ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2026.1835.285

AUTOMATED IDENTIFICATION OF INACCURATE CITATION BY SIG-GATE FUSION ADAPTION

Zahraa Yahya Mahdi Al-Mayali[1*], Salam Al-augby[2]

[1*]*College of Computer Science and Information Technology, Wasit University, Kut, Iraq.*
e-mail: zmahadi@uowasit.edu.iq, orcid: https://orcid.org/0009-0007-0595-5948
[2]*Faculty of Computer Science and Mathematics, Kufa University, Kufa, Iraq.*
e-mail: salam.alaugby@uokufa.edu.iq, orcid: https://orcid.org/0000-0001-8247-9497

Received: January 06, 2026; Revised: February 20, 2026; Accepted: April 09, 2026; Published: May 29, 2026

SUMMARY

Citation network analysis is increasing rapidly nowadays. However, the most important challenge is addressing inaccurate citation detection, where many unreasonable citations happen today, such as when the content of the cited paper is not relevant to the reference paper (citing paper), reciprocity and small cycles, which mean that author A cites author B and author B cites author A (soviet citation). This can also happen in small groups or within institution concentration, such that authors are affiliated with the same institution. Inconsistency in the venue or scope is another significant issue. All these situations are studied in this paper, whereas other papers focus on one or two situations, by combining three complementary evidence blocks : (i) text features captured from the relations between the citing paper and the references using SBERT cosine and BM25 ranks, (ii) graph features extracted from citation –author network such as (PageRank, clustering, reciprocity counts, same-institution reference share and author overlap) and (iii) metadata feature extraction, such as (length,/ size, author affiliations and temporal context). These features are selected using ensample classification explanations (SHAP) to get stable features only. Our proposed method presents SiG-Gate, which is a similarity adaptive fusion head for identifying inaccurate citations by mutually leveraging text semantics, citation graph structure, and metadata features. Learning weights in the proposed SiG-Gate are adjusted by calibrating the involvement of each instance level, by participating calibrated unimodal outputs with explicit indicators of cross-modal disagreement, resulting in decisions that are both stronger and more interpretable. On stratified cross-validation, the approach attains accuracy = 0.946, F1-score = 0.91, precision =0.90 and AUC=0.96. The results indicate that consistency-aware, convergent evidence is essential for reliable and accurate citation detection.

Key words***:*** *citation network, graph network analysis, inaccurate citation, metadata analysis, semantic similarity, SHAP.*

INTRODUCTION

Throughout the fundamental stone of scholarly communication is citations. They acknowledge prior contributions, are responsible for providing supporting evidence, and help form new knowledge in specific areas of scientific investigation. Scientific knowledge is constructed from both the wide-ranging structure and individual studies that shape the academic integrity. Citation accuracy supports the authority and transparency of research, whereas inaccurate, confusing, or immaterial citations distort the

transmission of knowledge and destabilize trust in the academic evaluation system [1].

Recently, the volume of scientific publications has rapidly increased, strengthening this challenge. The propagation of large-scale data warehouses and open access platforms has dramatically expanded the amount of existing publications, particularly in computer science and related fields. While this progress enables broader knowledge sharing, it also increases the risk of citation inaccuracies, whether stemming from misunderstanding, human error, or deliberate manipulation [2].

Citation inaccuracies can take many forms, including references that are unconnected to the claims they are meant to support or are contained largely to influence scholarly metrics. Such practices may distort the interpretation of key indicators and devalue the integrity of individual studies [3].

Subsequently, citation accuracy represents a core component of academic honesty, not only a procedural requirement. There is an urgent need for identifying citation inaccuracies. Accordingly, there is an essential need for automated detection methods. Principally, it is nearly impossible to manually verify the reliability of citations by auditors or peer reviewers, as the volume of publications grows rapidly [4].

To address this issue, several major viewpoints can be leveraged. The first approach is based on semantic similarity assessment, which estimates the contextual association between the citing paper and the source work. The second approach is based on citation network analysis, which simulates the relational structure among scholarly networks and builds patterns for citation behavior, indicating irregular pattern. The third approach is based on metadata of the papers, the journals, and publishers, where this branch may give weak accuracy if used alone; it needs to be combined with other baselines to provide greater validity in identification. Another approach is based on transformers architectures, which may use one branch or combine multiple branches that enable richer cross modal connections and more clarity for citation behavior [5].

This paper introduces an automated approach that integrates both network analysis and semantic similarity to identify inaccurate citations across all disciplines under study. The main goal is to enhance the reliability of scholarly communication by providing scalable tools that measure the integrity of citations, evaluate support practices, and strengthen the foundations of research trustworthiness.

The proposed method's novelty lies in SiG-Gate, a similarity-guided fusion module that aligns unimodal posteriors through calibration, derives cross-modal agreement cues, and learns sample-specific gates to up- or down-weight modalities at test time. This instance-conditioned fusion goes beyond fixed early or late-fusion schemes and off-the-shelf stacking, providing built-in interpretability and improved robustness when classes are highly imbalanced.

The paper is organized into seven primary sections. Section 1 provides the Introduction. Section 2 defines the Research Questions and Problem. Section 3 discusses the Study Objective and Contributions. Section 4 presents the Literature Survey and Background, which integrates related work and an overview of metadata and textual approaches. Section 5 describes the Datasets and Methods, including dataset preparation, the Reference Bank, and the proposed SiG-Gate methodology. Section 6 reports the Results and Discussion, covering the experimental performance and ablation studies. Finally, Section 7 concludes the paper and offers suggestions for future work. While the original introductory text listed eleven sections, the revised manuscript now aligns with this streamlined seven-section format.

RESEARCH QUESTIONS AND PROBLEM

Evaluating research quality by increasingly reliable citation-based metrics has made the integrity of citations a critical concern for the scientific community [6]. While citations are expected to form accurate connections between scholarly works, in practice, many references fail to achieve this objective. Inaccurate citations may occur unintentionally or deliberately, through misinterpretation, or oversight, or in cases where an author attempts to manipulate the indicators of bibliometric. These attempts weaken the credibility of the output research, distort the academic record, and pose challenges for fair evaluation of researchers, journals, and institutions [7].

Against this situation, the core of the research problem can be addressed and stated in this work as follows:

How can inaccurate citations in any discipline's literature be automatically identified through the integration of citation network analysis, metadata analysis and semantic similarity?

To instruct and guide this investigation, the following Research Questions (RQs) are formulated:

(1) RQ1: What are the most common forms of citation inaccuracy observed in any field studied in this paper?

(2) RQ2: How can the proposed methods detect structural anomalies that indicate potentially inaccurate references?

(3) RQ3: How effective is the approach that integrates citation network analysis with semantic similarity and metadata, compared to using either method in isolation?

By formulating and addressing these questions, this work aims to provide a new methodological contribution to automate the identification of inaccurate citations, in addition to offering practical insights for the improvement of scholarly evaluation systems.

Study Objective and Contributions

The main contribution of the study can be summarized as follows:

(1) Conceptual Clarification: Establishing a framework to define and classify various forms of inaccurate citations in scholarly communication, outlining clear characteristics, and distinguishing between unintentional errors and deliberate manipulations [8].

(2) Network Analysis Integration: Examining the capacity of citation network structures-for instance, in-degree, out-degree, and community detection- to expose anomalous patterns indicative of inaccurate referencing [9].

