Scientometrics
https://doi.org/10.1007/s11192-026-05808-6

# Understanding prolonged retraction lag among retracted biomedical papers: an explainable machine-learning analysis

**Zhaoqi Peng[1,2] · Mengfan Zhao[3] · Hao Wang[1,2] · Alex Jie Yang[1,2] · Sanhong Deng[1,2]**

Received: 29 March 2026 / Accepted: 1 September 2026
© Akadémiai Kiadó Zrt 2026

**Abstract**
Retractions are vital for scientific self-correction, yet prolonged lags between publication and retraction allow flawed papers to persist in the literature and distort future research through citations. Here, we applied explainable machine learning to identify the key factors associated with prolonged retraction lags in biomedical papers. Using 13,685 biomedical papers that were ultimately retracted, we classify whether the observed publication-to-retraction interval exceeded a practically meaningful three-year threshold, distinguishing between relatively short and prolonged retraction lag. LightGBM achieved the highest F1-score and performance comparable to Random Forest and XGBoost in discrimination and calibration; therefore, it was selected as the primary model for SHAP-based interpretation. SHAP analysis results show that papers published by Hindawi were associated with relatively short retraction lag, a pattern that may reflect differences in publication volume, post-publication scrutiny, and research-integrity oversight environments rather than superior correction efficiency. SJR score showed a nonlinear association with prolonged retraction lag, while moderate levels of retraction reason complexity were generally associated with a higher predicted likelihood of prolonged retraction lag. The academic age Gini coefficient shows an inverted U-shaped association with retraction lag, with moderate levels associated with higher predicted likelihoods of prolonged lag. Prolonged retraction lags typically resulted from the cumulative interaction of multiple factors, whereas relatively short retraction lags were more often characterized by a simpler and more concentrated predictive structure. These results highlight the importance of institutional publishing contexts in shaping retraction timeliness and provide practical insights for research-integrity governance. These associations are conditional on eventual retraction and should be interpreted as retrospective evidence about variation in retraction timeliness, rather than as an early-warning model for the broader population of published papers.

✉ Alex Jie Yang
alexjieyang@outlook.com

✉ Sanhong Deng
sanhong@nju.edu.cn

[1] School of Information Management, Nanjing University, Nanjing 210023, China

[2] International Joint Informatics Laboratory, Nanjing University, Nanjing 210023, China

[3] School of Information Resource Management, Renmin University of China, Beijing 100872, China

Published online: 11 September 2026

**Keywords** Retraction lag · Explainable machine learning · SHAP · Research integrity

## Introduction

Scientific progress depends on the cumulative and self-correcting nature of knowledge. Researchers must be able to trust the integrity of the published literature, yet this trust is increasingly challenged by the growing number of article retractions (Li et al., 2025). Retractions serve as a critical mechanism for removing erroneous or fraudulent findings from the scientific record. However, the effectiveness of retractions depends not only on whether a paper is retracted, but also on how quickly the process occurs. Substantial lags, often spanning multiple years, allow flawed studies to continue circulating, accumulate citations, and potentially influence subsequent research and public policy (Malkov et al., 2023; Szilagyi et al., 2022).

The scale of the problem has become alarming. Retractions have risen sharply in recent years, with tens of thousands of articles retracted globally in 2024 alone (Lendvai & Sasvári, 2025). While this increase partly reflects improved detection of misconduct, it also highlights persistent weaknesses in the self-correction mechanisms of science (Abalkina et al., 2025). During the interval between publication and retraction, problematic papers can exert outsized influence, especially since citation rates typically peak within the first two to three years after publication. This "citation contamination" undermines the reliability of the biomedical literature and erodes public confidence in science.

Existing research has largely focused on the causes and consequences of retractions themselves, documenting that misconduct such as fabrication, falsification, and plagiarism accounts for the majority of cases (Fang et al., 2012; Hwang et al., 2023; Ioannidis et al., 2025). Studies have also examined honest errors, the limited reach of official retraction notices, and the role of platforms like Retraction Watch and social media in disseminating information about problematic papers (Amiri et al., 2024; Andersen & Wray, 2019; Casadevall et al., 2014; De Cassai et al., 2022; Haunschild & Bornmann, 2021; Khan et al., 2022; Xu et al., 2023). More recently, attention has turned to the career impacts on authors, particularly early-career researchers (Memon et al., 2025). However, far less is known about the temporal dynamics of the retraction process itself. Retraction lag, defined as the time between publication and formal retraction, remains poorly understood despite its direct implications for the speed and effectiveness of scientific self-correction.

The mechanisms driving retraction lag are complex and multifaceted. They include the inherent difficulty of detecting concealed misconduct, the challenges posed by large-scale "paper mills," fragmented responsibility in collaborative teams, and the rapid evolution of new forms of fraud enabled by generative AI (Abalkina et al., 2025; Bar-Ilan & Halevi, 2018; Madhugiri et al., 2021; Walsh et al., 2019; Wang & Chen, 2025). Traditional statistical approaches have struggled to capture the multidimensional interactions among publishing systems, journal characteristics, team structures, and retraction reason complexity.

In this study, we address these gaps by developing a predictive and interpretable machine-learning framework to identify the key drivers of prolonged retraction lag in biomedical papers. Drawing on 13,685 retracted articles from the Retraction Watch Database matched with PubMed metadata, we model retraction lag as a binary outcome distinguishing relatively short from prolonged (exceeding three years). We systematically compare multiple machine-learning algorithms and employ SHAP (SHapley Additive exPlanations) to provide transparent, mechanism-level insights into how factors such as publisher, SJR

score, retraction reason complexity, and team characteristics interact to prolong or accelerate the retraction process. Importantly, this study focuses exclusively on papers that were ultimately retracted. Therefore, our analysis does not explain why problematic research in general survives in the literature, but rather why some already retracted papers remain in the scientific record longer before retraction occurs.

This work contributes to the literature in three main ways. First, this study shifts the focus from whether a paper is retracted to whether a retracted paper is likely to experience a prolonged retraction lag, offering a multidimensional analysis of factors associated with prolonged retraction lag among already retracted biomedical papers. Second, by integrating explainable machine learning, it moves beyond black-box predictions to reveal nonlinear effects and interaction structures that traditional models cannot easily uncover. Third, it provides retrospective empirical evidence on institutional and organizational patterns associated with variation in retraction timeliness among completed retraction cases.

The remainder of the paper is organized as follows. Sect. "Literature review" reviews the relevant literature on retraction lag and explainable machine learning. Sect. "Data and methods" describes the data sources, variable construction, and methodological framework. Sect. "Results and discussion" presents the empirical results, including model performance and detailed SHAP-based interpretability analyses. Sect. "Conclusions" concludes with implications for research integrity governance and directions for future work.

## Literature review

### Studies on retraction lag

In recent years, the phenomenon of article retractions has attracted increasing scholarly attention. Existing studies mainly focus on the causes, characteristics, and academic impacts of retractions. A broad consensus is that traditional forms of scientific misconduct, such as fabrication, falsification, and plagiarism, remain the primary reasons for retraction (Hwang et al., 2023; Ioannidis et al., 2025). Empirical evidence shows that misconduct accounts for more than two-thirds of retracted publications in PubMed (Fang et al., 2012). In contrast, honest errors, such as experimental mistakes, analytical flaws, and irreproducible results, constitute another major category of retractions (Andersen & Wray, 2019; Casadevall et al., 2014). The effectiveness of retractions is also shaped by information dissemination mechanisms. Official retraction notices often have limited reach, which can result in low awareness among researchers (De Cassai et al., 2022). In contrast, unofficial platforms such as Retraction Watch play an important role in accelerating the spread of retraction information (Xu et al., 2023). Social media platforms, including Twitter and Mendeley, are also commonly used to disseminate information about retracted articles (Khan et al., 2022). The sentiment associated with such discussions is often negative (Amiri et al., 2024). This suggests that social media can help draw attention to problematic publications and may even serve as an early warning signal (Haunschild & Bornmann, 2021).

Retractions also have notable career consequences. Memon et al. (2025) find that nearly half of authors in STEM fields exit academic publishing after experiencing a retraction. This effect is more pronounced among early-career researchers, female authors, and those with lower academic impact. These findings suggest that retraction lag is influenced not only by investigation processes but also by differences in information dissemination and recognition.

Against this background, some studies have begun to examine the temporal dimension of retractions, commonly referred to as retraction lag. Ghorbi et al. (2021) report that the average retraction lag for Iranian authors is approximately 591 days, with some cases extending up to 28 years. More recently, Joëts and Mignon (2026) demonstrate that retraction lag exhibits substantial heterogeneity across publications and disciplines, highlighting the persistence of "zombie papers" that continue to influence the scientific record long before formal retraction. However, the mechanisms behind retraction lag are complex. They involve not only individual misconduct but also institutional and technological factors.

First, scientific misconduct is often concealed, which makes detection and investigation time-consuming. In recent years, “paper mills” have emerged as a major driver of large-scale and organized academic fraud. These entities produce fraudulent manuscripts in bulk and sell authorship, thereby severely contaminating the research ecosystem. Although some of these papers can be identified through features such as image manipulation, plagiarism, and fake peer review, their scale and evolving tactics continue to challenge journal screening systems and prolong the retraction process (Abalkina et al., 2025).

Second, team structure may influence how quickly problems are detected. Walsh et al. (2019), drawing on the concept of “pathogenic organization” argue that highly specialized teams may increase the concealment of errors due to fragmented responsibility and weak oversight. However, empirical findings are mixed. Madhugiri et al. (2021) report no significant relationship between the number of authors and retraction lag. In contrast, Bar-Ilan and Halevi (2018) find that retraction time tends to increase with larger author teams. Wang and Chen (2025) further show that team size is associated with error types. Medium-sized teams are more likely to produce data-related errors, while no clear relationship is observed between collaboration patterns and honest error types.

Beyond team size, the international composition of a collaboration and the publisher through which a paper appears have also been linked to retraction timing. Zhang et al. (2020) show that the extent of international collaboration among coauthors of retracted papers is systematically associated with retraction timing, consistent with the additional coordination that cross-border teams require to resolve a case. Similarly, Mhamdi (2026), analyzing retracted publications involving Saudi Arabia, reports distinct retraction patterns for internationally collaborative papers, suggesting that differences in institutional oversight and cross-national coordination may affect the efficiency of retraction investigations. At the organizational level, Azoulay et al. (2017) document that the institutional response to retraction, including the incentives and resources that publishers and journals devote to correcting the scientific record, varies substantially across publishing outlets. Collectively, these findings indicate that retraction lag is not solely an individual issue but is also shaped by collaborative organization, international coordination, and the publishing environment.

Finally, the misuse of generative artificial intelligence (GAI) and randomly generated content (RGC) has lowered the barriers to ghostwriting, fabricated text, and synthetic research materials (Kwon, 2024; Thorp, 2023). Although detection methods and editorial screening systems are continuously improving, technological advances often outpace existing regulatory frameworks and integrity safeguards. This structural gap between innovation and oversight means that emerging forms of misconduct may further increase the complexity of retraction investigations and contribute to prolonged retraction lags.

To sum up, existing studies provide substantial evidence for understanding article retractions, primarily focusing on their causes, structural characteristics, and consequences. Most research addresses a binary question, namely whether a paper is retracted, and commonly employs logistic or linear models to examine retraction risk and lag time. While

these approaches offer strong interpretability, they are limited in capturing multidimensional features, nonlinear relationships, and complex interactions.

### Explainable machine learning and its applications

With the widespread use of machine learning models in complex prediction tasks, their "black-box" nature has become a major barrier to theoretical interpretation and policy application. Explainable Machine Learning (XML) aims to make model predictions more transparent and understandable, enabling humans to interpret and evaluate how models arrive at particular outputs (Marcinkevičs & Vogt, 2023). From a cognitive perspective, explainability emphasizes the alignment between model decision logic and human reasoning. This is typically achieved by revealing feature contributions, variable interactions, and decision paths, thereby improving the traceability of predictions.

