MSTS Methods in Science and Technology Studies

futuristech PT. Teknologi Futuristik Indonesia

Received: July 15, 2026; Revised: September 8, 2026; Accepted: September 10, 2026; Published: September 12, 2026; Current version: September 12, 2026. e-ISSN: 3123-4232
Digital Object Identifier 10.64539/msts.v2i2.2026.570

Article

A Machine-Learning Model for Identifying Non-Credible Scholarly Journals

Fouad Abdulameer Salman[1,*], Bakhtawar Baluch[2], Zuriana Abu Bakar[3], Assane Lo[2]

[1] Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad 10070, Iraq; e-mail: fouad.a@coadec.uobaghdad.edu.iq (F. A. Salman).
[2] School of Engineering, University of Wollongong in Dubai, Dubai 20183, United Arab Emirates; e-mail: bakhtawarbaluch@uowdubai.ac.ae (B. Baluch), assanelo@uowdubai.ac.ae (A. Lo).
[3] Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus 21300, Malaysia; e-mail: zuriana@umt.edu.my (Z. A. Bakar).
* Correspondence Author

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Abstract: The rapid growth of scholarly publishing has increased the need of reliably identify credible and non-credible journal websites. Evaluating these entities requires close, cautious, thorough, and at times skeptical scrutiny by researchers, institutions, and funding agencies. However, existing automated detection approaches appear to be customized for specific aspects of the criteria and do not effectively detect non-credible entities. Therefore, in this paper, we come up with a comprehensive dataset comprising 4,174 manually verified journal websites. We classify the dataset into two categories: credible journals and non-credible journals, including predatory, commercial, and hijacked journals. We also represent 36 new features derived from different resources of the journal websites. We used these new features to train and test the proposed model to improve accuracy. XGBoost achieved a precision of 0.9778, recall of 0.9601, F1-score of 0.9689, and accuracy of 97.86%, outperforming both GBC and Random Forest in the experimental results. Feature-group experiments show that Web-Based System and Academic Integrity features substantially improve overall classification performance, while HTML-based characteristics provide strong discriminatory power. The proposed approach shows the potential of integrating technical, structural, and scholarly-verification signals to support automated journal-credibility assessment and give researchers an efficient way to identify potentially non-credible scholarly venues.

Keywords: Predatory journals; Legitimate journals; Journal credibility assessment; Machine learning; XGBoost.

1. Introduction

The rapid expansion of scholarly publishing has created a practical need to distinguish credible journals from journal websites associated with predatory, commercial, or hijacked activity. These activities or categories fall under the broader term non-credible journals. Predatory refers to the extent to which an organization exploits academics and researchers for profit through high processing charges to provide open access to their articles, rather than meeting publishing standards [1].

Predatory publishers continuously email large numbers of people as part of their standard procedure. Moreover, they accept a large volume of manuscripts and favour fast-track publication for fees, while credible journals charge much lower fees. In some cases, they ask for publication fees and then collect them, but in the end those articles don't appear on the website as promised. Conversely, a few people consciously publish articles in those journals that don't follow standard procedures and are not "prey" [2]. Apparently, these non-credible journals claim to follow the rules and regulations of bona fide journals and meet peer-review standards. Still, in practice, they don't comply with their own rules and regulations and accept articles without adhering to the peer-review standard.

For this, reliable organizations, such as the Commit-

tee on Publication Ethics (COPE), the International Committee of Medical Journal Editors (ICMJE), or the World Association of Medical Editors (WAME), sought to figure out standard policies, such as documenting the journal content, management of potential conflicts of interest, handling of errata, and transparency of journal processes and policies, including fees [3]. Also, the Iraqi Higher Education Ministry is not behind this issue, which obliges it to adopt Beall's lists as guidance for researchers in accrediting scholarly publishers and journals. It also classified the black list of journals into three categories: hijacked, predatory, and commercial journals [4].

However, the journals listed as hijacked are fake websites that imitate credible journals and request article submissions to collect article fees from researchers deceived into believing the journal is legitimate. Commercial journals are also predatory entities, but the Iraqi Higher Education Ministry defines them according of its criteria [4]. Yet researchers, authors, peer reviewers, and editors must also identify non-credible journals. Therefore, any research work that is not thoroughly examined should not be included in the registry. Therefore, this article proposes a machine-learning model for *i*-faking to automatically identify non-credible scholarly entities using sound logic and multiple artificial-intelligence approaches of prevent shortcomings.

This paper is categorized as Follows: Section 2 covers the review of the detection approaches for scholarly non-credible entities, and Section 3 elucidates the dataset construction. Section 4 proposes the methodology of the detection model. Section 5 presents the empirical results of the proposed model. Section 6 presents the limitations of this work. Finally, Section 7 covers the conclusion and future directions.

2. Detecting Approaches of the Scholarly Non-Credible Entities

In an attempt to help scholars identify non-credible venues among credible ones, Jeffrey Beall was the first to publish the blacklist of deceptive open access (OA) journals back in 2010 [5]. Afterwards, many other authors listed it, and the dubious journals list was updated. When the list was published back in 2017 [5], Beall faced continued harassment and threats, leading him to take it down. Some unknown scholars published a list containing 1462 dubious journals, 152 of which were added since Beall's original version [5]. This situation opens the door for scholars to advance journal criteria to uncover non-credible venues. Still, there is a gap in its criteria. As a result, the lists are often criticized for their overlap [6].

Scholars have attempted to rectify the differences between blacklists and white-listed journals by creating frameworks to assist researchers and funders in detecting non-credible venues. These efforts lie in two parts. One is for "frameworks," and the other is for "lists" [6]. A major part of "frameworks" depends on the standards or principles [7]. A few of them convert these criteria into metrics or measurements [8], while algorithms are used for the rest [9]. Though these "frameworks" lack information on their development, validation, and reliability, they have been criticized [7]. For this reason, few researchers have interpreted the existing frameworks and guidelines [10]. Furthermore, a few researchers try their best to solve the problem automatically instead of using non-automated approaches [11]. To identify non-credible journals, Laine and Winker [12] designed a procedure based entirely on Beall's standards and procedures. Directory of Open Access Journals (DOAJ) and the "Think. Check. Submit" initiative (see Figure 1).

They redesigned their existing algorithm because existing approaches are not fully reliable. Their mechanism combines all the assessed criteria, requiring researchers to count violations manually. In a related effort, McCann and Polacsek [13] reviewed 28 articles from 2018 on non-credible open-access publishing in nursing and midwifery. From these, they constructed a framework with criteria to guide nursing researchers, which includes verifying OA journal and publisher integrity in DOAJ and OASPA, as well as evaluating the venue's website content, title, and publication history. The non-credible rate metric, derived from 14 of Beall's criteria, categorizes questionable venues as non-credible (exhibiting predatory practices) or credible, based on a weighted sum of violations. Their case study illustrated the metric, though the rationale for the criteria selection was not provided. Reference [14] criticized this approach for mixing quality indicators with self-sufficient criteria. Multiple studies have since analyzed such frameworks, highlighting overlapping criteria and the risk of false positives (labeling the authorization of non-credible journals) and false negatives (the reverse). Olivarez *et al*. [15] assessed Beall's list with an independent three-person panel, arguing it has a preference for non-OA journals as compared to OA journals. Moreover, they used Thomson Reuters' Web of Science JCR to collect data on various journals. They assessed the journals that were part of 80 non-OA and the seven OA—respected journals in information and library science with a median age of 32 years—and discovered that both types could be considered non-credible by Beall's standards. They concluded that Beall's criteria are subjective and biased against OA journals. Thereafter, a study by da Silva *et al*. [16] examines various terms such as sensitivity, specificity, likelihood ratio, and pervasiveness to evaluate Cabell, Beall, and Crawford's criteria. The examination part discloses the fact that there is a substantial rate of positiveness in Beall's list. In contrast, the blacklist by Cabell's and the OA list by Crawford's gray disclose a limited (fewer false) positive rate.

