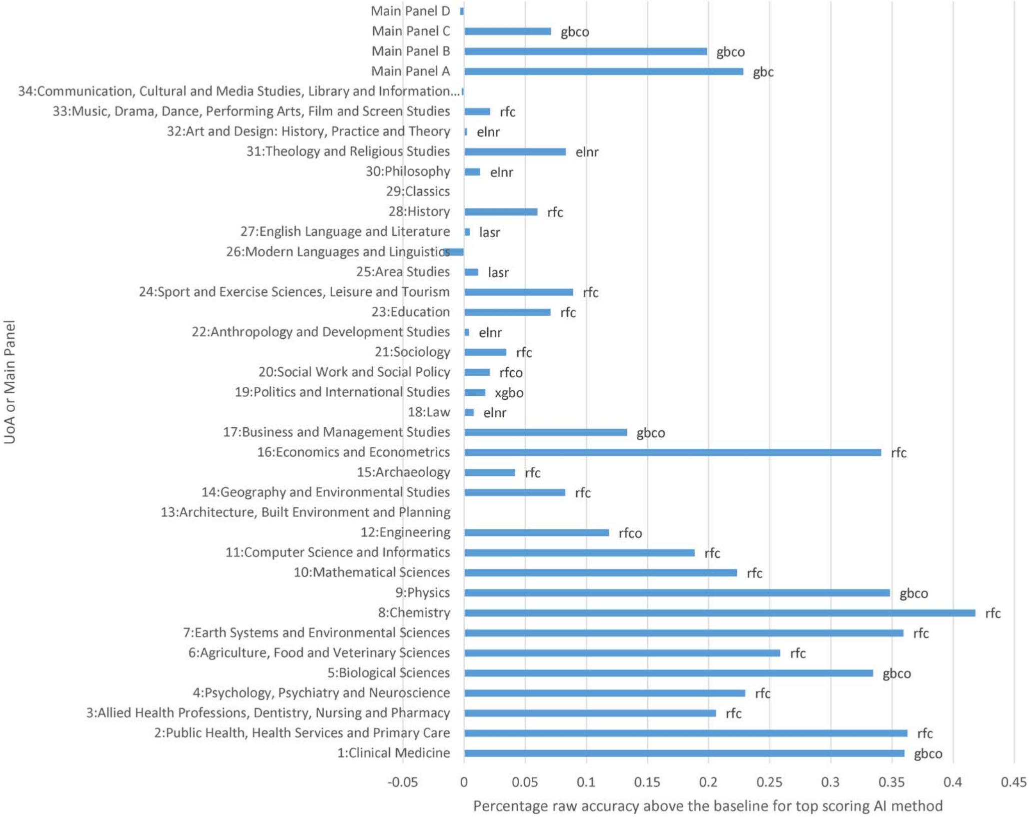

Figure 1. The percentage accuracy above the baseline for the most accurate machine learning method, trained on **50%** of the 2014–18 Input Set 3: Bibliometrics + journal impact + text, after excluding articles with shorter than 500-character abstracts **and excluding duplicate articles within each UoA**. The accuracy evaluation was performed on the articles excluded from the training set. No models were built for Classics due to too few articles. Average across 10 iterations.

3. RESULTS

3.1. RQ1, RQ2: Primary Machine Learning Prediction Accuracy Tests

The accuracy of each machine learning method was calculated for each year range (2014, 2015, 2016, 2017, 2018, 2019, 2020, 2014–18, 2014–20), separately by UoA and Main Panel. The results are reported as accuracy above the baseline (accuracy − baseline)/(1 − baseline), where the baseline is the proportion of articles with the most common score (usually 4* or 3*). Thus, the baseline is the accuracy of always predicting that articles fall within the most common class. For example, if 50% of articles are 4* then 50% would be the baseline and a 60% accurate system would have an accuracy above the baseline of (0.6 − 0.5)(1 − 0.5) = 0.2 or 20%. The results are reported only for 2014–18 combined, with the graphs for the other years available online, as are graphs with 10% or 25% training data, and graphs for Input Set 1 alone and for Input Sets 1 and 2 combined (Thelwall et al., 2022). The overall level of accuracy for each individual year from 2014 to 2018 tended to be similar, with lower accuracy for 2019 and 2020 due to the weaker citation data. Combining 2014 to 2018 gave a similar level of accuracy to that of the individual years, so it is informative to focus on this set. With the main exception of UoA 8 Chemistry, the accuracy of the machine learning methods was higher with 1,000 inputs (Input Set 3) than with 9 or 10 (Input Sets 1 or 2), so only the results for the largest set are reported.

Six algorithms tended to have similar high levels of accuracy (rfc, gbc, xgb, or ordinal variants) so the results would be similar but slightly lower overall if only one of them had been used. Thus, the results slightly overestimate the practically achievable accuracy by cherry-picking the best algorithm.

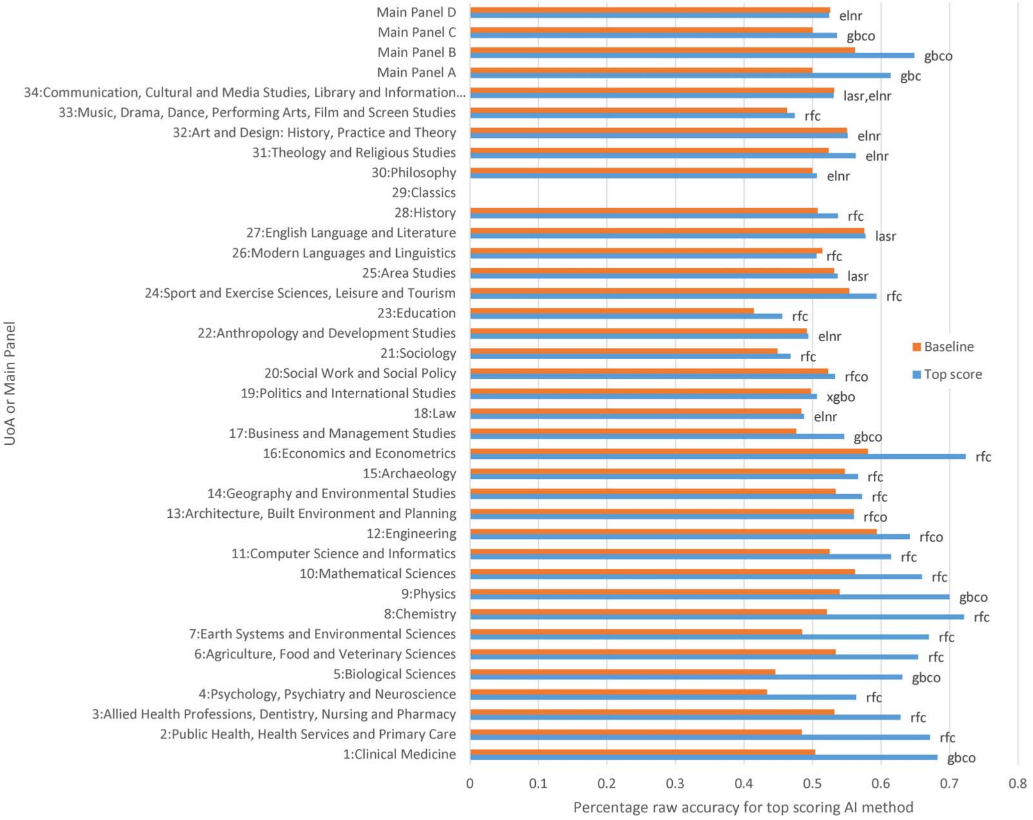

Figure 2. As for the previous figure but showing raw accuracy.

