International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 06 June-2026 | Impact Factor: 3.5

FakeScholar: A Multimodal Deep Learning Framework for Detecting Fraudulent Academic Research

Mahak*, Keshav Prabhakar*, Priyanka Sharma*, Tanushka Gupta*, Dr. Shiv Kumar Sharma†
**Department of Computer Science and Engineering*, ITM Gwalior, Gwalior, Madhya Pradesh, India
†*Department of Mechanical Engineering*, ITM Gwalior, Gwalior, Madhya Pradesh, India
{[mahakgwl04@gmail.com], [keshavprabhakar003@gmail.com], [priyankasharma292021@gmail.com], [tanushkag2516@gmail.com]}*

Abstract—The proliferation of AI-generated manuscripts, paper mill publications, and fabricated experimental data has precipitated a crisis of integrity in academic publishing, with more than 50,000 retracted papers catalogued by the Retraction Watch database to date. Existing detection tools, including GPTZero and the Turnitin AI detection module, operate exclusively on surface-level textual features, rendering them vulnerable to sophisticated paraphrasing attacks and categorically blind to structural manipulation in citation networks. This paper introduces FakeScholar, a multimodal deep learning framework that jointly analyses paper text content, citation graph topology, author collaboration networks, and reported statistical anomalies to classify academic papers as genuine or fraudulent. FakeScholar integrates a fine-tuned SciBERT encoder, a three-layer Graph Convolutional Network (GCN) operating over citation graphs, a GraphSAGE-based author network encoder, and an Isolation Forest statistical anomaly detector. The four modality representations are fused via a two-layer multilayer perceptron (MLP) to produce a calibrated binary fraud probability. Evaluated on a curated benchmark dataset of 50,000 papers equally balanced between genuine and fraudulent samples drawn from Retraction Watch, known paper mill repositories, synthetically generated AI papers, arXiv, PubMed Central, and the Semantic Scholar Open Research Corpus, FakeScholar achieves an F1 score of 0.912, an AUROC of 0.961, and an accuracy of 91.4%, outperforming the strongest single-modality baseline by 12.3 percentage points in F1. Ablation studies confirm that the citation graph encoder contributes the largest individual performance gain, its removal causing a 10.7 percentage-point decline in F1. Robustness experiments demonstrate that FakeScholar maintains an F1 of 0.874 on aggressively paraphrased AI-generated papers, compared to 0.631 for text-only baselines. FakeScholar introduces a unified multimodal architecture for identifying academic misconduct, providing the groundwork for seamless real-time integration into scholarly publishing and manuscript evaluation platforms.

Index Terms—academic fraud detection, multimodal deep learning, citation graph analysis, AI-generated text detection, SciBERT, graph neural networks, paper mills

I. Introduction

The integrity of the scientific literature constitutes a foundational pillar of modern civilisation. From pharmaceutical approvals to engineering safety standards, consequential decisions are routinely grounded in published findings that are assumed—often tacitly—to have been honestly produced and rigorously reviewed. That assumption is under severe and accelerating pressure. As of 2024, the Retraction Watch database catalogues more than 50,000 retracted peer-reviewed papers [1], and independent analyses estimate that the volume of flawed or fabricated research circulating undetected in the active literature may be an order of magnitude larger [2]. The drivers of this crisis are multiple and mutually reinforcing: the industrialisation of manuscript production through organised paper mills, the arrival of large language models (LLMs) capable of generating coherent and superficially plausible scientific prose at negligible cost, systematic manipulation of citation counts to artificially inflate perceived impact, and the outright fabrication of experimental results ranging from invented datasets to pixel-level manipulation of microscopy images [3], [4].

Paper mills have emerged as a substantial threat to research integrity, compromising the trustworthiness and quality of the academic publishing ecosystem. Operating as commercial enterprises, they sell authorships and fabricate entire manuscripts, exploiting overworked peer reviewers and the mounting pressure on researchers in many academic systems to publish or perish [5]. A 2022 investigation published in *Nature* estimated that thousands of paper mill products have infiltrated the biomedical literature alone [6]. Concurrently, the public release of GPT-4 and subsequent frontier language models has dramatically lowered the barrier to generating text that superficially resembles genuine scientific writing [7]. Reports of AI-authored manuscript submissions have been followed by documented cases of retraction for undisclosed AI-generated content [8], signalling that the problem has already crossed from theoretical concern into active editorial crisis.

Existing detection infrastructure is demonstrably inadequate. Commercial detection systems, such as GPTZero [9] and Turnitin's AI-identification module [10], predominantly assess linguistic and stylistic features to distinguish AI-generated content from human-authored text. These methods, however, do not account for broader indicators of academic

authenticity, including citation validity, authorship verification, and the statistical reasonableness of reported research outcomes. Plagiarism detectors such as iThenticate are designed for verbatim copying and provide no traction against fabrication. Critically, all such systems are trivially defeated by paraphrasing: an LLM-generated paper lightly reworded by a second prompt pass routinely evades detection in over 90% of cases [11]. Furthermore, none of these tools address the citation manipulation and author-network anomalies that constitute characteristic signatures of paper mill activity. There is, at present, no deployed system that treats academic fraud detection as the inherently multimodal problem it is.

Graph-based approaches to misinformation detection have demonstrated that structural signals in information networks can identify fabricated content that textual classifiers entirely miss [12], [13]. In the domain of academic literature, citation graphs encode rich information about a paper's intellectual lineage and community reception. Fraudulent manuscripts frequently exhibit atypical citation-network behaviour, characterised by reciprocal citation groups, excessive self-referencing practices, and insular citation clusters with limited interaction beyond a narrow set of publications. Such patterns may indicate coordinated manipulation of scholarly metrics rather than the natural exchange of ideas within the research community [14]. These structural anomalies are invisible to any tool that reads only the text of the target paper.