The above-mentioned analyses reviewed the framework for the limited number of studies based on au-

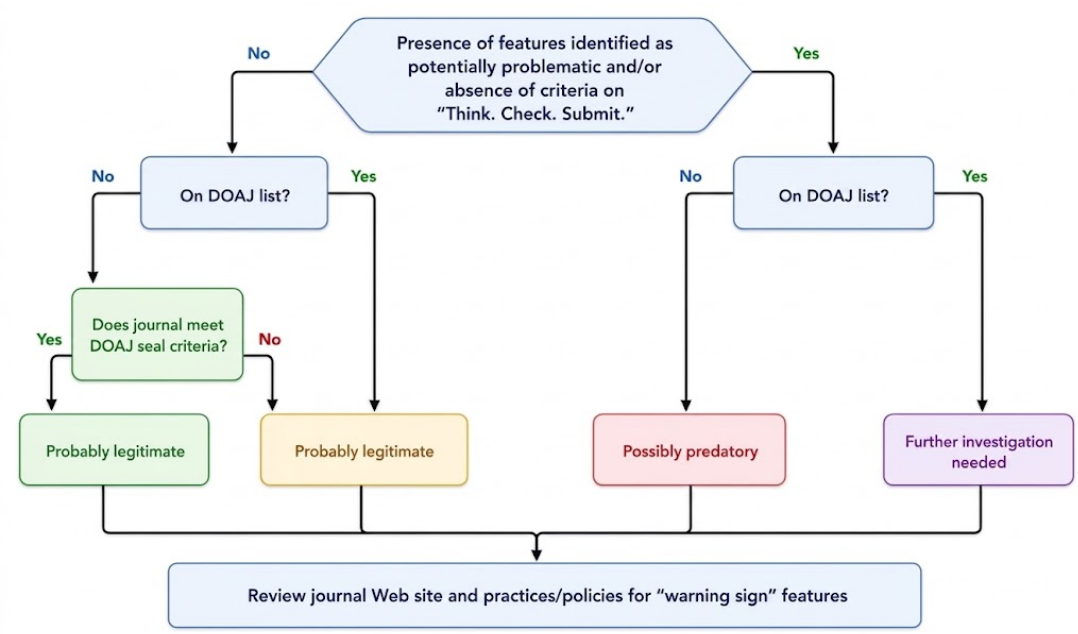

Figure 1. Predatory journals algorithm.

tomatic detection. In contrast, most non-credible venues rely on manual detection. Moreover, out of the three automatic detection methods, two are based on non-credible venues and one on predatory venues. To further investigate automatic detection, Shahri *et al*. [17] explored various decision tree algorithms to identify non-credible venues. They trained these algorithms on 104 journal websites, out of which 45 journals lie in the category of non-credible, whereas 59 lie in the authorized category. The best performance was achieved by employing the algorithm named Random Tree, which provided a 0% error. When evaluating this model on 10 journal websites, it reported an error rate of 10%. For identifying the non-credible journals, nine characteristics were employed, those are age of the domain, the rank of domain in search engines, the journal website of countries that are included, objective and coverage, the number of invalid links, volume of articles in a year that are published, uniformity among the country journals and its server, volume of invalid links, and the character "-" that are used in Uniform Resource Locator (URL).

In addition, another study by Dadkhah *et al*. [18] tested different decision-tree algorithms to identify journals in the non-credible category. A total of 84 journals were checked, of which 56 belong to the hijacked category and the remaining 28 are authorized journals. They manually retrieved 12 characteristics to identify hijacked journals and achieved a 2.53% rate, demonstrating strong performance when using the random tree. The illustrative tree consisted of five characteristics: full-text accessibility, researchers' country, level in browsing results, domain lifespan, and accessibility to past issues. Moreover, they performed analyses on the journals by employing the tree algorithms, out of which 10 were hijacked, and 15 were authorized, yielding an error rate of 12%.

A study by Adnan *et al*. [19] conducted on non-credible venues by employing automatic detection, in which the problem is defined categorically and implements three conventional machine-learning techniques: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB). A total of 200 journals are collected from Beall's and DOAJ lists, out of which 100 are non-credible, and the remaining 100 are authorized journals. Characteristics of two classes were utilized: heuristics-based (indicators such as email domains or editorial board transparency, manually selected from Beall's list) and text-based (words and phrases quantified via TF-IDF after preprocessing). Experiments with both types used information gain to determine the optimal feature set size for each classifier, and results were validated using 10-fold cross-validation. SVM achieved the best results for its classes. For heuristic characteristics, it achieved 98% (F1-score), and for the text-based characteristics, it achieved 96%. The approach meets Beall's criterion, which, as previously noted, has been criticized as subjective and error-prone. It stands to reason, however, that all approaches in the literature appear to be customized for specific aspects of the criteria. The issue is that these approaches did not effectively detect non-credible entities. Therefore, utilizing model-based comprehensive features will be more successful.

3. Dataset

In this section, we present the process dataset, including its content and structure.

3.1. Construction

The main objectives of this paper are, first, to propose a binary class model that is insightful for automatically distinguishing between credible Journals and non-

Table 1. The proportion of the dataset's categories.

Category	Non-Credible Journals			Credible Journals (Q1)
	Hijacked Journals n (%)	Predatory Journals n (%)	Commercial Journals n (%)	Journals n (%)
Computer Science	31 (24.2%)	179 (11.7%)	68 (17.2%)	160 (7.5%)
Engineering	32 (25.0%)	351 (22.9%)	69 (17.4%)	397 (18.7%)
Medical & Health	12 (9.4%)	220 (14.4%)	37 (9.3%)	512 (24.2%)
Business & Management	19 (14.8%)	135 (8.8%)	45 (11.4%)	67 (3.2%)
Social Sciences	8 (6.2%)	177 (11.6%)	44 (11.1%)	37 (1.7%)
Education	4 (3.1%)	86 (5.6%)	13 (3.3%)	46 (2.2%)
Arts & Humanities	7 (5.5%)	155 (10.1%)	27 (6.8%)	251 (11.8%)
Environmental Science	3 (2.3%)	49 (3.2%)	19 (4.8%)	31 (1.5%)
Mathematics	3 (2.3%)	43 (2.8%)	18 (4.5%)	127 (6.0%)
Physics & Chemistry	2 (1.6%)	37 (2.4%)	27 (6.8%)	188 (8.9%)
Agriculture	2 (1.6%)	31 (2.0%)	15 (3.8%)	270 (12.7%)
Multidisciplinary	5 (3.9%)	67 (4.4%)	14 (3.5%)	34 (1.6%)
Total	128 (100%)	1,530 (100%)	396 (100%)	2,120 (100%)
		2,054 (100%)		**2,120 (100%)**

credible journals, accompanied by valid justification; second, to construct an annotated dataset encompassing non-credible entities that cover the perspective of Iraqi institutions; and third, to evaluate the model and measure its effectiveness. Although data on unpredictable publishers and journals are publicly available, they are distributed across multiple websites, which can impede researchers and stakeholders from accessing the necessary information and potentially cause confusion. Conversely, providing a unified dataset facilitates researchers and developers in developing or comparing techniques that rely on otherwise uncollected resources. Therefore, we have developed a standardized, current dataset appropriate for several approaches utilising the extensive collection of resources outlined below.

