

Predicting retracted research

Aaron HA Fletcher* and Mark Stevenson*

*School of Computer Science, The University of Sheffield, Regent Court,
Sheffield, S1 4DP, United Kingdom.

Contributing authors: ahafletcher1@sheffield.ac.uk;
mark.stevenson@sheffield.ac.uk;

Abstract

Background: Retractions undermine the scientific record’s reliability and can lead to the continued propagation of flawed research. This study aimed to (1) create a dataset aggregating retraction information with bibliographic metadata, (2) train and evaluate various machine learning approaches to predict article retractions, and (3) assess each feature’s contribution to feature-based classifier performance using ablation studies.

Methods: An open-access dataset was developed by combining information from the Retraction Watch database and the OpenAlex API. Using a case-controlled design, retracted research articles were paired with non-retracted articles published in the same period. Traditional feature-based classifiers and models leveraging contextual language representations were then trained and evaluated. Model performance was assessed using accuracy, precision, recall, and the F1-score

Results: The Llama 3.2 base model achieved the highest overall accuracy. The Random Forest classifier achieved a precision of 0.687 for identifying non-retracted articles, while the Llama 3.2 base model reached a precision of 0.683 for identifying retracted articles. Traditional feature-based classifiers generally outperformed most contextual language models, except for the Llama 3.2 base model, which showed competitive performance across several metrics.

Conclusions: Although no single model excelled across all metrics, our findings indicate that machine learning techniques can effectively support the identification of retracted research. These results provide a foundation for developing automated tools to assist publishers and reviewers in detecting potentially problematic publications. Further research should focus on refining these models and investigating additional features to improve predictive performance.

Trial registration: Not applicable.

Keywords: Retraction prediction, Machine Learning, Scientific Publishing

1 Background

Retracting scientific articles is essential for safeguarding the integrity of the research record, but the growing number of retractions also reveals weaknesses in peer review and editorial oversight [1–4]. Determining the extent of retractions is complicated by “stealth retractions”, in which notices are omitted [5], and journals face a persistent tension between preventing the publication of flawed work and ensuring timely dissemination of results [6]. Although retracted research can still be useful—alerting the community to invalid findings or spurring new investigations—this utility depends on the clarity of its retracted status, which is often inconsistently handled. Unchecked, problematic work can damage authors’ reputations [7], tarnish journals [8, 9], and undermine domain integrity [10].

Once published, retracted papers can continue to influence discourse if their invalidation is overlooked. Avenell et al. demonstrated how just 12 misconduct-tainted clinical trials were repeatedly cited in systematic reviews and guidelines, substantially altering or obscuring conclusions [11]. Schneider et al. found that 96% of direct citations to a retracted 2008 clinical trial did not acknowledge its retraction [12], while Hsiao and Schneider showed that only 5.4% of citing contexts across 7,813 retracted papers reflected the retraction [13]. Even high-profile examples, such as the discredited vaccine-autism paper, accrue citations that rarely probe the retraction’s specific invalidations [14]. Although flawed data do not generally spread through secondary citations [15], the persistence of direct citations underscores the need for consistent, visible retraction notices. While recent advances, such as the CrossRef API directly integrating the Retraction Watch database into their metadata [16], more approaches are needed to increase the visibility of retracted works.

Retractions commonly stem from honest errors or from misconduct such as fabrication, plagiarism, or falsified authorship, but the relative prevalence of each is disputed. For example, Steen [2] attributed 73.5% of PubMed retractions to honest mistakes, whereas Moylan and Kowalczyk [17] reported that 76% of retractions in BioMed Central journals were due to misconduct—suggesting that clarifying a paper’s retraction status is more straightforward than classifying its cause. Paper mills pose a growing challenge: they mass-produce fraudulent articles and often share textual or methodological signatures [18–20]. Studies indicate an increase in retractions tied to paper mills, frequently concentrated in specific countries and article types [21–23], hinting that metadata features—like author origin, affiliation, or study design—could aid retraction prediction.

Despite the rising awareness of retracted research’s harms and the role of journal gatekeeping, there has been limited investigation into machine learning models to detect future retractions. Related work has classified *why* a paper was withdrawn [24], but not whether it *should* be retracted in the first place. This study aimed to (1) create a dataset aggregating retraction information with bibliographic metadata, (2) train and evaluate various machine learning approaches to predict article retractions, and (3) assess each feature’s contribution to feature-based classifier performance using ablation studies.

Table 1 Journal Exclusion Criteria.

Criteria	Description
CrossRef	If journal was not included in CrossRef’s journal title list.
Works Count	If work count (<i>based on OpenAlex API</i>) - total retraction count (<i>based on Retraction Watch Dataset</i>) < Sample Size (1)
Retraction Count	If journal total retractions < 5 (<i>determined by the Retraction Watch dataset</i>).