Table 3. Article-level Pearson correlations between machine learning predictions with 50% used for training and actual scores for articles 2014–18, following Strategy 1 (averaged across 10 iterations). L95 and U95 are lower and upper bounds for a 95% confidence interval

Data set	Articles 2014–18	Predicted at 50%	Pearson correlation	L95	U95
1: Clinical Medicine	7,274	3,637	0.562	0.539	0.584
2: Public Health, Health Services & Primary Care	2,855	1,427	0.507	0.467	0.545
3: Allied Health Professions, Dentistry, Nursing & Pharmacy	6,962	3,481	0.406	0.378	0.433
4: Psychology, Psychiatry & Neuroscience	5,845	2,922	0.474	0.445	0.502
5: Biological Sciences	4,728	2,364	0.507	0.476	0.536
6: Agriculture, Food & Veterinary Sciences	2,212	1,106	0.452	0.404	0.498
7: Earth Systems & Environmental Sciences	2,768	1,384	0.491	0.450	0.530
8: Chemistry	2,314	1,157	0.505	0.461	0.547
9: Physics	3,617	1,808	0.472	0.435	0.507
10: Mathematical Sciences	3,159	1,579	0.328	0.283	0.371
11: Computer Science & Informatics	3,292	1,646	0.382	0.340	0.423
12: Engineering	12,511	6,255	0.271	0.248	0.294
13: Architecture, Built Environment & Planning	1,697	848	0.125	0.058	0.191
14: Geography & Environmental Studies	2,316	1,158	0.277	0.223	0.329
15: Archaeology	371	185	0.283	0.145	0.411
16: Economics and Econometrics	1,083	541	0.511	0.446	0.571
17: Business & Management Studies	7,535	3,767	0.353	0.325	0.381
18: Law	1,166	583	0.101	0.020	0.181
19: Politics & International Studies	1,595	797	0.181	0.113	0.247
20: Social Work & Social Policy	2,045	1,022	0.259	0.201	0.315
21: Sociology	949	474	0.180	0.091	0.266
22: Anthropology & Development Studies	618	309	0.040	−0.072	0.151
23: Education	2,081	1,040	0.261	0.203	0.317
24: Sport & Exercise Sciences, Leisure & Tourism	1,846	923	0.265	0.204	0.324
25: Area Studies	303	151	0.142	−0.018	0.295
26: Modern Languages and Linguistics	630	315	0.066	−0.045	0.175
27: English Language and Literature	424	212	0.064	−0.071	0.197
28: History	583	291	0.141	0.026	0.252
29: Classics	56	0	–	–	–
30: Philosophy	426	213	0.070	−0.065	0.203

Table 3. (*continued*)

Data set	Articles 2014–18	Predicted at 50%	Pearson correlation	L95	U95
31: Theology & Religious Studies	107	53	0.074	−0.200	0.338
32: Art and Design: History, Practice and Theory	665	332	0.028	−0.080	0.135
33: Music, Drama, Dance, Performing Arts, Film & Screen Studies	350	175	0.164	0.016	0.305
34: Communication, Cultural & Media Studies, Library & Information Management	583	291	0.084	−0.031	0.197

Articles 2014–18 in most UoAs could be classified with above baseline accuracy with at least one of the tested machine learning methods, but there are substantial variations between UoAs (Figure 1). There is not a simple pattern in terms of the types of UoA that are easiest to classify. This is partly due to differences in sample sizes and probably also affected by the variety of the fields within each UoA (e.g., Engineering is a relatively broad UoA compared to Archaeology). Seven UoAs had accuracy at least 0.3 above the baseline, and these are from the health and physical sciences as well as UoA 16: Economics and Econometrics. Despite this variety, the level of machine learning accuracy is very low for all Main Panel D (mainly arts and humanities) and for most of Main Panel C (mainly social sciences).

Although larger sample sizes help the training phase of machine learning (e.g., there is a Pearson correlation of 0.52 between training set size and accuracy above the baseline), the UoA with the most articles (12: Engineering) had only moderate accuracy, so the differences between UoAs are also partly due to differing underlying machine learning prediction difficulties between fields.

Table 4. Institution-level Pearson correlations between machine learning predictions with 50% used for training and actual scores for articles 2014–18, following Strategy 1 (averaged across 10 iterations) and aggregated by institution for UoAs 1–11 and 16

UoA	Actual vs. machine learning predicted average score	Actual vs. machine learning predicted total score
1: Clinical Medicine	0.895	0.998
2: Public Health, Health Services and Primary Care	0.906	0.995
3: Allied Health Professions, Dentistry, Nursing & Pharmacy	0.747	0.982
4: Psychology, Psychiatry and Neuroscience	0.844	0.995
5: Biological Sciences	0.885	0.995
6: Agriculture, Food and Veterinary Sciences	0.759	0.975
7: Earth Systems and Environmental Sciences	0.840	0.986
8: Chemistry	0.897	0.978
9: Physics	0.855	0.989
10: Mathematical Sciences	0.664	0.984
11: Computer Science and Informatics	0.724	0.945
16: Economics and Econometrics	0.862	0.974

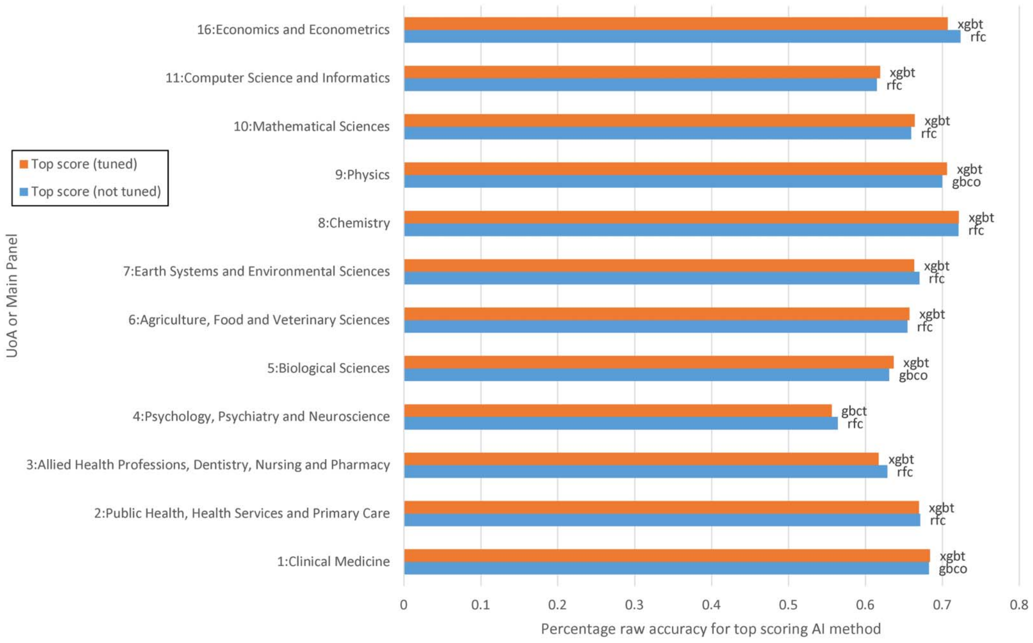

Figure 3. The percentage accuracy for the most accurate machine learning method with and without hyperparameter tuning (out of the main six), trained on **50%** of the 2014–18 articles and **Input set 3: bibliometrics + journal impact + text; 1,000 features in total**. The most accurate method is named.