The dataset comprises data collected from official websites. This process can be challenging because some recent web technologies hinder the extraction of the required information [20]. To establish the operational labels before feature extraction, we sourced a list of entities and their domains from, for example, the SCImago website (limited to Q1 journals). of adopted Q1 journals to build a stable dataset and avoid frequent updates due to journal accreditation changes and cancellations, since Q1 journals have a significantly lower rate of removal from accreditation. The Research and Development Department portal on the official website of the Iraq Higher Education Ministry. These sources provided an updated compilation of commercial and predatory scholarly journals.

Additionally, we obtained entities from Beall's list and other websites recommended by the Iraqi Ministry. The dataset primarily contains entity names, domains, and ISSN values. Ultimately, the dataset comprises more than 4000 entities exactly 4,174 journal websites as targets to support benchmarking of automated journal-credibility assessment (see Table 1).

3.1.1. Class Definitions

Table 1 provides a summary of the two primary classes. The label source is significant, as the classes were not determined exclusively by website features. Credible journals were selected from SCImago Q1. Non-credible journals included predatory and commercial journals from the Iraqi institutional portal, hijacked journals from Beall's List, and other sources recommended by the Ministry. As a result, the two labels are likely to inherit the scope, inclusion criteria, and historical biases present in the source lists. Accordingly, the model should be interpreted as learning the classification defined by these sources, rather than representing an absolute or source-independent measure of journal quality.

3.1.2. Schema of the Dataset

Each record in the dataset represents a journal website and includes the source label, identifiers such as the journal name, domain, and ISSN, and extracted website resources. The six resource groups are characterized by 36 numerical or binary features, as detailed in Table 2: URL (8), SSL Certificate (2), HTTP Headers (6), HTML (8), Web-Based System (7), and Academic Integrity (5). The feature schema is explicit and reproducible, with clear feature names, codes, and extraction descriptions.

3.1.3. Statistical Description of the Dataset

Figure 2 shows that the dataset includes 4,174 journal websites, divided into two main categories: 2,120 credible journals (50.79%) and 2,054 non-credible journals (49.21%). The dataset also contains 1,530 predatory journals (36.66%), 396 commercial journals (9.49%), and 128 hijacked journals (3.07%). The percentages presented in Table 1 are calculated for each class individually and do not represent percentages of the entire dataset. The descriptive statistics provide essential context for interpreting aggregate accuracy and F1-score. In addition to class

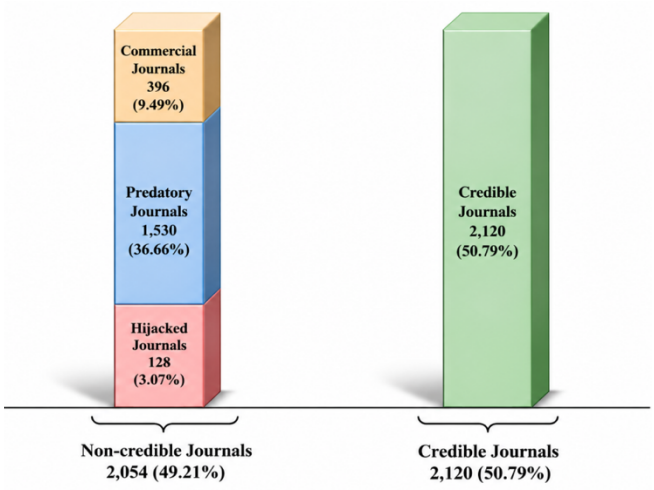

Figure 2. Distribution of journal websites in the dataset.

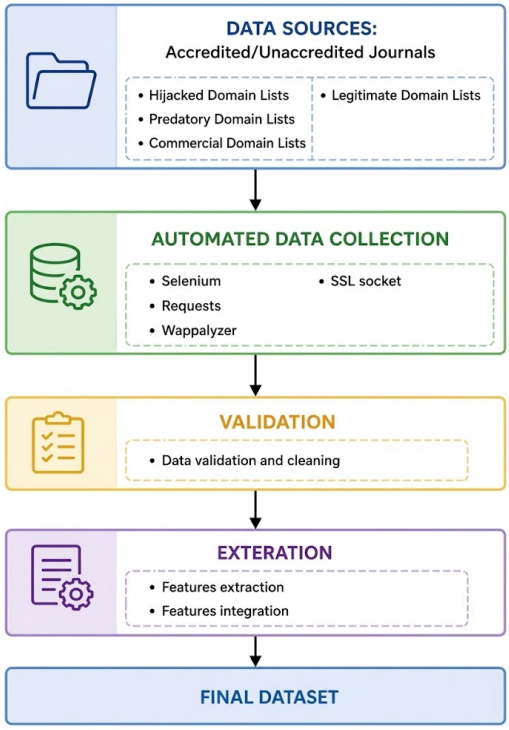

Figure 3. The collection of the data steps.

distribution, the dataset includes 12 subject-area categories, ensuring representation from a range of scholarly disciplines. Engineering is the largest subject area with 849 instances, followed by Medical & Health with 781. Incorporating disciplinary diversity mitigates the risk that the classification model identifies patterns unique to a single academic field.

3.2. Collected information

The pool of resources is gathered to obtain significant data, such as context markup language (HTML), Uniform Resource Identifier (URI), and screenshot data that other researchers previously adopted. In addition, the resources have been expanded to derive the latest attributes and optimize performance (see Figure 3).

These features work together to provide different types of evidence. URL and HTML features show the words used and the content of the page. Web-Based System and SSL features reveal the technical setup. HTTP Headers highlight security settings, while Academic Integrity features point to publishing standards and scholarly checks. The decisive resources that are gathered are described below:

3.2.1. URL

In the existing research [21], URLs were employed and serve as a principal resource for similar tasks, such as identifying fake websites [22]. To extract various features, the URL is examined along with its parts, as illustrated in Figure 4. Each listed portion corresponds to a specific topic along with its associated features, demonstrated by the number following the topic. For instance, the features derived from NLP (5) denote a total of five features that lie in the topic. Those features are derived from the input URL.

- Number of digits in the domain [22]: The concerning fact is that non-credible URLs have more digits than authenticated URLs [23]. Those digits are countable as they lie in their domain.
- Domain and subdomain length [22]: In the context of academic organizations, they use brand names within a short domain to introduce themselves to interested parties, such as researchers and postgraduate students. Additionally, for subdomains, the organizations design their names to generate a coherent and logical website structure. To what extent a URL exists with respect to its domain and subdomain, we quantify it in terms of its character.
- NLP features [22]: From the URL, a total of five metrics for the word item are extracted. To achieve these words, the URL was split using ASCII characters. Among the NLP-derived features are the number of words, the average word length, and the standard deviation. The lengths of the longest words and also shortest words are listed. Notably, non-credible journals often use misleading terms such as “International,” “Global,” “Advanced,” “World,” and “Innovative” to appear credible.