(3) Semantic Similarity Application: Investigating how the similarity of the text between citing contexts and referenced documents can be used as a citation accuracy indicator [10] [11].

(4) Framework Development: Devising and implementing a hybrid model approach that combines graph-based features with semantic alignment measures and metadata extraction features, producing SiG-Gate, which is a similarity guided adaptive fusion head that learns per-instance modality weights from calibrated scores and cross-modal inconsistency, in order to strengthen a more robust detection system. This head produces a transparent diagnostic via gate-weight analysis and case studies, with a stratified and statistically grounded evaluation protocol that includes cross-validation and confidence intervals.

RELATED WORK

Liu et al. introduce ACTION, defining and formalizing "anomalous citations" within a framework of a heterogeneous academic graph that consists of (paper–author–venue) relations. This framework merges the content signals with a rich network structure to identify suspicious links at scale. While the heterogeneous formulation remains strong, the method is largely structure-centric and devotes less attention to the alignment of semantics between the citing statement and the referenced content. Additionally, the publicly available details on edge-level gold labels and fully reproducible benchmarks are limited, deterring the direct comparisons focused specifically on citation inaccuracies. [12]. Ye et al. propose an automated citation auditing system, which is an AI procedure that verifies each reference across multiple databases without assuming correctness. This method achieved a 91.7% verification rate, effectively identifying fabricated and retracted citations. However, it relies on structured metadata and rules-based validation, lacking adaptive multimodal learning. [13]

Rifat et al. extended context-aware recommendation systems by incorporating textual inputs (e.g., rhetorical or auxiliary fields) while preserving the graph components, leading to incremental gains over

text-only baselines. These contributions remain tied to recommendation metrics and do not evaluate performance against edge-level inaccuracy labels. These improvements could reflect a better retrieval rather than enhancing the identification of mis-citations within manuscripts. [14]

Bandy et al. analyzed the impact of large language models on citation inflation and hallucination in scientific text generation. The authors reported that 23% of generated citations were imagined with over half lacking verifiable identifiers or accurate metadata. Their proposed method includes a set of rule-based filters for DOI and venue validation, achieving around 88% accuracy in detecting fake citations. While valuable in identifying overt fabrication, this approach does not generalize to subtle misuses or contextual mismatches between citing and cited content. [15]

Shen et al. introduced CiteGuard, a hybrid system that uses SBERT-based sentence embeddings and citation context alignment to verify citation authenticity. Their model, trained with an XGBoost classifier, attained 91.2% accuracy and 90.5% F1, demonstrating strong performance in identifying contextual inconsistencies. However, it is highly dependent on access to full-text citing documents and omits structural signals such as citation networks or author reputation. [16]

BACKGROUND OVERVIEW

Previous work on inaccurate citation detection is fragmented, focusing either on semantic similarity or graph network analysis, with limited attention to metadata analysis. In contrast, the proposed method combines the three baselines together. The following is a brief background overview of each baseline.

Metadata Approach

Many features carried by metadata may be weak but complementary signals, such as venue, year, authorship, patterns, affiliations, overlaps, and reference counts that are hard to capture by text alone or with a graph. [17] Employing ensemble learning, which combines diverse learners (e.g. linear and tree-based) to produce these heterogeneous signals, enhances generalization by reducing variance and bias relative to a single model. In citation checking, this means that small but effective and possibly noisy indicators (e.g., an abnormal venue or topic, or an unusual concentration within an institution) can be aggregated into a stronger decision rule.

The second necessity is auditability. SHAP provides additive feature attributions based on cooperative game theory, allowing for per-decision explanations and global ranking of influential metadata fields [18]. SHAP could be employed to help reviewers weigh the decisions, which may depend on believable factors (e.g., venue scope, publication age), rather than fraudulent articles. This will support sensitivity checks by probing how attributions shift under different assumptions.

Text Approach

In this field, many methods can be used to align the semantics between citing papers and their references. Sentence- BERT(SBERT) produces compact sentence or document embeddings; cosine similarity approximates topical fit while remaining robust to paraphrase [19]. Primarily, this phase acts as a semantic layer aimed at minimizing false positives. First, the BM25 method plays a lexical matching role by retrieving sources whose token distributions are most similar to the query (title/abstract/keywords), yielding a strong sparse baseline, especially under scarce training data circumstances in specialized fields where the terminology is fixed [21]. In addition, its simple vector-space foundations support interpretability and efficient large-scale retrieval [20]. Finally, fuzzy matching (e.g., Levenshtein distance) mitigates noisy metadata such as spelling variants, transliteration differences, or truncated titles by aligning near-duplicates before scoring, thereby improving both recall and record linkage quality [22].

Unification of Classification

The unification of classification is a model that concatenates graph, metadata, and textual features across

channel collaborations where moderate semantic mismatch becomes suspicious only when unified with structural anomalies or venue/context standards. Logistic regression provides a calibrated linear baseline and clear, coefficient-level interpretability for quick checks and threshold setting [23]. Non-linear ensembles address higher-order effects. Random Forests average many decorrelated trees to capture interactions with robust out-of-bag validation and resistance to overfitting in tabular data [24]. Gradient-boosted trees (XGBoost) achieve state-of-the-art performance on mixed feature types by utilizing additive trees, shrinkage, and regularization, often yielding the strongest accuracy–interpretability trade-off when paired with post-hoc SHAP explanations [25]. In practice, one trains all three, performs cross-validated model selection, and retains SHAP to document which text, metadata, and graph signals most consistently drive decisions across specialties.

The evidence on how to identify an erroneous citation is scattered, which mainly considers single elements such as semantic similarity or graph network analysis, overlooking the useful pieces of evidence that are contained in metadata. The structure-focused models of the past usually can offer adequate consideration to semantic correspondence between the citation sentences and the content that they refer to. Moreover, most existing approaches are based on hard and rule-driven verification or predetermined fusion plans instead of learning based on instances and conditioned by instances. Authoritative, edge-level, gold label, benchmarks and reproducible datasets are also notably missing to determine different types of citation inaccuracies, specifically. Majority of the research undertakings deal with one or two individual citation problems, including the topic mismatch instead of offering a multifaceted examination of multifaceted issues such as reciprocity rings, institutional concentration as well as venue inconsistency.

DATASETS AND METHODS

This paragraph presents the datasets used in the paper and the proposed method

Preparing Dataset

The dataset used in this proposed approach undergoes many preparation stages.

First Dataset

In the first dataset, real-world publications from Scopus data for the year 2025 in the computer science discipline are collected from globally leading universities such as Harvard, Oxford, and MIT, according to QS World University Rankings and Times Higher Education (THE). Because no authoritative benchmark of inaccurate citation exists, we built this dataset and manually labeled it. We contrast two label corpora: Label0 (Real), where references are kept as is from the sources before inclusion, resolving DOIs and reconciling title/author/year/venue against OpenAlex/Crossref, leaving any conflicting records; and Label1 (Inaccurate), where references are controlled, and replacements are added for a randomly selected subset of citing papers, with the same number of injected links per paper as the original ones.