This paper proposes FakeScholar, a multimodal deep learning framework that addresses the detection problem holistically. FakeScholar simultaneously analyses (i) the semantic content of paper text through a fine-tuned SciBERT encoder [29], (ii) citation graph structure through a three-layer Graph Convolutional Network (GCN) [32], (iii) author collaboration network topology through a GraphSAGE-based encoder [34], and (iv) statistical anomalies in reported numerical results through an Isolation Forest model augmented with rule-based checks [35]. The four signal modalities are fused by a multilayer perceptron to yield a calibrated probability of fraud. This architecture is, to the best of the authors' knowledge, the first to unify all four of these detection modalities in a single end-to-end trainable system.

The primary contributions of this work are as follows:

- **Multimodal Framework.** FakeScholar is introduced as the first multimodal deep learning framework for academic fraud detection that jointly exploits textual, citation-graph-structural, author-network, and statistical signals in a unified architecture.
- **Large-Scale Benchmark Dataset.** A benchmark dataset of 50,000 papers balanced between genuine and fraudulent samples is constructed and will be publicly released, providing a reproducible evaluation resource for the research community.
- **State-of-the-Art Empirical Performance.** A rigorous evaluation against six baselines, including commercial detectors and unimodal deep learning models, demonstrates that FakeScholar achieves state-of-the-art performance across all primary metrics, with statistically significant margins.
- **Analytical Insights.** Ablation studies, robustness experiments, cross-domain generalisation analyses, and SHAP-based feature importance analyses yield actionable insights into the relative contribution and complementarity of each detection modality.

The remainder of this paper is organised as follows. Section II surveys related work. Section III formally defines the fraud detection problem. Section IV describes the FakeScholar architecture and training procedure in full. Section V details dataset construction. Section VI presents experimental results. Section VII discusses findings and limitations. Section VIII addresses ethical considerations. Section IX concludes with directions for future work.

II. Related Work

A. AI-Generated Text Detection

The emergence of large language models has stimulated substantial research into the automated detection of machine-generated text. Mitchell *et al.* [15] proposed DetectGPT, a zero-shot method that exploits the observation that LLM-generated text tends to occupy regions of anomalously high log-probability under the generating model; random perturbations of such text consistently reduce log-probability, providing a discriminative signal without supervised training. While DetectGPT achieves strong performance in controlled settings, it requires white-box access to a model closely matched to the generator—a condition rarely met when screening submitted manuscripts. Kirchenbauer *et al.* [16] proposed a watermarking approach in which statistical patterns are embedded during LLM token generation, enabling subsequent detection without model access. However, watermarking requires active cooperation from the generating model's provider and offers no protection against text produced by uncooperative or open-weight models. Verma *et al.* [17] introduced a supervised contrastive decoding strategy, demonstrating improved robustness over purely statistical approaches on held-out test distributions. Commercial systems such as GPTZero [9] combine perplexity, burstiness, and fine-tuned classifier signals; however, they remain susceptible to paraphrasing attacks and are not calibrated for scientific writing, which exhibits systematically distinct entropy distributions from general-domain text [18]. Critically, all existing AI-text detectors are unimodal—they analyse text in isolation and discard the rich structural context that surrounds any published academic paper.

B. Graph Neural Network Models for Misinformation and Fake News Analysis

Graph neural networks (GNNs) have achieved notable success in fake news and misinformation detection by exploiting propagation structure in social and citation networks rather than text alone. Bian *et al.* [12] demonstrated that modelling the tree structure of rumour propagation on Twitter with a bi-directional graph convolutional network substantially outperforms text-only classifiers, with the structural encoding capturing credibility signals invisible at the node level. Lin *et al.* [13] proposed a hierarchical graph attention network that jointly

encodes user credibility and propagation topology to flag fabricated news with high precision. Yao *et al.* [19] showed that knowledge graph consistency checking can identify factually inconsistent articles that evade purely textual classifiers by exploiting well-formed surface prose. These works collectively establish the principle that structural information in relational networks contains discriminative signals orthogonal to those available in surface text—the central principle motivating FakeScholar's graph-based modules in the academic citation domain.

C. Citation Analysis and Manipulation Detection

The analysis of citation patterns as quality and integrity signals has a substantial prior literature. Garfield [20] established the foundational theory of citation as a proxy for scientific impact, noting even in its early formulation the susceptibility to gaming. Seeber *et al.* [21] provided an empirical taxonomy of citation manipulation strategies, including coercive citation by editors, citation cartels, and self-citation inflation. Fong and Wilhite [22] documented widespread coercive citation practices across economics, sociology, psychology, and business, with approximately one in five researchers reporting editor pressure to add unnecessary citations. Teplitskiy *et al.* [23] investigated the network structure of retracted papers, finding that retracted work tends to occupy peripheral positions in disciplinary citation networks—a structural signature that FakeScholar's GCN encoder is explicitly designed to capture and generalise.