Existing approaches are generally classified into intrinsically interpretable models and post hoc explanation methods (Yao & Ni, 2023). Intrinsically interpretable models achieve transparency through structural constraints. Common examples include linear regression, logistic regression, decision trees, and rule-based models. These models embed interpretability into their structure by limiting complexity or imposing assumptions on data distributions (Molnar, 2020). However, their expressive power is limited. They often perform poorly when dealing with multidimensional and nonlinear problems compared to ensemble or deep learning models. In contrast, post hoc explanation methods are designed for complex and high-performance models, such as XGBoost, random forests, and neural networks. These methods provide external explanations using model-agnostic or model-specific techniques. They enhance transparency while maintaining predictive accuracy and have become the dominant approach in recent research (Guidotti et al., 2018).

Several widely used techniques have been developed in this area, including LIME, PDP, ICE, and SHAP. LIME (Local Interpretable Model-Agnostic Explanations), proposed by Ribeiro et al. (2016), explains individual predictions by fitting a local surrogate model around a specific instance. It is flexible and model-agnostic, but its results may be sensitive to local sampling. Partial Dependence Plots (PDP), introduced by Friedman (2001), show the average effect of a feature on the predicted outcome. PDP is widely used to analyze nonlinear relationships and interaction effects. However, it only reflects average behavior. To address this limitation, Goldstein et al. (2015) proposed Individual Conditional Expectation (ICE), which visualizes prediction changes at the individual level and captures heterogeneity across samples.

Among these methods, SHAP (SHapley Additive exPlanations), developed by Lundberg and Lee (2017), provides a unified framework based on Shapley values from cooperative game theory. SHAP quantifies the marginal contribution of each feature to a prediction and supports both local and global interpretation. The later development of TreeSHAP improves computational efficiency for tree-based models. Compared with other methods, SHAP satisfies desirable theoretical properties such as local accuracy, consistency, and missingness. It can effectively capture nonlinear relationships and feature interactions, achieving a balance between theoretical rigor and practical stability.

As machine learning becomes increasingly applied in information science and scientometrics, the demand for model interpretability continues to grow. Previous studies have used machine learning to predict citation counts (Akella et al., 2021), citation importance (Aljohani et al., 2021), researchers' h-index (Huang et al., 2022), and high-value patents (Choi et al., 2021). However, traditional predictive models often lack transparency in

their decision-making processes, which limits their applicability in policy and governance contexts.

To address, this issue, an increasing number of studies have adopted the SHAP (SHapley Additive exPlanations) framework. By quantifying the marginal contribution of individual features, SHAP enables both global and local interpretability, providing a powerful tool for uncovering underlying mechanisms in complex systems. For example, Yao and Ni (2023) employ SHAP to reveal the nonlinear effects of knowledge accumulation and technological complexity on patent grant probability. Dong et al. (2024) apply SHAP to identify key driving features in scientific writing, while Liu and Hu (2024) demonstrate that publisher characteristics and journal scale play significant roles in predicting retractions. More recently, Liu et al. (2025) utilize explainable machine learning to identify determinants of retracted biomedical papers, highlighting the importance of publisher attributes, email domains, country, and journal publication volume.

Overall, explainable machine learning not only enhances predictive performance but also bridges the gap between data-driven modeling and theoretical interpretation by revealing latent structural relationships. However, despite these advances, predictive frameworks specifically targeting retraction lag remain underdeveloped. This limitation is particularly evident in contexts involving the interaction of multiple dimensions, such as team structure, knowledge characteristics, and publishing environments, where the underlying mechanisms of retraction lag are still insufficiently understood.

To address this gap, this study integrates data from authoritative sources, including the Retraction Watch Database and PubMed, and adopts an explainable machine learning framework to construct a multidimensional model for predicting prolonged retraction lag. The aim is to identify key determinants and uncover the institutional and social mechanisms shaping retraction lag, thereby providing both quantitative tools and policy-relevant insights for research integrity governance.

## Data and methods

### Research framework

To systematically identify the key factors associated with prolonged retraction lag and develop an interpretable predictive model, this study established an integrated research framework combining machine learning prediction with interpretability analysis. The overall research process comprises three phases: data collection and feature selection, model construction and performance evaluation, and SHAP-based interpretability analysis. The research framework is illustrated in Fig. 1.

First, in the data collection stage, this study integrates retracted publication data from the Retraction Watch database covering the period from 2000 to 2024. These data are further matched with journal impact indicators, author team characteristics, and publisher information to construct a multidimensional dataset. The raw data are then processed through deduplication, missing value treatment, and outlier detection. Based on the research objectives, a comprehensive feature system is developed. This includes variables such as team size, author attributes, institutional and international collaboration scale, disciplinary diversity and interdisciplinarity, SJR score, and the complexity of retraction reasons. In addition, correlation analysis is conducted to examine linear relationships among

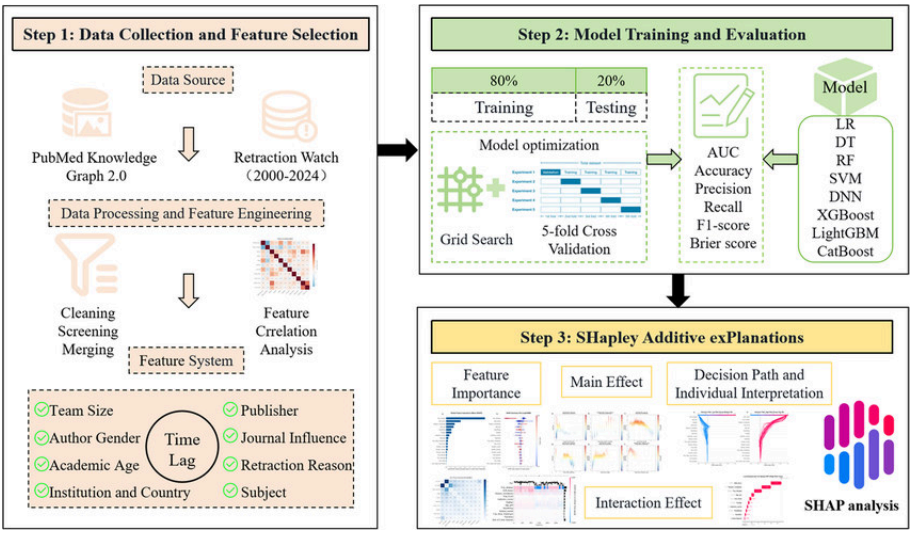



**Fig. 1** Research framework

features and to detect potential multicollinearity, ensuring the robustness of the feature system.

Second, in the modeling stage, multiple machine learning algorithms are employed to predict the categories of retraction lag. These include logistic regression, decision trees, random forests, support vector machines, deep neural networks, and three gradient boosting methods: XGBoost, LightGBM, and CatBoost. These models represent different learning paradigms. Logistic regression serves as a linear baseline to test interpretable relationships among variables. Decision trees and random forests are tree-based models that capture nonlinear relationships and interaction effects, while also improving robustness. Support vector machines enable nonlinear classification through kernel functions in multidimensional spaces. Deep neural networks extract complex feature representations through multiple nonlinear transformations. The three gradient boosting models demonstrate strong predictive performance in handling structured and multidimensional data (Table 1).

Model parameters are optimized using 5-fold cross-validation and grid search. Model performance is evaluated using multiple metrics, including AUC, Precision, Recall, F1-score, Accuracy, and Brier score. In addition, ROC curves, precision–recall curves, and confusion matrices are used for visualization. Based on these evaluations, the model with the best overall performance is selected as the base model for subsequent interpretation.

Finally, in the model interpretation stage, this study applies the SHAP (Shapley Additive Explanations) method to analyze the decision mechanisms of the selected model. Interpretation is conducted at multiple levels. At the global level, feature importance is assessed by ranking the mean absolute SHAP values, allowing the identification of key factors influencing retraction lag. At the structural level, feature interaction strength matrices and SHAP heatmap clustering are used to reveal interactions and underlying contribution structures among variables. At the group level, SHAP decision paths are compared between relatively short and prolonged retraction lag samples to examine differences in feature contribution patterns. At the individual level, local explanations are used to analyze typical cases and uncover the mechanisms behind prolonged retraction lag.

**Table 1** Conceptual framework and measurement of model features

| Conceptual dimension | Variable | Operational definition |
|---|---|---|
| Publishing and oversight environment | Publisher | Publisher responsible for the journal in which the article appeared |
| | SJR score | SCImago Journal Rank (SJR) of the publishing journal |
| Retraction-case and investigation complexity | Reason complexity | Number and diversity of retraction reasons recorded in the Retraction Watch Database for each retracted paper |
| Team structure | Team size | Total number of authors |
| | Average academic age | Mean academic age of all authors |
| | Academic age Gini coefficient | Gini coefficient of authors' academic ages |
| | Gender entropy | Shannon entropy calculated from authors' gender composition |
| | Institution count | Number of distinct author affiliations |
| | Country count | Number of countries represented by the author affiliations |
| Knowledge and disciplinary structure | Subject count | Number of subject categories assigned to the publication |
| | Rao-Stirling diversity | Rao-Stirling diversity index calculated from disciplinary categories |

By combining machine learning prediction with explainable analysis, this study not only identifies papers with prolonged retraction lag, but also systematically uncovers the key structural factors that drive differences in retraction duration. This approach provides empirical evidence and data support for understanding the institutional and social drivers of the self-correcting process in science.

## Data sources and preprocessing

The data used in this study were obtained from the Retraction Watch Database[1](RW) and the PubMed Knowledge Graph 2.0[2](PKG 2.0). All data were collected on December 13, 2025. The RW database compiles retraction records from multiple sources, including public databases (e.g., Google Scholar), publisher platforms (e.g., SpringerLink and ScienceDirect), major journals (e.g., Nature and Science), research misconduct investigation agencies, and social media. It is widely regarded as one of the most comprehensive repositories of retracted publications (Brainard, 2018). The database provides rich metadata, including title, discipline, institution, journal, publisher, country, authors, article type, publication date, retraction date, type of retraction notice, access status, and reasons for retraction. PKG 2.0 integrates metadata from more than 36 million publications indexed in PubMed, covering information such as titles, authors, affiliations, keywords, and citations (Xu et al., 2025). Compared with the original PubMed dataset, PKG 2.0 incorporates author disambiguation data from sources such as Author-ity and Semantic Scholar. It also applies deep neural network techniques to identify and normalize author identities, addressing the issue of sparse and ambiguous author information in PubMed. PKG 2.0 further utilizes tools such as Genderize[3] and Gender Guesser[4] to probabilistically infer gender and race based on names. This approach has been widely adopted in prior research (Karimi et al., 2016; Mueller & Stumme, 2016). It provides a feasible way to examine how individual characteristics within author teams influence retraction lag.

To construct a high-quality dataset of retracted articles, a rigorous data preprocessing procedure was applied. First, the RW database categorizes post-publication updates into retractions, corrections, and expressions of concern. In this study, a record is defined as a "retraction" if the article has been temporarily or permanently removed from publication. Records labeled as corrections, where the article remains published but has been modified, are excluded. "Expressions of concern" are also excluded, as they represent editorial notices rather than confirmed retractions. Second, considering key phases in the digital transformation of scholarly publishing-namely, the digitization of traditional journals (2000–2010), the rise of open access (2010–2020), and the expansion of preprints after 2020-as well as the completeness and update cycle of PKG 2.0, this study restricts both publication and retraction dates to the period from 2000 to 2024. Finally, the RW and PKG 2.0 datasets are merged using PMID as a key identifier. The resulting dataset includes essential variables such as article title, discipline, publisher, country, retraction reasons, publication and retraction dates, number of authors and gender composition,

[1] The Retraction Watch Database [Internet]. New York: The Center for Scientific Integrity. 2018. ISSN:2692–4579. Available from: http://retractiondatabase.org/.

[2] PubMed Knowledge Graph 2.0 [Dataset]. Figshare. Available from: https://figshare.com/s/05158da6355d6199d3b7.

[3] Genderize.io [Internet]. Available from: https://genderize.io/our-data.

[4] Gender Guesser [Internet]. Available from: https://gender-guesser.com/.

institutional affiliations, citation counts, references, journal name, and SJR scores. Additional data cleaning steps are performed to ensure data quality. Missing values in continuous variables-such as team size, number of institutions and countries, and retraction lag are examined, and outliers are removed. Specifically, extreme cases with team size greater than 100 or retraction lag exceeding 30 years are excluded. Missing values in categorical variables are labeled as "unknown". After these procedures, the final dataset contains 13,685 retracted records.