We use four reproducible ways: (i) topic mismatch (low title/ abstract similarity and low concept overlap), (ii) reciprocity and short cycles, meaning author A cites author B and author B cites author A (Soviet citation), counting 3-4 cycles, (iii) within same institution (high same affiliation share among cited authors), (iv) venue/field inconsistency (mismatched journals/conference fields). For every injection, probable metadata is stabilized and the recipe is logged (corruption type, fields changed, magnitude, and random seed), addressing self-referential concerns. All Real items undergo external registry checks. A stratified sample of injected cases is manually inspected as a sanity check, and splits are cited in the paper to prevent leakage. Table 1 shows the number of documents for each university and the injection and splitting processes. The dataset has (46) unified attributes fetched from the Scopus database as metadata information. The dataset includes the following fields: Authors, Author full names, Author(s) ID, Title, Year, Source title, Volume, Issue, Art. No., Page start, Page end, Page count, Cited by, DOI, Link, Affiliations, Authors with affiliations, Abstract, Author Keywords, Index Keywords, Funding Details, References, Correspondence Address, Editors, Publisher, Conference name,

Conference date, Conference location, ISSN, ISBN, CODEN, PubMed ID, Language of Original Document, Abbreviated Source Title, Document Type, Publication Stage, Open Access, Source, and EID. The textual attributes (Title, Abstract, Author Keywords, Index Keywords) are used for the text mining approach, other attributes are used for metadata extraction features, and all attributes are used for building a graph network and extracting features.

Table 1. Datasets summary

University	Split	Label0 (Real)	Label1 (Inaccurate)	Total Papers
Oxford	Train (80%)	771	771	1542
	Test (20%)	193	193	386
Harvard	Train (80%)	460	460	920
	Test (20%)	115	115	230
MIT	Train (80%)	55	54	109
	Test (20%)	13	14	27

Second Dataset

The second dataset (References Bank) is built by extracting all the references from the datasets and collecting the textual attributes from global papers websites such as OpenAlex, Semantic Scholar, Europe PMC, Crossref, Lens, CORE, arXiv API, DOAJ, Springer (Meta/OpenAccess), and Unpaywall in an automatic way using Algorithm (1) "Building References Bank", which pulls the textual attributes for the references to increase the textual features to the references. Table 2 shows the number of extracted references from the standard dataset (Harvard, Oxford, and MIT).

Table 2. References number

	No. of papers	No. of references
Standard dataset (Harvard, Oxford, MIT)	3214	89393

Proposed Methods

The proposed method detects anomalous citations using metadata-driven attributes, textual similarity, and graph-based citation structure. The first stage in the methodology begins by pre-processing and normalizing large amounts of bibliometric scholarly research metadata. Following that, multi-modal features are extracted; this feature extraction process undergoes feature selection using ensemble classification (SHAP explainability). After that, the method trains ensemble models, applies Platt scaling for calibration, and tunes thresholds to maximize predictive accuracy. Finally, a comparison is built between the text-based approach and the graph-based approach. Figure 2 illustrates the workflow of the proposed methodology. The proposed method steps are carefully defined in Algorithm (2): Fake Reference Detection Model, where the detailed steps are explained in the following approaches.

Algorithm 1. Building References Bank

Input:

- Folder F containing {f_1, f_2, f_3, ……f_n} where n is the number of files, and each f_n contains paper metadata for each discipline
- Global websites requirements if available for websites (OpenAlex, Semantic Scholar, Europe PMC, Crossref, Lens, CORE, arXiv API, DOAJ, Springer, Unpaywall)

Output: for each disciplines D: CSV file of pulled references

Step 1: Data Preparation and Preprocessing

- Load data and split columns (DOI, Title, Reference) and preprocess for each file f_j in disciplines d_i

- Create two files one for success pulled and second for failed pulled for each file f_j in the disciplines d_i

Step 2: Searching and Pulling

- For each file f_j in disciplines d_i search in websites (OpenAlex, Semantic Scholar, Europe PMC, Crossref, Lens, CORE, arXiv API, DOAJ, Springer, Unpaywall) and stop at the first success per reference
- Add success pulled reference to success file and unsuccessful one to failed file

Step 3: Output Query File

- Print output query file with percentage of success pulled references

The data acquisition pipeline of Building References Bank works as an automated system as shown in Algorithm 1. It uses local files to identify paper identifiers such as DOIs and titles, and makes queries to global scholarly databases such as OpenAlex, Crossref, and Semantic Scholar in a systematic manner. Through a stop-at-first-success reasoning and keeping two logs: successful and unsuccessful retrievals, it guarantees the establishment of a strong and metadata-enriched repository that is used further to conduct an analysis.

Metadata Feature Engineering Stage

The metadata Feature engineering represents the first stage of the proposed method to extract the first block of features such as (Page count, Title Length, Abstract Length, Author Keyword Len., Authors Count, Affiliation Count) as shown in equation (1). Following that, a training process is applied using XGBoost on metadata features only, and measured preliminary feature importance is measured using ensemble classification with SHAP explainability.

$$f^{meta}(p) = \{PageCount; AbstactLen; TitleLen ; AuthKWLen; AffilsCount; IndxKWLen; AuthorsCount\} \quad (1)$$

$$A_j = \frac{1}{N}\sum_{n=1}^{N} |SHAP_n(f_j)| \quad (2)$$

Calculate mean absolute SHAP values for each fold on the validation set and average them across folds to obtain a robust rank score denoted (SHAPn), where features were ranked by SHAPn and the top subset was selected via cross-validated F1maximization (alternative: 90% cumulative importance or stability selection) as shown in equation (2). This stage produced a squeezed, leakage-free metadata feature set for subsequent classification.

Textual Similarity Feature Engineering Stage

This stage starts with building a text string for each paper containing (Title, Abstract, Author Keyword, Index Keyword) into one text denoted by (ti). A semantic vector is constructed for each text (ti) using SBERT to convert the text to an embedding, as shown in equation (3).

$$e_j = \text{SBERT}(t_i) \quad (3)$$

Where ti is the textual feature, a mapping process is applied to each cited references with the Reference Bank to encode the reference text vector denoted by (tr) and then measure the similarity of each paper- reference pair using SBERT cosine similarity (semantic similarity) as in equation (4) and BM25 similarity (lexical similarity) to get the semantic from equation (5) and the lexical understanding from

equation (6) for the paper content following that calculate the average of both similarities to become input for the classifier.

$$S^{SBERT}(p,r)=\frac{e_p;e_r}{||e_p||\,||e_r||} \quad (4)$$

$$\Delta\mu\ \text{SBERT} = \mu\ (\text{sBERTorg}) - \mu\ (\text{sBERTnow}) \quad (5)$$

$$\Delta\mu\ \text{BM25} = \mu\ (\text{sBM25org}) - \mu\ (\text{sBM25now}) \quad (6)$$

Where *p* for paper and *r* for references.

Graph and Author-Relation Features Stage

A directed graph G (V, E) is built, where each node is denoted by (v) and an edge (u→v) means paper (u) cites paper (v). Figure 1 shows part of the building directed graph, where papers are represented and edges represent the citations between documents and references. The many features extracted from the building directed graph table 3 show the extracted features and their explanations for each feature. Extracting graph features added significant value to the detection of anomalies in citations because it focuses on the suspicious relationships between authors from the same institution within author groups, as detected by citation rings, which complements other text features. A clear pattern of citations becomes apparent when using a directed graph network.