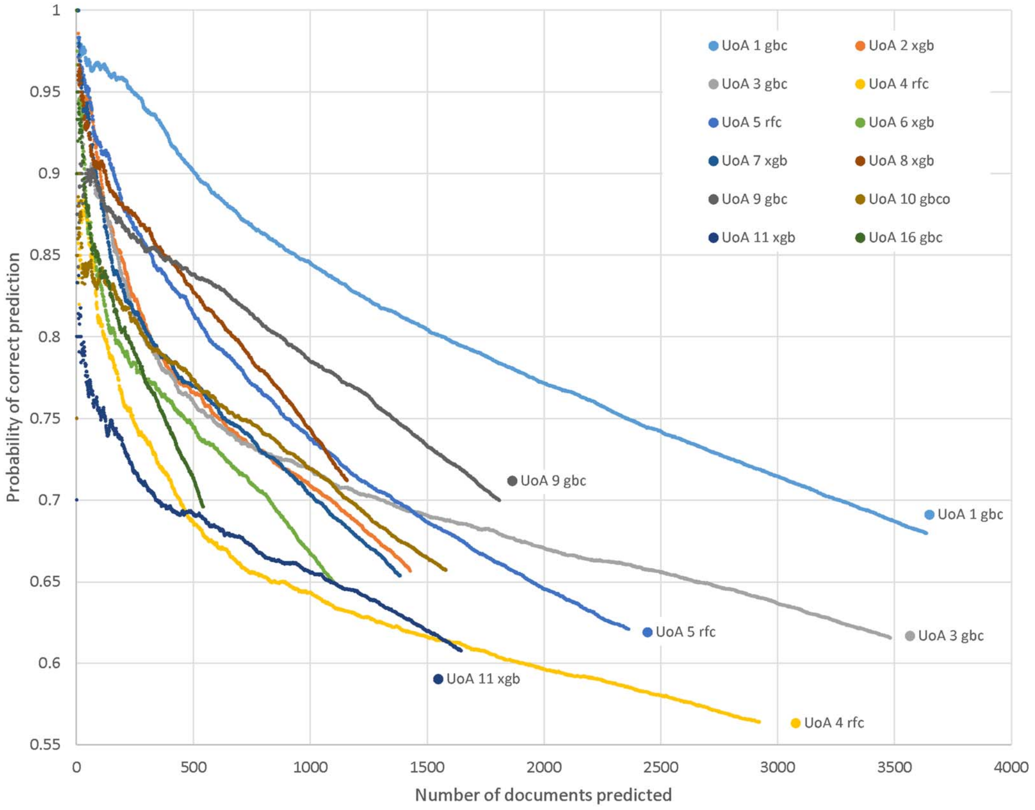

Figure 4. Probability of a machine learning prediction (best machine learning method at the 85% level, trained on 50% of the data 2014–18 with 1,000 features) being correct against the number of predictions for UoAs 1–11, 16. The articles are arranged in order of the probability of the prediction being correct, as estimated by the AI. Each point is the average across 10 separate experiments.

The individual inputs were not tested for predictive power but the three input sets were combined and compared. In terms of accuracy on the three sets, the general rule for predictive power was bibliometrics < bibliometrics + journal impact < bibliometrics + journal impact + text. The differences were mostly relatively small. The main exceptions were Chemistry (bibliometrics alone is best) and Physics (bibliometrics and bibliometrics + journal impact + text both give the best results) (for details see Thelwall et al., 2022).

The most accurate UoAs are not all the same as those with highest accuracy above the baseline because there were substantial differences in the baselines between UoAs (Figure 2). The predictions were up to 72% accurate (UoAs 8, 16), with 12 UoAs having accuracy above 60%. The lowest raw accuracy was 46% (UoA 23). If accuracy is assessed in terms of article-level correlations, then the machine learning predictions always positively correlate with the human scores, but at rates varying between 0.0 (negligible) to 0.6 (strong) (Table 3). These correlations roughly match the prediction accuracies. Note that the correlations are much higher when aggregated by institution, reaching 0.998 for total institutional scores in UoA 1 (Table 4).

Hyperparameter tuning systematically searches a range of input parameters for machine learning algorithms, looking for variations that improve their accuracy. Although this marginally increases accuracy on some UoAs, it marginally reduces it on others, so has little difference overall (Figure 3). The tuning parameters for the different algorithms are in the Python code (https://doi.org/10.6084/m9.figshare.21723227). The same architectures were used for the tuned and untuned cases, with the tuning applying after fold generation to simulate the situation that would be available for future REFs (i.e., no spare data for separate tuning).

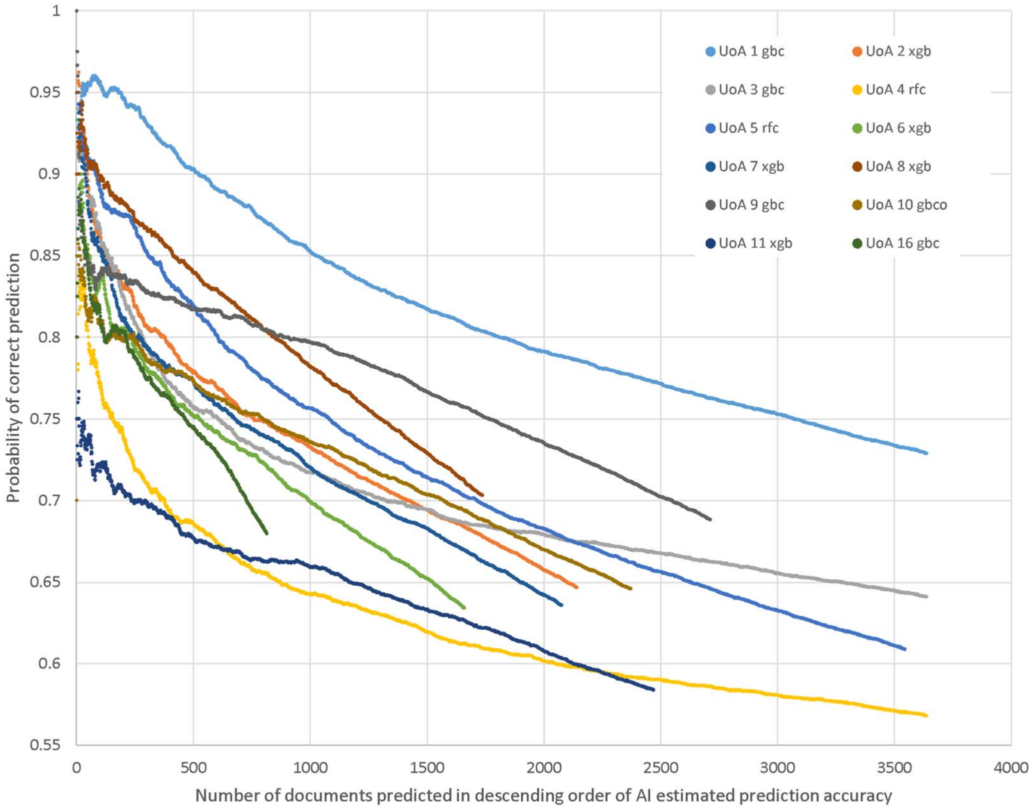

Figure 5. As in Figure 4, but trained on 25% of the data.

3.2. RQ3: High Prediction Probability Subsets

The methods used to predict article scores report an estimate of the probability that these predictions are correct. If these estimates are not too inaccurate, then arranging the articles in descending order prediction probability can be used to identify subsets of the articles that can have their REF score estimated more accurately than for the set overall.

The graphs in Figure 4 for the UoAs with the most accurate predictions can be used to read the number of scores that can be predicted with any given degree of accuracy. For example, setting the prediction probability threshold at 90%, 500 articles could be predicted in UoA 1. The graphs report the accuracy by comparison with subpanel provisional scores rather than the machine learning probability estimates. The graphs confirm that higher levels of machine learning score prediction accuracy can be obtained for subsets of the predicted articles. Nevertheless, they suggest that there is a limit to which this is possible. For example, no UoA can have substantial numbers of articles predicted with accuracy above 95% and UoA 11 has few articles that can be predicted with accuracy above 80%.

If the algorithm is trained on a lower percentage of the articles, then fewer scores can be predicted at any high level of accuracy, as expected (Figure 5).