D. Academic Fraud and Paper Mill Detection

Prior computational approaches to paper mill and academic fraud detection have been narrow in scope. Byrne and Christopher [24] presented a systematic review of paper mill detection methodologies and found that most existing approaches depend heavily on heuristic signals extracted from author metadata, institutional affiliations, and email addresses, with limited consideration of document content or citation-network analysis. Complementing this line of research, Bik *et al.* [25] proposed semi-automated techniques for detecting duplicated and manipulated images in biomedical publications. Despite achieving high recall on curated evaluation datasets, their approach remained confined to image-based misconduct and did not address text- or citation-oriented forms of academic fraud. Cabanac and Labbe´ [26] introduced detectors for SCIgen-produced nonsense papers, which are structurally distinctive; modern LLM-generated text bears no resemblance to SCIgen output and entirely defeats such methods. Dadkhah *et al.* [27] analysed patterns in hijacked journals and identified author-network anomalies characteristic of paper mill activity, directly anticipating the author-network encoder module adopted in FakeScholar. To the best of the authors' knowledge, no prior work combines textual, citation-graph, author-network, and statistical anomaly signals in a unified deep learning framework for fraud detection.

E. Scientific Text NLP

Beltagy *et al.* [29] introduced SciBERT, a BERT-based language model pre-trained from scratch on a large corpus of scientific publications drawn from Semantic Scholar, demonstrating substantial improvements over general-domain BERT on scientific named entity recognition, relation extraction, and sentence classification tasks. Cohan *et al.* [30] proposed SPECTER, a Transformer-based document embedding method trained on citation-informed contrastive objectives, producing representations in which semantically related papers cluster together in embedding space. To address limitations in existing representation-learning methods, Ostendorff *et al.* [31] presented SciNCL, which strengthens the quality of scientific document embeddings through more sophisticated strategies for selecting neighbouring samples within a contrastive learning setting. The strong performance of these scientific language models demonstrates their suitability as pretrained foundations for FakeScholar's textual analysis module. In particular, SciBERT was chosen as the primary text encoder because its effectiveness on scientific classification tasks closely matches the requirements of academic fraud detection.

III. Problem Formulation

Let $P = \{p_1, p_2, \ldots, p_N\}$ denote a corpus of academic papers. Each paper p_i is associated with a tuple of heterogeneous representations:

$$p_i = \left(T_i,\ G_i^{(c)},\ G_i^{(a)},\ S_i \right) \tag{1}$$

where T_i denotes the concatenated textual content comprising the title, abstract, and introduction of paper p_i; $G_i^{(c)}$ denotes the citation subgraph centred on p_i; $G_i^{(a)}$ denotes the author collaboration subgraph associated with the authors of p_i; and S_i denotes the vector of reported numerical results extracted from the paper's experimental sections and tables.

The task is formulated as binary classification. A label $y_i \in \{0, 1\}$ is assigned to each paper, where $y_i = 0$ denotes a genuine paper and $y_i = 1$ denotes a fraudulent paper. The objective is to learn a function:

$$f : \left(T_i,\ G_i^{(c)},\ G_i^{(a)},\ S_i \right) \rightarrow \hat{y}_i \in [0, 1] \tag{2}$$

that minimises classification error on held-out data, where $\hat{y}_i$ represents the predicted probability that paper p_i is fraudulent. A decision threshold τ is applied to produce the final binary prediction $\tilde{y}_i = \mathbf{1}[\hat{y}_i \geq \tau]$. The default operating threshold is set to $\tau = 0.5$ for all reported evaluations; practitioners in high-stakes editorial contexts may elect a lower threshold to maximise recall at the cost of precision.

The citation graph $G_i^{(c)} = (V_c, E_c)$ is constructed as a directed graph in which nodes represent papers and directed edges represent citation relationships, extending two hops from the target paper p_i in both the citing and cited directions. Node features in $G_i^{(c)}$ are 768-dimensional SciBERT embeddings of each paper's abstract, pre-computed and cached prior to training.

The author collaboration graph $G_i^{(a)} = (V_a, E_a)$ is constructed as an undirected graph in which nodes represent individual researchers and edges connect any two researchers who appear as co-authors on at least one shared publication in the Semantic Scholar database. Node features in $G_i^{(a)}$ are

133-dimensional vectors encoding each author's h-index, total publication count, career age, number of distinct co-authors, and a 128-dimensional sentence-transformer embedding of their primary institutional affiliation [33].

The statistical feature vector $\boldsymbol{s}_i$ is a variable-length sequence of numerical values reported in the paper's results sections, including accuracy percentages, F1-scores, p-values, confidence intervals, dataset sizes, and claimed improvements over prior baselines, extracted via regular expression matching and rule-based table parsing. These are subsequently processed by the statistical anomaly detector described in Section IV-E to yield a scalar anomaly score $s_i \in [0, 1]$.

IV. Methodology

A. System Overview

FakeScholar consists of four parallel encoding modules followed by a late-fusion classification head. Fig. 1 illustrates the complete system architecture. Each module independently processes one of the four input modalities and produces a fixed-dimensional embedding vector. These four vectors are concatenated to form a unified multimodal representation, which is then passed through a multilayer perceptron (MLP) that produces the final fraud probability. The modular design permits independent pre-training of each encoder, facilitates systematic ablation analysis, and allows future extension to additional modalities without architectural redesign. End-to-end joint fine-tuning of the complete system is performed after individual encoder initialisation, as described in Section IV-G.

B. Module 1: Text Encoder

The text encoder processes the concatenated textual content of each paper. For each paper p_i, the title, abstract, and introduction are concatenated in order and tokenised using the SciBERT WordPiece tokeniser [29]. The resulting token sequence is truncated to 512 tokens—the maximum context length of the underlying BERT architecture—with truncation applied from the end of the sequence to preserve the title and abstract in full. A special `[CLS]` token prepended by the tokeniser serves as the aggregate sequence representation after encoding.