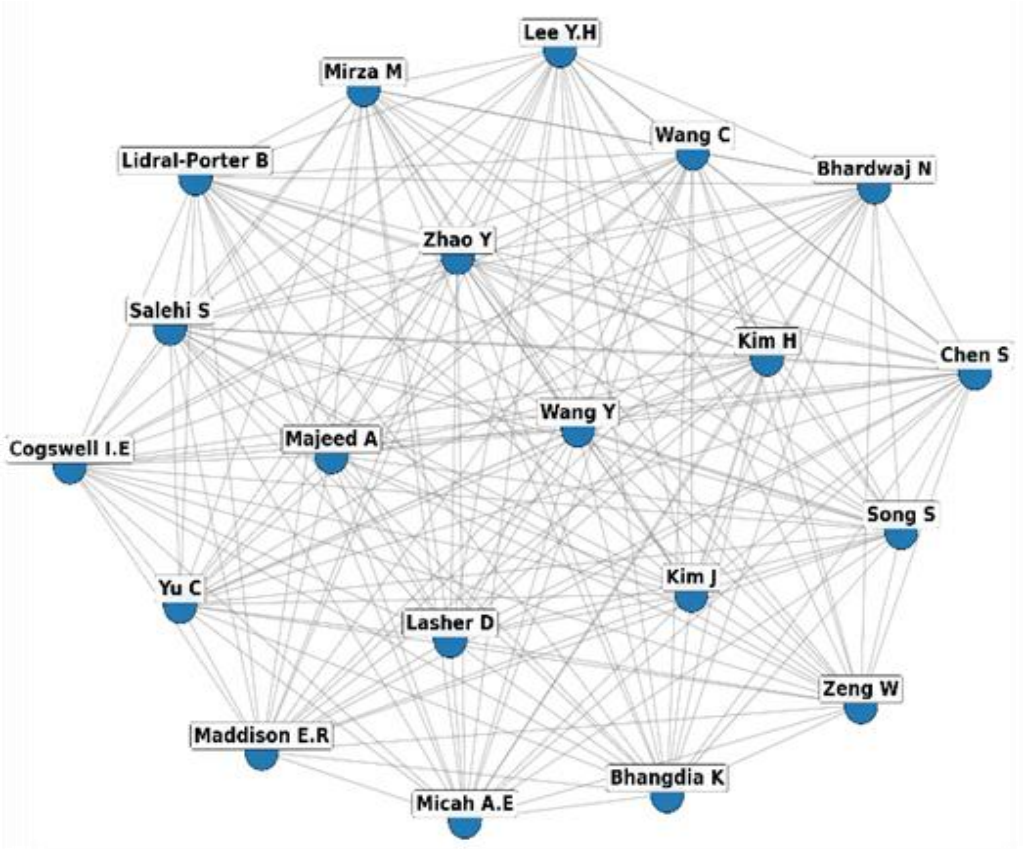

Figure 1. Sub-direction graph author citation network

Table 3. Feature usage

Feature	Equation	Usage
In- degree	$d_{\text{in}}(r_i) = \lvert \{p_i : (p_i, r_j) \in E\} \rvert$ (7)	In equation (7) this feature shows how many papers cite node v, a basic popularity- impact signal.
Out- degree	$d_{\text{out}}(r_j) = \lvert \{r_k : (r_j, r_k) \in E\} \rvert$ (8)	In equation (8) this feature shows how many papers node v cites, a basic referencing signal.
PageRank	$\text{PR}(r_j) = (1-\alpha) + \alpha \sum_{x \in E} \frac{d_{\text{out}}(x)}{\text{PR}(x)}$ (9)	In equation (9) this feature measures citations, where a citation from a well-connected paper counts more than one from a rarely cited paper.
Clustering coefficient	$C_v^{\boldsymbol{dir}} = \frac{((A+\boldsymbol{A^T})^3)_{vv}}{2K_2^{\boldsymbol{in+out}}(K_v^{\boldsymbol{in+out}}-\boldsymbol{1})-2K_v^{\leftrightarrow}}$ (10) Where A directed adjacency matrix of node v $K_v^{\boldsymbol{in+out}}$ represent degree of node v $K_v^{\leftrightarrow}$ represent the number of mutual links adjacency to v $((A+\boldsymbol{A^T})^3)_{vv}$ represent the counts number of the direction aware triangles through v	In equation (10) this feature represents the local triangle density, measuring how densely neighbors of v cite each other. High values mean v sits in tightly inter-citing community, indicating linked citation communities.
Reciprocity	$K_v^{\leftrightarrow} = \sum_{u=1}^{U} A_{vu} A_{uv}$ (11) $r_v = \frac{K_v^{\leftrightarrow}}{\max(1; K_v^{out})}$ (12) Where $A_{vu} = 1$ if v→u $K_v^{\leftrightarrow}$ counts mutual links (both v→u and u→v) $K_v^{out} = \sum_{u=1}^{U} A_{vu}$	İn equation (11) and equation (12) this feature represents the number of mutual citations around. A high fraction of reciprocity can indicate citation loops within suspicions groups. It detects anomalous citations by frequent mutual citation patterns (citation rings), representing relations among authors.
Institutional homogeneity	SI(v)= $\frac{\lvert \{u \in OutNeighbor(v) : inst(u) = inst(v)\} \rvert}{\lvert OutNeighbor(v) \rvert}$ (13) Where $OutNeighbor(v)$ set of nodes cited by v	In equation (13) this feature measures the number of v's citations that go to the same institution. High values suggest within- institution concentration and group bias, where many citations point to the same institution.
Author overlap (Jaccard)	$J(A_u; A_v) = \frac{\lvert A_u \cap A_v \rvert}{\lvert A_u \cup A_v \rvert}$ (14)	In equation (14) this feature represents the value of the set of authors on paper v where high value represents self/co-author citation noting value equal 1 represent identical author list.

Modelling and Ensemble

Starting with the feature preparation process in this stage, standardization is applied on the training set to avoid data leakage. Diverse learner are fitted to reduce variance and improve robustness, and their outputs are calibrated to obtain comparable probabilities per modality. Logistic regression, random forest, and XGBoost are applied, then the ensemble probability is computed in the following equation (15):

$$\text{P} = \frac{1}{M} \sum_{m=1}^{M} P^m (M \in \{2; 3\} \qquad (15)$$

The proposed SiG-Gate replaces the fixed average with a tiny gating network that receives the three calibrated probabilities plus agreement signals (consistency/ inconsistency), using equation (17) and (18). SiG-Gate returns three weights (summing to 1); the final probability is their weighted average, allowing the model to upweight agreeing modalities and downweight outliers.

Finally, adapt threshold τ applied for all models, including SiG-Gate, to maximize the F1 score, which is defined in equation (16).

$$\tau = argmax \frac{2.Precision(\tau).Recall(\tau)}{Precision(\tau) + Recall(\tau)} \quad (16)$$

$$\text{incons} = \frac{1}{3}(|P_t - P_g| + |P_t - P_m| + |P_g - P_m|) \quad (17)$$

$$\text{cons} = 1 - \text{incons} \quad (18)$$

Explainability

SHAP explainability is applied to the final XGBoost model (trained on the whole training set with the pre-processing stack). On the test report, the global SHAP summary for the top 30 features with a leg on each side of text similarity, graph structure, and checked metadata. Checking the behavior of the model, how it moves above/ below the operating threshold τ. Ablations and validity checks guiding a measured ablation comparing text approach, graph approach, metadata approach, text- graph approach, text-metadata approach, and graph- metadata approach under a fixed evaluation protocol (Accuracy, Precision, Recall, F1, ROC-AUC, AP with PR/ROC curves). Integrity checks include:

(1) Key-split sanity: ensuring no paper key overlap across train/ test.
(2) Leakage audit: excluding availability of fields (e.g., NonNullRatio, HasDOI/ISSN, source flags).
(3) Threshold robustness: re-estimating τ on bootstrapped validation folds and reporting metric variability.