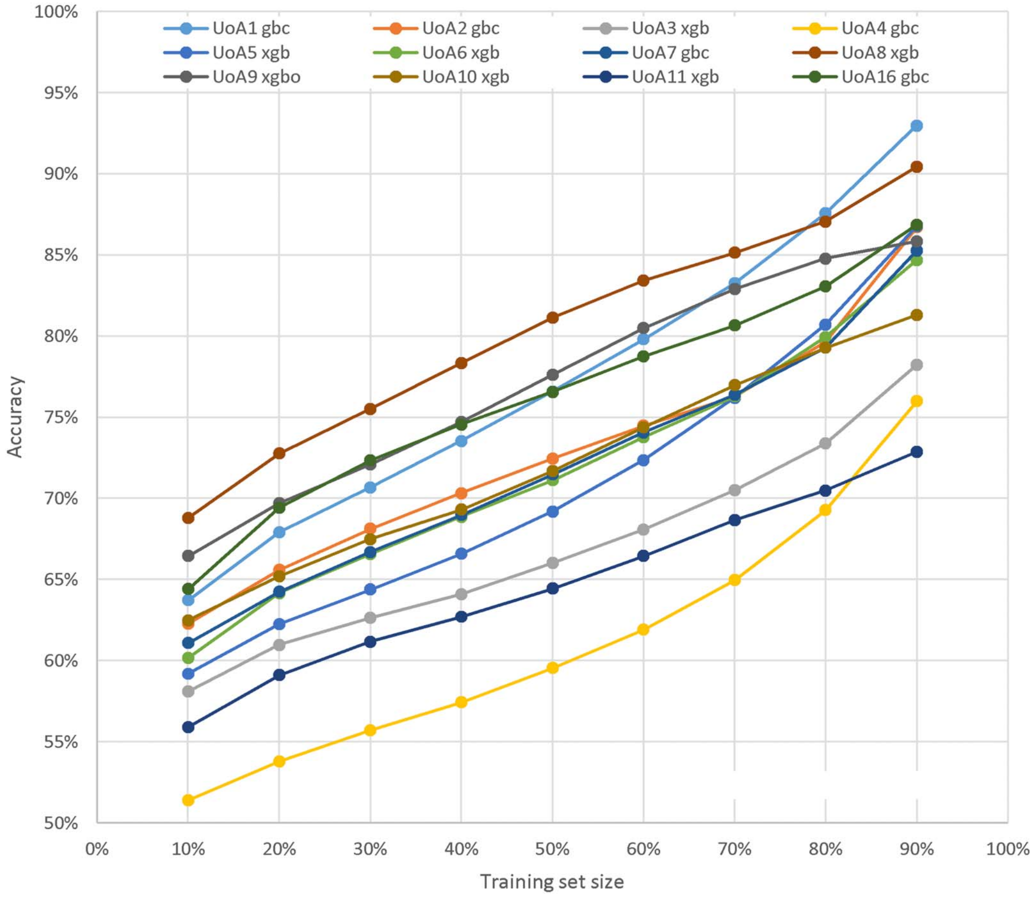

Figure 6. Active learning on UoAs 1–11, 16 showing the results for the machine learning method with the highest accuracy at 90% and 1,000 input features. Results are the average of 40 independent full active learning trials.

3.3. Active Learning Summary

The active learning strategy, like that of selecting high prediction probability scores, is successful at generating higher prediction probability subsets (Figure 6). Active learning works better for some UoAs relative to others, and is particularly effective for UoAs 1, 4, and 5 in the sense that their accuracy increases faster than the others as the training set size increases. The success of the active learning strategy on a UoA depends on at least two factors. First, UoAs with fewer articles will have less data to build the initial model from, so will be less able to select useful outputs for the next stage. Second, the UoAs that are more internally homogeneous will be able to train a model better on low numbers of inputs and therefore benefit more in the early stages.

Active learning overall can predict more articles at realistic thresholds than the high prediction probability strategy (Table 5). Here, 85% is judged to be a realistic accuracy threshold as a rough estimate of human-level accuracy. In the 12 highest prediction probability UoAs, active learning identifies more articles (3,688) than the high prediction probability strategy (2,879) and a higher number in all UoAs where the 85% threshold is reached. Active learning is only less effective when the threshold is not reached.

Table 5. The number of articles that can be predicted at an accuracy above **85%** using active learning or high prediction probability subsets in UoAs 1–11,16. Overall accuracy includes the human scored texts for eligible and ineligible articles

UoA	Human scored articles	Human scored articles (%)	Active learning accuracy (%)	Machine learning predicted articles	High prediction probability articles
1: Clinical Medicine	5,816	80	87.6	1,458	952*
2: Public Health, Health Serv. & Primary Care	2,565	90	86.7	290	181
3: Allied Health Prof., Dentist., Nurs. Pharm.	6,962	100	–	0	163
4: Psychology, Psychiatry & Neuroscience	5,845	100	–	0	66
5: Biological Sciences	4,248	90	86.8	480	308
6: Agriculture, Food & Veterinary Sciences	2,212	100	–	0	86
7: Earth Systems & Environmental Sciences	2,484	90	85.3	284	142
8: Chemistry	1,617	70	85.1	697	402*
9: Physics	3,249	90	85.9	368	362
10: Mathematical Sciences	3,159	100	–	0	86
11: Computer Science & Informatics	3,292	100	**	0	29
16: Economics & Econometrics	972	90	86.9	111	102
Total				**3,688**	**2,879**

* 25% training set size instead of 50% training set size because more articles were predicted.

– The 85% active learning threshold was not reached.

3.4. RQ4: HEI-Level Accuracy

For the U.K. REF, as for other national evaluation exercises, the most important unit of analysis is the institution, because the results are used to allocate funding (or a pass/fail decision) to institutions for a subject rather than to individual articles or researchers (Traag & Waltman, 2019). At the institutional level, there can be nontrivial score shifts for individual institutions, even with high prediction probabilities. UoA 1 has one of the lowest average score shifts (i.e., change due to human scores being partly replaced by machine learning predictions) because of relatively large institutional sizes, but these are still nontrivial (Figure 7). The score shifts are largest for small institutions, because each change makes a bigger difference to the average when there are fewer articles, but there is also a degree of bias, in the sense that institutions then benefit or lose out overall from the machine learning predictions. The biggest score shift for a relatively large number of articles in a UoA (one of the five largest sets in the UoA) is 11% (UoA 7) or 1.9% overall (UoA 8), considering 100% accuracy for the articles given human scores (Table 6). Although 1.9% is a small percentage, it may represent the salaries of multiple members of staff and so is a nontrivial consideration. The institutional score shifts are larger for Strategy 1 (not shown).

Bias occurs in the predictions from active learning, even at a high level of accuracy. For example, in most UoAs, larger HEIs, HEIs with higher average scores, and HEIs submitting