2 Methods

This study used existing data from open-access catalogues, databases, machine learning models and closed/open-sourced [Large Language Models \(LLMs\)](#). The research design used was a retrospective observational study with a case-control approach.

2.1 Dataset Construction

Publicly available data from two online databases were used to construct the dataset (Retraction Watch and OpenAlex). Retraction Watch is a human-validated retraction dataset and is compiled from various sources, including journal databases, institutional reports, social media, and direct tips [25, 26]. Although not exhaustive due to unannounced or “stealth” retractions [5], it provides partial metadata for some retracted articles, such as title, journal, publisher, and author. OpenAlex is an open-access online catalogue of academic publications, similar to Scopus and Web of Science, which aggregates data from multiple sources and releases monthly updates. These resources were combined to create a single dataset suitable for predicting article retractions [27].

A set of retracted articles were identified using the Retraction Watch dataset. Only articles and review works were considered. Conference papers were excluded due to a mass retraction of conference papers undertaken by the Institute of Electrical and Electronic Engineers between 2009 and 2011 (having retracted over 10,000 such papers in the past two decades) [28] and because there is no process to retract papers from many conference venues. Retractions were limited to a 20 year period from 2000 to 2020 due to the lack of retracted works before this date, the median post-publication time to retraction being 1.8 years [22] and the increased use of natural language technologies subsequent to this period. Information from OpenAlex (API queried on 24/07/2024) was also used to filter out some works. A full list of exclusion criteria for journals and articles are shown in Tables 1 and 2.

Another set of non-retracted articles was formed. For each retracted article, another article was randomly sampled that was published in the same year as the retracted article, did not meet the works exclusion criteria outlined in Table 2, was not included in the retraction watch dataset and whose OpenAlex API flag of ‘*is_retracted*’ was False.

Articles containing keywords strongly indicating that it has been retracted were also excluded, using the following list of keywords: “retraction”, “retracted”, “retract”, “retractionwatch”, “retraction watch”, “removed”, “withdrawn”, “withdrawal”, “withdraw”, “retracted article” and “article”.

Table 2 Work Exclusion Criteria.

Criteria	Description
Retracted Works	If Retraction Watch parameter ' <i>ArticleType</i> ' not in {Research Article, Conference Abstract/Paper, Clinical Study, Review Article, Case Report, Meta-Analysis}
Retracted Works/Non-retracted works	If OpenAlex ' <i>source</i> ' not in {Conference, Journal} and ' <i>type</i> ' not in {article, review}
English Language	Work excluded if OpenAlex API ' <i>language</i> ' value not 'en'
ISSN Data	If OpenAlex API ' <i>issn</i> ' value not available
OpenAlex ID	If OpenAlex API ' <i>id</i> ' value not available
Article Type	If OpenAlex API ' <i>type</i> ' value not in {article, review, conference-paper}
Publication Year	If ' <i>publication_year</i> ' OpenAlex API value not available
Publication Year Minimum	If ' <i>publication_year</i> ' OpenAlex API value < 2000
Publication Year Maximum	If ' <i>publication_year</i> ' OpenAlex API value > 2020
Reformulated Abstract Length	If < 5 words
Unretracted works's whose title contained	" <i>retraction</i> ", " <i>retraction:</i> ", " <i>withdrawn</i> ", " <i>correction</i> ", " <i>erratum</i> ", " <i>retracted</i> ", " <i>withdrawal</i> ", " <i>conclusion</i> ", " <i>editorial</i> ", " <i>contributions</i> ", " <i>commentary</i> ", " <i>contributors</i> ".
Retracted and unretracted works if abstract or title contained the words	" <i>elsevier</i> ", " <i>notice</i> ", " <i>editor</i> ", " <i>editors</i> ", " <i>publisher</i> ".

The following features were extracted for each work (retracted and non-retracted): Abstract Inverted Index, Publication Date, Primary Topic, First Author, Institution, Citation Count, First Author Countries, Is Retracted Flag and Article Type.

Both sets (retracted articles and non-retracted articles) were combined and balanced through undersampling, resulting in a total of 9,028 pieces of research, with equal numbers of retracted and non-retracted. It was divided into training (64%), validation (16%) and test sets (20%). Group sizes were chosen with nested split using the Pareto principle. A balanced dataset was used, as the distribution of retracted works to non-retracted works across all works is highly imbalanced. This imbalance potentially leads to models learning a priori class distributions rather than learning from the features provided.

All textual fields were preprocessed by converted to lowercase, and eliminating non-ASCII characters, special punctuation, and numbers. The title and abstract were then combined. The Publication Date feature was converted solely to its year in YYYY format.

2.2 Classifier Creation

Multiple approaches to machine learning classification were trained: Feature-based and LLMs (Encoder-based and Decoder-based) classifiers.