Evaluation Protocol

Reporting on the test set, many measurements were taken to evaluate the model. Table 4 lists all the evaluation metrics used.

Table 4. Measurement summary

Measure	Equation	Variables
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$ (19)	In equation (19) TP true positive, TN true negative, FP false positive and FN false negative
Precision	$\frac{TP}{TP+FP}$ (20)	In equation (20) Of the items predicted positive items, the fraction that are truly positive.
Recall	$\frac{TP}{TP+FN}$ (21)	In equation (21) Of the truly positive items the fraction correctly found
F_1 score	$2 \times \frac{precision \times Recall}{precision+Recall}$ (22)	In equation (22) Harmonic mean of precision and recall, balance both
ROC-AUC	$AUC = \int_0^1 TPR(FPR)d(FPR)$(23)	In equation (23) Area under ROC curve (TPR vs. FPR) equivalently probability a random positive scores higher than a random negative. FPR = $\frac{FP}{FP+TN}$, TPR=Recall
Conf. matrix	TP FP FN TN	2*2 count table of predicted columns vs. actual rows classes

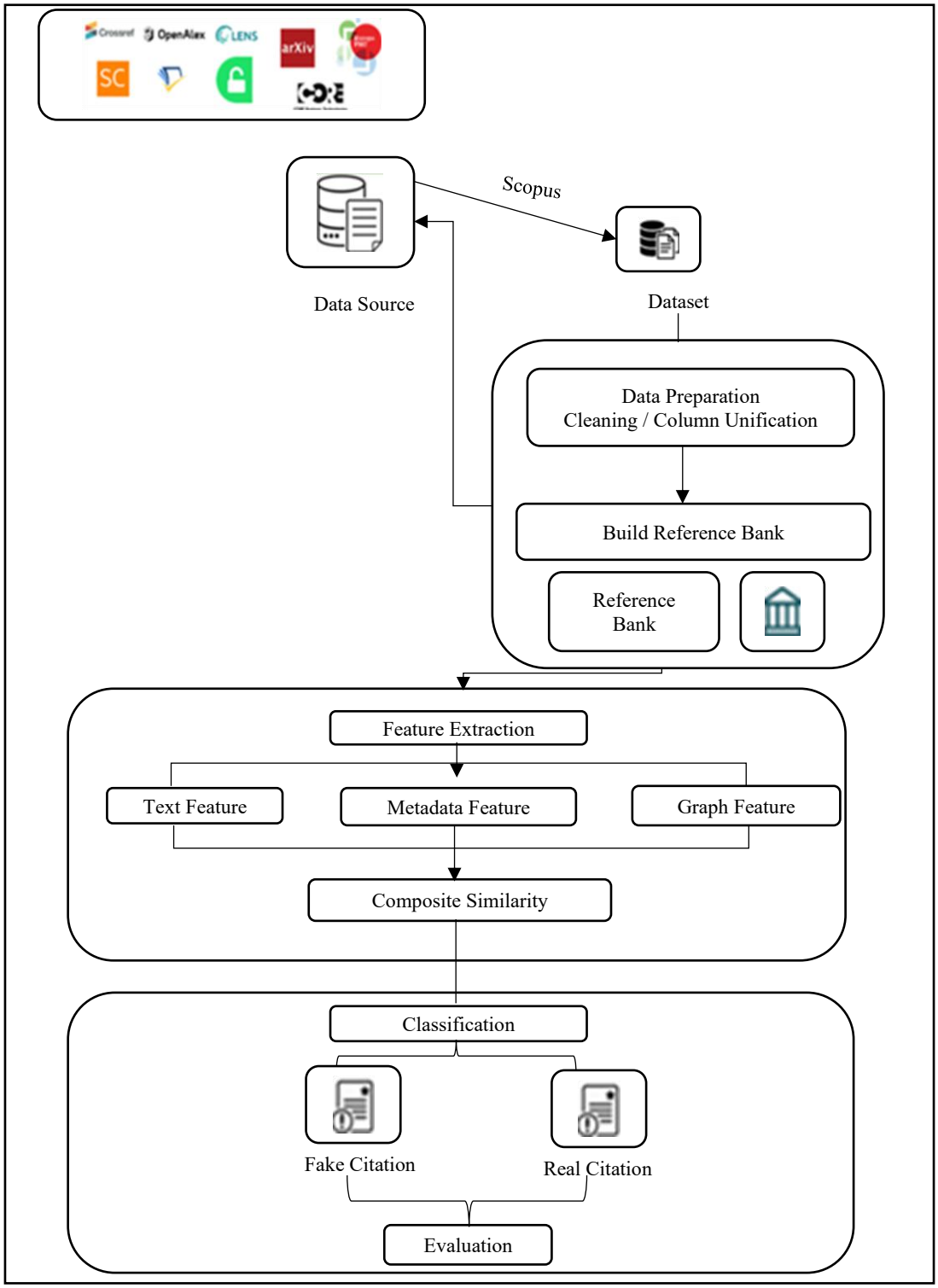

Figure 2. Methodology diagram

Algorithm 2. Fake References Detection Model

Input:

- Let the set of files F={f_1, f_2, f_3, ……f_k} where k is the number of files, and each f contains papers metadata for each discipline.
- Let the set of papers P={p_1,p_2,p_3,…….p_n} where n is the number of papers in each disciplines file f

- Let the set of the references $R=\{r_1,r_2,r_3,\ldots\ldots r_m\}$

Output: $y_{ij} \in \{0,1\}$ where 1=fake and 0=real

Step 1: Data preprocessing and References Association

- Preprocess and standardize the dataset
- Divide dataset into training and testing data
- Each paper p_i associated with a subset of references where $R_i \subseteq R$ to form citation pair (p_i,r_i)

Step 2: Metadata Feature Extraction

- Extract Metadata feature (f_{meta}) from equation (1) for a paper p
- Apply XGBoost on training set to rank metadata feature by SHAP as illustrated in equation (2)

Step 3: Text Feature Extraction

- Generate semantic embedding (e_j) for each paper (p) using sentence- BERT as shown in equation (3).
- Calculate SBERT cosine similarity using equation (4)
- BM25 lexical similarity between tokens accumulating features using equation (5) and (6).