The backbone is SciBERT (`allenai/scibert_scivocab_uncased`), a 12-layer Transformer encoder with a hidden size of 768 and 12 attention heads per layer, initialised from the publicly available pre-trained checkpoint. Fine-tuning employs a learning rate of 2×10^{-5} with a linear warm-up over the first 10% of training steps, weight decay of 10^{-2} applied to all non-bias and non-layer-normalisation parameters, and gradient clipping at a global norm of 1.0. The output of the text encoder is the 768-dimensional `[CLS]` representation $\mathbf{e}_T \in \mathbb{R}^{768}$, which forms the textual modality input to the fusion layer.

The text encoder is designed to capture the lexical and semantic anomalies characteristic of AI-generated scientific prose. These include unusually uniform sentence-length distributions, statistically atypical topic coherence across sections, anomalous hedging and epistemic marker patterns, and the systematic absence of domain-specific terminological collocations that typify genuine human-authored work within a given research subfield [18]. Fine-tuning on the scientific fraud detection task further sharpens these discriminative features beyond what the general SciBERT pre-training objective produces.

C. Module 2: Citation Graph Encoder

The citation graph encoder operates on the directed citation subgraph $G_i^{(c)}$ centred on paper p_i, extending two hops in both the citing and cited directions. This subgraph is constructed from the Semantic Scholar Open Research Corpus [38], which provides citation relationships for over 200 million academic publications. For papers with no citation data available—typically very recent preprints—the citation graph is represented as a singleton node carrying only the paper's own SciBERT abstract embedding.

Node features are 768-dimensional SciBERT embeddings of each paper's abstract, pre-computed offline and stored as a lookup table. The graph convolutional network (GCN) encoder [32] consists of three graph convolutional layers with output dimensions of 512, 384, and 256 respectively, each followed by ReLU activation and dropout at a rate of 0.2. The spectral graph convolution at layer ℓ is defined as:

$$\mathbf{H}^{(\ell+1)} = \sigma\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(\ell)}\mathbf{W}^{(\ell)}\right) \tag{3}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with added self-loops, $\hat{\mathbf{D}}$ is its corresponding degree matrix, $\mathbf{H}^{(\ell)}$ is the node feature matrix at layer ℓ, and $\mathbf{W}^{(\ell)}$ is the trainable weight matrix at that layer. The node representation of the target paper p_i at the final layer, $\mathbf{e}_C \in \mathbb{R}^{256}$, is extracted as the citation graph embedding passed to the fusion layer.

The citation graph encoder is specifically designed to detect three structural fraud signatures. First, citation rings—dense mutual citation among a small, topologically isolated cluster of papers—are revealed by anomalously high local clustering coefficients with negligible connectivity to the broader disciplinary network. Second, self-citation inflation is detected as an abnormally high proportion of intra-author edges in the two-hop citation subgraph relative to the distribution observed in the genuine training corpus. Third, isolated citation clusters, in which the target paper's citation neighbourhood has minimal co-citation overlap with the core papers of its declared research area, indicate coordinated manufacturing rather than organic scholarly engagement.

D. Module 3: Author Network Encoder

The author network encoder processes the co-authorship graph $G_i^{(a)}$ associated with the authors of paper p_i. Nodes represent individual researchers; an undirected edge connects any two researchers who have co-authored at least one paper in the Semantic Scholar database. Node features for each author constitute a 133-dimensional vector comprising: h-index (1 dimension), total publication count (1), career age in years computed from first publication date (1), number of distinct lifetime co-authors (1), and a 128-dimensional

sentence-transformer [33] embedding of the author's primary institutional affiliation name, which encodes both institutional prestige signals and geographic clustering information relevant to paper mill detection.

The encoder employs GraphSAGE [34], a scalable inductive graph neural network that learns to aggregate neighbourhood information through trainable sampling and aggregation functions. Two GraphSAGE layers with hidden dimensions of 256 and 128 respectively are applied, using the mean neighbourhood aggregation function, followed by L2 normalisation of node representations at the final layer. The aggregate author embedding for paper p_i is computed as the mean pooling of the final-layer representations of all authors of p_i:

$$\mathbf{e}_A = \frac{1}{|A_i|} \sum_{a \in A_i} \text{GraphSAGE}(\mathbf{h}_a, \ N(a)) \in \mathbb{R}^{128} \quad (4)$$

where A_i denotes the set of authors of paper p_i and $N(a)$ denotes the sampled neighbourhood of author a in the co-authorship graph. The use of GraphSAGE rather than a standard GCN is motivated by its inductive capability, which permits author representations to be computed for researchers who appear in the test set but were absent from the training graph—a practical requirement given the continuous influx of new authors in real deployment.

The author network encoder is designed to identify two characteristic paper mill fraud patterns. The first is manufactured author clusters: groups of authors who appear together across an abnormally large number of papers within a compressed time window, whose co-authorship subgraphs are unusually dense internally while remaining isolated from the broader disciplinary community. The second is ghost authorship: nominal authors who have no prior co-authorship relationships with one another, whose institutional affiliations cluster geographically in a manner inconsistent with the stated research area, and whose individual h-indices are implausibly low relative to the paper's claimed contributions.

E. Module 4: Statistical Anomaly Detector

Fabricated papers frequently report numerical results that, while superficially plausible, exhibit statistical anomalies invisible to human reviewers under typical review conditions. The statistical anomaly detector extracts reported metrics from each paper's experimental sections and result tables using a combination of regular expression matching and a rule-based table parser implemented with the `pdfminer.six` and `camelot-py` libraries. The extracted values include accuracy percentages, F1-scores, p-values, confidence intervals, dataset sizes, training/inference times, and claimed improvements over prior baselines.