Feature-based classifiers were selected that have demonstrated effectiveness for text-classification tasks: Gradient Boosting, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), Random Forest, Multilayer Perceptron (MLP) and Decision Trees [29, 30]. A Super Learner model, an ensemble approach with multiple machine learning models, was also utilised [28]. Several LLMs known for strong classification performance was selected. Contrasting encoder-based pre-trained models were included: Bidirectional Encoder Representations from Transformers (BERT) (“bert-base-uncased”), trained on a broad dataset, and BioBERT (“dmis-lab/biobert-base-cased-v1.2”), pre-trained on biomedical data. Contrasting decoder-based LLMs were selected, differentiated by their fine-tuning: Large Language Model Meta AI (Llama) 3.2 (“unsloth/Llama-3.2-3B-bnb-4bit”), its instruction-tuned variant (“unsloth/Llama-3.2-3B-Instruct-bnb-4bit”), Gemma 2 (“unsloth/gemma-2-9b”), and its instruction-tuned variant (“unsloth/gemma-2-9b-it-bnb-4bit”). Commercially available LLMs were also evaluated: GPT-4o mini [31] and Claude 3.5 sonnet [32].

Different model architectures necessitated different input formats. Inputs for feature-based classifiers were formed into a single vector representation: Numerical features (Publication Year and Citation Count) were min-max normalised to scale between 0 and 1. Categorical features (First Author’s country) were one-hot encoded based on the training data subset. All other text features were represented using term weights produced by Best Matching 25 [33] with $k = 2$, $b = 0.3$ and no maximum vocabulary length specified. All features were then concatenated to produce the single vector representation. Inputs for the encoder-based LLM models consisted of the features (Text, Primary Topic, First Author, First Author Country, Cited By Count and Publication Year) concatenated with [SEP] tokens separating each feature - as shown in Figure 1. Decoder-based LLMs used the prompt template illustrated in Figures 2 and 3.

All models were trained using the training dataset, with early stopping being determined using the validation dataset to prevent overfitting. An expectation of this was the commercial models (i.e. GPT-4o mini and Claude 3.5 sonnet), where a zero-shot approach was used. LLM fine-tuning was conducted using supervised fine-tuning for a maximum of 10 epochs. Only the pooling layer and the classification head were updated with encoder-based models. Decoder-based LLMs were fine-tuned using the input format and prompt template illustrated in Figures 2 and 3. With these, a low-order rank adaptation approach was used, with output vocabulary restricted to “yes” and “no” tokens. During testing, these models were evaluated by providing the input text and the question without the label to assess their ability to predict the retraction status independently, with a softmax of the logits for the “yes” and “no” output tokens forming the model’s prediction.

All models output a binary classification label; “0” denoting if a piece of work is retracted and “1” not. Models were evaluated using the testing dataset. From this, model performance was measured using standard metrics for classification problems. Accuracy is the proportion of instances correctly classified as either retracted or non-retracted. Precision, recall and F1 scores are computed individually for the retracted and non-retracted classes and then averaged.

INPUT_TEXT Formation

```
{title + abstract} [SEP] {primary topic}  
[SEP] {first author} [SEP] {first author country}  
[SEP] {cited by count} [SEP] {publication year}
```

Fig. 1 Input format for encoder-based models.

INPUT_TEXT Formation

```
Text: {title + abstract} Primary Topic: {primary topic}  
First Author: {first author}  
First author country: {first author country}  
Cited by count: {cited by count}  
Publication Year: {publication year}
```

Fig. 2 Input format for decoder-based models.

Prompt Template

```
Here is a research article:  
{INPUT_TEXT}  
  
Is this paper retracted?  
The correct answer is: {LABEL}
```

Fig. 3 Prompt template used for structuring input data during model fine-tuning.

2.3 Ablation

To evaluate the relative contribution of each feature to classifier performance on all feature-based classifiers, an ablation study was performed. For each feature, an “ablated” version of the dataset was created by removing that feature from all training instances while leaving the remaining features intact. Every feature-based classifier was then re-trained from scratch using the ablated dataset, and evaluated on the same test set employed for the full-featured models.

For instance, when ablating the *Abstract* feature, all tokens originating from the abstract were excluded, but the *Title*, *Primary Topic*, *Publication Year*, *First Author*, *First Author’s Country*, and *Citation Count* features were retrained. This procedure was repeated for each of the remaining features in turn. The resulting evaluation metrics (accuracy, precision, recall, and F1) were then compared against the “full-feature” baselines to quantify the performance drop caused by removing that feature. Lower scores in the ablated setting indicate a more critical feature, as its removal impairs model performance more severely.