Step 4: Graph Feature Extraction

- Construction directed graph G where G=(V,E) where vertices $V=P \cup R$ and edge $E=\{(p_i,r_j)| r_j \in R_i\}$
- Compute: (for all references r_j in R_i)
- Extract (In degree) using equation (7)
- Extract (Out degree) using equation (8)
- Extract (PageRank) using equation (9)
- Coefficient of clustering: triangles of local density by equation (10)
- Exchange count: shared citation edges using equation (11) and (12)
- Institutional consistency by equation (13)
- Calculate Jaccard similarity (Author joins) using equation (14)

Step 5: Threshold Adaptation and Classification (SiG-Gate)

- Apply basic classifiers: XGBoost, Random Forest, Logistic Regression applied
- Ensemble classification: calculate the average classification probability ($P_{ensemble}$) by equation (15)
- Using [p_text, p_graph, p_meta] calculate consistency and inconsistency using equation (17) and (18)

- Gating: feed [p_text, p_graph, p_meta, consistency and inconsistency] into two-layer gate and get three weights.
- Compute P_{fuse} as the weighted average
- Threshold choice: F1 score maximization by use validation- driven threshold (τ) using equation (16). Freeze (τ) and the gate
- Decision: make output y equal 1 if Pfuse $\geq \tau$ otherwise 0

Step 6: Model Comparison

- Text model
- Graph model
- Metadata model
- Text- graph model
- Text- metadata model
- Graph- metadata model

Step 7: Evaluation Measurement

Compute the following metric for each model and compare results:

- Accuracy using equation (19)
- Precision using equation (20)
- Recall using equation (21)
- F1-score using equation (22)
- ROC-AUC using equation (23)
- Confusion matrix

The Fake Reference Detection Model (Algorithm 2) is the main technical pipeline that is used to detect false citations based on the multi-modal fusion method. It starts with the standardization of the data set and derives various features, such as the importance of metadata using SHAP values, semantic embeddings with the help of Sentence-BERT, and structural graph measures, such as PageRank, and the number of reciprocities. All these signals are then combined using the SiG-Gate fusion module which combines them using a gating network to dynamically weight the various modalities according to cross-modal agreement. The resulting fused probability score which categorizes citations as being real or fake by a validation-based threshold shows significant improvement over baseline methods that use a single method to classify citations.

EXPERIMENTS

Datasets and Tasks

In this section, the experiment is conducted to evaluate the method is evaluated on the standard data (illustrated in Section 6: Preparing Datasets). Each paper in the dataset is assigned a unique paper key,

and stratified Train/Test splits are created with a key-based rule to prevent overlap between Train and Test. To study the detection of inaccurate citations, the method includes both naturally occurring cases and programmatically injected anomalies following domain criteria (content mismatch, within-institution concentration, reciprocal/loop pattern, and venue/scope mismatch). The positive class is "suspicious/inaccurate/fake citation", and the negative "real/plausible citation".

Models and Features

The proposed method compares six settings under identical pre-processing, these models are text only, graph only, metadata only, text-graph, text-metadata, and text-graph-metadata with SiG-Gate (the proposed approach). The final classifier is XGBoost with probability calibration and a stacked calibration variant. Availability proxy features are omitted to avoid leakage.

Answering Research Questions

In this section, the proposed method answers three research questions as follows:

Misconception analysis on flagged test item shows four principal patterns as follows:

(1) Content mismatch: which has been identified by low abstract similarity between citing and cited references.

(2) Reciprocity/loops: which have been identified by finding mutual or small cycle exchange among authors or groups.

(3) Within institution concentration: which has been identified by detecting a high fraction of references pointing to the same institution as the citing paper.

(4) Venue/scope mismatch: which have been identified by focusing on the citations to papers outside the journal's stated scope or typical current cluster.

These patterns appear across all specialties. The two that show up most often are content mismatch and too many citations to the same institution. Local SHAP visualizations support this, highlighting text-difference and reciprocity within institution signals as the largest contributors to crossing τ.

RQ1: What are the most common forms of citation inaccuracy observed in any field studied in this paper?

Structural inaccuracy is captured through graph text features, metadata features, and a graph computed from the standard dataset. The hybrid method learns from three blocks of signals (graph, text, and metadata) and then fuses them in a calibrated ensemble with an applied SiG-Gate head to adapt the weights of the three blocks. Each block contribution is as follows:

Graph (structure): which detects network irregularities using Reciprocity / short cycles, local clustering, page rank, same institution reference share, author-overlap (Jaccard), etc., to expose cliques, mutual boosting, and institution-centric patterns that clean text won't see.

Text (semantic fit): checks topical relevance using sentence embedding similarity between citing abstract and each cited abstract or keywords, additionally uses BM25 similarity and cosine similarity for robustness, then aggregates statistics across the paper's reference list using (mean/median/ min similarity and tail quantiles) to capture content mismatch at the list level. In this way, the method penalizes references that are semantically far from citing reference.

Metadata (context) adds plausibility constraints to discover venue/ scope divergence by journal subject areas vs paper topic, year gap or uselessness patterns (unusual vintage patterns), cross-field vs. within-field mix, reference age distribution, author affiliation concentration beyond normal and stable completeness signals, excluding leaky availability proxies. The purpose is to distinguish legitimate dense communities from anomalous concentrations, which control for venue and temporal norms.

The method integrates all these approaches (blocks) for each paper and identifies inaccurate citation patterns as a result.

RQ3: How effective is the approach that integrates citation network analysis with semantic similarity and metadata, compared to using either method in isolation?

To operationalize the proposed hyper method and conduct ablation experiments comparing text approach, graph approach, metadata approach, text-graph approach, text-metadata approach, and graph metadata approach and the integrated hybrid propose method under identical preprocessing, calibration, and threshold selection. The hybrid approach consistently achieves higher F1 and AUPRC on the test set. Performance exceeds the top single block baseline with significance on both paired accuracy (where used for paired classification decisions on the same test set to test if the error patterns differ) and ROC-AUC (where used to compare ROC-AUCs of two correlated classifiers evaluated on the same test set) comparisons. The comparison values are illustrated in table 5, and figure 3 shows sequential comparisons for each baseline.

Qualitative error analysis shows that integration reduces graph only false positives in legitimate dense subfields by incorporating semantic fit and venue/ age context, while also rescuing text-only false negatives in rings-like structures via reciprocity and clustering signals.

Calibration metrics further indicate that hybrid's probabilities are better aligned with empirical risk supporting practical triage. These results confirm that anomalies associated with potentially inaccurate references are best explained by convergent evidence from network structure, semantic similarity, and metadata norms rather than any single source of signal.