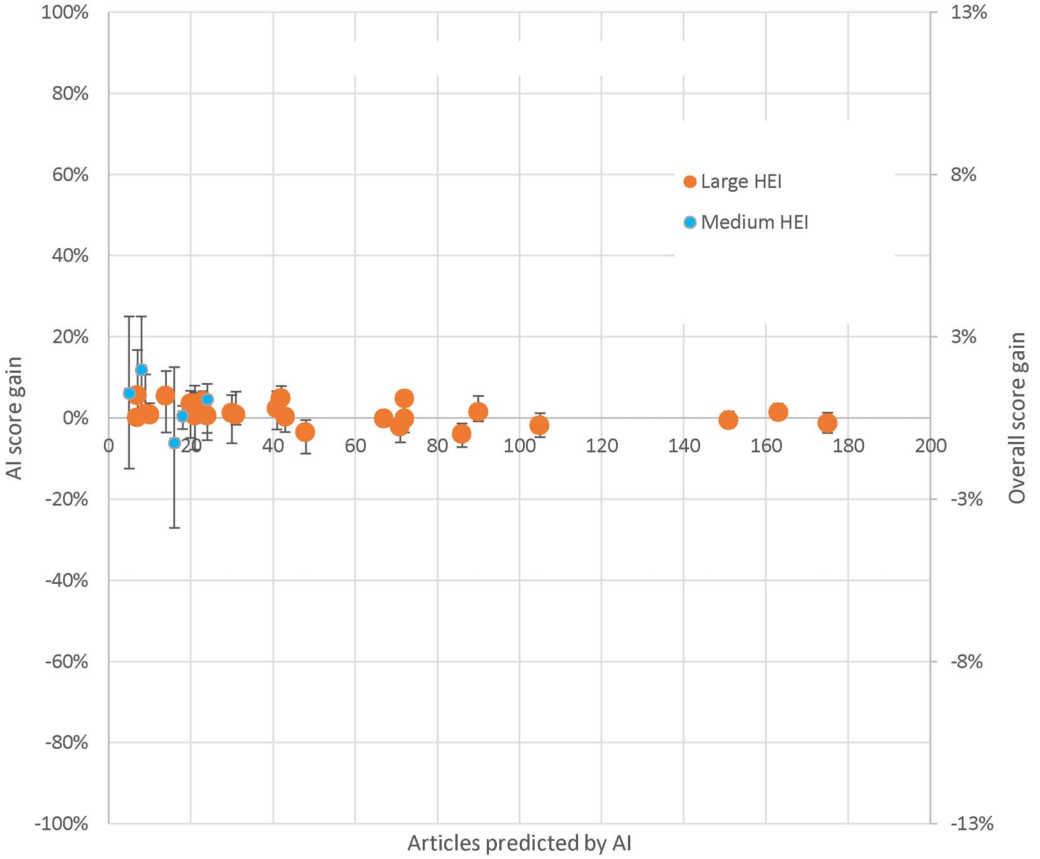

Figure 7. The average REF machine learning institutional score gain on UoA 1: Clinical Medicine for the most accurate machine learning method with active learning, stopping at 85% accuracy on the 2014–18 data and **bibliometric + journal + text inputs, after excluding articles with shorter than 500 character abstracts**. Machine learning score gain is a financial calculation (4* = 100% funding, 3* = 25% funding, 0–2* = 0% funding). The *x*-axis records the number of articles with predicted scores in one of the iterations. The right-hand axis shows the overall score gain for all REF journal articles, included those that would not be predicted by AI. Error bars indicate the highest and lowest values from 10 iterations.

Table 6. Maximum average machine learning score shifts for five largest Higher Educational Institution (HEI) submissions and for all HEI submissions with active learning with an 85% threshold. The same information for the largest machine learning score shifts rather than the average score shifts. Overall figures include all human coded journal articles

UoA or Panel	Human scores (%)	Max. HEI av. score shift (overall) (%)	Max. top 5 HEIs av. score shift (overall) (%)	Max HEI largest score shift (overall) (%)	Max. top 5 HEIs largest score shift (overall) (%)
1: Clinical Medicine	80	12 (1.5)	1.9 (0.2)	27 (3.4)	5.4 (0.7)
2: Public Health, H. Services & Primary Care	90	27 (1.7)	13 (0.8)	75 (4.7)	16 (1.0)
3: Allied Health Prof., Dentist Nurs Pharm	100				
4: Psychology, Psychiatry & Neuroscience	100				
5: Biological Sciences	90	63 (3.9)	7.3 (0.5)	75 (4.7)	10 (0.6)
6: Agriculture, Food & Veterinary Sciences	100				
7: Earth Systems & Environmental Sciences	90	32 (2.0)	11 (0.7)	75 (4.7)	16 (1.0)
8: Chemistry	70	11 (2.1)	10 (1.9)	75 (14)	14 (2.6)
9: Physics	90	10 (0.6)	3.7 (0.2)	75 (4.7)	10 (0.6)
10: Mathematical Sciences	100				
11: Computer Science & Informatics	100				
16: Economics and Econometrics	90	35 (2.2)	5.1 (0.3)	75 (4.7)	19 (1.2)

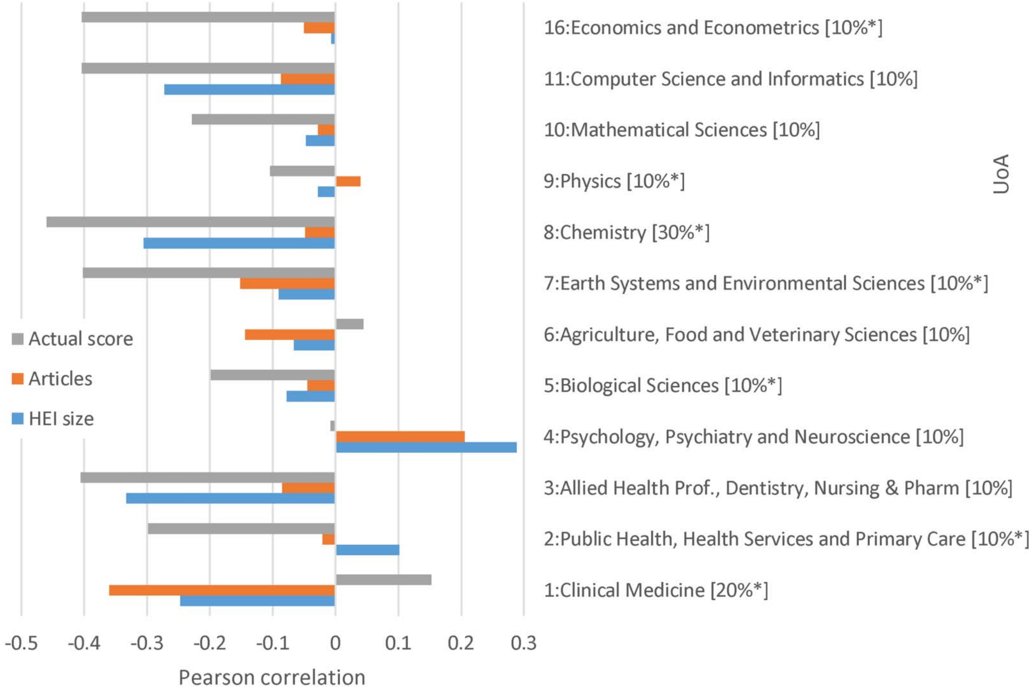

Figure 8. Institution-level Pearson correlations between institutional size (number of articles submitted to REF) or submission size (number of articles submitted to UoA) or average institutional REF score for the UoA and average REF machine learning institutional score gain on UoA 1: Clinical Medicine to UoA 16: Economics and Econometric for the most accurate machine learning method with active learning, stopping at 85% accuracy on the 2014–18 data and **bibliometric + journal + text inputs, after excluding articles with shorter than 500 character abstracts**. Captions indicate the proportion of journal articles predicted, starred if the 85% accuracy active learning threshold is met.

more articles to a UoA tend to be disadvantaged by machine learning score predictions (Figure 8). This is not surprising because, other factors being equal, high-scoring HEIs would be more likely to lose from an incorrect score prediction. This is because they would have a higher proportion of top scoring articles (which would always be downgraded by errors). Similarly, larger HEIs tend to submit more articles and tend to have higher scores.

3.5. RQ5: Accuracy on Scopus Broad Fields

If the REF articles are organized into Scopus broad fields before classification, then the most accurate machine learning method is always gbco, rfc, rfco, or xgbo. The highest accuracy