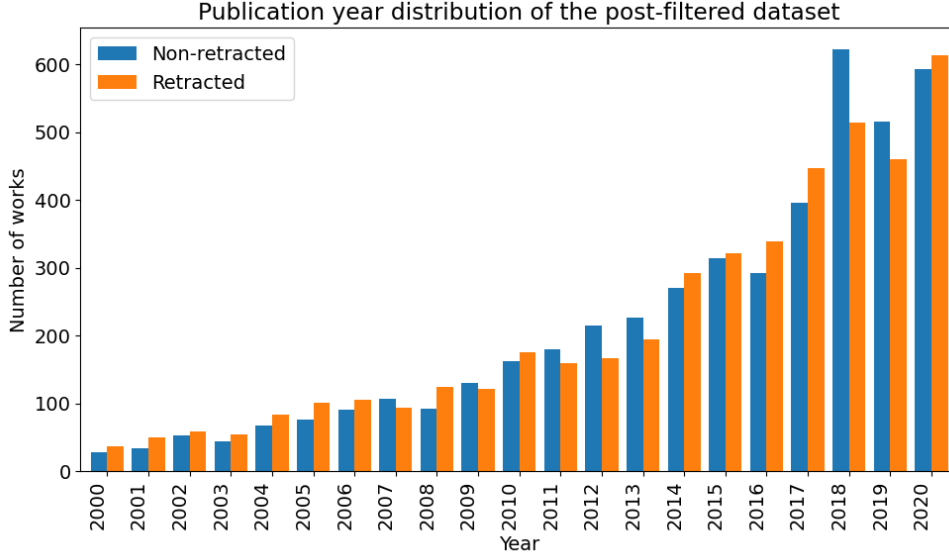


Fig. 4 Publication year distribution for retracted and non-retracted works.

3 Results

3.1 Dataset Characteristics

Each year’s works distribution is shown in Figure 4. It can be seen that the number of included works associated with each year increases over time, reflecting the trend of retractions increasing over time in the original Retraction Watch dataset.

In the generated dataset, 7.54% of the articles reported as retracted in Retraction Watch were not marked as retracted by OpenAlex, possibly because OpenAlex’s metadata is derived from multiple input sources. This discrepancy further illustrates the difficulty of identifying retracted research since it may not be labelled as such. This discrepancy has been recently directly addressed with CrossRef integrating the Retraction Watch database into the metadata returned from their API [16].

Analysis of correlations between journal features revealed two notable findings:

1. A weak, significant positive correlation between the work count log and the retraction count log (Pearson correlation coefficient 0.065, p-value < 0.05). This seems counterintuitive, as more retractions are likely to occur given more publications, and hence, a strong positive correlation would be present. This finding could indicate that journals that publish fewer works are less proactive at detecting potential retractions or that publishing research that will be retracted is more complicated within journals with greater work output, presumably due to increased scrutiny of these works.
2. A strong negative correlation between the retraction count and log of h-index (Pearson correlation coefficient -0.656, p-value < 0.05). This relationship is expected. The

h-index, a widely used measure of a journal’s productivity and impact, is based on its most cited papers. Typically when a work is considered for inclusion in a journal or conference, a peer reviewer is tasked with subjecting that research to the scrutiny of others who are experts in the same field [34]. This reviewer is sourced from academics who review for many reasons (primarily altruistic), such as keeping up with the latest developments, building associations with journals, and demonstrating a commitment to the scientific field [35]. Importantly, time available to review is a finite resource [36]. It is likely that more reviewers are available for greater h-index journals. Publishing venues with higher h-index values potentially have a more rigorous peer review process, authors are more diligent when submitting to these journals, or higher-quality journals attract better-quality research.

3.2 Classifier Performance

Results for all classifiers are presented in Table 3, showing performance for both the retracted and not retracted classes. The highest-scoring approaches for each metric are highlighted in bold. Commercial models were excluded from further analysis as all commercial models responded that no research was retracted within the testing dataset. This was thought to be due to the safety restrictions implemented within these models, which prevented responses that could be considered problematic [37].

All models outperformed random guessing (i.e. 0.5 as this is a binary classification task), although the improvement varies considerably between models. The highest accuracy (0.682) is achieved by Llama 3.2-base, although accuracy scores overall are generally higher for more traditional feature-based approaches such as gradient boost, SVM, XGBoost, and Random Forest achieved superior precision compared to the more modern contextually aware LLMs.

Regarding the retracted class, SVM achieved the highest precision (0.690) and Llama 3.2-base the highest recall (0.683). Interestingly, both instruction-tuned decoder-based LLMs (Gemma 2-instruct and Llama 3.2-instruct) also achieve high recall for the retracted class but this is achieved by predicting retracted for the majority of instances, as demonstrated by the very low recall for the non-retracted class. This could be due to instruction tuning, as they are trained to be more cautious and risk-averse, indicating that instruction-tuned models might not be suitable for this type of classification task.

These findings establish baseline results using the dataset.