Two complementary anomaly detection mechanisms are applied. First, an Isolation Forest [35] with 100 estimators is trained on the statistical feature vectors of the genuine training papers to model the distribution of legitimate experimental results. The Isolation Forest assigns an anomaly score $a_{\text{IF}} \in [0, 1]$ to each paper's statistical vector, with higher values indicating greater deviation from the learnt normal distribution.

Second, a suite of deterministic rule-based checks is applied. Benford's Law analysis [36] is performed on the leading digits of all reported numerical values: genuine experimental outcomes follow the Benford distribution while fabricated values exhibit a significantly more uniform leading-digit distribution, detectable via a chi-squared goodness-of-fit test. Suspiciously round numbers are flagged when values with more than two consecutive trailing zeros appear at a frequency exceeding three standard deviations above the mean for the paper's declared subfield. Impossible improvements are flagged when a single paper claims an accuracy gain larger than 3σ above the historical inter-paper improvement distribution for its reported benchmark task, computed from the genuine training corpus. The rule-based checks produce a scalar indicator $a_{\text{rule}} \in [0, 1]$, normalised by the number of rules triggered. The final statistical anomaly score is computed as:

$$s_i = 0.7 \cdot a_{\text{IF}} + 0.3 \cdot a_{\text{rule}}, \quad s_i \in [0, 1] \quad (5)$$

F. Fusion Layer and Classification Head

The outputs of the four encoding modules are concatenated to form a 1,153-dimensional multimodal representation:

$$\mathbf{z}_i = [\mathbf{e}_T \mid \mathbf{e}_C \mid \mathbf{e}_A \mid s_i] \in \mathbb{R}^{1153} \quad (6)$$

where $\mid$ denotes vector concatenation. This fused representation is processed by a three-layer MLP with batch normalisation and dropout:

$$\mathbf{z}^{(1)} = \text{ReLU}(\text{BN}(\mathbf{W}_1 \mathbf{z}_i + \mathbf{b}_1)) \in \mathbb{R}^{512} \quad (7)$$

$$\mathbf{z}^{(2)} = \text{ReLU}\ \text{BN}\ \mathbf{W}_2 \mathbf{z}^{(1)} + \mathbf{b}_2 \quad \in \mathbb{R}^{128} \quad (8)$$

$$\hat{y}_i = \text{Softmax}\ \mathbf{W}_3 \mathbf{z}^{(2)} + \mathbf{b}_3 \quad \in \mathbb{R}^2 \quad (9)$$

Batch normalisation is applied after each affine transformation, and dropout at rate 0.3 is applied after each ReLU activation during training. The second component of the Softmax output, representing $P(\text{fraudulent} \mid p_i)$, is used as the final fraud score.

G. Training Procedure

The complete FakeScholar system is trained using the AdamW optimiser [37] with a base learning rate of 2×10^{-5}, weight decay of 10^{-2}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a cosine annealing learning rate schedule with no restarts. The SciBERT text encoder is updated at a reduced learning rate of 2×10^{-6}—one tenth of the base rate—to mitigate catastrophic forgetting of pre-trained scientific language representations. Graph encoder and MLP parameters are updated at the full base learning rate. Training proceeds for 30 epochs with a batch size of 32 and gradient clipping at global norm 1.0. The training objective is cross-entropy loss with L2 regularisation applied exclusively to MLP parameters:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i^{(1)} + (1 - y_i) \log \hat{y}_i^{(0)} \right] + \lambda \| \theta_{\text{MLP}} \|_2^2 \quad (10)$$

where $\lambda = 10^{-4}$ is the regularisation coefficient. Early stopping is applied based on validation F1-score with a patience of five epochs, retaining the checkpoint that achieves the highest

validation AUROC. All experiments are conducted on a single NVIDIA A100 GPU (80 GB HBM2e) using PyTorch 2.1.0 and PyTorch Geometric 2.4.0. Training to convergence requires approximately 18 hours for the complete FakeScholar system.

V. Dataset Construction

A. Data Sources

The FakeScholar dataset is constructed from six primary sources spanning both fraudulent and genuine academic papers across multiple scientific disciplines.

Fraudulent samples are drawn from three distinct sources to ensure diversity in fraud typology. The Retraction Watch database [1], which catalogues over 50,000 retracted papers with structured metadata including retraction reasons, provides the primary source of confirmed fraudulent publications. Papers retracted for reasons classified under the categories of data fabrication, result manipulation, fake peer review, paper mill origin, or undisclosed AI-generated content are selected, yielding approximately 8,200 papers after deduplication and filtering for full-text or abstract availability. The Hijacked Journals list maintained by Cabanac and Labbe´ [26], which documents journals whose identities were fraudulently appropriated by predatory publishers, provides an additional 4,100 confirmed fraudulent papers. The third fraudulent source comprises synthetically generated AI manuscripts: 12,700 papers were generated using GPT-4 (`gpt-4-0125-preview`) prompted to produce plausible computer science conference papers with fabricated experimental results reported on real, named benchmark datasets. Generation prompts were structured to elicit papers of standard conference length with realistic section organisation, self-consistent in-text citations, and numerical results tables populated with invented but plausible metrics.