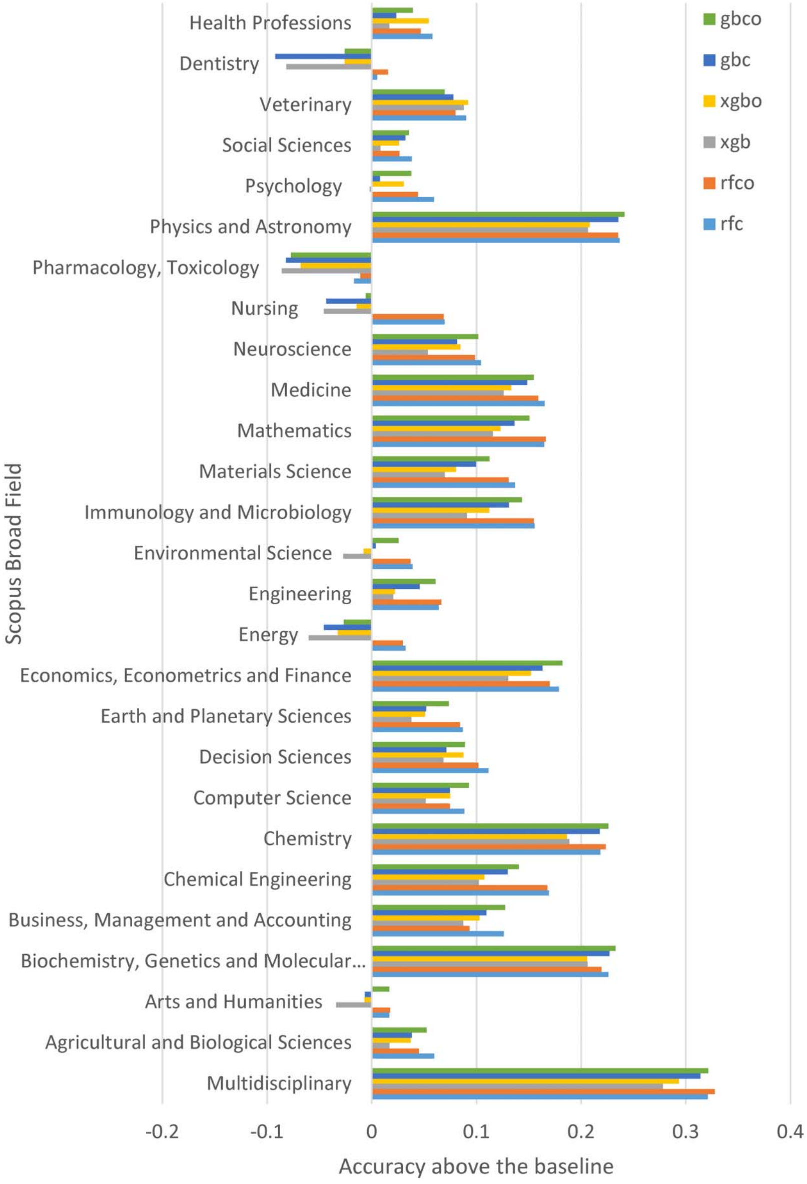

Figure 9. The percentage accuracy above the baseline on Scopus broad fields for the three most accurate machine learning methods and their ordinal variants, trained on **50%** of the 2014–18 Input Set 3: Bibliometrics + journal impact + text, after excluding articles with shorter than 500-character abstracts, zero scores, or duplicate within a Scopus broad field.

above the baseline is generally much lower in this case than for the REF fields, with only Multidisciplinary having accuracy above the baseline above 0.3, with the remainder being substantially lower (Figure 9). The lower accuracy is because the Scopus broad fields are effectively much broader than UoAs. They are journal based rather than article based and journals can be allocated multiple categories. Thus, a journal containing medical engineering articles might be found in both the Engineering and the Medicine categories. This interdisciplinary, broader nature of Scopus broad fields reduces the accuracy of the machine learning methods, despite the field-normalized indicators used in them.

4. DISCUSSION

The results are limited to articles from a single country and period. These articles are self-selected as presumably the best works (1 to 5 per person) of the submitting U.K. academics over the period 2014–2020. The findings used three groups (1*–2*, 3*, 4*) and finer grained outputs (e.g., the 27-point Italian system, 3–30) would be much harder to predict accurately because there are more wrong answers (26 instead of 2) and the differences between scores are smaller. The results are also limited by the scope of the UoAs examined. Machine learning predictions for countries with less Scopus-indexed work to analyze, or with more recent work, would probably be less accurate. The results may also change in the future as the scholarly landscape evolves, including journal formats, article formats, and citation practices. The accuracy statistics may be slightly optimistic due to overfitting: running multiple tests and reporting the best results. This has been mitigated by generally selecting strategies that work well for most UoAs, rather than customizing strategies for UoAs. The main source of overfitting is probably machine learning algorithm selection, as six similar algorithms tended to perform well and only the most accurate one for each UoA is reported.

The results generally confirm previous studies in that the inputs used can generate above-baseline accuracy predictions, and that there are substantial disciplinary differences in the extent to which article quality (or impact) can be predicted. The accuracy levels achieved here are much lower than previously reported for attempts to identify high-impact articles, however. The accuracy levels are also lower than for the most similar prior study, which predicted journal thirds as a simple proxy for article quality, despite using less training data in some cases and a weaker set of inputs (Thelwall, 2022). Collectively, this suggests that the task of predicting article quality is substantially harder than the task of predicting article citation impact. Presumably this is due to high citation specialties not always being high-quality specialties.

Some previous studies have used input features extracted from article full texts for collections of articles where this is easily available. To check whether this is a possibility here, the full text of 59,194 REF-submitted articles was supplied from the core.ac.uk repository of open access papers (Knoth & Zdrahal, 2012) by Petr Knoth, Maria Tarasiuk, and Matteo Cancellieri. These matched 43.3% of the REF articles with scores and strategy 1 was rerun with this reduced set, using the same features with added word counts, character counts, figure counts, and table counts extracted from the full text, but accuracy was lower. This was probably partly due to the full texts often containing copyright statements and line numbers as well as occasional scanning errors, and partly due to the smaller training set sizes. Tests of other suggested features (supplementary materials, data access statements) found very low article-level correlations with scores, so these were not included.

The practical usefulness of machine learning predictions in the REF is limited by a lack of knowledge about the reliability of the human scores. For example, if it were known that reviewing team scores agreed 85% of the time (which was the best, but very crude estimate from the REF data, as in Section 3.2.1 of Thelwall et al., 2022) then machine learning predictions that are at least 85% accurate might be judged acceptable. Nevertheless, assessing this level of accuracy is impossible on REF data because each score is produced by two reviewers only after discussion between themselves and then the wider UoA group, supported by REF-wide norm referencing. Because of this, comparing the agreement between two human reviewers, even if REF panel members, would not reveal the likely agreement rate produced by the REF process. The REF score agreement estimate of 85% mentioned above was for articles in a UoA that seemed to have "accidentally" reviewed multiple copies of the same articles, giving an apparently natural experiment in the consistency of the overall REF process.

This article has not considered practical issues, such as whether those evaluated would attempt to game a machine learning prediction system or whether it would otherwise lead to undesirable behavior, such as targeting high-impact journals or forming citation cartels. These are important issues and led to the recommendation in the report that this article is derived from that even the technically helpful advisory system produced should not be used because of its possible unintended consequences (Thelwall et al., 2022, p. 135). Thus, great care must be taken over any decision to use machine learning predictions, even for more accurate solutions than those discussed here.

It is unfortunate that the REF data set had to be destroyed for legal and ethical reasons, with all those involved in creating, managing, and accessing REF output scores being required to confirm that they had permanently deleted the data in 2022. Although a prior study used journal impact calculations to create an artificial data set for similar experiments (Thelwall, 2022), its design precludes the use of journal-related inputs, which is a substantial drawback, as is the use of journals as proxy sources of article quality information. It does not seem possible to create an anonymized REF data set either (for REF2028) because article titles and abstracts are identifying information and some articles probably also have unique citation counts, so almost no information could be safely shared anonymously without the risk of leaking some REF scores.

5. CONCLUSION

The results show that machine learning predictions of article quality scores on a three-level scale are possible from article metadata, citation information, author career information, and title/abstract text with up to 72% accuracy in some UoAs for articles older than 2 years, given enough articles for training. Substantially higher levels of accuracy may not be possible due to the need for tacit knowledge to understand the context of articles to properly evaluate their contributions. Although academic impact can be directly assessed to some extent through citations, robustness and originality are difficult to assess from citations and metadata, although journals may be partial indicators of these in some fields and original themes can in theory be detected (Chen, Wang et al., 2022). The tacit knowledge needed to assess the three components of quality may be more important in fields (UoAs) in which lower machine learning prediction accuracy was attained. Higher accuracy may be possible with other inputs included, such as article full text (if cleaned and universally available) and peer reviews (if widely available).

The results suggest, somewhat surprisingly, that Random Forest Classifier and the Gradient Boosting Classifier tend to be the most accurate (both classification and ordinal variants) rather than the Extreme Gradient Boosting Classifier. Nevertheless, xgb is the most accurate for some UoAs, and especially if active learning is used.

If high levels of accuracy are needed, small subsets of articles can be identified in some UoAs that can be predicted with accuracy above a given threshold through active learning. This could be used when phased peer review is practical, so that initial review scores could be used to build predictions and a second round of peer review would classify articles with low-probability machine learning predictions. Even at relatively high prediction probability levels, machine learning predictions can shift the scores of small institutions substantially due to statistical variations and larger institutions due to systematic biases in the machine learning predictions, such as against high-scoring institutions.