## Variable measurement

### Dependent variable

Retraction lag refers to the interval between the publication of a paper and its subsequent retraction, calculated as the number of days between the publication date and retraction date. The distribution of retraction lag in our dataset is highly right-skewed, ranging from 0 to more than 8,000 days. The mean retraction lag is 1,363 days (approximately 3.7 years), whereas the median is 806 days (approximately 2.2 years), indicating substantial heterogeneity in the timing of retractions. Rather than modeling the exact duration of retraction, this study focuses on identifying factors associated with prolonged retraction lag. Accordingly, we formulate the problem as a binary classification task by distinguishing relatively short and prolonged retraction lag.

A threshold of three years (1,095 days) is adopted as the primary cutoff. This choice is motivated by considerations of scientific knowledge diffusion and research integrity. The purpose of this study is to distinguish papers that are retracted after a relatively limited period of exposure from those that remain in the scientific record long enough to exert broader influence on subsequent research. Prior studies suggest that the first several years following publication represent a critical period for knowledge dissemination, citation accumulation, and incorporation into later studies. Papers that remain unretracted beyond this stage may have greater opportunity to shape subsequent scientific activity and propagate potentially problematic findings. Therefore, from a research-integrity perspective, a three-year threshold provides a practically meaningful distinction between relatively short and prolonged retraction lags. The binary target variable is defined as follows:

$$\mathrm{Y} = \begin{cases} 1, \ \textit{Retraction lag time} > \ 1095 \ \textit{days} \\ 0, \ \textit{Retraction lag time} \ \leq \ 1095 \ \textit{days} \end{cases} \tag{1}$$

After classification, papers with prolonged retraction lags account for 40.5% of the dataset, while the remaining 59.5% are classified as relatively short lag cases. This distribution is sufficiently balanced for supervised machine-learning analysis. The binary formulation is determined by the substantive estimand of this study rather than by an assumption that classification is generally superior to continuous-duration or survival models. Specifically, our primary target is the conditional probability that the observed retraction lag exceeds a practically meaningful exposure horizon among papers that were ultimately retracted,

$$P(T > 1{,}095 days | X, R = 1) \tag{2}$$

where T denotes the observed publication-to-retraction interval and R=1 denotes eventual retraction. This estimand differs from the expected duration E(T|X, $R$=1) estimated by conventional duration regression and from the time-specific retraction hazard h(t|X) typically

examined in survival analysis. Critically, this estimand is defined only within the sub-population of papers that were ultimately retracted ($R = 1$); it does not identify $P(R = 1|X)$, the probability that an arbitrary published paper is problematic or will eventually be retracted, nor the retraction hazard for the broader population of published papers. Accordingly, the model characterizes variation in retraction timeliness among completed retraction cases and should not be interpreted as an early-warning or screening tool for currently published, not-yet-retracted papers.

Threshold-based classification is adopted for three reasons. First, the research question concerns whether problematic findings remained in the literature beyond a substantively defined exposure window, rather than the prediction of the exact number of days to retraction. Crossing this threshold has a direct governance interpretation because it distinguishes relatively limited exposure from prolonged persistence in the scientific record. Second, the observed lag distribution is strongly right-skewed and contains complex nonlinearities and interactions. Tree-based classification allows these patterns to be examined without imposing a linear functional form on exact duration. Third, the present dataset contains only papers for which the retraction event was ultimately observed and does not include the unretracted or right-censored source population. Consequently, the principal advantage of survival analysis in handling censoring cannot be fully utilized in the current design.

Nevertheless, dichotomization inevitably discards within-category timing information. We therefore do not present the classification framework as a replacement for continuous-duration analysis. Instead, it addresses a distinct, policy-oriented question concerning prolonged lag. Sensitivity analyses using one-year and five-year thresholds are reported in Appendix 1, and a continuous-duration analysis is provided as an additional robustness assessment in Appendix 2.

### Predictive variables

Predictor selection was guided by a four-dimensional conceptual framework rather than by the isolated availability of individual variables. First, the publishing and oversight environment captures institutional differences in editorial governance, journal visibility, research-integrity capacity, and external scrutiny, and is represented by publisher affiliation and SJR score. Second, retraction-case and investigation complexity reflects the amount and diversity of evidence and coordination potentially involved in resolving a case, and is represented by retraction reason complexity. Third, team structure captures the allocation of responsibility, hierarchical structure, and coordination burden within and across research organizations, and includes team size, average academic age, academic-age inequality, gender composition, institution count, and country count. Fourth, knowledge and disciplinary structure captures the cognitive breadth and disciplinary boundaries across which potential problems must be evaluated, and is represented by subject count and Rao-Stirling diversity. Together, these dimensions reflect institutional, procedural, organizational, and cognitive mechanisms that may be associated with variation in retraction lag.

Team size refers to the number of authors of a retracted paper and serves as a core indicator of the scale of scientific collaboration. Larger research teams generally involve more complex collaborative governance, increasing communication costs, responsibility allocation, and cross-author coordination when research concerns arise. These organizational characteristics may influence both the investigation process and the time required to reach a retraction decision.

Gender entropy is derived from Shannon entropy and measures the uncertainty and balance of gender composition within a team. Team diversity may influence communication patterns, oversight mechanisms, and collective decision-making processes. Although evidence regarding gender composition and retraction outcomes remains limited, diversity-related characteristics may be associated with how research problems are detected and addressed. Let $p_g$ denote the proportion of gender g in the team, and k represent the number of gender categories (typically k=2, i.e., male and female). When $p_{male} = p_{female} = 0.5$, entropy reaches its maximum value of ln(2), indicating perfect balance.

$$\text{Gender Entropy} = -\sum_{g=1}^{k} p_g \ln\left(p_g\right) \tag{3}$$

Average academic age measures the average academic experience of team members. Teams with greater average academic age may possess more accumulated research experience and stronger familiarity with publication norms. Such experience may facilitate the recognition and handling of problematic research, although it may also be associated with more complex professional networks and institutional procedures. Let $Age_i$ denote the academic age of the i member, and N the team size. This indicator reflects the overall seniority structure of the team.

$$\text{Average Age} = \frac{1}{N}\sum_{i=1}^{N} \text{Age}_i \tag{4}$$

Academic age Gini coefficient measures inequality in academic seniority within a team. Differences in academic seniority may affect the distribution of authority and responsibility within research teams. Teams with greater disparities in academic age may exhibit more hierarchical structures, which could influence both internal oversight and the investigation process when research concerns arise. Let $x_i$ and $x_j$ denote the academic ages of members i and j, and $\overline{x}$ the mean academic age. The value ranges from 0 to 1. Higher values indicate greater disparity in academic age and stronger generational heterogeneity, while lower values indicate a more homogeneous structure.

$$\text{AgeGini} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|X_i - X_j\right|}{2N^2\overline{X}} \tag{5}$$

Institution counts refer to the number of distinct institutions affiliated with the authors of a retracted paper. Multi-institution collaborations require coordination across different organizational structures and administrative procedures. When research-integrity concerns emerge, a larger number of participating institutions may increase communication costs and delay the completion of investigations.

Country counts refer to the number of distinct countries represented by the authors. International collaborations introduce differences in regulatory frameworks, institutional policies, languages, and administrative practices. These additional coordination requirements may increase the complexity of cross-border investigations and contribute to prolonged retraction lag.

Subject counts measure the number of subject labels assigned to a retracted paper. These labels, derived from the RW database, represent research areas in which the paper is most likely to be cited or searched. Papers spanning multiple subject areas may attract attention from more diverse scientific communities while also requiring expertise from multiple

disciplines during investigation. Such disciplinary breadth may simultaneously increase the likelihood of problem detection and the complexity of evidence evaluation.

Rao-Stirling diversity is based on the weighted product of the distribution of references across disciplines and a disciplinary distance matrix. Let $p_i$ and $p_j$ denote the proportions of references in disciplines i and j, and $d_{ij}$ the cognitive distance between them. The index reaches higher values when the distribution is balanced and disciplinary differences are substantial. It captures the extent to which a paper integrates knowledge across fields. Interdisciplinary research integrates knowledge from multiple domains and may therefore be subject to more heterogeneous evaluation standards. Potential problems may be more difficult to identify when expertise is distributed across disciplinary boundaries, potentially affecting the timing of retraction decisions.

$$\text{Rao - Stirling} = \sum_i \sum_j P_i P_j d_{ij} \tag{6}$$

Publisher refers to the publishing organization responsible for the journal in which a retracted article was published. As an institutional characteristic of the publication venue, publisher affiliation may reflect differences in editorial governance, integrity oversight, and investigation capacity. These differences can shape how research concerns are processed and may therefore be associated with variation in retraction lag. Since publisher is a high-cardinality categorical variable with fragmented and non-standardized naming (e.g., "Elsevier" vs. "Elsevier-Cell Press"), direct use of raw data would lead to sparsity and diluted effects. To address, this issue, a standardization strategy based on corporate ownership is applied. Sub-brands are mapped to their parent companies (e.g., "Springer-Nature Publishing Group" and "Springer-BioMed Central" are unified as "Springer"). Given the high concentration of the global publishing market, a "top-tier grouping" approach is further adopted. The six largest publishers-Hindawi, Elsevier, Wiley, Springer, PLoS, and Taylor & Francis-account for 77% of retraction cases in the dataset. Although Hindawi was acquired by Wiley in 2021, it is treated as an independent category due to its distinct retraction patterns. All remaining publishers are grouped into "Other Publishers." To meet machine learning requirements and avoid artificial ordinal assumptions, the processed publisher variable is encoded using one-hot encoding, generating binary dummy variables. This approach reduces dimensionality while allowing the model to estimate differences in institutional oversight and governance capacity across major publishing systems.

SJR score is measured using the SCImago Journal Rank, obtained directly from the PKG 2.0 dataset. As a widely used indicator of journal prestige and visibility, SJR score may reflect differences in editorial resources, external scrutiny, and institutional capacity for research-integrity oversight. Journals with different levels of prestige may therefore exhibit systematic differences in the efficiency with which allegations of misconduct or error are investigated and resolved.

Reason complexity is operationalized as the number and diversity of retraction reason labels associated with a single paper in the Retraction Watch Database. Rather than representing an intrinsic attribute of the publication itself, this variable characterizes the complexity of the completed retraction case and serves as a proxy for the scope of investigation required before a retraction decision is reached. Cases involving multiple concerns often require sequential evaluation of different types of evidence, consultation among authors, institutions, editors, and publishers, and resolution of overlapping ethical or procedural issues, thereby increasing coordination costs and potentially prolonging retraction lag. Because this information is derived from the final retraction record, it is interpreted as a retrospective explanatory feature rather than an ex ante predictor available before retraction.