3.3 Ablation Analysis

The importance of individual features to the feature-based classification models was explored by conducting an ablation study on all input features. Datasets were created for each feature by permuting the data to exclude that feature and then averaging the evaluation metrics (F1 score, precision, recall, accuracy) across all models for each ablation. Lower scoring metrics indicate a greater contribution to the performance of a classifier.

Several observations on the ablation of features can be made given the results reported in Table 4.

Table 3 Retraction classifier performance results.

Model	Acc.	Non-Retracted			Retracted		
		P	R	F1	P	R	F1
Logistic Regression	0.638	0.638	0.647	0.642	0.639	0.630	0.635
Decision Tree	0.568	0.570	0.574	0.572	0.568	0.564	0.566
Random Forest	0.666	0.648	0.731	0.687	0.689	0.601	0.642
SVM	0.671	0.655	0.725	0.688	0.690	0.616	0.651
XGBoost	0.665	0.654	0.705	0.679	0.678	0.624	0.650
AdaBoost	0.631	0.619	0.684	0.650	0.645	0.577	0.609
Super Learner	0.669	0.661	0.699	0.680	0.678	0.640	0.659
MLP	0.655	0.650	0.675	0.663	0.660	0.634	0.647
Gemma 2-base	0.553	0.615	0.292	0.396	0.534	0.816	0.645
Gemma 2-instruct	0.529	0.730	0.098	0.173	0.515	0.963	0.671
BERT	0.609	0.612	0.602	0.607	0.606	0.616	0.611
BioBERT	0.608	0.598	0.668	0.631	0.621	0.548	0.582
Llama 3.2-base	0.682	0.686	0.674	0.680	0.678	0.689	0.683
Llama 3.2-instruct	0.535	0.714	0.121	0.208	0.518	0.951	0.671

Table 4 Ablation performance metrics: lowest scoring ablations are in bold.

Model	Acc.	Non-Retracted			Retracted		
		P	R	F1	P	R	F1
Abstract	0.655	0.670	0.612	0.638	0.645	0.678	0.669
Citation Count	0.648	0.659	0.611	0.634	0.638	0.684	0.660
First Author	0.649	0.662	0.609	0.634	0.639	0.689	0.663
First Author Countries	0.649	0.663	0.604	0.632	0.638	0.693	0.664
Primary Topic	0.644	0.658	0.602	0.628	0.634	0.687	0.659
Publication Year	0.641	0.656	0.594	0.622	0.630	0.688	0.657
Title	0.648	0.657	0.618	0.636	0.641	0.678	0.658

The publication year proved to be the most crucial feature, with its ablation resulting in the lowest scores across all metrics. This suggests that temporal information plays a more significant role in classification than detailed textual content, which is logical given the increase in publications and the corresponding increase in retractions. This temporal information component also highlights the challenge that traditional classification approaches potentially face in this area; given that publication year is such a strong signal, its presence might eclipse other valuable contributions from other features. The Primary Topic feature also demonstrated substantial importance, producing the second-lowest scores when ablated. Reduction in performance when First Author Countries are ablated provides some indication of the likelihood that a work will be retracted, supporting previous findings [38].

Contrary to what might be intuitively expected, the abstract, despite being the longest and most detailed textual component, emerged as the least influential feature across all evaluation metrics. When ablated, it yielded the highest average scores for

accuracy (0.655), precision (0.657), recall (0.655), and F1 score (0.654), indicating its removal had the least negative impact on model performance. This counterintuitive finding regarding the abstract’s limited influence could be attributed to several factors. First, structured metadata features (like publication date and primary topic) may provide more consistent and unambiguous signals for classification compared to the potentially noisy and variable nature of abstract text. Second, there might be considerable information redundancy between the abstract and other textual features like the title, making its individual contribution less distinctive.

4 Discussion

Machine learning approaches can successfully identify retracted papers using a created open-access dataset.

One of the potential applications of the classifier described above is as a tool during the peer review process, in much the same way that text similarity tools are often used to identify potential plagiarism. The required level of precision or recall would depend on how the tools would be used. If used as a screening tool to flag potentially problematic papers for additional review, a high recall would be preferable to avoid missing articles that are subsequently retracted. However, if used as a check which a submitted article must pass then high precision would be necessary to avoid the suppression of valid research. The performance of the models reported above, while promising, indicates that identification of retracted articles is not a trivial prediction task and may not be sufficient for some purposes. The decision regarding the involvement of systems to detect potential retractions within the peer review process is ultimately the choice of publishers.

The automatic prediction of potential retractions also raises ethical concerns. Predictive models, such as the ones described here, can introduce bias thereby raising potential fairness issues [39, 40].