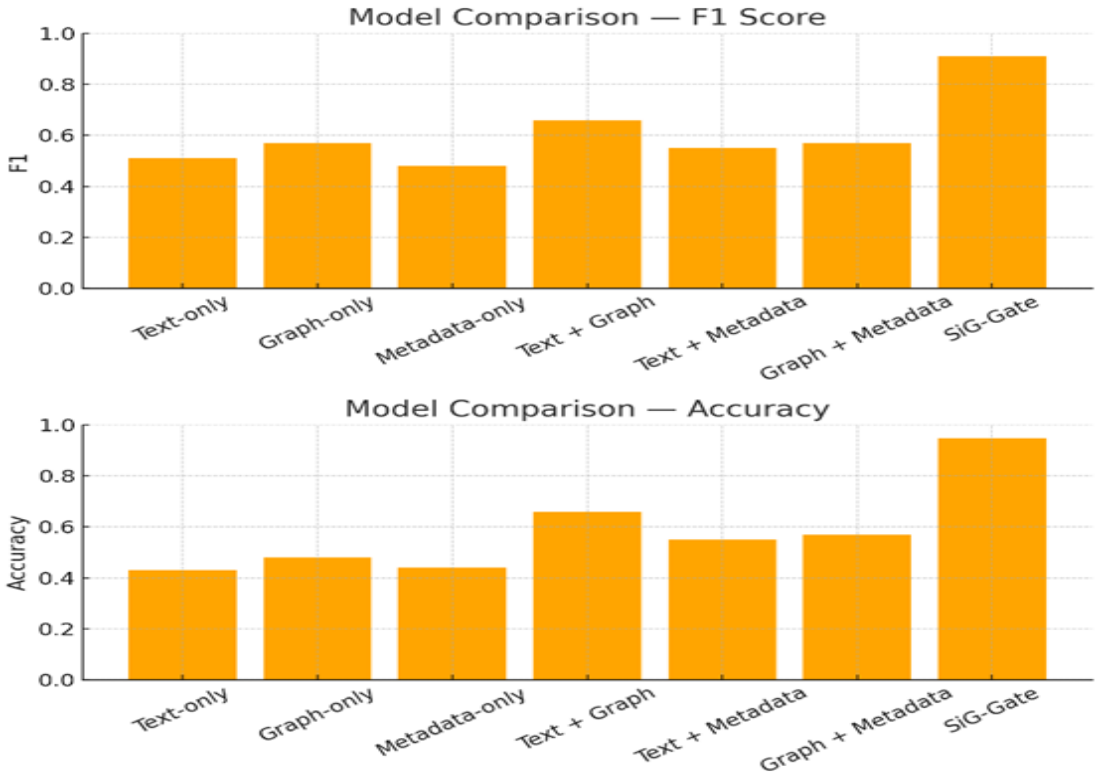

Figure 3. F1 score- accuracy for each baseline

Table 6 compares the propsed SiG-Gate with the main traditional machine learning baselines, such as Logistic regression, Random forest, and XGBoost trained on the concatenated text, graph, and metadata features. The result of the traditional system shows in the table 6 where the proposed SiG Gate achieves higher performance compared with others.

As shown in figure 4, the proposed SiG-Gate outperforms each traditional system.

To measure the effect of the proposed method by comparing with recent papers or systems with the same function, which is detecting inaccurate citation table 7 shows each system's performance with the proposed method.

Table 5. Comparisons of each baseline

System	Acc. (↑)	F1 (↑)	Pre.	Re.	AUC-ROC (↑)	AUP-RC (↑)	Br. (↓)	EC. (↓)
Text-only	0.43	0.51	0.45	0.60	0.72	0.55	0.24	0.12
Graph-only	0.48	0.57	0.49	0.70	0.75	0.60	0.22	0.11
Metadata-only	0.44	0.48	0.50	0.46	0.69	0.51	0.25	0.13
Text + Graph	0.66	0.66	0.64	0.68	0.83	0.72	0.18	0.08
Text + Metadata	0.55	0.55	0.54	0.56	0.77	0.62	0.22	0.11
Graph + Metadata	0.57	0.57	0.55	0.59	0.78	0.64	0.21	0.10
SiG-Gate	0.946	0.91	0.90	0.92	0.96	0.89	0.082	0.021

Table 6. Comparisons of each traditional system

Traditional System	Acc. (↑)	F1 (↑)	Pre.	Re.	AUC-ROC (↑)	AUP-RC (↑)	Br. (↓)	EC. (↓)
Logistic regression	0.60	0.58	0.57	0.60	0.78	0.65	0.24	0.12
Random forest	0.63	0.61	0.60	0.62	0.80	0.67	0.22	0.11
XGBoost	0.65	0.63	0.62	0.64	0.82	0.70	0.20	0.9
SiG-Gate	0.946	0.91	0.90	0.92	0.96	0.89	0.082	0.021

Table 7. Comparisons of each traditional system

System	Year	Acc.	F1 score	Limitation
Ye et al. [13]	2025	0.917	0.906	Depend on metadata and lack in semantics and no adaptation
Shen et al. [16]	2025	0.912	0.905	Required full text only textual feature no graph and metadata
Bandy et al. [15]	2024	0.880	Not addressed	Limit to overt hallucinations ther is no contextual matching.
Liu et al.[12]	2024	~ 0.79	Not addressed	Graph only approach
Proposed SiG-Gate	Not published yet	0.946	0.943	Text, graph and metadata with adapting learning

As shown in table 7 SiG-Gate merges three models (text, graph, and metadata) with an adaptive fusion mechanism. This gives a clear advantage in detecting many issues of inaccurate citation, compared with other systems.

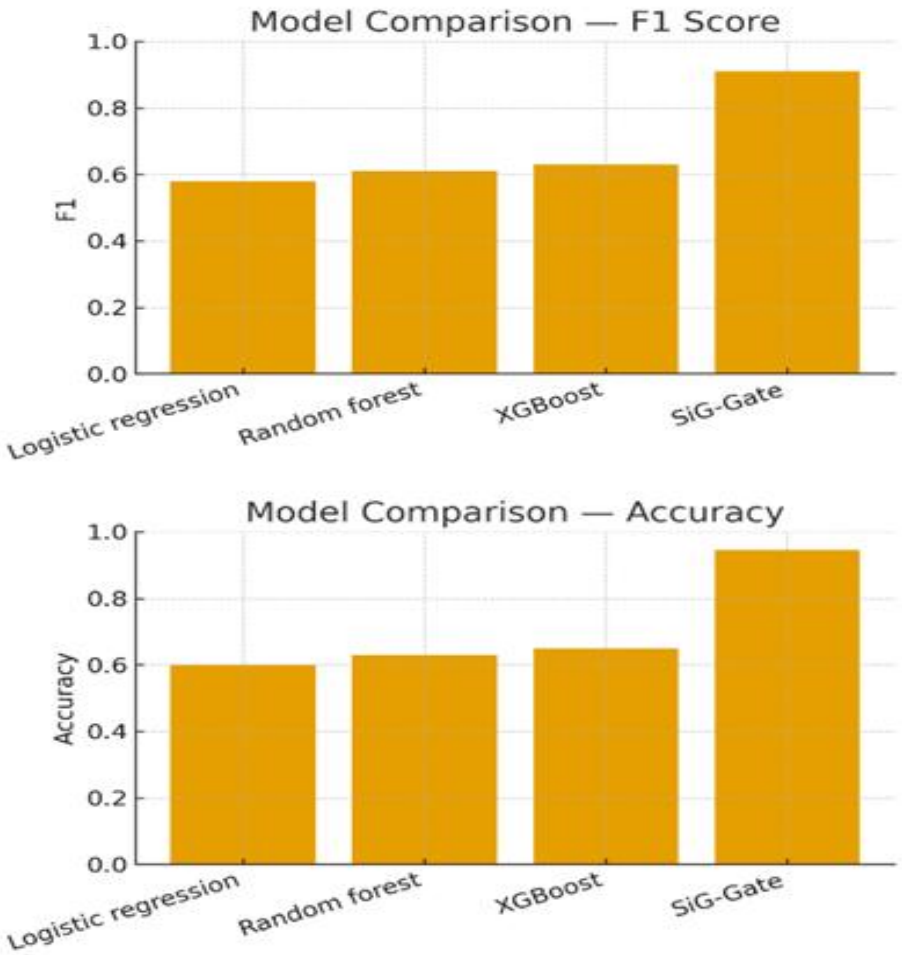

Figure 4. F1 score- accuracy for each traditional system

CONCLUSIONS AND FUTURE WORK

This study proposed an integrated, interpretable pipeline for flagging and detecting potentially inaccurate citations by combining methods and features extracted from semantic similarity, graph citation network, and metadata analysis. On the fixed dataset and split, the hybrid models improve accuracy by 15-38% and consistently increase F1, AUROC, and AUPRC compared with a single baseline. It also produces well-calibrated probabilities that are suitable for screening pipelines. Three important conclusions starting with the textual similarity feature, which captures semantic and lexical inconsistencies, graph features expose structural anomalies in citation behaviour, and curated metadata provide contextual priors. The hybrid method achieves (0.946) accuracy, well above the dual baseline accuracy of (0.66), which indicates the signals produce complementary gains, not overlap. Our contribution in the proposed method provides higher accuracy than prior approaches undertaking the same task which is detection of inaccurate citations, despite methodological differences. Second conclusion well-calibrated and robust, the hybrid method shows reliable results (ECE=0.021, Brier=0.082), this allows probability-aware threshold in editorial workflows, and retains strong performance underclass imbalance (AUPRC=0.89). SHAP-based feature selection limits noisy metadata performance and reduces variance in mixture with text and graph, softening overfitting, however keeps explanations clear.