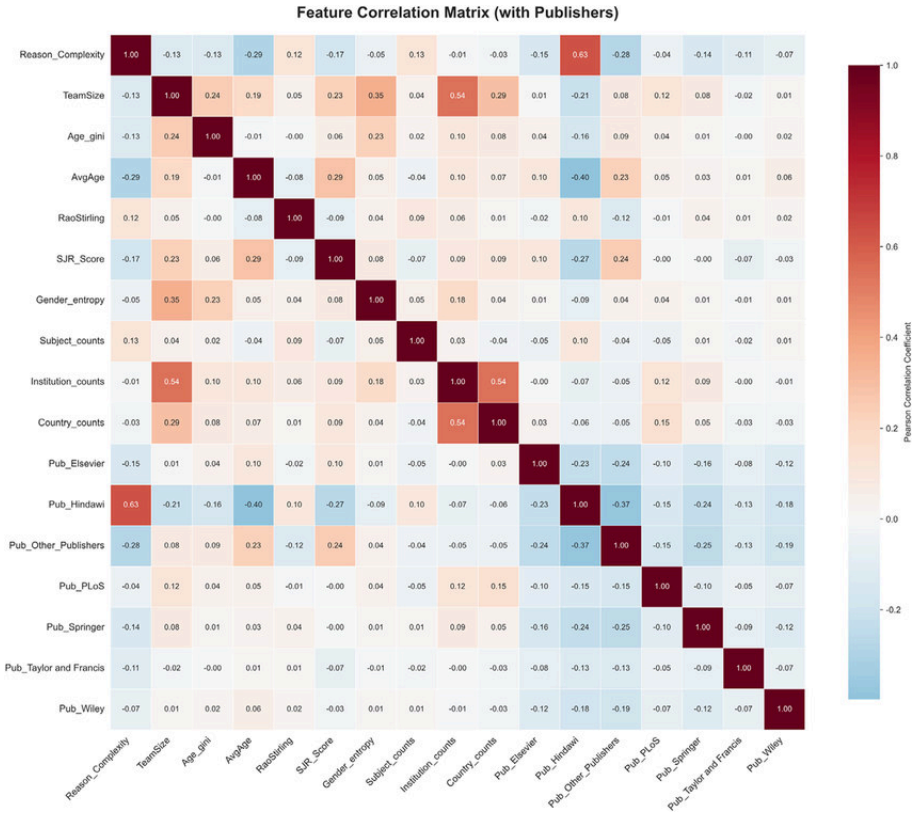



**Fig. 2** Variable correlation Heatmap

In summary, this study constructs a multidimensional feature system based on four complementary conceptual dimensions: publishing and oversight environment, retraction-case and investigation complexity, team structure, and knowledge and disciplinary structure. Together, these dimensions capture institutional, procedural, organizational, and cognitive mechanisms that may contribute to variation in retraction lag. Publishing-related variables (publisher and SJR score) characterize differences in editorial governance and oversight capacity. Reason complexity represents the complexity of completed retraction cases and serves as a retrospective indicator of investigation burden. Team-level variables describe collaborative organization, demographic composition, and cross-institutional coordination, while knowledge-structure variables characterize disciplinary breadth and interdisciplinarity. Collectively, this framework provides an integrated perspective for examining the determinants of prolonged retraction lag.

## Results and discussion

### Correlation analysis

Before constructing the predictive models, a Pearson correlation analysis was conducted on all numerical variables, along with the one-hot encoded publisher variables, to examine the relationships among potential explanatory factors (see Fig. 2). Although tree-based machine learning models can partially mitigate multicollinearity, identifying correlations among variables helps avoid redundant attribution to the same underlying constructs during model interpretation.

Most variables show low correlation coefficients, with no strong correlations observed. This suggests that the feature system has good statistical independence. At a more detailed level, institution counts are moderately correlated with both team size and country counts ($r = 0.54$). This indicates that larger research teams are often associated with broader institutional participation and greater international collaboration.

Apart from these structural relationships, correlations among other variables are generally weak. This implies that team structure, knowledge-related features, and publishing environment characteristics are relatively independent from a statistical perspective. Notably, the Hindawi publisher variable shows a positive correlation with retraction reason complexity ($r = 0.63$), suggesting that retraction patterns may vary across publishers within the sample.

Overall, the correlation analysis indicates that there is no severe multicollinearity in the dataset. This provides a solid foundation for subsequent machine learning modeling and SHAP-based interpretation.

### Model training and evaluation

To capture potential nonlinear relationships between retraction lag and its influencing factors, this study constructs machine learning classification models. The dependent variable is the category of retraction lag, while the predictive features are drawn from publishing and oversight environment, investigation complexity, team structure, knowledge and disciplinary structure. Publishing and oversight features include publisher type and SJR score; investigation complexity is represented by reason complexity; team structure features include team size, average academic age, academic-age inequality, gender entropy, institution count, and country count; and knowledge-structure features include subject count and Rao-Stirling diversity.

During model training, the dataset is split into training and testing sets with a ratio of 8:2. To improve robustness, five-fold stratified cross-validation is applied. Grid search is used to systematically optimize key hyperparameters. The parameter grid is defined based on commonly used ranges for classification tasks of similar scale. The optimal parameter settings for each model are reported in Table 2.

To comprehensively evaluate the performance of different algorithms in identifying retraction lag, this study conducts a systematic comparison based on the optimal parameter settings. Multiple evaluation tools are employed, including ROC curves, precision-recall (P-R) curves, confusion matrices, and the Brier Score. The results are presented in Figs. 3 and 4. The ROC curve illustrates the trade-off between the true positive rate and the false positive rate, capturing the model's overall ability to distinguish between classes. The area under the curve (ROC-AUC) provides a quantitative measure of discriminative

**Table 2** Optimal parameter values for each model

| Model | Parameter | Optimum | Model | Parameter | Optimum |
|---|---|---|---|---|---|
| XGBoost | colsample_bytree | 0.8 | Random Forest | max_depth | 20 |
| | learning_rate | 0.01 | | min_samples_leaf | 1 |
| | max_depth | 11 | | n_estimators | 200 |
| | n_estimators | 500 | DecisionT ree | max_depth | 10 |
| | subsample | 0.8 | | min_samples_leaf | 10 |
| LightGBM | colsample_bytree | 0.8 | Logistic Regression | C | 10 |
| | learning_rate | 0.05 | | penalty | l2 |
| | n_estimators | 200 | SVM | C | 0.68 |
| | num_leaves | 50 | | gamma | auto |
| | subsample | 0.8 | | kernel | rbf |
| CatBoost | depth | 6 | DNN | activation | tanh |
| | iterations | 500 | | alpha | 0.0019 |
| | l2_leaf_reg | 3 | | hidden_layer_sizes | (50,) |
| | learning_rate | 0.05 | | —— | —— |

performance. In contrast, the P-R curve evaluates the relationship between precision and recall across different thresholds. It is particularly informative for identifying prolonged retraction lag samples in classification tasks with moderately imbalanced data.

Confusion matrices provide a detailed view of prediction outcomes by showing how each model classifies relatively short and prolonged retraction lag samples. The Brier Score evaluates the accuracy of probabilistic predictions by measuring the deviation between predicted probabilities and actual outcomes. A lower Brier Score indicates better calibration and higher prediction accuracy.

In terms of overall discriminative performance, the models show similar patterns in the ROC curves (Fig. 3). Among them, Random Forest, XGBoost, and LightGBM achieve the highest ROC-AUC values of 0.857, 0.857, and 0.856, respectively. These results are clearly superior to those of other models, indicating strong ability to distinguish between prolonged and relatively short retraction lag samples. The P-R curves in Fig. 3 further highlight performance differences. Ensemble tree-based models outperform the linear baseline model. Random Forest and XGBoost both achieve an average precision (AP) of 0.777, followed by LightGBM (AP = 0.775) and CatBoost (AP = 0.771).

The confusion matrices in Fig. 3 provide a more detailed view of model performance at specific classification thresholds. LightGBM achieves the highest F1-score (0.734). It correctly identifies 1,166 relatively short retraction lag samples and 910 prolonged retraction lag samples. At the same time, it misclassifies 464 prolonged samples as relatively short and 197 relatively short samples as prolonged. In comparison, although Random Forest and XGBoost perform similarly in overall discrimination, they are slightly more conservative in identifying prolonged retraction lag samples.

From the perspective of probabilistic prediction, the Brier Score further evaluates the consistency between predicted probabilities and actual outcomes. This metric measures the mean squared error between predicted probabilities and true labels. It reflects both classification performance and probability calibration. In Fig. 4, the lightly shaded background represents histograms of predicted probabilities. The calibration curves generally show a

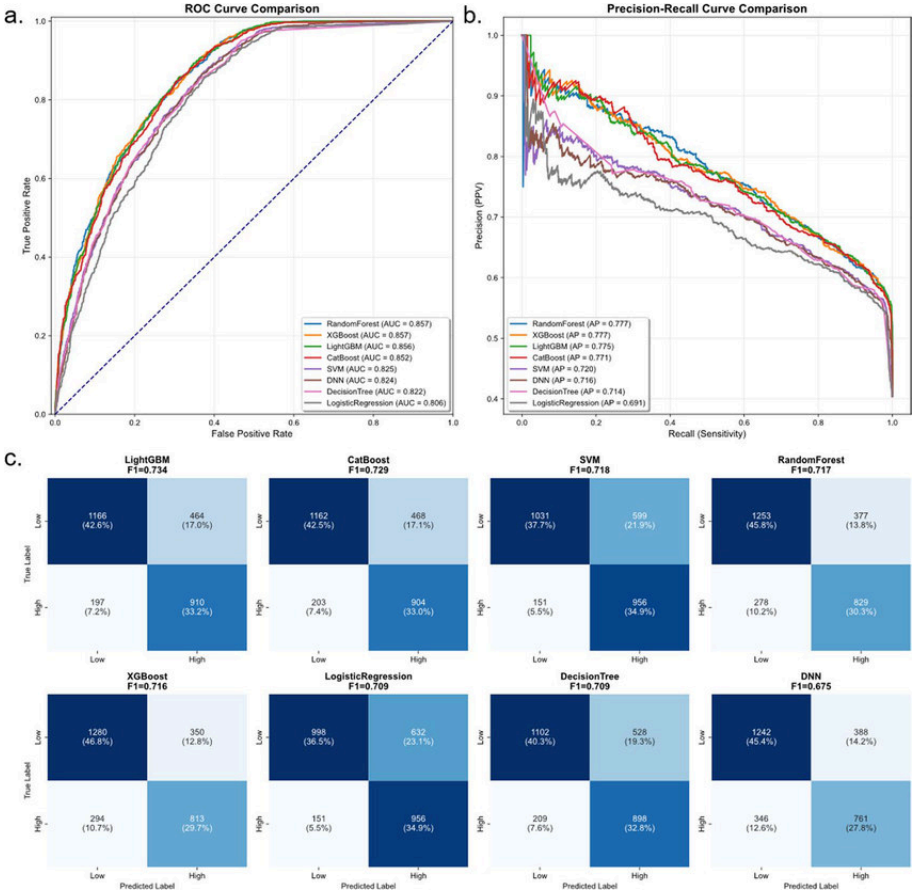



**Fig. 3** ROC and PR Curve, confusion matrices

monotonic increasing trend, indicating good alignment between predicted and observed probabilities.

Among all models, Random Forest achieves the lowest Brier Score (0.1503), suggesting slightly better calibration performance. LightGBM (0.1539) and XGBoost (0.1511) show comparable results, indicating that ensemble models perform consistently well in probability estimation. In contrast, models such as SVM, decision trees, and logistic regression have higher Brier Scores, reflecting larger prediction errors. It should be noted that the difference in Brier Scores between Random Forest and LightGBM is small and has limited impact on overall performance.

Considering overall classification ability, risk identification performance, and probability stability, as well as LightGBM's advantages in handling complex nonlinear relationships and multidimensional feature interactions, this study ultimately selects LightGBM as the base model for subsequent SHAP-based interpretability analysis.

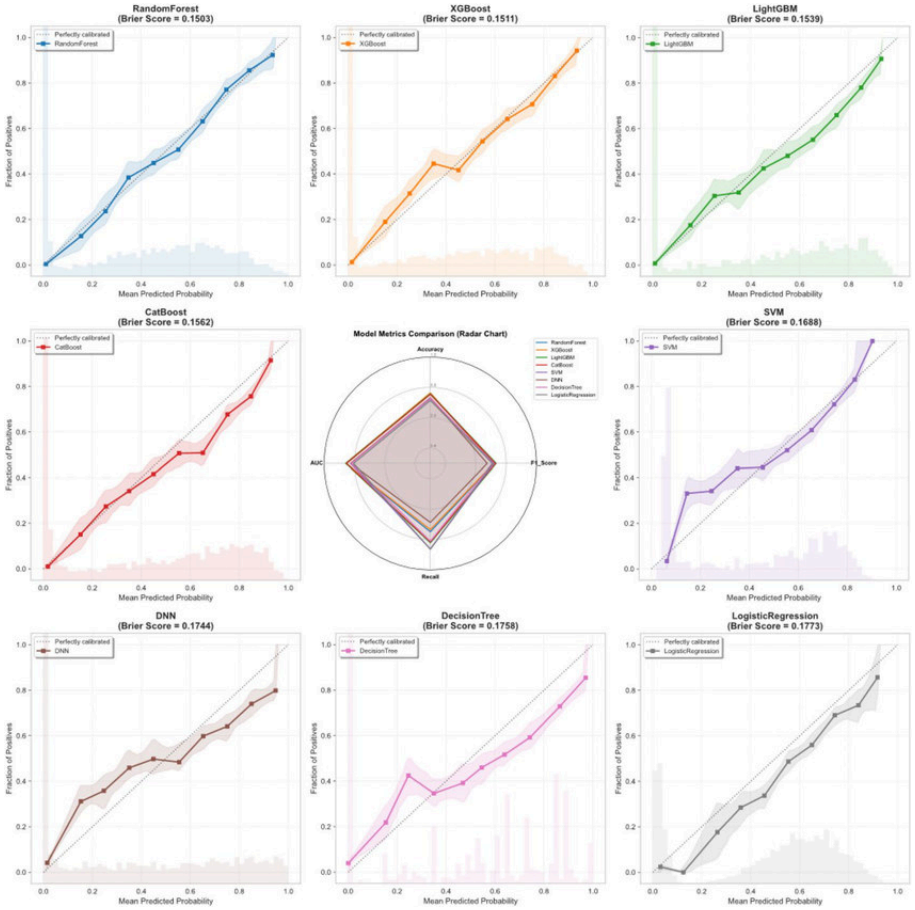



**Fig. 4** Brier score

## SHAP analysis

To examine the factors underlying the model predictions, LightGBM was selected as the final model based on its superior overall performance and stability across evaluation metrics. SHAP based interpretability analysis was then applied to this model.