Such biases can unfairly penalise the groups more likely to be identified as producing research that will be retracted (e.g., first authors from particular locations) while benefiting those it is less likely to identify. This could introduce inductive bias into investigations, potentially leading to unforeseen consequences in the scientific publishing landscape, such as influencing which research questions are investigated and which methodologies are applied. In addition, authors may attempt to report results in ways that avoid detection by these models, potentially leading to self-censorship or overly cautious reporting of results. Conversely, bad actors with knowledge of these models may exploit that information to avoid detection, potentially facilitating the dissemination of invalid results.

An important consideration is how best to apply these models in practice. While machine learning classifiers can highlight publications at higher risk of retraction, final decisions on whether a paper should be investigated or retracted must rest with human experts—editors, reviewers, and domain specialists. For example, automated models flag potential anomalies in medical and clinical contexts, but the ultimate judgment requires expert oversight [41, 42]. Similarly, the classifiers reported here are

intended to aid decision-making rather than stand-alone arbiters of scientific validity. A fully automated retraction process is not desirable, nor is it necessarily the duty of model developers to initiate or recommend retraction investigations on every flagged paper. Instead, these outputs can be a starting point for further human-led scrutiny. This workflow ensures that any potential reasons for retraction—which may be multifaceted and not always captured by the model—are carefully examined. It also prevents the undue penalisation of authors, institutions, or countries that might otherwise be overrepresented due to biases in the training data. By maintaining a robust human-in-the-loop process, publishers and editorial boards can leverage model predictions ethically and effectively to uphold the reliability of the scientific record.

5 Limitations

This study has several notable limitations. The study design relied on a single data source, the Retraction Watch database, which provides valuable but incomplete coverage. The presence of “stealth retractions”, wherein papers are removed without official notice or may not be reported to Retraction Watch, creates the potential for missed or under-detected retractions. Additionally, it was retrospective, using data from 2000 to 2020, which limits the ability to assess the models’ real-time or prospective effectiveness in detecting erroneous work at publication. Theoretical limitations exist within the model choice, as they capture correlational rather than causal relationships, potentially leading to false positives or negatives, as using these patterns can misrepresent the underlying reasons for retractions. Data was sampled from 2000 to 2020, which would not represent more recent changes in retracted works. Since 2020, there have been innovative natural language generation models that could potentially increase the count of retracted works. Features that are not fully representative of a piece of research were used. Due to copyright restrictions, abstracts and metadata were used rather than full-text articles. Indicators of methodological errors or unsupported conclusions might appear in the main text and not the title and abstract, potentially reducing the reliability of our retraction-prediction metrics.

Additionally, some LLM may have been partly trained on the same corpus used to develop or validate our dataset, inflating their performance scores. This issue does not affect purely feature-based approaches but undermines the reliability of LLM-derived results. Features such as the first author’s country or institution may reflect systemic biases in scientific publishing rather than genuine predictors of flawed work. Such biases risk penalising authors from certain regions or affiliations if used in editorial decision-making. Models may overfit to spurious textual or demographic correlations in the training data, leading to unjustified flags or missed detections when applied to new, diverse datasets.

6 Conclusions

This research demonstrates the potential of machine learning approaches in predicting retracted articles, contributing to efforts aimed at enhancing the integrity of scientific publication. By creating a novel open-source dataset that combines information from

the Retraction Watch database and the OpenAlex API, a resource for future investigations in this area has been contributed. Our dataset encompasses 9,028 articles published between 2000 and 2020, evenly divided between retracted and non-retracted works, and includes a variety of features such as abstracts, citation metrics, and author information.

Experiments showed that, with the exception of the recently released Llama 3.2 base model, traditional feature-based classifiers, such as gradient boosting machines and SVMs, outperformed contextual language models like BERT, BioBERT, and Gemma in terms of precision. The best-performing model achieved a precision of 0.690, indicating that while machine learning techniques hold promise, there remains a need for significant improvement before they can be effectively integrated into the peer review process. The ablation study highlighted the importance of the publication year, primary topic and the first author’s country in predicting retractions, aligning with previous findings that suggest certain demographics may be more prone to retractions due to various factors.

6.1 Future work

There is potential for the approaches described here to be extended by making use of additional information with the potential to assist in the identification of retracted research. For example, the citation network of references to a paper and the references within the paper itself may provide useful information. In addition, the models described here analysed abstracts, but analysis of the full text itself could potentially allow models to evaluate flaws in methodology, result synthesis or false conclusions. Finally, analysis of the full author list of an article could reveal patterns of collaboration or even help to identify potential paper mills.

7 List of Abbreviations

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The dataset supporting the conclusions of this article is available in the Predicting Article Retractions repository [43].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

Authors' contributions

AF contributed to this research's conception, design analysis, data interpretation, and submission drafting. MS contributed to this research's conception, design analysis, data interpretation, and submission drafting. All authors read and approved the final manuscript.