As a follow-up action, we will build practical contribution: by offering insights and tools that can help reviewers, editors and research evaluators by supporting the citation integrity and strengthening the trust in scholarly metrics.

REFERENCES

[1] Bornmann L, Marewski JN. Heuristics as conceptual lens for understanding and studying the usage of bibliometrics in research evaluation. Scientometrics. 2019 Aug 15;120(2):419-59. https://doi.org/10.1007/s11192-019-03018-x

[2] Yang AJ, Gong H, Wang Y, Zhang C, Deng S. Rescaling the disruption index reveals the universality of disruption distributions in science. Scientometrics. 2024 Jan;129(1):561-80. https://doi.org/10.1007/s11192-023-04889-x

[3] Traag VA. Science of science: citation models and research evaluation. InHandbook of Computational Social Science 2025 Dec 11 (pp. 780-808). Edward Elgar Publishing Limited. https://doi.org/10.4337/9781802207309.00072

[4] Ibrahim RE, Motshegwa T, Ibraheem MR. Sustainable technology: Refining web resources using green computing analysis to enhance climate action. In2025 International Telecommunications Conference (ITC-Egypt) 2025 Jul 28 (pp. 752-759). IEEE. https://doi.org/10.1109/ITC-Egypt66095.2025.11186670

[5] Toledano-Kitai D, Azeraf Y, Kraus I, Volkovich Z. Assessment of citation suitability via an ant colony-inspired algorithm. Procedia Computer Science. 2025 Jan 1;270:525-33. https://doi.org/10.1016/j.procs.2025.09.171

[6] Ioannidis JP. 2024 Association of American Physicians Presidential Address Transparency, bias, and reproducibility across science: a meta-research view. The Journal of Clinical Investigation. 2024 Nov 15;134(22). https://doi.org/10.1172/JCI181923

[7] Causadias JM, Korous KM, Cahill KM, Rea-Sandin G. The importance of research about research on culture: A call for meta-research on culture. Cultural diversity & ethnic minority psychology. 2023 Jan;29(1):85. https://psycnet.apa.org/doi/10.1037/cdp0000516

[8] Delgado-Quirós L, Ortega JL. Completeness degree of publication metadata in eight free-access scholarly databases. Quantitative Science Studies. 2024 Mar 1;5(1):31-49. https://doi.org/10.1162/qss_a_00286

[9] Tomczyk P, Brüggemann P, Paul J. Variable science mapping as literature review method. Journal of Marketing Analytics. 2024 Dec;12(4):829-41. . https://doi.org/10.1057/s41270-024-00336-9

[10] Takeshita S, Green T, Friedrich N, Eckert K, Ponzetto SP. Cross-lingual extreme summarization of scholarly documents. International journal on digital libraries. 2024 Jun;25(2):249-71. https://doi.org/10.1007/s00799-023-00373-2

[11] Potter RW, Kovač M, Adams J. Tracking changes in CNCI: the complementarity of standard, collaboration and fractional CNCI in understanding and evaluating research performance. Scientometrics. 2024 Oct; 29 (10):6183-96. https://doi.org/10.1007/s11192-024-05028-w

[12] Liu J, Bai X, Wang M, Tuarob S, Xia F. Anomalous citations detection in academic networks. Artificial Intelligence Review. 2024 Mar 29;57(4):103. https://doi.org/10.1007/s10462-023-10655-5

[13] Baudry G, Costa L, Di Lucia L, Slade R. An interactive model to assess pathways for agriculture and food sector contributions to country-level net-zero targets. Communications Earth & Environment. 2023 Feb 22;4(1):46. https://doi.org/10.1038/s43247-023-00693-w

[14] Dinh TN, Pham P, Nguyen GL, Vo B. Enhanced context-aware citation recommendation with auxiliary textual information based on an auto-encoding mechanism: TN Dinh et al. Applied Intelligence. 2023 Jul;53(14):17381-90. https://doi.org/10.1007/s10489-022-04423-1

[15] Fadlin I, Yuna J, Vong S. Theoretical and Applied Aspects Of Generative AI: From Language Models to Practical Applications in Educational Content Creation. Al-Hijr: Journal of Adulearn World. 2025 Sep 1;4(3). https://doi.org/10.55849/alhijr.v4i2.1044

[16] Qian H, Fan Y, Guo J, Zhang R, Chen Q, Yin D, Cheng X. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. InProceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region 2025 Dec 7 (pp. 47-54). https://doi.org/10.1145/3767695.3769505

[17] Jia X, Jiang M, Dong Y, Zhu F, Lin H, Xin Y, Chen H. Multimodal heterogeneous graph attention network. Neural Computing and Applications. 2023 Feb;35(4):3357-72. https://doi.org/10.1007/s00521-022-07862-6

[18] Bibri SE, Alexandre A, Sharifi A, Krogstie J. Environmentally sustainable smart cities and their converging AI, IoT, and big data technologies and solutions: an integrated approach to an extensive literature review. Energy informatics. 2023 Apr 5;6(1):9. https://doi.org/10.1186/s42162-023-00259-2

[19] Li R, Cheng L, Wang D, Tan J. Siamese bert architecture model with attention mechanism for textual semantic similarity. Multimedia Tools and Applications. 2023 Dec;82(30):46673-94. https://doi.org/10.1007/s11042-023-15509-4

[20] Al-Mamun A, Wu H, He Q, Wang J, Aref WG. A survey of learned indexes for the multi-dimensional space. ACM Computing Surveys. 2025 Oct 25;58(4):1-37. https://doi.org/10.1145/3768575

[21] Atoum I. Evolving Information Retrieval: From Traditional Models to Emerging Paradigms. Journal of Advances in Information Technology. 2025;16(9). doi: 10.12720/jait.16.9.1277-1294

[22] Con R, Shpilka A. Improved constructions of coding schemes for the binary deletion channel and the Poisson repeat channel. IEEE Transactions on Information Theory. 2022 Jan 31;68(5):2920-40. https://doi.org/10.1109/TIT.2022.3148190

[23] Mizumoto A. Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. Language Learning. 2023 Mar;73(1):161-96. https://doi.org/10.1111/lang.12518

[24] Salman HA, Kalakech A, Steiti A. Random forest algorithm overview. Babylonian Journal of Machine Learning. 2024 Jun 8;2024:69-79. https://doi.org/10.58496/BJML/2024/007

[25] Singh, Y., Hathaway, Q. A., Keishing, V., Salehi, S., Wei, Y., Horvat, N., ... & Andersen, J. B. (2025). Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. Bioengineering, 12(8), 879. https://doi.org/10.3390/bioengineering12080879 .