Although it would be possible to use the models built for the current paper in the next REF (probably 2027), their accuracy could not be guaranteed. For example, the text component would not reflect newer research topics (e.g., any future virus) and future citation data would not be directly comparable, as average citation counts may continue to increase in future years in some specialties.

Finally in terms of the technical side, unless other machine learning approaches work substantially better than those tried here, it seems clear that machine learning prediction of scores is irrelevant for the arts and humanities (perhaps due to small article sets), most of the social sciences, and engineering, and is weak for some of the remaining areas. It would be interesting to try large language models such as ChatGPT for journal article quality classification, although they would presumably still need extensive scored training data to understand the task well enough to perform it.

From the practical perspective of the REF, the achievable accuracy levels, although the highest reported, were insufficient to satisfy REF panel members. This also applied to the prediction by probability subsets, despite even higher overall accuracy (Thelwall et al., 2022). Moreover, the technically beneficial and acceptable solution of providing machine learning predictions and their associated prediction probabilities to REF assessors to support their judgments in some UoAs for articles that were difficult for the experts to classify was not recommended in the main report due to unintended consequences for the United Kingdom (emphasizing journal impact despite initiatives to downplay it: Thelwall et al., 2022). In the U.K. context, substantially improved predictions seem to need more understanding of how peer review works, and close to universal publication of machine readable clean full text versions of articles online so that full text analysis is practical. Steps towards these would therefore be beneficial and this might eventually allow more sophisticated full text machine learning algorithms for published article quality to be developed, including with deep learning if even larger data sets can be indirectly leveraged. From ethical and unintended consequences perspectives, however, the most likely future REF application (and over a decade in the future) is to support reviewers' judgments rather than to replace them.

ACKNOWLEDGMENTS

This study was funded by Research England, Scottish Funding Council, Higher Education Funding Council for Wales, and Department for the Economy, Northern Ireland as part of the Future Research Assessment Programme (https://www.jisc.ac.uk/future-research-assessment-programme). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

AUTHOR CONTRIBUTIONS

Mike Thelwall: Methodology, Writing—original draft, Writing—review & editing. Kayvan Kousha: Methodology, Writing—review & editing. Paul Wilson: Methodology, Writing—review & editing. Meiko Makita: Methodology, Writing—review & editing. Mahshid Abdoli: Methodology, Writing—review & editing. Emma Stuart: Methodology, Writing—review & editing. Jonathan Levitt: Methodology, Writing—review & editing. Petr Knoth: Data curation, Methodology. Matteo Cancellieri: Data curation.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This study was funded by Research England, Scottish Funding Council, Higher Education Funding Council for Wales, and Department for the Economy, Northern Ireland as part of the Future Research Assessment Programme (https://www.jisc.ac.uk/future-research-assessment-programme). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

DATA AVAILABILITY

Extended versions of the results are available in the full report (https://cybermetrics.wlv.ac.uk/ai/). The raw data was deleted before submission to follow UKRI legal data protection policy for REF2021. The Python code is on Figshare (https://doi.org/10.6084/m9.figshare.21723227).

REFERENCES

Abramo, G., Cicero, T., & D'Angelo, C. A. (2014). Are the authors of highly cited articles also the most productive ones? *Journal of Informetrics*, *8*(1), 89–97. https://doi.org/10.1016/j.joi.2013.10.011

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

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific 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. 3615–3620). https://doi.org/10.18653/v1/D19-1371

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, *115*(19), 4887–4890. https://doi.org/10.1073/pnas.1719557115, PubMed: 29686094

Bonaccorsi, A. (2020). Two decades of experience in research assessment in Italy. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.27

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

Chen, Y., Wang, H., Zhang, B., & Zhang, W. (2022). A method of measuring the article discriminative capacity and its distribution. *Scientometrics*, *127*(3), 3317–3341. https://doi.org/10.1007/s11192-022-04371-0

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

CoARA. (2022). *The agreement on reforming research assessment*. https://coara.eu/agreement/the-agreement-full-text/

de Moya-Anegon, F., Guerrero-Bote, V. P., López-Illescas, C., & Moed, H. F. (2018). Statistical relationships between corresponding authorship, international co-authorship and citation impact of national research systems. *Journal of Informetrics*, *12*(4), 1251–1262. https://doi.org/10.1016/j.joi.2018.10.004

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, *7*(4), 861–873. https://doi.org/10.1016/j.joi.2013.08.006

Fairclough, R., & Thelwall, M. (2022). Questionnaires mentioned in academic research 1996–2019: Rapid increase but declining citation impact. *Learned Publishing*, *35*(2), 241–252. https://doi.org/10.1002/leap.1417

Fox, C. W., & Paine, C. T. (2019). Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecology and Evolution*, *9*(6), 3599–3619. https://doi.org/10.1002/ece3.4993, PubMed: 30962913

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

Gershoni, A., Ishai, M. B., Vainer, I., Mimouni, M., & Mezer, E. (2018). Positive results bias in pediatric ophthalmology scientific publications. *Journal of the American Association for Pediatric Ophthalmology and Strabismus*, *22*(5), 394–395. https://doi.org/10.1016/j.jaapos.2018.03.012, PubMed: 30077820

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

Haffar, S., Bazerbachi, F., & Murad, M. H. (2019). Peer review bias: A critical review. *Mayo Clinic Proceedings*, *94*(4), 670–676. https://doi.org/10.1016/j.mayocp.2018.09.004, PubMed: 30797567

HEFCE. (2015). The Metric Tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the independent Review of the Role of Metrics in Research Assessment and Management). Higher Education Funding Council for England. https://www.dcscience.net/2015_metrictideS2.pdf

Hemlin, S. (2009). Peer review agreement or peer review disagreement: Which is better? *Journal of Psychology of Science and Technology*, *2*(1), 5–12. https://doi.org/10.1891/1939-7054.2.1.5

Herrera, A. J. (1999). Language bias discredits the peer-review system. *Nature*, *397*(6719), 467. https://doi.org/10.1038/17194, PubMed: 10028961

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a, PubMed: 25903611

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

Hu, Y. H., Tai, C. T., Liu, K. E., & Cai, C. F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics*, *14*(1), 101004. https://doi.org/10.1016/j.joi.2019.101004

Jackson, J. L., Srinivasan, M., Rea, J., Fletcher, K. E., & Kravitz, R. L. (2011). The validity of peer review in a general medicine journal. *PLOS ONE*, *6*(7), e22475. https://doi.org/10.1371/journal.pone.0022475, PubMed: 21799867

Jones, S., & Alam, N. (2019). A machine learning analysis of citation impact among selected Pacific Basin journals. *Accounting & Finance*, *59*(4), 2509–2552. https://doi.org/10.1111/acfi.12584

Jukola, S. (2017). A social epistemological inquiry into biases in journal peer review. *Perspectives on Science*, *25*(1), 124–148. https://doi.org/10.1162/POSC_a_00237

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., … Schwartz, R. (2018). A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers)* (pp. 1647–1661). https://doi.org/10.18653/v1/N18-1149

Kitayama, S. (2017). *Journal of Personality and Social Psychology*: Attitudes and social cognition [Editorial]. *Journal of Personality and Social Psychology*, *112*(3), 357–360. https://doi.org/10.1037/pspa0000077, PubMed: 28221091

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

Knoth, P., & Zdrahal, Z. (2012). CORE: Three access levels to underpin open access. *D-Lib Magazine*, *18*(11/12). Retrieved from https://oro.open.ac.uk/35755/. https://doi.org/10.1045/november2012-knoth

Kousha, K., & Thelwall, M. (2022). Artificial intelligence technologies to support research assessment: A review. *arXiv*, arXiv:2212.06574. https://doi.org/10.48550/arXiv.2212.06574

Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney, W. M. (2010). Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLOS ONE*, *5*(4), e10072. https://doi.org/10.1371/journal.pone.0010072, PubMed: 20386704

Larivière, V., & Costas, R. (2016). How many is too many? On the relationship between research productivity and impact. *PLOS ONE*, *11*(9), e0162709. https://doi.org/10.1371/journal.pone.0162709, PubMed: 27682366