Open Access

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

- [1] Steen, R.G.: Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of Medical Ethics* **37**(4), 249–253 (2011) <https://doi.org/10.1136/jme.2010.040923>
- [2] Steen, R.G.: Retractions in the scientific literature: do authors deliberately commit research fraud? *Journal of Medical Ethics* **37**(2), 113–117 (2011) <https://doi.org/10.1136/jme.2010.038125>
- [3] Steen, R.G., Casadevall, A., Fang, F.C.: Why has the number of scientific retractions increased? *PLoS ONE* **8**(7), 68397 <https://doi.org/10.1371/journal.pone.0068397>
- [4] Kühberger, A., Streit, D., Scherndl, T.: Self-correction in science: The effect of retraction on the frequency of citations. *PloS One* **17**(12), 0277814 <https://doi.org/10.1371/journal.pone.0277814>
- [5] Teixeira Da Silva, J.A.: Silent or stealth retractions, the dangerous voices of the unknown, deleted literature. *Publishing Research Quarterly* **32**(1), 44–53 <https://doi.org/10.1007/s12109-015-9439-y>
- [6] Perera, R., Nand, P.: Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature. *Computing and Informatics* **36**(1), 1–32 (2017) <https://doi.org/10.4149/cai.2017.1.1>
- [7] Kearney, M., Downing, M., Gignac, E.A.: Research integrity and academic medicine: the pressure to publish and research misconduct. *Journal of Osteopathic Medicine* **124**(5), 187–194 (2024) <https://doi.org/10.1515/jom-2023-0211>

- [8] Caporale, C., Zagarella, R.M.: Ethics and Integrity in Academic Publishing. In: Congiunti, L., Lo Piccolo, F., Russo, A., Serio, M. (eds.) *Ethics in Research*, pp. 53–69. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-24060-7_5. Series Title: UNIPA Springer Series. https://link.springer.com/10.1007/978-3-031-24060-7_5 Accessed 2025-01-24
- [9] Collaborative Working Group from the conference “Keeping the Pool Clean: Prevention and Management of Misconduct Related Retractions”: RePAIR consensus guidelines: Responsibilities of Publishers, Agencies, Institutions, and Researchers in protecting the integrity of the research record. *Research Integrity and Peer Review* **3**(1), 15 (2018) <https://doi.org/10.1186/s41073-018-0055-1>. Accessed 2025-01-24
- [10] Grey, A., Avenell, A., Klein, A.A., Byrne, J.A., Wilmshurst, P., Bolland, M.J.: How to improve assessments of publication integrity. *Nature* **632**(8023), 26–28 (2024) <https://doi.org/10.1038/d41586-024-02449-8>
- [11] Avenell, A., Stewart, F., Grey, A., Gamble, G., Bolland, M.: An investigation into the impact and implications of published papers from retracted research: systematic search of affected literature. *BMJ Open* **9**(10) (2019) <https://doi.org/10.1136/bmjopen-2019-031909>
- [12] Schneider, J., Ye, D., Hill, A.M., Whitehorn, A.S.: Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics* **125**(3), 2877–2913 (2020) <https://doi.org/10.1007/s11192-020-03631-1>
- [13] Hsiao, T.-K., Schneider, J.: Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies* **2**(4), 1144–1169 (2021) <https://doi.org/10.1162/qss.a.00155>
- [14] Heibi, I., Peroni, S.: A qualitative and quantitative analysis of open citations to retracted articles: the wakefield 1998 et al.’s case. *Scientometrics* **126**(10), 8433–8470 (2021) <https://doi.org/10.1007/s11192-021-04097-5>
- [15] Vet, P.E., Nijveen, H.: Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal nature. *Research Integrity and Peer Review* **1**(1), 3 (2016) <https://doi.org/10.1186/s41073-016-0008-5>
- [16] Rittman, M.: Retraction Watch Retractions Now in the Crossref API. <https://doi.org/10.13003/692016>. <https://www.crossref.org/blog/retraction-watch-retractions-now-in-the-crossref-api/> Accessed 2025-02-18
- [17] Moylan, E.C., Kowalczyk, M.K.: Why articles are retracted: a retrospective cross-sectional study of retraction notices at biomed central. *BMJ Open* **6**(11) (2016)