3.2.2. HTML

For HTML, the target users can see the content in a browser that provides essential information. Based on it, the subsequent features are derived:

- Domain in the title: In fraudulent websites, the main point of concern is that there lies a discrepancy between the domain name and brand [24]. Authentic journals and websites position themselves in the form of domain and title through the brand name, whereas fraudulent websites cannot do so because those domains are already

registered. credible journals typically include the exact journal or publisher name in their website titles, whereas non-credible journals use inconsistent or misleading titles. This feature checks whether the journal domain matches the page title. We also count the frequency of the journal name in the extracted HTML text, as credible websites consistently reference their journal or publisher name throughout the content.

- Unrealistic Publication Volume: Some journals publish an unusually large number of papers within short publication cycles. In the analyzed cases, journals published around 70 papers per monthly issue, totaling nearly 800 per year. Due to the large volume of publication, the impact and reliability of peer review are at risk. The reason for all this is the lack of time for the publication to be properly reviewed.
- Localized Authorship Ratio: According to data on the Iraqi Higher Education Ministry website, non-credible publishers influence their authors from a single country [4]. In the course of review, one of the highlighted points was that in one journal issue there were 265 publications in one year, out of which 255 were Iraqi. For many other journal issues, a similar trend was identified. Whereas international authorized journals create a diverse environment not only for authors but also for Institutions, as well as for those who want research partnerships. Published research papers, Author segments of HTML, and metadata provide the exact affiliations of authors along with their country of origin.
- Editorial board: Authorized journals present an evident editorial board that contains authentic academic affiliations. Conversely, non-credible journals consistently display fabricated, insufficient, untraceable information about the editorial board. These attributes are important in authenticating any data related to the editorial board [25].
- Peer Review Information: This attribute assesses the clarity and accuracy of the journal's policy. Authorized or authentic journals provide their peer review process in a systematic and evident way. Conversely, non-credible journals provide an unclear overview, such as fast publication or promised publication [25].
- Contact Information Quality: The significance of this attribute is that it inspects the quality of contact information for all journals. Authorized journals can provide the official email addresses of institutions associated with credible domains. non-credible journals, on the other hand, provide emails associated with Gmail or Yahoo. This type of attribute and its consistency with the contact information provided are being analyzed [25].
- Archive Availability: Reliable journals maintain accessible archives of previous issues and published papers. non-credible journals frequently provide incomplete or fabricated archives. We verify the accessibility and consistency of archival content. The function of get_text() is being extracted from the BeautifulSoup4 library to remove all unnecessary parts from HTML containing (links and tags). Moreover, the text is turned into lowercase, and then the occurrences count for the domain [26].

3.2.3. Web-Based System

To strengthen the appearance, extensibility, and effectiveness of websites, web technologies are employed by the developers. To identify the technologies, Wappalyzer12 is used, which can classify around 1950 technologies organized under 71 segments developed from CSS, HTML, or JavaScript code using fingerprints. This enhances the algorithm's effectiveness and its resistance to attempts to circumvent the process, especially for attackers who modify HTML. Further, the following features have been taken from the Wappalyzer report [27].

- Number of Technologies: Authorize publishers usually employ advanced web technologies for security, indexing, and content management. non-credible publishers often use minimal, poorly maintained infrastructure [24].
- Citation and Indexing Instability: Some journals indexed in Scopus may still exhibit non-credible publishing characteristics despite their indexing status. For instance, in 2024, one of the journals had a 2.3 citation index; afterwards, in 2025, it reduced to 0. This type of fluctuation in citation index highlights poor and doubtful publication. Hence, any Scopus indexing should not be regarded as trustworthy if it has these issues [28].
- Low Citation Performance: Specific Scopus-indexed journals show minimal citation metrics. For example, a few of the journals contain a citation index of 0.4 for the two consecutive years (2024 and 2025). In academia, a minimal citation count indicates limited scientific impact along with minimal visibility. The get_text() form of the function is utilized to extract data from the BeautifulSoup4 library:

$$Average\ Citation = \frac{\sum_{i=1}^{n} C_i}{n} \quad (1)$$

The citation metrics could be compiled instantly from Web of Science (Clarivate Analytics) and Scopus APIs. Moreover, to determine authentici-

Table 2. The description of the resources.

No	Resources	Description	Re.
1	**URL**	It stands for the website's unique identifier. The organization name is included in the domain of its academic websites.	[21]
2	**HTML**	It represents a website's source code, which consists of HTML, Cascading Style Sheets (CSS) , and JavaScript code	[24]
3	**Web based system**	It is an online platform used to serve academic, scientific, and industry publishers by managing and hosting. These kinds of platforms support the editorial lifecycle from submission to peer review and open-access publication.	[27]
4	**SSL certificate**	For website security, the Secure Sockets Layer (SSL) certificates take initiatives to secure the encryption of data along with a robust framework that ensures the security for its clients	[29]
5	**Academic Integrity Features**	Academic Integrity features are concerned with the evaluation of the academic reputation and publishing integrity of the journal	[30]
6	**HTTP headers**	Hypertext Transfer Protocol (HTTP) headers are used to initiate the first tier of security for a website. Further, it controls client activity, which consists of a vast range of headers that enhance the security of a website; for instance, HSTS (HTTP Strict Transport Security) fulfills the criteria for clients that employ HTTPS to connect to the website, along with X-Frame-Options, which specifies the approved sources to load the page in frames.	[31]

ty, the journal management system is being examined: the publishing channels that are acknowledged are being assessed, particularly Open Journal Systems (OJS), as testimony of authorized journal management. The analytics system is also being reviewed: Authorized journals employ this system to enhance the journal site and user experience, whereas non-credible publishers tend to use it inadequately [28].

- Security Technologies: We analyze the presence of security technologies such as reCAPTCHA and X-Frame-Options. Vulnerable security systems often demonstrate substandard or fake websites [32].

3.2.4. SSL Certificate

For a website's core protection, an SSL certificate enables end-to-end encryption and helps maintain all shared data confidential. Credible academic journals generally maintain valid SSL certificates to secure communication. Among the certificate features, the two features have been named as binary and SSL certificate, indicating whether the certificate is valid (1) or invalid (0). The listed number of names in the certificate: For subdomains or various websites, a single SSL certificate is sufficient. The brands validate the corresponding certificate for their entire websites. The domains as well as subdomains listed in the SSL certificate are counted via analyzing the certificate's "alternative names" domain [29].

3.2.5. Academic Integrity Features

Academic Integrity features assess indicators related to scholarly verification and the publishing process [30]. These steps include checking indexing claims, verifying ISSNs and DOIs, confirming consistent publication frequency and reviewing the journal's scope. These features may offer valuable insights, but they also present a significant risk of label leakage because some source lists rely on similar scholarly-verification criteria. Indexing-claim verification: Compare the claimed indexing status with the relevant official database, provided such verification resources are available. Consider this feature label-derived if it was constructed using the same database status as the original class label.

- ISSN validation: We compare the reported ISSN with the official ISSN registry. If ISSN validity has already been verified during source-based checks, this feature may be considered label-derived.
- DOI availability and validation: We verify article DOI information with official registration records to confirm the source and support source-based labeling.
- Publication-frequency consistency: We compare the planned publication schedule with what was actually published. Any differences can offer helpful insights, though they may also result from special issues, journal changes, or updates to editorial policy.
- Scope breadth: The topical scope of the journal is evaluated for excessive breadth or insufficient disciplinary coherence. This criterion functions as a contextual indicator and should not be interpreted as evidence of illegitimacy in the case of multidisciplinary journals.