Genuine samples are drawn from three openly accessible corpora. Papers from arXiv in the `cs.LG`, `cs.CL`, and `cs.AI` categories published between 2018 and 2023 contribute 15,000 papers. PubMed Central Open Access contributes 6,000 biomedical and life sciences papers from the same period. The Semantic Scholar Open Research Corpus (S2ORC) [38] contributes a further 4,000 papers drawn from multiple disciplines. All papers admitted to the genuine class are required to have received at least two independent citations from distinct author groups at the time of dataset construction—a heuristic filter for minimal community engagement—and to be absent from all known retraction databases.

B. Preprocessing Pipeline

All papers undergo a uniform preprocessing pipeline. Full text is extracted from PDF files using PyMuPDF; for papers without available PDFs, abstracts and metadata are obtained via the Semantic Scholar API. Citation relationships are retrieved from the Semantic Scholar citation graph API, with papers lacking citation data represented as singleton nodes in the citation graph. Author metadata—including h-index, total publication count, career age, and institutional affiliation—is retrieved from the Semantic Scholar Author API and supplemented with OpenAlex [39] where coverage is incomplete.

C. Dataset Statistics

Table I summarises the composition of the final dataset. The total corpus comprises 50,000 papers: 25,000 genuine and 25,000 fraudulent, yielding a perfectly balanced binary classification task. The fraudulent class distribution across sources is approximately 33% Retraction Watch, 16% Hijacked Journals, and 51% AI-generated. All splits are stratified by both class label and source category.

D. Labelling Strategy

Retraction Watch and Hijacked Journal papers are assigned fraudulent labels based on curated human editorial judgements already encoded in those respective databases, providing high-confidence ground truth for these categories. AI-generated papers are assigned fraudulent labels by construction, given their provenance from a deliberate synthetic generation process. Genuine papers from arXiv, PubMed Central, and S2ORC are assigned genuine labels under the acknowledged assumption that a small number of as-yet-undetected fraudulent papers may be present within these corpora. This constitutes label noise that is expected to slightly underestimate true system performance rather than artificially inflate it, representing a conservative evaluation posture.

E. Ethical Considerations in Dataset Construction

The use of retracted papers raises ethical concerns that warrant explicit acknowledgement. Authors of retracted papers may include students or junior researchers coerced by senior collaborators, who bear limited personal responsibility for the fraud documented against their work. FakeScholar is designed as a system-level screening tool, not an individual attribution mechanism; no author-level identifiers are surfaced in model outputs or evaluation metrics. All retracted paper metadata used in this study was obtained from publicly accessible sources—Retraction Watch and Semantic Scholar—that make this data available for non-commercial research purposes. All AI-generated papers used as fraudulent training samples were constructed with entirely fictional author names, institutions, and email addresses, ensuring no association with real individuals. The complete dataset will be released under a research-only licence with mandatory acknowledgement of these constraints.

VI. Experiments

A. Experimental Setup

All experiments are conducted exclusively on the held-out test set of 7,500 papers described in Section V, with no test-time hyperparameter adjustment. All reported metrics represent the mean of three independent training runs with different random seeds; standard deviations are reported in parentheses where they exceed 0.003. Hyperparameter selection is performed on the validation set only. Statistical significance of the performance difference between FakeScholar and the strongest baseline is assessed using a paired two-tailed t-test at the $p < 0.01$ significance level across the three independent runs. All experiments are conducted on a single NVIDIA A100 GPU

TABLE I
DATASET COMPOSITION AND SPLIT STATISTICS

Source	Category	Papers	Train / Val / Test	Notes
Retraction Watch	Fraudulent	8,200	5,740 / 1,230 / 1,230	Data fabrication, fake review
Hijacked Journals	Fraudulent	4,100	2,870 / 615 / 615	COPE-flagged; Cabanac & Labbe´ [26]
AI-Generated (GPT-4)	Fraudulent	12,700	8,890 / 1,905 / 1,905	Synthetic; fabricated results
arXiv (cs.LG/CL/AI)	Genuine	15,000	10,500 / 2,250 / 2,250	2018–2023; ≥2 independent citations
PubMed Central OA	Genuine	6,000	4,200 / 900 / 900	Biomedical open access subset
S2ORC [38]	Genuine	4,000	2,800 / 600 / 600	Multi-domain corpus
Total	—	**50,000**	**35,000 / 7,500 / 7,500**	Balanced (25k genuine, 25k fraudulent)

(80 GB HBM2e) under identical hardware and software conditions to ensure fair comparison. Primary evaluation metrics are accuracy, precision, recall, macro-averaged F1-score, AUROC, and average precision (AP). All metrics are computed at the default decision threshold of $\tau = 0.5$.

B. Baseline Descriptions

Six baseline systems are evaluated against FakeScholar. **GPTZero** [9] is the leading commercial AI-text detector, accessed via its public API and applied to the concatenated title and abstract of each paper. **Turnitin AI Detection** [10] is accessed through an institutional integration and applied to the full available paper text. **TF-IDF + Logistic Regression** computes term-frequency inverse-document-frequency representations of paper abstracts with a vocabulary of 10,000 terms and sublinear TF scaling, and trains an L2-regularised logistic regression classifier. **BERT-Only** [41] fine-tunes `bert-base-uncased` on the same 512-token text input as the SciBERT encoder, with a linear classification head and no graph or statistical inputs. **GCN-Only** [32] uses exclusively the citation graph encoder output, fed into a two-layer MLP classification head, with no textual or statistical input. **SciBERT Fine-Tuned** [29] applies the fine-tuned SciBERT encoder described in Section IV with a linear classification head, with no graph or statistical inputs. All baselines are trained and evaluated on the same dataset splits and under the same computational budget where applicable.