SHAP quantifies the marginal contribution of each feature to individual predictions across different feature combinations, providing a theoretically grounded framework for interpreting complex machine learning models. In this study, the model predicts the outcome $Y$, where $Y$=1 denotes prolonged retraction lag and $Y$=0 denotes relatively short retraction lag. A positive SHAP value indicates that a feature shifts the model output toward $Y$=1, whereas a negative value indicates a shift toward $Y$=0. Because SHAP values are calculated from the model's predicted probability of prolonged retraction lag, they are interpreted as increasing or decreasing the predicted likelihood of prolonged retraction lag, rather than as causing longer or shorter retraction lag. Thus, SHAP describes the model's decision process rather than a causal mechanism.

SHAP values should also be distinguished from statistical significance or causal importance. A feature's SHAP based importance reflects its contribution to model predictions and predictive performance, but does not by itself establish a statistically significant or causal relationship with retraction lag. Accordingly, the following analyses distinguish between predictive importance, referring to the extent to which a feature contributes to model predictions, and explanatory importance, referring to the extent to which its contribution can be interpreted through a stable and substantively plausible pattern. Cases where these dimensions diverge are interpreted with appropriate caution.

## Feature importance

The SHAP summary plot provides an intuitive visualization of the contribution and direction of different variables in the model's prediction process. For global importance analysis, SHAP values are first computed for each feature in every sample. The mean absolute SHAP value across all samples is then calculated to measure the overall contribution of each feature. A higher value indicates a stronger influence on the model's classification decisions.

The results are shown in Fig. 5. The bar chart (a) presents the global ranking of feature importance, while the summary plot (b) further illustrates the direction and magnitude of feature effects across samples. In the plot, each point represents an individual sample. The x-axis shows the SHAP value, and the y-axis corresponds to specific features. The color indicates the feature value, with red representing higher values and blue representing lower values.

As shown in Fig. 5a, features from the publishing and oversight environment dimension dominate the importance ranking. The publisher indicator for Hindawi has the highest mean absolute SHAP value, followed by SJR score, indicating that publisher and journal characteristics make the largest contributions to the model's classification of retraction lag.

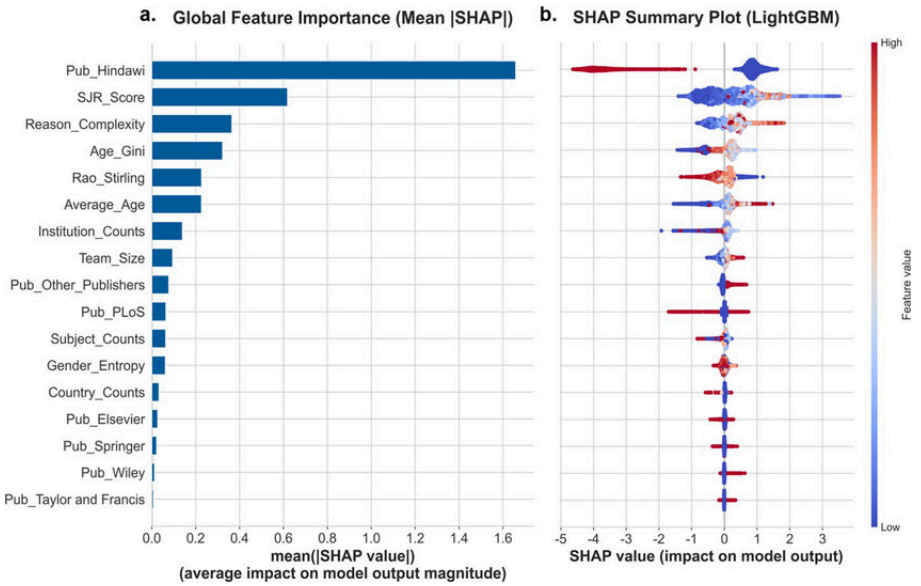



**Fig. 5** Feature importance ranking

Reason complexity ranks third, suggesting that the complexity of retraction cases is also an important factor. Academic age Gini coefficient and average academic age show relatively high importance, followed by the number of institutions and team size. Rao-Stirling diversity, representing the knowledge and disciplinary structure dimension, also makes a notable contribution, although its importance is lower than that of the leading publishing and team related features.

Further insights can be obtained from the SHAP summary distribution in Fig. 5b. The Hindawi publisher indicator shows the clearest directional pattern among all features: high values, represented by red points, are concentrated mainly in the negative SHAP region, whereas low values are more concentrated in the positive region. This indicates that papers published by Hindawi are more strongly associated with lower predicted probabilities of prolonged retraction lag. SJR score shows a comparatively broader distribution, with higher values tending to extend toward positive SHAP values. Reason complexity also exhibits a generally positive association, although its effect is less pronounced. By comparison, the publisher indicators for Springer, Elsevier, Wiley, and Taylor & Francis are concentrated close to zero, indicating relatively limited contributions to model output.

The prominent contribution of the Hindawi publisher indicator should also be interpreted in light of its distribution in the dataset, as Hindawi accounts for 3,583 retracted papers in the sample. Its strong SHAP contribution may reflect differences in publication volume, post publication scrutiny, and research integrity oversight. Periods of heightened attention to particular publishing environments may also facilitate the screening, reporting, and investigation of problematic articles, potentially contributing to shorter observed retraction lags. Therefore, the relatively short retraction lag associated with Hindawi papers should not be interpreted as evidence of superior editorial self correction, but rather as an association that may reflect broader differences in the publishing and oversight environment and warrants further investigation.

Academic age Gini coefficient and average academic age show generally positive effects, with higher values tending to be associated with positive SHAP values. The number of institutions and team size have smaller overall contributions and more dispersed SHAP distributions. Rao-Stirling diversity shows a relatively weak and mixed directional pattern, with higher values appearing on both sides of zero, indicating that its association with retraction lag is less pronounced and less consistent than that of the leading predictors. The number of subject categories, gender entropy, and the number of countries also have relatively small mean absolute SHAP values, with their SHAP distributions largely centered near zero.

Overall, retraction lag is primarily associated with the publishing and oversight environment, followed by retraction-case and investigation complexity and selected characteristics of team structure. The knowledge and disciplinary structure dimension contributes less to the model's overall predictive importance, although Rao-Stirling diversity remains among the more influential features within this dimension.

To evaluate the robustness of the findings, we further examined alternative definitions of prolonged retraction lag using one-year and five-year thresholds (Appendix 1). As expected, some variation in the relative ranking of individual features was observed across threshold specifications, reflecting differences in how prolonged retraction lag is defined. Nevertheless, the overall explanatory structure remained highly consistent. Key predictors, including publisher-related variables, SJR score, retraction reason complexity, and selected team-structure characteristics, consistently ranked among the more influential features across models. This pattern indicates that the identified associations are not specific to the three-year cutoff but reflect broader characteristics of retraction lag.

This conclusion is further reinforced by the continuous duration analyses reported in Appendix 2, where a lognormal AFT model yielded directionally concordant coefficients for the leading predictors and a correlated importance ranking across frameworks (Spearman $\rho = 0.556$, $p = 0.025$), while independently reproducing the nonlinear SHAP patterns under a continuous outcome specification. Accordingly, the three-year threshold is retained as the primary analytical setting because it provides a practically meaningful distinction between relatively short and prolonged exposure of problematic research while yielding a balanced classification structure for machine-learning analysis.

## Feature main effects

In the classification of retraction lag, feature variables not only differ in their overall importance but also exhibit distinct nonlinear relationships with model outputs. To further examine the main effects of key variables, this study employs SHAP dependence plots to characterize the relationship between feature values and their corresponding SHAP values.

Compared with traditional partial dependence plots, SHAP dependence plots are based on the Shapley value decomposition framework. They capture the true marginal contribution of a single feature while controlling for the effects of other variables, thereby providing a more accurate representation of complex nonlinear relationships. Since SHAP dependence plots are most informative when a feature exhibits a relatively stable marginal pattern, this analysis focuses on representative variables with interpretable nonlinear trends. Other variables contribute to model predictions but display highly dispersed, interaction-dominated, or non-monotonic distributions, making isolated main-effect interpretation less informative. Therefore, this study focuses on representative features with identifiable patterns, as shown in Fig. 6. Overall, variables such as SJR score, retraction reason complexity, academic age Gini coefficient, average academic age, Rao-Stirling diversity, and institution count all show clear nonlinear relationships with model predictions.

First, SJR score exhibits a nonlinear relationship with the model output. At relatively low SJR values, SHAP values are widely dispersed across both positive and negative

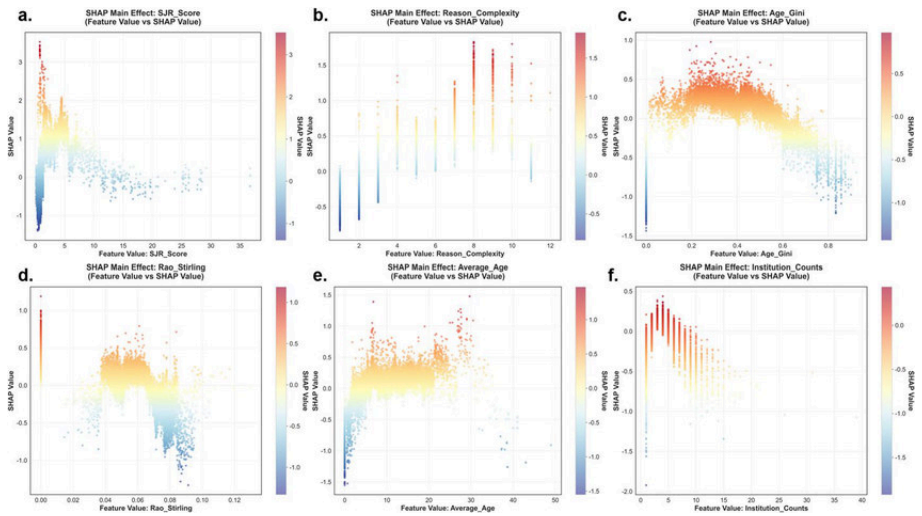



**Fig. 6** Feature main effects

regions. As SJR score increases from approximately 1 to 5, SHAP values become increasingly positive, indicating that higher SJR scores within this range tend to increase the model prediction of prolonged retraction lag. At higher SJR values, however, SHAP values gradually decline toward zero and in some cases become negative, suggesting that the positive association weakens and may reverse at higher levels.

Second, retraction reason complexity shows a generally positive relationship with the prediction of prolonged retraction lag. At lower levels of reason complexity, SHAP values are predominantly negative, indicating a greater contribution toward predicting relatively short retraction lag. As the number of retraction reasons increases, SHAP values progressively rise and become predominantly positive, suggesting that papers involving more complex retraction reasons are more likely to be classified as having prolonged retraction lag. At the highest levels, however, the SHAP values become more dispersed, with some observations falling into the negative region. Thus, the positive relationship is most evident across the central range of reason complexity rather than being strictly monotonic at extreme values.