3.2.6. HTTP header

HTTP headers have a significant impact on web security. Appropriate layout minimizes the attack surface while maintaining a protective environment for both server and user. The analysis by [31] shows the role of HTTP headers in detecting malicious websites, with an

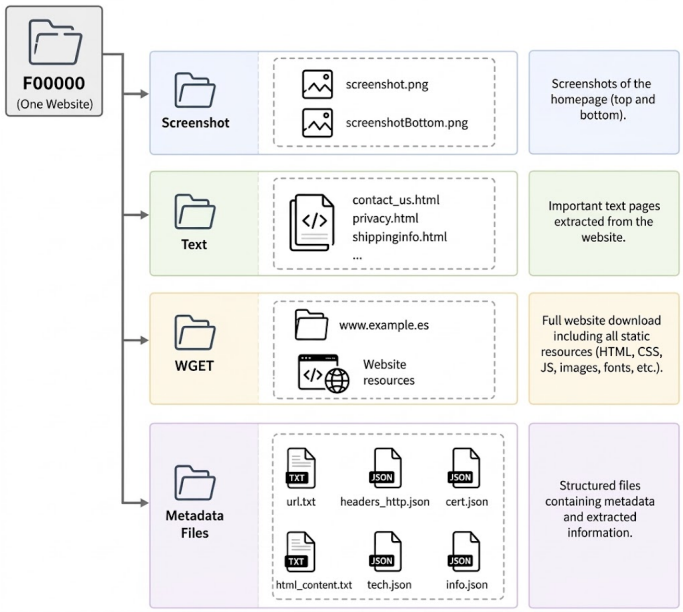

Figure 4. Directory structure of the obtained data.

estimated 91.05% reliability. In this research, four security-related HTTP headers Content Security Policy, X Frame Options, and Strict Transport Security are employed, along with the appropriate headers used by [31], Expect CT and Cache-Control, to distinguish the most effective layout for website detection. Below is a summary of the proposed resources.

3.3. Collection process

To build a dataset covering two class lists and collect website resources, the authors developed a Python 3 application to automatically and efficiently structure all the available resources. To accomplish the process smoothly, resources based on libraries and also APIs are utilized to gather the needed data. Figure 4 depicts the collection steps. Starting with a list of domains obtained from the mentioned services. The lists based on credible journal samples were collected over two months, and non-credible journal samples were collected over three months and updated at a 15-day interval, prompting immediate execution of the collection system to avoid spending time on seized or empty domains. As of May 2026, a limited number of websites have extended the process for non-credible journal list collection. Next, we visited each domain and loaded its content in a supervised browsing context. Utilizing Selenium WebDriver, URLs, HTML, and the remaining terms and conditions were extracted. We obtained HTTP headers using the requests library, Wappalyzer captured domain results, and the SSL socket retrieved certificate data. The unauthorized user format was eliminated.

4. Methodology of the Proposed Model

In this paper, six group features were used which are: URL (U), HTML (H), Web-Based System (W), SSL Certificate (S), Academic Integrity (A), and HTTP Headers (P). This feature pattern aims to highlight the variation among credible and non-credible journals. To optimize performance, we select the most effective features and propose a new feature set to improve the existing website algorithm. We use a dedicated Python 3 module to process each group and save the results as structured JSON. We transform the extracted features into numerical vectors and combine them to form a comprehensive feature vector for each website.

Let *Uv*, *Hv*, *Wv*, *Sv*, *Av*, and *Pv* represent the feature vectors corresponding to URL, HTML, Web-Based System, SSL Certificate, Academic Integrity, and HTTP Header, respectively. The feature vector is defined as $Fv = [Uv, Hv, Wv, Sv, Av, Pv]$. Each website is represented by a 36-dimensional vector, forming a feature matrix with 36 rows and 4,174 columns, consisting of 36 collected features from the selected resources, as shown in Table 3.

5. The Results of the Experimentation

5.1. Experiments setup

In this work, we used AMD Ryzen 5 3600X, an RTX 5060 Ti with 8GB of memory, and 16 GB of DDR4 RAM. Scikit-learn 1.3 and Python 3 were used to implement the experiments and build the machine learning models. Eight [33], [35] classification algorithms: (1) Naïve Bayes (NB), (2) Gradient Boosting Classifier (GBC), (3) k-Nearest Neighbour (kNN), (4) Random Forest (RF), (5) AdaBoost (ADA), (6) Logistic Regression (LR), (7) eXtreme Gradient Boosting (XGBoost), and (8) Support Vector Machines (SVM) were engaged to compare and test the level of performance these algorithms on the design features. The most suitable hyper parameters identified by a 5-fold cross-validated grid search procedure were used to train classifiers. Using scikit-learn's Standard Scaler, we scaled the feature vector throughout the whole training set before applying the resulting scaler to the test set. Lastly, we used k-fold cross-validation to evaluate classifier performance with the following parameters: random_state = 42, shuffle = True, and k = 5. The specified hyper parameters for the proposed model are shown in Table 4.

5.2. Evaluation metrics

We reported F1-score, precision, recall, and accuracy based on results obtained from five-fold cross-validation methods. TP stands for true positives, or the number of counterfeit journal websites that were accurately classified. The term "false positives" (FP) describes the quantity of authentic samples that were incorrectly labeled as non-credible. The number of valid samples that were successfully classified is shown by TN, or true negatives. Lastly, FN stands for false negatives, which indicate how many non-credible journal websites are mistakenly identified as authentic ones.

Table 3. Representation of the model training configuration.

Code	Group	Feature	Description
U1.1	**URL**	length of the standard word	Standard deviation of the words length
U1.2		The longest word	In the URL, the longest word length
U1.3		The average length of words	In the URL, the average words length
U1.4		The shortest word	In the URL, the shortest length word
U1.5		Counting the raw words	The words within the URL
U2.1		The length of domain	characters in the domain name
U2.2		The length of sub-domain	characters in the sub-domain
U3.1		Counting the digit domain	digits in the domain name
S1	**SSL**	has_cert	The validation of the SSL certificate
S2		n_name	domain names are registered in the SSL certificate
P1	**HTTP**	content-security-policy	Website defines CSP header
P2		strict-transport-security	HSTS is implemented in the website
P3		x-content-type-options	*Nosniff* directive is set
P4		x-frame-options	Use *deny* or *sameorigin* directives
P5		cache-control	Website does not use *post-check* outdated directive
P6		expect-ct	Header is configured in the website
NH1.1	**HTML**	Compare domain string with page title using NLP/string similarity	Whether the journal domain matches the journal name/title
NH1.2		Email/domain analysis	Quality and legitimacy of contact information
NH1.3		dominant_country_authors / total_authors	Proportion of authors from the dominant country
NH2.1		Count article URLs/articles per issue/year	Number of papers published within a defined period
NH2.2		Crawl archive URLs and count valid records	Availability and consistency of previous issues/articles
NH2.3		text.count(journal_name)	Number of times the journal/publisher name appears in HTML text
NH3.1		HTML/NLP extraction + validation	Presence/verification of editorial board information
NH3.2		Rule/NLP-based scoring	Degree of transparency of peer-review policy
NW1.1	**Web-Based**	Wappalyzer	Number of technologies detected on the website
NW1.2	**System**	Wappalyzer categories	Number/types of technology categories detected
NW1.3		Wappalyzer/HTML detection	Presence of analytics technologies
NW1.4		Technology detection	Presence of OJS or another recognized system
NW2.1		Calculate year-to-year variation	Variation in citation/indexing metrics over time
NW2.2		Average/median citation metric	Persistently low citation performance
NW3		Detect reCAPTCHA, X-Frame-Options, etc.	Presence of security mechanisms
NA1.1	**Academic**	Compare claims with official database	Whether claimed indexing is actually verified
NA1.2	**Integrity**	Query/validate against ISSN registry	Validity of the journal's ISSN
NA1.3		Crossref/DOI validation	Whether article DOIs exist and are valid
NA2		Compare expected vs. actual frequency	Consistency between stated and actual publishing schedule
NA3		NLP/subject-category analysis	Degree to which journal scope is broad/unrelated

Table 4. The machine learning parameters for the proposed methods.