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing & Management*, *47*(2), 300–308. https://doi.org/10.1016/j.ipm.2010.09.005

Li, J., Sato, A., Shimura, K., & Fukumoto, F. (2020). Multi-task peer-review score prediction. In *Proceedings of the First Workshop on Scholarly Document Processing* (pp. 121–126). https://doi.org/10.18653/v1/2020.sdp-1.14

Li, M., Xu, J., Ge, B., Liu, J., Jiang, J., & Zhao, Q. (2019a). A deep learning methodology for citation count prediction with large-scale biblio-features. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (pp. 1172–1176). IEEE. https://doi.org/10.1109/SMC.2019.8913961

Li, S., Zhao, W. X., Yin, E. J., & Wen, J.-R. (2019b). 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

Mattsson, P., Sundberg, C. J., & Laget, P. (2011). Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position. *Scientometrics*, *87*(1), 99–105. https://doi.org/10.1007/s11192-010-0310-9

Medoff, M. H. (2003). Editorial favoritism in economics? *Southern Economic Journal*, *70*(2), 425–434. https://doi.org/10.1002/j.2325-8012.2003.tb00580.x

Morgan, R., Hawkins, K., & Lundine, J. (2018). The foundation and consequences of gender bias in grant peer review processes. *Canadian Medical Association Journal*, *190*(16), E487–E488. https://doi.org/10.1503/cmaj.180188, PubMed: 29685908

PLOS. (2022). Criteria for publication. https://journals.plos.org/plosone/s/criteria-for-publication

Prins, A., Spaapen, J., & van Vree, F. (2016). Aligning research assessment in the Humanities to the national Standard Evaluation Protocol Challenges and developments in the Dutch research landscape. In *Proceedings of the 21st International Conference on Science and Technology Indicators—STI 2016* (pp. 965–969).

Qian, Y., Rong, W., Jiang, N., Tang, J., & Xiong, Z. (2017). Citation regression analysis of computer science publications in different ranking categories and subfields. *Scientometrics*, *110*(3), 1351–1374. https://doi.org/10.1007/s11192-016-2235-4

REF2021. (2019). *Index of revisions to the 'Guidance on submissions' (2019/01)*. https://www.ref.ac.uk/media/1447/ref-2019_01-guidance-on-submissions.pdf

Ross, J. S., Gross, C. P., Desai, M. M., Hong, Y., Grant, A. O., … Krumholz, H. M. (2006). Effect of blinded peer review on abstract acceptance. *Journal of the American Medical Association*, *295*(14), 1675–1680. https://doi.org/10.1001/jama.295.14.1675, PubMed: 16609089

Settles, B. (2011). From theories to queries: Active learning in practice. In *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010* (pp. 1–18).

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

Tan, J., Yang, C., Li, Y., Tang, S., Huang, C., & Zhuang, Y. (2020). Neural-DINF: A neural network based framework for measuring document influence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6004–6009). https://doi.org/10.18653/v1/2020.acl-main.534

Tennant, J. P., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. *Research Integrity and Peer Review*, *5*, 6. https://doi.org/10.1186/s41073-020-00092-1, PubMed: 32368354

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. (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

Thelwall, M., Allen, L., Papas, E. R., Nyakoojo, Z., & Weigert, V. (2021). Does the use of open, non-anonymous peer review in scholarly publishing introduce bias? Evidence from the F1000Research post-publication open peer review publishing model. *Journal of Information Science*, *47*(6), 809–820. https://doi.org/10.1177/0165551520938678

Thelwall, M., & Fairclough, R. (2015). Geometric journal impact factors correcting for individual highly cited articles. *Journal of Informetrics*, *9*(2), 263–272. https://doi.org/10.1016/j.joi.2015.02.004

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., … Levitt, J. (2022). Can REF output quality scores be assigned by AI? Experimental evidence. *arXiv*, arXiv:2212.08041. https://doi.org/10.48550/arXiv.2212.08041

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

Thelwall, M., & Sud, P. (2016). National, disciplinary and temporal variations in the extent to which articles with more authors have more impact: Evidence from a geometric field normalised citation indicator. *Journal of Informetrics*, *10*(1), 48–61. https://doi.org/10.1016/j.joi.2015.11.007

Thelwall, M., & Wilson, P. (2016). Does research with statistics have more impact? The citation rank advantage of structural equation modeling. *Journal of the Association for Information Science and Technology*, *67*(5), 1233–1244. https://doi.org/10.1002/asi.23474

Traag, V. A., & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, *5*, 29. https://doi.org/10.1057/s41599-019-0233-x

van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review and project selection: A case study in the social and behavioral sciences. *Research Evaluation*, *18*(4), 273–288. https://doi.org/10.3152/095820209X475360

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601–1615. https://doi.org/10.1007/s11192-013-1154-x

Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy*, *48*(5), 1260–1270. https://doi.org/10.1016/j.respol.2019.01.002

Wen, J., Wu, L., & Chai, J. (2020). Paper citation count prediction based on recurrent neural network with gated recurrent unit. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (pp. 303–306). IEEE. https://doi.org/10.1109/ICEIEC49280.2020.9152330

Wessely, S. (1998). Peer review of grant applications: What do we know? *Lancet*, *352*(9124), 301–305. https://doi.org/10.1016/S0140-6736(97)11129-1, PubMed: 9690424

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford, UK: Oxford University Press.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., & Hill, S., (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. London, UK: HEFCE. https://doi.org/10.4135/9781473978782

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, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, *75*, 171–212. https://doi.org/10.1613/jair.1.12862

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

APPENDIX: INPUT FEATURES CONSIDERED BUT NOT USED

The set of inputs used is not exhaustive because many others have been proposed. We excluded previously used inputs for the following reasons: peer review reports (Li, Zhao et al., 2019b) because few are public; topic models built from article text (Chen & Zhang, 2015) because this seems unnecessarily indirect given that article topics should be described clearly in abstracts; citation count time series (Abrishami & Aliakbary, 2019) due to not being relevant enough for quality prediction; citation network structure (Zhao & Feng, 2022), as this was not available and is not relevant enough for quality prediction; and language (Su, 2020), because most U.K. articles are English. Other excluded inputs, and corresponding reason for exclusion, were: funding organization (Su, 2020), because funding is very diverse across the REF and the information was not available; research methods and study details (Jones & Alam, 2019), because full text was not available for most articles; semantic shifts in terms (Tan, Yang et al., 2020), because this was too complex to implement in the time available (it would require network calculations on complete Scopus data, not just UK REF data, and including years from before the REF) and the limited evidence that it works on different data sets so far, although it seems promising; altmetrics (Akella, Alhoori et al., 2021) because these can be manipulated; and specific title features, such as title length or the presence of colons (van Wesel, Wyatt, & ten Haaf, 2014), because these seem too superficial, minor, and with varied results.

The most important omission was SciBERT (Beltagy et al., 2019). SciBERT converts terms into 768 dimensional vectors that are designed to convey the sense of words in the contexts in which they are used, learned from a full text scientific document corpus. We could have used SciBERT vectors as inputs instead of unigrams, bigrams, and trigrams, replacing 768 of them with SciBERT vectors and retaining the remaining 222 dimensions (out of the 990 available for text features) for journal names selected by the feature selection algorithm. SciBERT gives good results on many scientific text processing tasks and may well have generated slight improvements in our results. We did not use it for our primary experiments because 768 dimensional vectors are nontransparent. In contrast we could (and did: Thelwall et al., 2022) analyze the text components of the text inputs to understand what was influential, finding writing styles and methods names, which gave important context to the results. For example, the journal style features led to abandoning an attempt to create more “responsible” (in the sense of Wilsdon et al., 2015) solutions from text and bibliometrics, ignoring journal names and impact information, because abstract text was being leveraged for journal information. We had intended to repeat the study with SciBERT (and deep learning experiments) after the main set but ran out of time because most of the two-months data access we were given was needed for data cleaning and matching, troubleshooting, and testing different overall strategies. It is not certain that SciBERT would improve accuracy, as specific terms such as “randomized control trial” and pronouns were powerful, and these may not be well captured by SciBERT.