Third, academic age Gini coefficient displays a clear inverted U shaped relationship. At very low values, SHAP values are predominantly negative, indicating a tendency toward relatively short retraction lag. As the age Gini coefficient increases, SHAP values rise and become predominantly positive at moderate levels, suggesting that moderate differences in academic seniority are associated with a higher probability of prolonged retraction lag. At higher levels of age inequality, SHAP values decline and become predominantly negative, indicating that this positive association weakens and may reverse when seniority inequality becomes pronounced.

In contrast, average academic age shows a rapid increase followed by stabilization. At low values, SHAP values rise quickly from negative to near-zero or positive levels, indicating that younger teams are more likely to be associated with relatively short retraction lag. As average academic age increases further, the variation in SHAP values diminishes, suggesting a limited marginal effect at higher levels.

Fourth, Rao-Stirling diversity also exhibits a nonlinear relationship with model output. At very low levels, SHAP values are relatively dispersed but include a substantial proportion of positive values. At moderate levels, SHAP values are predominantly positive, indicating a greater contribution to the prediction of prolonged retraction lag. As Rao-Stirling diversity increases further, SHAP values decline and become predominantly negative at higher levels. This suggests that moderate levels of disciplinary diversity are more strongly associated with prolonged retraction lag, whereas the contribution becomes weaker or reverses at higher levels.

Fifth, institution count shows a generally negative relationship with the prediction of prolonged retraction lag. At low institution counts, SHAP values are predominantly positive, indicating a greater tendency toward prolonged retraction lag. As the number of institutions increases, SHAP values gradually decline and become increasingly negative, suggesting that greater institutional participation is associated with a higher tendency toward relatively short retraction lag. However, observations at higher institution counts are relatively sparse and dispersed, so this relationship should be interpreted with caution.

In summary, the SHAP main-effect analysis reveals that the influence of these variables on retraction lag is predominantly nonlinear and varies across value ranges: academic age Gini coefficient and Rao-Stirling diversity follow inverted U-shaped patterns, in which only moderate levels are associated with prolonged lag; SJR score and reason complexity show a largely monotonic direction that partially reverses at extreme values; average academic age follows a rising-then-saturating pattern; and institution count shows a weak, largely

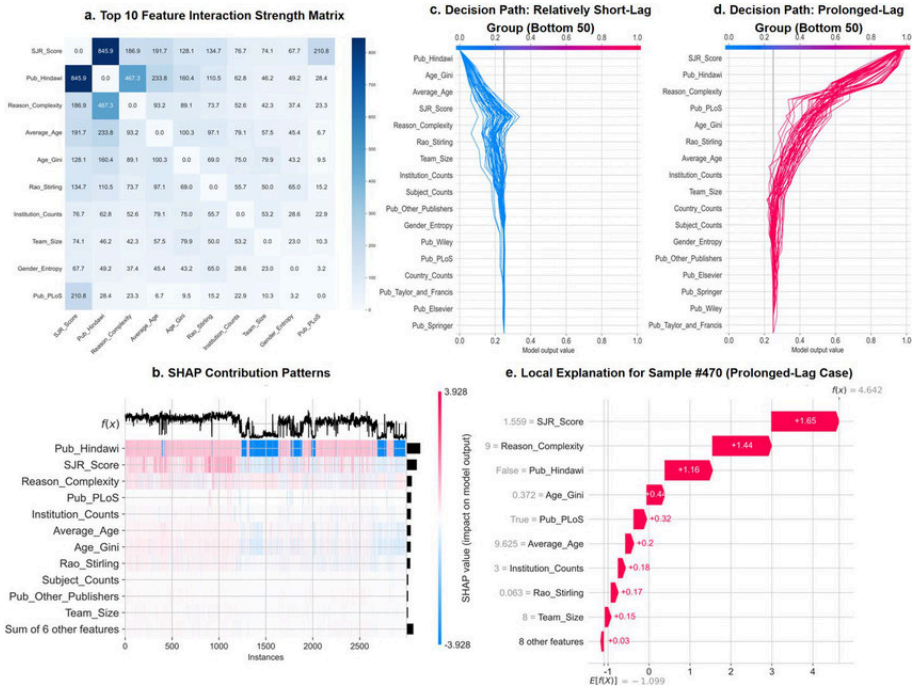



**Fig. 7** Feature interaction strength matrix, decision path, contribution patterns and individual explanation

unstable relationship. This diversity of functional forms underscores the multidimensional and dynamic nature of the mechanisms underlying retraction lag.

Moreover, the lack of a clear marginal pattern for certain variables should not be interpreted as evidence of limited relevance. Some variables remain measured contributors to model predictions but exhibit unstable or interaction-dominated effect patterns that a single dependence plot cannot capture; Sect. "Feature interaction and structural patterns" examines these patterns directly through SHAP interaction and clustering analysis. Consequently, predictive importance and interpretable marginal effects are treated as related but distinct dimensions of feature contribution throughout this study.

## Feature interaction and structural patterns

To uncover the interaction mechanisms among features, this study first conducts a SHAP interaction effect analysis. SHAP interaction values extend the Shapley value framework by quantifying the joint contribution of two features to the prediction, after removing their individual main effects. Specifically, interaction values are computed using the shap_interaction_values function of TreeExplainer. For each sample, an Nfeatures × Nfeatures interaction matrix is generated. The diagonal elements represent the main effects of individual features (i.e., their standalone SHAP values), while the off-diagonal elements (i, j) capture the interaction effects between feature i and feature j, satisfying symmetry. This approach helps identify synergistic or suppressive relationships between features and provides deeper insights into the model's decision-making process.

Figure 7a presents the top 10 feature pairs ranked by SHAP interaction strength. Larger values indicate stronger interaction effects in the model. The interaction between SJR score and the Hindawi publisher indicator reaches 845.9, ranking first among all feature pairs. This indicates that the joint contribution of these two features to model predictions is substantially larger than their individual contributions considered separately. The interaction between the Hindawi publisher indicator and retraction reason complexity (467.3), as well as that between SJR score and reason complexity (186.9), also shows high interaction strength. These results suggest that the contribution of retraction reason complexity to the model varies across different publishing environments and journal impact levels. In addition, interactions such as the Hindawi publisher indicator and average academic age (233.8), the Hindawi publisher indicator and the age Gini coefficient (160.4), and SJR score and Rao-Stirling diversity (134.7) indicate that team and knowledge structure characteristics may have complementary contributions under specific conditions. By contrast, most other feature pairs exhibit relatively low interaction strength, suggesting that their contributions are primarily captured by main effects rather than strong interactions. Overall, the results indicate that the prediction of retraction lag depends mainly on a limited number of key features and their combinations, rather than widespread interactions across the full set of predictors.

To further identify the synergistic structure among features, this study applies SHAP heatmap clustering to perform sample clustering analysis, results are shown in the Fig. 7b. This method performs hierarchical clustering on the sample-by-feature SHAP value matrix, enabling the identification of common feature contribution patterns and similarities among samples.

The results show clear stratification in feature contribution structures. Among all variables, the Hindawi publisher indicator, SJR score, and retraction reason complexity exhibit large variations in SHAP values and strong color contrasts. These features serve as the primary sources of sample differentiation. In some cases, they also display coordinated variation patterns, suggesting potential joint effects on model predictions. Moreover, these three features show partial horizontal alignment in their color patterns, indicating that they may act in combination to influence model outputs for certain groups of samples.

In contrast, average academic age, the academic age Gini coefficient, and the Rao-Stirling diversity, exhibit relatively smooth color gradients. Their SHAP values vary within a narrower range and do not form clear segmented patterns across samples. This suggests that, although these features contribute to model predictions, their marginal role in distinguishing between samples is relatively limited, functioning more as supplementary factors.

Variables such as institution counts, team size, and some publisher indicators show near-zero SHAP values across most samples, appearing as uniformly light colors at the bottom of the heatmap. This indicates that their contributions are generally neutral and have limited influence on model outputs. They also do not form distinct clustering patterns, further confirming their relatively weak discriminative power.

Taken together, the interaction and clustering results clarify why some variables lack a stable marginal pattern in Sect. "Feature main effects". Variables with identifiable main effects, namely SJR score, reason complexity, academic-age Gini, average academic age, and Rao-Stirling diversity, also participate in the strongest pairwise interactions, chiefly with the Hindawi publisher indicator and SJR score, indicating that their influence combines a genuine average direction with meaningful context-dependence on the publishing and oversight environment dimension. By contrast, team size, institution count, and minor publisher categories show consistently limited contribution across the main-effect, interaction, and clustering analyses, indicating that their weak marginal patterns reflect low overall

relevance rather than concealed interaction effects. Gender entropy shows a comparable pattern of limited and unstable contribution; consistent with this, the continuous-duration robustness analysis in Appendix 2 identifies gender entropy as one of the few predictors whose direction differs from the classification-based SHAP result, suggesting that its role is sensitive to the specific modeling framework rather than reflecting a stable, generalizable effect. We therefore limit our interpretive claims about these variables to their conditional, model-based contribution rather than asserting a universal directional relationship.

### Feature decision path and individual explanation

Building on the identification of group-level structures, this study further employs SHAP decision plots to analyze the decision processes for samples in different retraction lag categories, as illustrated in Fig. 7c-d. The results indicate substantial differences in feature contribution pathways between relatively short and prolonged groups. For relatively short retraction lag samples, model outputs are largely concentrated within a lower value range. The magnitude of SHAP contributions remains relatively small, and the decision paths exhibit a high degree of consistency. Among the features, the Hindawi publisher indicator consistently demonstrates a pronounced negative contribution, representing the feature that most strongly pushes the model's prediction toward relatively short retraction lag. In contrast, the contributions of SJR score and reason complexity remain comparatively limited within this group.

For prolonged retraction lag samples, the structural pattern differs markedly. Model outputs increase progressively through the cumulative effects of multiple features, resulting in more dispersed decision paths and stronger additive contributions. Core variables, including SJR score, the Hindawi publisher indicator, and reason complexity, push the model's cumulative prediction substantially toward prolonged retraction lag at the early stages of the decision process. These effects are subsequently reinforced by team structure and knowledge structure variables, such as average academic age and Rao-Stirling diversity, which further amplify the upward trend in predicted prolonged retraction lag. Overall, the results suggest that elevated prolonged retraction lag arises from the combined influence of multiple key determinants, rather than reflecting the influence of any single dominant feature.

To further elucidate the model's decision-making process at the individual level, a single case was randomly selected from the correctly classified prolonged retraction lag samples identified by the final LightGBM model. The corresponding SHAP waterfall plot is presented in Fig. 7e. This case is provided solely for illustrative purposes, demonstrating how the model aggregates and weighs feature contributions to generate a prolonged retraction lag prediction for an individual observation. It is important to note that this example is not intended to be statistically representative of the broader sample, and no inferences regarding population-level patterns or model behavior should be drawn from this single case alone.

Among them, SJR score and reason complexity emerge as the features that most strongly push the prediction toward prolonged retraction lag, followed by the combined effects of variables such as the Hindawi publisher indicator and academic age Gini coefficient. Although several features exhibit negative contributions, their magnitudes remain relatively small and exert limited influence on the overall prediction. This case illustrates that, under prolonged retraction lag conditions, model predictions are typically shaped by the synergistic amplification of multiple key variables, rather than being triggered by a

single extreme feature. This finding is consistent with the results obtained from the global and group-level analyses.

## Conclusions

This study integrates machine learning prediction with the SHAP interpretability framework to uncover the multidimensional drivers and structural characteristics underlying retraction lag. The findings provide new empirical evidence for understanding institutional differences in the scientific correction process, and offer methodological support and decision references for academic publishing governance.

The results suggest that prolonged retraction lag is associated with the combined influence of publishing environment, team structure, and content-related characteristics rather than any single factor. Among these dimensions, publishing-environment variables consistently exhibit substantial predictive importance. Importantly, the influence of publisher-related variables should not be interpreted as a direct assessment of publishing environment or performance. Instead, these variables capture differences in the institutional contexts in which problematic papers are published, scrutinized, and corrected.

The prominent contribution of the Hindawi variable, together with the comparatively weaker effects observed for Springer, Elsevier, and several other major publishers, suggests that retraction lag is not uniformly distributed across publishing environments. Rather than reflecting superior correction mechanisms, relatively short retraction lags may also emerge in publishing contexts that experience elevated levels of post-publication scrutiny or large-scale integrity investigations. Therefore, publisher-related effects are better understood as indicators of heterogeneous publication and oversight environments than as evidence of better or worse publisher performance. Collectively, these findings highlight the important role of institutional publishing contexts in understanding variation in retraction timeliness across retracted papers.