No	Classifier	Abbr	Hyper-parameter	Value
1	Random Forest	RF	n_estimators	100
2	Gradient Boosting Classifier	GBC	learning_rate	0.1
			max_depth	3
3	EXtreme Gradient Boosting	XGBoost	n_estimators	100
			learning_rate	0.01
			max_depth	5
			subsample	0.8
			colsample_bytree	1.0
4	Support Vector Machines	SVM	C	100
			gamma	0.001
5	k-Nearest Neighbour	KNN	n_neighbors	3
6	Logistic Regression	LR	C	0.1
7	AdaBoost	ADA	n_estimators	43
8	Naive Bayes	NB	var_smoothing	1e-9

Table 5. Testing results of machine learning classifiers.

No	Algorithm	Precision	Recall	F1-Score	Accuracy (%)
1	XGBoost	0.9778	0.9601	**0.9689**	97.86
2	GBC	0.9751	0.9619	0.9686	97.73
3	RF	0.9865	0.9483	0.9671	97.59
4	SVM	0.9622	0.9602	0.9611	97.19
5	LR	0.9564	0.9641	0.9587	97.23
6	KNN	0.9538	0.9372	0.9463	96.42
7	ADA	0.9385	0.9516	0.9438	96.36
8	NB	0.9324	0.9438	0.9317	94.97

Precision:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (2)$$

Recall:

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (3)$$

F1:

$$F1_c - 2\frac{Precision_c \text{ x } Recall_c}{Precision_c + Recall_c} \quad (4)$$

Accuracy:

$$Accuracy = \frac{\sum_c TP_c}{N} \quad (5)$$

5.3. Evaluations of resources for non-credible entities detection

The key goal of this effort is to develop a fruitful machine-learning model that can identify non-credible scholarly entities to help the academic community identify suspicious journals that could lead to deceit.

In the first step, we tested the optimized classification performance using the entire set of the designed features. As shown in Table 5, the feature results are: F1 Score (0.9689) for XGBoost, which had the best classification performance; RF (0.9671); and GBC (0.9686). This performance represents a value level of detecting non-credible journal websites, in which the XGBoost algorithm identified the entire sample set with an accuracy of 97.86%. The best two performances barely differed from one another, whereas Random Forest behaved differently, increasing precision by 0.0087 and decreasing recall by as much as 0.0118. It is therefore recommended for models that require increased sensitivity to these statuses.

5.3.1. Experiment 1

To determine the most effective resource for classifying fraudulent websites of open-access journals, the algorithms' performance with multi-input resources was assessed. As shown in Figure 5, the findings for the Random Forest, GBC, and XGBoost classifiers are displayed with the highest values. On the whole set of specified features, the XGBoost method fared better than the other algorithms (0.9689 F1-Score). The results hardly diminished when algorithms were trained and tested without the URL group. This demonstrates how poorly the suggested URL characteristics operate. A significant factor is the structure of non-credible journal website URLs, which appear authentic because scammers avoid using combo-squatting, or extended subdomains to deceive users. Further, when we conducted tests on non-credible journal websites, excluding the HTML feature group, we observed the lowest performance. This finding underscores the significance of these features for this research. The RF model achieved an F1-Score of 0.9526, representing the highest performance and only 0.0145 lower than its score with the full dataset. In comparison, XGBoost exhibited a decrease of 0.0224 in F1-Score. Performance declined significantly when the Web-Based System and Academic Integrity were excluded, indicating their importance in achieving higher performance. Lastly, the performance of the algorithm was not significantly affected by removing SSL or HTTP headers, and they might be disregarded if those resources are not available.

5.3.2. Experiment 2

The purpose of the second experiment is to show how well the proposed feature sets perform when used independently of the other groups (see Figure 6). With F1-scores of 0.9544 for XGBoost, 0.9548 for the Gradient Boosting Classifier, and 0.9567 for Random Forest, the eight HTML features achieved the best results. Seven of the eight features relate to the domain name, while HTML features require the full URL. As a result, it is advised for models with limited resources. These findings validate the eight designed features in this category. Their reliance on other parties is the primary disadvantage; as a consequence, they cannot function if any of the services are unavailable. For XGBoost and GBC, the seven parameters from the Wappalyzer technology report yielded F1-scores of 0.8329 and 0.8339, respectively, which are still appropriate for a general-purpose model. In this experiment, The HTTP header, SSL, and URL feature set yielded suboptimal results for several reasons. Primarily, the identical domain names of credible and non-credible journal websites mean that the URL set lacks discriminative information. Although keyword lists

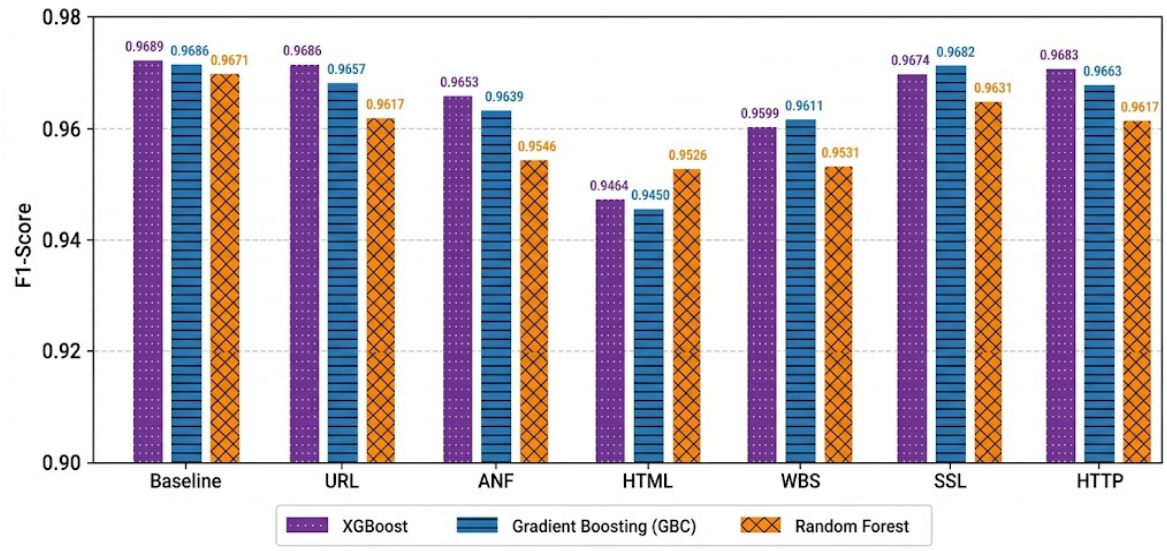

Figure 5. Outcomes of the top-performing algorithms when one of the resources is removed.