<https://doi.org/10.1136/bmjopen-2016-012047>

- [18] Hvistendahl, M.: China’s publication bazaar. *Science* **342**(6162), 1035–1039 (2013) <https://doi.org/10.1126/science.342.6162.1035>
- [19] Liu, X., Chen, X.: Journal retractions: Some unique features of research misconduct in china. *Journal of Scholarly Publishing* **49**(3), 305–319 (2018) <https://doi.org/10.3138/jsp.49.3.02>
- [20] Tian, M., Su, Y., Ru, X.: Perish or publish in china: Pressures on young chinese scholars to publish in internationally indexed journals. *Publications* **4**(2) (2016) <https://doi.org/10.3390/publications4020009>
- [21] Candal-Pedreira, C., Ross, J.S., Ruano-Ravina, A., Egilman, D.S., Fernández, E., Pérez-Ríos, M.: Retracted papers originating from paper mills: cross sectional study. *BMJ* **379**, 071517 (2022) <https://doi.org/10.1136/bmj-2022-071517>
- [22] Gaudino, M., Robinson, N.B., Audisio, K., Rahouma, M., Benedetto, U., Kurlansky, P., Fremes, S.E.: Trends and Characteristics of Retracted Articles in the Biomedical Literature, 1971 to 2020. *JAMA Internal Medicine* **181**(8), 1118–1121 (2021) <https://doi.org/10.1001/jamainternmed.2021.1807>
- [23] Byrne, J.A., Christopher, J.: Digital magic, or the dark arts of the 21st century—how can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS Letters* **594**(4), 583–589 (2020) <https://doi.org/10.1002/1873-3468.13747>
- [24] Rao, D., Young, J., Dietterich, T., Callison-Burch, C.: WithdrarXiv: A Large-Scale Dataset for Retraction Study. *arXiv*. <https://doi.org/10.48550/arXiv.2412.03775> . <http://arxiv.org/abs/2412.03775> Accessed 2025-02-18
- [25] Retraction Watch. The Center for Scientific Integrity, New York (2018). <https://retractionwatch.com/> Accessed 2024-06-14
- [26] Retraction Watch Database User Guide (2024). <https://retractionwatch.com/wp-content/uploads/2023/12/Building-The-Database.pdf> Accessed 2024-06-14
- [27] Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv* (2022). <https://doi.org/10.48550/arXiv.2205.01833>
- [28] Van Noorden, R.: More than 10,000 research papers were retracted in 2023 — a new record. *Nature* **624**(7992), 479–481 (2023) <https://doi.org/10.1038/d41586-023-03974-8>
- [29] Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv.* **34**(1), 1–47 (2002) <https://doi.org/10.1145/505282.505283>

- [30] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L.: A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.* **13**(2) (2022) <https://doi.org/10.1145/3495162>
- [31] OpenAI Platform. <https://platform.openai.com> Accessed 1/12/2024
- [32] Introducing Claude 3.5 Sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet> Accessed 11/12/2024
- [33] Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.* **3**(4), 333–389 (2009) <https://doi.org/10.1561/15000000019>
- [34] Banks, D.: Thoughts on Publishing the Research Article over the Centuries. *Publications* **6**(1), 10 (2018) <https://doi.org/10.3390/publications6010010>
- [35] Steer, P.J., Ernst, S.: Peer review - Why, when and how. *International Journal of Cardiology Congenital Heart Disease* **2**, 100083 (2021) <https://doi.org/10.1016/j.ijcchd.2021.100083>
- [36] Warne, V.: Rewarding reviewers – sense or sensibility? A Wiley study explained. *Learned Publishing* **29**(1), 41–50 (2016) <https://doi.org/10.1002/leap.1002>
- [37] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S.R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., Kaplan, J.: Constitutional AI: Harmlessness from AI Feedback (2022). <https://arxiv.org/abs/2212.08073>
- [38] Stretton, S., Bramich, N.J., Keys, J.R., Monk, J.A., Ely, J.A., Haley, C., Woolley, M.J., Woolley, K.L.: Publication misconduct and plagiarism retractions: a systematic, retrospective study. *Current Medical Research and Opinion* **28**(10), 1575–1583 (2012) <https://doi.org/10.1185/03007995.2012.728131>
- [39] Caton, S., Haas, C.: Fairness in machine learning: A survey. *ACM Comput. Surv.* **56**(7) (2024) <https://doi.org/10.1145/3616865>
- [40] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**(6) (2021) <https://doi.org/10.1145/3457607>
- [41] Prictor, M.: WHERE DOES RESPONSIBILITY LIE? ANALYSING LEGAL

AND REGULATORY RESPONSES TO FLAWED CLINICAL DECISION SUPPORT SYSTEMS WHEN PATIENTS SUFFER HARM **31**(1), 1–24 <https://doi.org/10.1093/medlaw/fvac022> . Accessed 2025-03-13

- [42] Funer, F., Liedtke, W., Tinnemeyer, S., Klausen, A.D., Schneider, D., Zacharias, H.U., Langanke, M., Salloch, S.: Responsibility and decision-making authority in using clinical decision support systems: an empirical-ethical exploration of german prospective professionals' preferences and concerns **50**(1), 6–11 <https://doi.org/10.1136/jme-2022-108814>
- [43] Anonymised Repository - Anonymous GitHub. <https://anonymous.4open.science/r/RetractionWatch/README.md> Accessed 2025-01-15