SJR score and reason complexity serve as important complementary dimensions. SJR score demonstrates a nonlinear, interval-dependent effect, with higher-score journals generally associated with prolonged retraction lag. Reason complexity shows a similar stage-dependent pattern, increasing the predicted risk of prolonged retraction lag within a moderate range, while this effect weakens and may reverse at the most extreme values. The academic age Gini coefficient follows an inverted U-shaped pattern. Moderate levels of seniority disparity are associated with prolonged retraction lag, whereas this effect diminishes under highly unequal conditions. The average academic age shows an initial increase followed by stabilization, with younger teams more likely to be associated with relatively short retraction lag.

Interaction analysis further reveals that institutional features of the publishing and oversight environment not only exert independent effects on retraction duration, but also shape the influence of other structural factors. In particular, the interaction structure formed by publisher, SJR score, and reason complexity plays a dominant role. The interaction between Hindawi and SJR score ranks highest among all feature pairs, indicating a strong synergistic effect. Similarly, the interaction between Hindawi and retraction reason complexity is highly pronounced, suggesting that the influence of retraction reasons is moderated by the publishing context. SHAP clustering and decision path analyses further indicate that prolonged retraction lag samples reflect the cumulative effects of multiple core variables,

whereas relatively short lag samples are characterized by a more consistent and relatively simple pattern of negative contributions.

The practical implications of this study should be interpreted within the boundaries of its sampling design. Because the dataset contains only papers that were ultimately retracted, the framework estimates variation in prolonged retraction lag conditional on eventual retraction. It does not estimate the probability that an arbitrary published paper is problematic, will be retracted, or will remain unretracted for a prolonged period. Accordingly, the model should not be used as an early-warning or screening system for the broader population of published papers. Within this boundary, the findings provide retrospective evidence for research-integrity governance. First, the identified publishing-environment patterns may support historical benchmarking of completed retraction processes and help generate hypotheses about institutional or procedural sources of delay. Second, the associations involving investigation complexity, collaboration structure, and disciplinary breadth may inform the design of clearer responsibility-allocation and cross-institutional coordination procedures. Third, the results may help publishers and integrity bodies identify recurring process characteristics associated with long completed cases, although the model has not been validated for prospective operational deployment.

More concretely, two influential predictors, reason complexity and publisher, illustrate both the relevance and boundaries of these findings. Because reason complexity is fully observable only after retraction, it should not be interpreted as an ex ante screening signal for individual manuscripts. Its practical value instead lies in retrospective process design: cases with higher reason complexity involved multiple concerns, suggesting that journals and research integrity offices may benefit from standardized, multi track investigation protocols, such as parallel rather than sequential review of data, authorship, and ethics related concerns when multiple allegations are identified. This would not predict which papers will ultimately be retracted, but could potentially improve the efficiency of investigations once multiple concerns emerge. For the publisher variable, our findings permit two non exclusive interpretations that cannot be distinguished by the present data: (1) publishers associated with shorter retraction lags, such as Hindawi in our sample, may have more streamlined retraction workflows, suggesting the value of sustained post publication monitoring rather than one time correction efforts; or (2) longer retraction lags may reflect more cautious, evidence intensive investigations. In the latter case, longer lag should not necessarily be interpreted as poorer publisher performance, but may indicate greater verification requirements and greater cumulative citation exposure while cases remain unresolved, warranting more prominent interim editorial notices, such as expressions of concern, where appropriate. We therefore suggest that publisher level associations be used primarily to inform internal process evaluation and benchmarking against peer publishers, rather than to rank publisher quality.

Several limitations should be acknowledged. First, the sample only includes retracted papers and therefore cannot explain why all problematic papers persist in the literature for extended periods. Second, the three-year threshold for defining long-lag retractions is a modeling choice; robustness checks using alternative thresholds, an alternative estimation sample, and a continuous-duration reformulation (Appendices A and B) consistently reproduced the importance and direction of the leading predictors, but some information loss relative to the full continuous distribution of retraction lag is inevitable. Third, SHAP values reflect predictive contributions rather than causal effects. Fourth, the publisher variable may confound multiple factors, including publication volume, retraction events, scrutiny intensity, and portfolio composition. Fifth, reason complexity may be a post-hoc variable and should not be simplistically applied for early warning purposes. Sixth, gender inference

based on names is subject to measurement errors and biases. Finally, the results are primarily derived from biomedical retracted papers; caution is warranted when generalizing the findings to other disciplinary contexts.

## Appendix

### Binary threshold and variable exclusion robustness analyses

To assess the sensitivity of the results to the choice of classification threshold, additional analyses were conducted using alternative cutoffs of one year and five years (Figs. 8 , 9). As expected, some variation in the relative ranking of individual features was observed across threshold specifications, reflecting differences in how prolonged retraction lag is defined. Nevertheless, the overall explanatory structure remained highly consistent. Key predictors from the publishing and oversight environment dimension (e.g., Pub_Hindawi and SJR score), the investigation-complexity dimension (retraction reason complexity), and the team structure dimension (e.g., academic-age Gini), consistently ranked among the most influential features across all models. These results suggest that while the relative importance of specific variables may vary under different operational definitions of retraction lag, the principal conceptual dimensions associated with prolonged retraction delays remain stable. Therefore, the main findings of this study are unlikely to be driven by the specific choice of the three-year threshold.

In addition, given the prominent contribution of the Pub_Hindawi variable in the primary model, an additional robustness analysis was conducted after removing this publisher indicator (Fig. 10). The results show that predictors from the investigation-complexity, publishing and oversight environment, team structure, and knowledge and disciplinary structure dimensions, reason complexity, SJR score, academic-age Gini, and Rao-Stirling

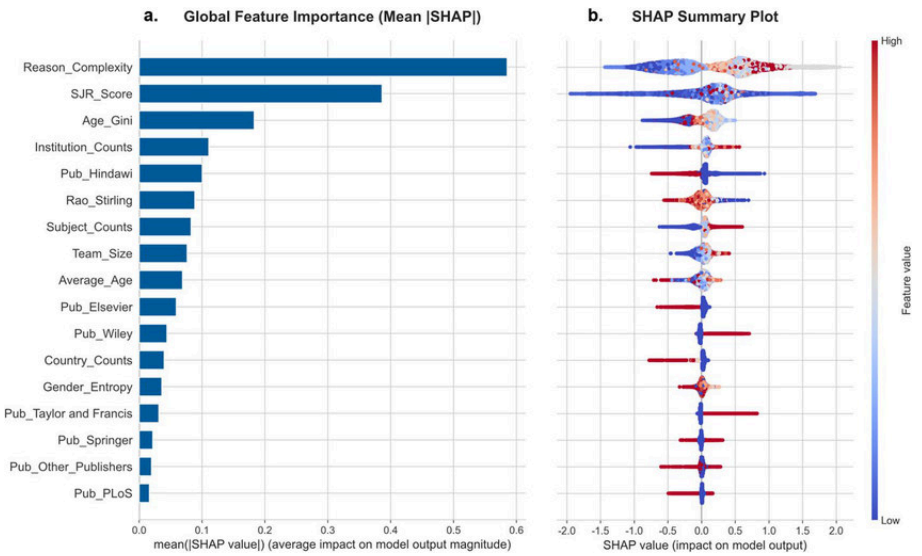



**Fig. 8** Feature importance with retraction lag 1-years

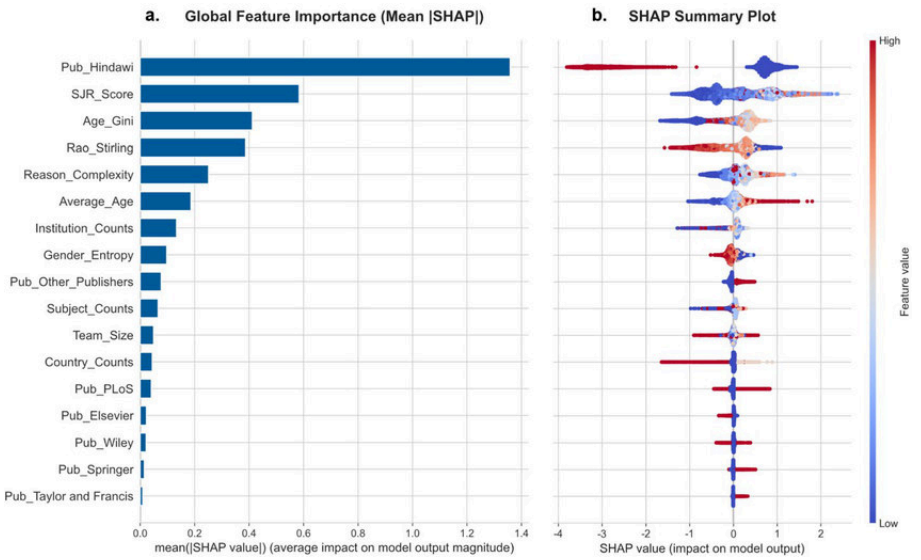



**Fig. 9** Feature importance with retraction lag 5-years

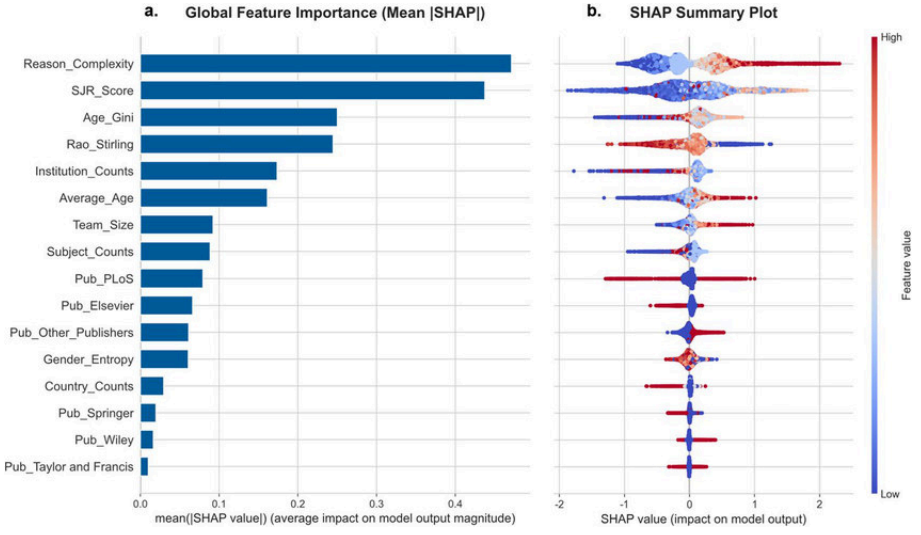



**Fig. 10** Feature importance after deleting Hindawi publisher variable

diversity, remained among the most influential predictors, and the overall ranking structure changed only moderately. This finding suggests that the explanatory framework is not solely dependent on the Hindawi effect and that the four conceptual dimensions identified in Sect. "Predictive variables" remain jointly relevant after excluding the most influential publisher-related variable.