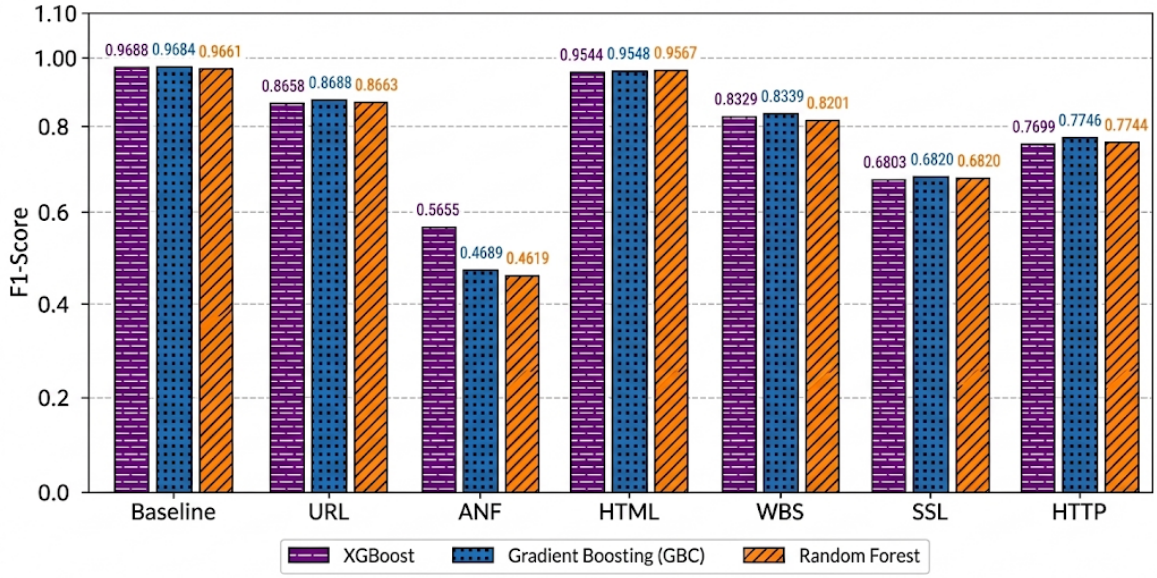

Figure 6. Outcomes of the top-performing algorithms individually.

could produce a language-dependent model, we avoided using them when designing features. Since there were only two parameters in the SSL set, the model's poor performance can be explained by the lack of data. Lastly, with an F1-Score of 0.7746 for the GBC method, for the preceding three sets, the HTTP Header set performed better than average. However, the primary issues with the GBC method are its high sensitivity and false-positive rate (recall of 0.9099 and precision of 0.6744).

6. Limitations

This study offers useful insights, but it also has some limitations that need to be recognized. Website content, technology, SSL certificates, indexing status, and citation details may change over time. As websites evolve, these changes can affect the model's performance. Second, several features depend on external services such as Wappalyzer and scholarly databases. We cannot quantify potential leakage from label-related Academic Integrity features because the original label-generation record and fold-level predictions are unavailable. Third, publisher-level overlap may lead to overly optimistic estimates if related journals appear in both training and validation sets. Therefore, the reported 97.86% accuracy reflects a promising internal result, not definitive real-world detection capability. Finally, journals from the same organization often share templates, HTML structures, technologies, SSL con-figurations, and contact methods. If related journals are distributed randomly across the folds, the results may appear artificially improved because the model could identify patterns unique to specific publishers. The dataset description does not include publisher identifiers; therefore, publisher-group splitting does not apply to the current results. A stronger reanalysis should group journals by publisher or domain family and keep groups intact between training and testing.

7. Conclusion and Future Directions

In this paper, the authors develop an intelligent model to detect non-credible journal websites by proposing a large number of features and using six resource groups to collect data from scholarly websites. The authors applied both conventional and novel features to extract data from the URL, the HTML front-end code, the Web-Based System, the SSL certificate, and Academic Integrity Features. The experimental results show that the proposed novel features improved performance across all groups. Creating a manually validated dataset with extensive data collection should help address the lack of publicly accessible data. 4174 journal websites from both non-credible and credible classes are included in this sample. This allows researchers to benchmark their work using a wide range of data from this dataset, regardless of the resources they employ. The results achieved the goal of the proposed model by detecting the websites of non-credible journals with high accuracy and precision. However, future research still has opportunities for improvement, particularly in the URL data collected. Further research needs to explore advanced learning techniques, including deep learning, to maximize the utility of these features and identify more complex patterns.

Furthermore, expanding the dataset is crucial for enhancing model reliability and generalization. Regularly adding new non-credible journal websites to the dataset keeps it current. This ongoing process helps detection models adapt to emerging threats and improves their real-world applicability.

8. Declarations

8.1. Author Contributions

Fouad Abdulameer Salman: Conceptualization, Methodology, Software, Writing - Original Draft; **Bakhtawar Baluch:** Formal analysis, Validation, Writing - Original Draft; **Zuriana Abu Bakar:** Writing - Review & Editing, Supervision; **Assane Lo:** Writing - Review & Editing, Supervision.

8.2. Institutional Review Board Statement

Not applicable.

8.3. Informed Consent Statement

Not applicable.

8.4. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

8.5. Acknowledgment

Not applicable.

8.6. Conflicts of Interest

The authors declare no conflicts of interest.

9. References

[1] A. M. S. Hinze, D. Stockemer, and T. Reidy, "The predation index: A tool to discover predatory journals," *PS: Political Science & Politics*, vol. 59, no. 1, pp. 158–164, 2026. https://doi.org/10.1017/S1049096525101145.

[2] F. A. do Carmo, S. O. Rezende, C. M. C. Paxiúba, and F. M. F. Lobato, "Predatory publishing in the academic landscape: Researchers' awareness and vulnerabilities," *Scientometrics*, vol. 131, no. 5, pp. 3029–3054, 2026. https://doi.org/10.1007/s11192-026-05614-0.

[3] J. Chigwada and P. Ngulube, "Use of artificial intelligence tools in the publishing process: Expectations from publishers through author guidelines," *Frontiers in Research Metrics and Analytics*, vol. 11, Art. no. 1740510, 2026. https://doi.org/10.3389/frma.2026.1740510.

[4] Research and Development Department, Iraqi Ministry of Higher Education and Scientific Research, "Journal of Research," 2026. [Online]. Available: https://jor.rdd.edu.iq.

[5] J. Beall, "Beall's list of predatory publishers 2016," *Scholarly Open Access*, 2016. [Online]. Available: https://web.archive.org/web/20170113114519/https://scholarlyoa.com/2016/01/05/bealls-list-of-predatory-publishers-2016/. [Accessed: Feb. 7, 2017].