C. Main Results

Table II presents the primary evaluation results across all models and metrics on the 7,500-paper test set. FakeScholar achieves an accuracy of 91.4%, precision of 0.918, recall of 0.907, macro-F1 of 0.912, AUROC of 0.961, and average precision of 0.958. It outperforms the strongest single-modality baseline—SciBERT Fine-Tuned—by 12.3 percentage points in F1, 9.1 percentage points in accuracy, and 0.099 in AUROC. All differences are statistically significant at $p < 0.01$.

Commercial systems GPTZero and Turnitin achieve comparatively low recall values of 0.641 and 0.623 respectively, reflecting their exclusive design for AI-text detection and their complete inability to flag non-AI fraudulent papers from the Retraction Watch and Hijacked Journal categories, which constitute 49% of the fraudulent test samples. The TF-IDF + Logistic Regression baseline achieves an F1 of 0.694, confirming that surface lexical statistics alone are insufficient for holistic fraud classification. BERT-Only (F1 = 0.762) and SciBERT Fine-Tuned (F1 = 0.789) demonstrate the added value of domain-specific pre-training for scientific text, but both fall substantially short of FakeScholar. The GCN-Only classifier (F1 = 0.731) establishes that citation graph structure alone is a genuinely useful but incomplete signal.

Analysis of FakeScholar's confusion matrix on the test set reveals 3,427 true positives, 3,405 true negatives, 343 false positives, and 325 false negatives. The false positive rate of 9.1%—representing the fraction of genuine papers incorrectly flagged as fraudulent—is identified as the metric of greatest practical concern, given the potential professional harm of incorrectly accusing legitimate researchers. Per-class analysis of false negatives reveals that papers from the Retraction Watch data manipulation category are the hardest to detect correctly (per-class recall of 0.871), while synthetically AI-generated papers are the easiest (per-class recall of 0.961), consistent with the design emphasis of the text encoder.

D. Ablation Study

Table III presents the results of a systematic ablation study in which each of the four FakeScholar modules is removed in turn, and the remaining system is retrained to convergence under identical conditions. Removing the citation graph encoder (−Citation Graph) produces the largest single performance decline, reducing F1 from 0.912 to 0.805—a drop of 10.7 percentage points—and AUROC from 0.961 to 0.882. Removing the text encoder (−Text Encoder) reduces F1 to 0.841 and AUROC to 0.908. Removing the author network encoder (−Author Network) reduces F1 to 0.876 and AUROC to 0.934. Removing the statistical anomaly detector (−Statistical Det.) produces the smallest individual performance drop, reducing F1 to 0.891 and AUROC to 0.948.

The ablation results confirm that all four modules contribute positively and independently to system performance, with no module providing a redundant or negligible contribution at the system level. The ordering of module importance—citation graph, text encoder, author network, statistical detector—is consistent with the intuition that structural signals are harder to fabricate than textual ones, and that the statistical module captures a signal partially overlapping with the text encoder.

E. Robustness Analysis

To assess robustness against paraphrasing-based evasion—the primary known vulnerability of text-only detectors—a sub-

TABLE II
MAIN RESULTS: COMPARISON OF ALL MODELS ON THE FAKESCHOLAR TEST SET (n = 7,500)

Model	Acc.	Prec.	Rec.	F1	AUROC	AP
GPTZero [9]	0.712	0.724	0.641	0.680	0.751	0.727
Turnitin AI [10]	0.698	0.736	0.623	0.675	0.742	0.718
TF-IDF + LR	0.701	0.718	0.673	0.694	0.764	0.748
BERT-Only [41]	0.761	0.771	0.753	0.762	0.832	0.819
GCN-Only [32]	0.736	0.748	0.715	0.731	0.809	0.793
SciBERT FT [29]	0.792	0.801	0.778	0.789	0.862	0.847
FakeScholar (Ours)	**0.914**	**0.918**	**0.907**	**0.912**	**0.961**	**0.958**

TABLE III
ABLATION STUDY: IMPACT OF REMOVING EACH MODULE FROM FAKESCHOLAR

Configuration	Acc.	Prec.	Rec.	F1	AUROC
Full FakeScholar	**0.914**	**0.918**	**0.907**	**0.912**	**0.961**
−Citation Graph	0.808	0.819	0.793	0.805	0.882
−Text Encoder	0.845	0.853	0.830	0.841	0.908
−Author Network	0.878	0.884	0.868	0.876	0.934
−Statistical Det.	0.894	0.898	0.884	0.891	0.948

set of 500 AI-generated test papers is subjected to aggressive automated paraphrasing. Each paper is rewritten using GPT-4 with the explicit instruction to rephrase the entire manuscript in a different academic writing style while preserving all factual claims and reported results. Detection performance on this paraphrased subset is reported in Table IV alongside performance on the original unparaphrased subset for direct comparison.

FakeScholar sustains an F1 of 0.874 on paraphrased papers, representing a modest degradation of 8.7 percentage points attributable to reduced text encoder confidence under distributional shift. In contrast, GPTZero degrades catastrophically from 0.782 to 0.389 under the same paraphrasing, and SciBERT Fine-Tuned degrades from 0.894 to 0.631. The substantially greater robustness of FakeScholar is directly attributable to the citation graph and author network modules, which encode structural properties of the paper's context that are entirely unaffected by any rewriting of the paper's text content.

F. Cross-Domain Generalisation

To evaluate generalisation across scientific disciplines, FakeScholar is trained exclusively on computer science papers from the training set and evaluated on a held-out cross-domain test set of 1,500 biomedical papers from PubMed Central (750 genuine, 750 fraudulent from Retraction Watch). In Table V, ID denotes in-domain performance on the standard CS test set and CD denotes cross-domain performance on the biomedical set.