**Table 3** Lognormal AFT model estimates for retraction lag

| Feature | Coef. (ln T) | Time Ratio exp(β) | *p*-value | SHAP concordant |
|---|---|---|---|---|
| Pub_Hindawi | −0.954 | 0.39 | <.001 | Yes |
| Reason_Complexity | 0.095 | 1.10 | <.001 | Yes |
| Pub_Elsevier | −0.391 | 0.68 | <.001 | Yes |
| Average_age | 0.019 | 1.02 | <.001 | Yes |
| SJR_Score | 0.031 | 1.03 | <.001 | Yes |
| Subject_Counts | 0.059 | 1.06 | <.001 | No |
| Rao_Stirling | −5.916 | 0.003 | <.001 | Yes |
| Country_Counts | −0.091 | 0.91 | <.001 | Yes |
| Team_Size | 0.016 | 1.02 | <.001 | Yes |
| Pub_PLoS | 0.118 | 1.13 | .004 | No |
| Pub_Springer | −0.049 | 0.95 | .119 (n.s.) | No |
| Age_Gini | 0.050 | 1.05 | .235 (n.s.) | Yes |
| Pub_Wiley | 0.036 | 1.04 | .315 (n.s.) | Yes |
| Gender_Entropy | 0.013 | 1.01 | .729 (n.s.) | No |
| Institution_Counts | −0.001 | 1.00 | .886 (n.s.) | Yes |
| Pub_Taylor and Francis | 0.007 | 1.01 | .859 (n.s.) | Yes |

Table 3 reports coefficients from the lognormal AFT model, estimated as OLS of log-transformed retraction lag with HC3 robust standard errors. Other Publishers is used as the reference category. A time ratio below 1 indicates a shorter expected retraction lag relative to the reference, whereas a ratio above 1 indicates a longer expected lag. "SHAP concordant" indicates whether the coefficient direction is broadly consistent with the dominant direction observed for the corresponding feature in the classification-based SHAP analysis in Sect. "Feature main effects". Because several SHAP relationships are nonlinear, concordance should be interpreted as directional consistency rather than equivalence between the two models. Rows are ordered by the magnitude of the standardized effect used for the rank comparison reported below

## Continuous duration robustness analysis

To assess whether the main findings are sensitive to the binarization of retraction lag, we estimated a continuous-duration model using the same sample and predictor set. Because all papers in the fully retracted sample have observed retraction dates and are therefore uncensored, a lognormal accelerated failure time (AFT) model can be estimated as an ordinary least squares regression of log-transformed retraction lag. The lognormal specification was adopted because of its tractable formulation for the observed duration distribution. Because the seven publisher categories are mutually exclusive and jointly exhaustive, including all seven categories together with an intercept would result in perfect multicollinearity. We therefore used Other Publishers as the reference category, consistent with the publisher classification in Sect. "Predictive variables". Diagnostic checks indicated no material multicollinearity among the sixteen predictors, and the coefficient estimates were consistent across the two estimation procedures used.

$$ln(T_i) = \beta_0 + \sum_k \beta_k X_{ik} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \tag{8}$$

where $T_i$ is the retraction lag (in days) for paper i, $X_{ik}$ is the value of predictor k (the sixteen features listed in Table3) for paper i, $\beta_0$ is the intercept, $\beta_k$ is the coefficient of predictor k, and $\varepsilon_i$ is the normally distributed error term.

Consistent with the publishing and oversight environment dimension, publisher-related variables show some of the strongest associations with continuous retraction lag. Relative to Other Publishers, papers published by Hindawi and Elsevier are associated with relatively shorter retraction lags, whereas papers published by PLoS are associated with prolonged retraction lags (Table 3). The particularly large negative coefficient for Hindawi is broadly consistent with its prominent contribution in the SHAP analysis, although the two estimates capture different aspects of model behavior. These results therefore provide additional evidence that the publishing environment is an important component of variation in retraction lag.

The continuous-duration results also show broad directional consistency with the classification-based analysis. Twelve of the sixteen predictors have coefficient directions that are broadly consistent with the dominant SHAP directions, while the remaining four show discordant directions. The standardized effect magnitudes are also correlated with SHAP-based importance rankings (Spearman $\rho$=0.556, $p$=.025). The four discordant predictors, namely subject count, gender entropy, PLoS, and Springer, have comparatively limited or framework-sensitive contributions and therefore do not materially alter the interpretation of the main findings. In particular, the difference observed for gender entropy is consistent with its relatively weak and unstable contribution in the classification analysis and suggests that its interpretation may depend on the outcome specification.

It is important to note that the continuous-duration model provides a complementary robustness assessment rather than an equivalent test of the classification results. The two approaches address related but distinct outcomes: the classification model estimates the likelihood that retraction lag exceeds the three-year threshold, whereas the continuous-duration model examines variation in the log-transformed observed lag among papers that were ultimately retracted. Agreement in coefficient directions and feature rankings therefore provides supporting evidence that the main patterns are not confined to the binary outcome specification, while differences for individual predictors are expected given the distinct estimands and functional forms.

Taken together, the threshold sensitivity analyses and the Hindawi-exclusion analysis in Appendix 1, together with the continuous-duration analysis in Appendix 2, provide complementary evidence for the robustness of the main findings. Across alternative thresholds, an alternative sample specification, and a continuous outcome formulation, the publishing and oversight environment remains an important dimension, while several team and knowledge structure variables retain broadly similar contributions. At the same time, some individual predictors vary in direction or relative importance across specifications, underscoring that their interpretation should remain conditional on the modeling framework. Overall, these robustness analyses support the stability of the study's broader explanatory structure without implying that every individual predictor exhibits an invariant effect across alternative specifications.

**Acknowledgements** This research was supported by the National Social Science Foundation of China (No. 25ATQ008), the National Natural Science Foundation of China (No. 724B2015), and the Nanjing University International Collaboration Initiative.

**Author contribution** Zhaoqi Peng: Writing-review and editing, Writing-original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Mengfan Zhao: Writing-review and editing, Supervision, Methodology, Data curation, Conceptualization. Hao Wang: Writing-review and

editing, Validation, Formal analysis. Alex Jie Yang: Writing-review and editing, Supervision, Methodology. Sanhong Deng: Writing-review and editing, Writing-original draft, Supervision, Methodology, Data curation, Conceptualization.

**Funding** National Social Science Foundation of China, No. 25ATQ008; National Natural Science Foundation of China, No. 724B2015.

**Data availability** No datasets were generated or analysed during the current study.

## References

Abalkina, A., Aquarius, R., Bik, E., et al. (2025). 'Stamp out paper mills': Science sleuths on how to fight fake research. *Nature, 637*(8048), 1047–1050.

Akella, A. P., Alhoori, H., Kondamudi, P. R., et al. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics, 15*(2), Article 101128.

Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2021). An in-text citation classification predictive model for a scholarly search system. *Scientometrics, 126*(7), 5509–5529.

Amiri, M., Yaghtin, M., & Sotudeh, H. (2024). How do tweeters feel about scientific misinformation: An infoveillance sentiment analysis of tweets on retraction notices and retracted papers. *Scientometrics, 129*(1), 261–287.

Andersen, L. E., & Wray, K. B. (2019). Detecting errors that result in retractions. *Social Studies of Science, 49*(6), 942–954.

Azoulay, P., Bonatti, A., & Krieger, J. L. (2017). The career effects of scandal: Evidence from scientific retractions. *Research Policy, 46*(9), 1552–1569.

Bar-Ilan, J., & Halevi, G. (2018). Temporal characteristics of retracted articles. *Scientometrics, 116*(3), 1771–1783.

Brainard, J. (2018). Rethinking retractions. *Science, 362*(6413), 390–393.

Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Sources of error in the retracted scientific literature. *FASEB Journal, 28*(9), 3847–3855.

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics, 126*(7), 5431–5476.

De Cassai, A., Geraldini, F., De Pinto, S., et al. (2022). Inappropriate citation of retracted articles in anesthesiology and intensive care medicine publications. *Anesthesiology, 137*(3), 341–350.

Dong, S., Mao, J., Ke, Q., et al. (2024). Decoding the writing styles of disciplines: A large-scale quantitative analysis. *Information Processing & Management, 61*(4), Article 103718.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *PNAS, 109*(42), 17028–17033.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 1189–1232.

Ghorbi, A., Fazeli-Varzaneh, M., Ghaderi-Azad, E., et al. (2021). Retracted papers by Iranian authors: Causes, journals, time lags, affiliations, collaborations. *Scientometrics, 126*(9), 7351–7371.

Goldstein, A., Kapelner, A., Bleich, J., et al. (2015). Peeking inside the black box: Visualizing statistical learning with ICE plots. *Journal of Computational and Graphical Statistics, 24*(1), 44–65.

Guidotti, R., Monreale, A., Ruggieri, S., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.

Haunschild, R., & Bornmann, L. (2021). Can tweets detect problems early in scientific papers? A case study of three retracted COVID-19/SARS-CoV-2 papers. *Scientometrics, 126*(6), 5181–5199.

Huang, S., Huang, Y., Bu, Y., et al. (2022). Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Information Processing & Management, 59*(2), Article 102799.

Hwang, S. Y., Yon, D. K., Lee, S. W., et al. (2023). Causes for retraction in the biomedical literature. *Journal of Korean Medical Science, 38*(41), Article e333.

Ioannidis, J. P. A., Pezzullo, A. M., Cristiano, A., et al. (2025). Linking citation and retraction data reveals the demographics of scientific retractions among highly cited authors. *PLoS Biology, 23*(1), Article e3002999.

Joëts, M., & Mignon, V. (2026). Slaying the undead: How long does it take to kill zombie papers? *Research Policy, 55*(2), Article 105401.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016, April). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In Proceedings of the 25th International conference companion on World Wide Web (pp. 53-54)

Khan, H., Gupta, P., Zimba, O., et al. (2022). Bibliometric and altmetric analysis of retracted COVID-19 articles. *Journal of Korean Medical Science, 37*(6), Article e44.

Kwon, D. (2024). AI-generated images threaten science: Here's how researchers hope to spot them. *Nature, 635*, 268–269.

Lendvai, G. F., & Sasvári, P. (2025). 'Wasted'research and lost citations: A scientometric assessment of retracted documents in Scopus between 2001 and 2024. *Journal of Information Science, 52*(3), 666–683.

Li, L., Wu, Y., & Wang, Y. (2025). Exposure of academic misconduct and universities' innovation output: Evidence from retractions in China. *Journal of Business Ethics, 200*(3), 649–668.

Liu, J., Wang, X., & Liang, X. (2025). Bibliometric feature identification and analysis of retracted papers in biomedicine: An interpretable machine learning perspective. *Information Processing & Management, 62*(5), Article 104176.

Liu, Z., & Hu, S. (2024). Predicting the fundraising performance of environmental crowdfunding projects: An interpretable machine learning approach. *Information Processing & Management, 61*(2), Article 103587.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Madhugiri, V. S., Nagella, A. B., & Uppar, A. M. (2021). An analysis of retractions in neurosurgery and allied clinical and basic science specialties. *Acta Neurochirurgica, 163*(1), 19–30.

Malkov, D., Yaqub, O., & Siepel, J. (2023). The spread of retracted research into policy literature. *Quantitative Science Studies, 4*(1), 68–90.

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13*(3), e1493.

Memon, S. A., Makovi, K., & Alshebli, B. (2025). Characterizing the effect of retractions on publishing careers. *Nature Human Behaviour, 9*, 1134–1146.

Mhamdi, R. (2026). The retraction problem in collaborations with Saudi Arabia: Evidence from 2014–2023. *Scientometrics, 131*(1), 425–441.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Mueller, J., & Stumme, G. (2016, August). Gender inference using statistical name characteristics in twitter. In *Proceedings of the the 3rd multidisciplinary international social networks conference on socialinformatics 2016, data science 2016* (pp. 1-8).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Szilagyi, I. S., Schittek, G. A., Klivinyi, C., et al. (2022). Citation of retracted research: a case-controlled, ten-year follow-up scientometric analysis of Scott S. Reuben's malpractice. *Scientometrics, 127*(5), 2611–2620.

Thorp, H. H. (2023). Generative approach to research integrity. *Science, 381*(6658), 587.

Walsh, J. P., Lee, Y. N., & Tang, L. (2019). Pathogenic organization in science: Division of labor and retractions. *Research Policy, 48*(2), 444–461.

Wang, D., & Chen, S. (2025). An empirical study of retractions due to honest errors: Exploring the relationship between error types and author teams. *Journal of Informetrics, 19*(1), Article 101600.

Xu, H., Ding, Y., Zhang, C., et al. (2023). Too official to be effective: An empirical examination of unofficial information channel and continued use of retracted articles. *Research Policy, 52*(7), Article 104815.

Xu, J., Yu, C., Xu, J., et al. (2025). Pubmed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *Scientific Data, 12*(1), 1018.

Yao, L., & Ni, H. (2023). Prediction of patent grant and interpreting the key determinants: an application of interpretable machine learning approach. *Scientometrics, 128*(9), 4933–4969.

Zhang, Q., Abraham, J., & Fu, H. Z. (2020). Collaboration and its influence on retraction based on retracted publications during 1978 to 2017. *Scientometrics, 125*(1), 213–232.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.