[6] L. Giray, "Identifying predatory journals: A practical guide for researchers," *Annals of Biomedical Engineering*, pp. 1–10, 2026. https://doi.org/10.1007/s10439-026-04231-5.
[7] W. M. B. Ateeq and H. S. Al-Khalifa, "Intelligent framework for detecting predatory publishing venues," *IEEE Access*, vol. 11, pp. 20582–20618, 2023. https://doi.org/10.1109/ACCESS.2023.3250256.
[8] D. J. Dunleavy, "Progressive and degenerative journals: On the growth and appraisal of knowledge in scholarly publishing," *European Journal for Philosophy of Science*, vol. 12, no. 4, Art. no. 61, 2022. https://doi.org/10.1007/s13194-022-00492-8.
[9] J. Wang, W. Halffman, and Y. H. Zhang, "Sorting out journals: The proliferation of journal lists in China," *Journal of the Association for Information Science and Technology*, vol. 74, no. 10, pp. 1207–1228, 2023. https://doi.org/10.1002/asi.24816.
[10] K. Swargiary, "The illusory index: Indian academia and the shadow of predatory and clone journals," ERA, U.S., 2025. https://books.google.co.id/books?id=EStiEQAAQBAJ.
[11] L. X. Chen, S. W. Su, C. H. Liao, K. S. Wong, and S. M. Yuan, "An open automation system for predatory journal detection," *Scientific Reports*, vol. 13, no. 1, Art. no. 2976, 2023. https://doi.org/10.1038/s41598-023-30176-z.
[12] C. Laine and M. A. Winker, "Identifying predatory or pseudo-journals," *Biochemia Medica*, vol. 27, no. 2, pp. 285–291, 2017. https://doi.org/10.11613/BM.2017.031.
[13] T. V. McCann and M. Polacsek, "False gold: Safely navigating open access publishing to avoid predatory publishers and journals," *Journal of Advanced Nursing*, vol. 74, no. 4, pp. 809–817, Apr. 2018. https://doi.org/10.1111/jan.13483.
[14] S. Eriksson and G. Helgesson, "Time to stop talking about 'predatory journals,'" *Learned Publishing*, vol. 31, no. 2, pp. 1–3, 2017. https://doi.org/10.1002/leap.1135.
[15] J. Olivarez, S. Bales, L. Sare, and W. vanDuinkerken, "Format aside: Applying Beall's criteria to assess the predatory nature of both OA and non-OA library and information science journals," *College & Research Libraries*, vol. 79, no. 1, p. 52, Jan. 2018. https://doi.org/10.5860/crl.79.1.52.
[16] J. A. Teixeira da Silva and P. Tsigaris, "Issues with criteria to create blacklists: An epidemiological approach," *The Journal of Academic Librarianship*, vol. 46, no. 1, Art. no. 102070, 2020. https://doi.org/10.1016/j.acalib.2019.102070.
[17] M. A. Shahri, M. D. Jazi, G. Borchardt, and M. Dadkhah, "Detecting hijacked journals by using classification algorithms," *Science and Engineering Ethics*, vol. 25, pp. 655–668, Apr. 2017. https://doi.org/10.1007/s11948-017-9914-2.
[18] M. Dadkhah, T. Maliszewski, and V. V. Lyashenko, "An approach for preventing the indexing of hijacked journal articles in scientific databases," *Behaviour & Information Technology*, vol. 35, no. 4, pp. 298–303, Apr. 2016. https://doi.org/10.1080/0144929X.2015.1128975.
[19] S. Adnan, S. Anwar, T. Zia, S. Razzaq, F. Maqbool, and Z. U. Rehman, "Beyond Beall's blacklist: Automatic detection of open access predatory research journals," in *Proc. IEEE 20th Int. Conf. High Performance Computing and Communications, IEEE 16th Int. Conf. Smart City, and IEEE 4th Int. Conf. Data Science and Systems (HPCC/SmartCity/DSS)*, Exeter, U.K., Jun. 2018, pp. 1692–1697. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00274.
[20] M. Maktabar, A. Zainal, M. A. Maarof, and M. N. Kassim, "Content based fraudulent website detection using supervised machine learning techniques," in *Advances in Intelligent Systems and Computing*, vol. 734, pp. 294–304, 2018. https://doi.org/10.1007/978-3-319-76351-4_30.
[21] K. Wu, S. Chou, S. Chen, C. Tsai, and S. Yuan, "Application of machine learning to identify counterfeit website," pp. 321–324, 2018. https://doi.org/10.1145/3282373.3282407.
[22] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, "A stacking model using URL and HTML features for phishing webpage detection," *Future Generation Computer Systems*, vol. 94, pp. 27–39, 2019. https://doi.org/10.1016/j.future.2018.11.004.
[23] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker, Jr., "Detecting fake websites: The contribution of statistical learning theory," *MIS Quarterly: Management Information Systems*, no. 3, pp. 435–461, 2010. https://doi.org/10.2307/25750686.
[24] Y. Ding, N. Luktarhan, K. Li, and W. Slamu, "A keyword-based combination approach for detecting phishing webpages," *Computers & Security*, vol. 84, pp. 256–275, 2019. https://doi.org/10.1016/j.cose.2019.03.018.
[25] N. A. Mazov and V. N. Gureev, "The editorial boards of scientific journals as a subject of scientometric research: A literature review," *Scientific and Technical Information Processing*, vol. 43, no. 3, pp. 144–153, 2016. https://doi.org/10.3103/S0147688216030035.
[26] M. Nazim and A. Ali, "An investigation of open access availability of library and

information science research," *DESIDOC Journal of Library & Information Technology*, vol. 43, no. 2, pp. 101–111, 2023. https://doi.org/10.14429/djlit.43.02.18580.

[27] M. Sánchez-Paniagua, E. Fidalgo, E. Alegre, and F. Jáñez-Martino, "Fraud detection in e-commerce: A comparative analysis of features to enhance machine learning models," *Electronic Commerce Research*, vol. 26, no. 2, pp. 2467–2502, 2026. https://doi.org/10.1007/s10660-025-10029-9.

[28] Z. Zhao, H. Wen, J. Ma, J. Zhan, T. Xu, Y. Wei, and Q. Zhang, "ResearchEVO: An end-to-end framework for automated scientific discovery and documentation," *arXiv preprint arXiv:2604.05587*, 2026. https://doi.org/10.48550/arXiv.2604.05587.

[29] Anti-Phishing Working Group (APWG), "Phishing activity trends report 2Q," 2021. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2021.pdf. [Accessed: Oct. 14, 2024].

[30] D. Mills, A. Branford, K. Inouye, N. Robinson, and P. Kingori, "'Fake' journals and the fragility of authenticity: Citation indexes, 'predatory' publishing, and the African research ecosystem," *Journal of African Cultural Studies*, vol. 33, no. 3, pp. 276–296, 2021. https://doi.org/10.1080/13696815.2020.1864304.

[31] J. M. Iv, D. Bhansali, M. Gratian, and M. Cukier, "A comprehensive evaluation of HTTP header features for detecting malicious websites," in *15th European Dependable Computing Conference (EDCC)*, pp. 75–82, 2019. https://doi.org/10.1109/EDCC.2019.00025.

[32] Y. Shibuya, K. Mwitondi, and S. Zargari, "Experimental analyses in search of effective mitigation for login cross-site request forgery," in *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, Cham, Switzerland: Springer International Publishing, 2020, pp. 233–266. https://doi.org/10.1007/978-3-030-35746-7_12.

[33] D. Singh, S. Singh, and N. Kumar, "Benchmarking credit card fraud detection models: A comprehensive evaluation across diverse datasets," *Computational Economics*, pp. 1–30, 2025. https://doi.org/10.1007/s10614-025-11234-2.

[34] E. Rudini and F. Ernawan, "Prediction of Alzheimer's dementia using soft voting ensemble learning with machine learning," *IJACI: International Journal of Advanced Computing and Informatics*, vol. 1, no. 1, pp. 48–55, 2025. https://doi.org/10.71129/ijaci.v1.i1.pp48-55.