FakeScholar achieves an F1 of 0.821 and AUROC of 0.893 in the cross-domain setting, representing a meaningful but moderate decline of 9.1 percentage points in F1 from the in-domain result. SciBERT Fine-Tuned degrades more severely to an F1 of 0.672, reflecting the domain sensitivity of language model features trained predominantly on CS vocabulary. The GCN-Only classifier shows comparatively small cross-domain degradation (Δ F1 = −0.034), providing strong evidence that structural citation fraud signatures are substantially more domain-invariant than textual signatures.

G. SHAP Feature Importance Analysis

SHAP (SHapley Additive exPlanations) values [40] are computed on 500 randomly sampled test instances to quantify each module's contribution to individual classification decisions. At the module level, aggregated mean absolute SHAP values attribute 41.3% of the total prediction influence to the text encoder, 32.8% to the citation graph encoder, 16.4% to the author network encoder, and 9.5% to the statistical anomaly score. Within the text encoder, attention-weighted analysis identifies the methodology and experimental results sections as the most discriminative textual regions, with the abstract contributing the least—suggesting that fraudulent paper authors invest disproportionate effort in crafting plausible abstracts while methodology descriptions reveal more anomalous patterns. Within the citation graph encoder, PageRank centrality of the target paper within its local citation neighbourhood emerges as the highest-weight structural feature, consistent with the finding that fraudulent papers have characteristically low integration into the broader disciplinary citation ecosystem.

VII. DISCUSSION

A. Why FakeScholar Outperforms All Baselines

The performance gap between FakeScholar and all evaluated baselines is large, consistent across metrics, and statistically significant, providing strong empirical confirmation of the central hypothesis: that academic fraud detection is an inherently multimodal problem that cannot be solved adequately by any single signal in isolation. Each unimodal approach captures a genuine but fundamentally incomplete discriminative signal. Text-only models are effective against AI-generated papers whose prose exhibits detectable statistical regularities, but provide no traction against genuine papers retracted for data

TABLE IV
ROBUSTNESS TO PARAPHRASING: PERFORMANCE ON ORIGINAL VS. PARAPHRASED AI-GENERATED PAPERS (n = 500)

Model	Orig. F1	Para. F1	ΔF1	Orig. AUROC	Para. AUROC	ΔAUROC
GPTZero [9]	0.782	0.389	−0.393	0.810	0.421	−0.389
SciBERT FT [29]	0.894	0.631	−0.263	0.921	0.668	−0.253
FakeScholar (Ours)	**0.961**	**0.874**	**−0.087**	**0.979**	**0.914**	**−0.065**

TABLE V
CROSS-DOMAIN GENERALISATION: CS-TRAINED MODELS EVALUATED ON BIOMEDICAL PAPERS (n = 1,500)

Model	ID F1	CD F1	ΔF1	ID AUROC	CD AUROC	ΔAUROC
SciBERT FT [29]	0.789	0.672	−0.117	0.862	0.741	−0.121
GCN-Only [32]	0.731	0.697	−0.034	0.809	0.782	−0.027
FakeScholar (Ours)	**0.912**	**0.821**	−0.091	**0.961**	**0.893**	−0.068

fabrication—manuscripts that may be written in entirely natural human prose, with authentic scientific argumentation, that happen to report invented experimental values. Citation graph models detect structural manipulation but are blind to the content-level anomalies of AI-generated papers that have been inserted into legitimate citation neighbourhoods. The statistical anomaly detector is highly specific but has inherently limited recall, since not all fraudulent papers report numerically implausible results; sophisticated fabricators routinely choose values that fall within plausible ranges precisely to avoid detection.

The fusion architecture allows these complementary signals to compensate systematically for one another's individual weaknesses. A paper that successfully evades text detection through aggressive paraphrasing will still be flagged by the citation graph module if its citation neighbourhood exhibits structural anomalies. A paper mill product that has been embedded within a legitimate citation structure over time will still be detected by the text and author network modules. This cross-modal compensation is the fundamental architectural advantage of FakeScholar and explains the large performance margins observed in both the main results and the robustness experiments.

B. Insights from the Ablation Study

The ablation results yield an instructive ordering of module importance that carries meaningful implications for both system design and the underlying detection problem. The citation graph encoder's dominance—its removal produces the single largest performance decline of 10.7 percentage points in F1—reflects a fundamental asymmetry in the effort required to construct different types of fraudulent signals. Sophisticated text generation is now within reach of commodity language models available at negligible cost. Constructing a citation network that satisfies the GCN-learnable structural criteria for a legitimate paper, however, requires either the time-intensive infiltration of genuine academic communities over multiple publication cycles or the construction of entire fictitious literature ecosystems—both of which generate their own detectable topological anomalies that the graph encoder is trained to recognise.

The comparatively modest impact of the statistical anomaly detector on ablation—a drop of only 5.0 percentage points in F1—reflects both its inherent specificity and its partial signal overlap with the text encoder. Papers that report statistically implausible results also tend to exhibit anomalous prose patterns in their methodology and results descriptions, meaning the text encoder partially captures the same underlying fraud signal through a different representational pathway. Nevertheless, the statistical module provides unique non-redundant value for a specific and practically important category: carefully written, stylistically convincing fraudulent papers whose only detectable flaw lies in the numerical implausibility of their claimed results—a category that all other modules fail to adequately cover.

C. Failure Case Analysis

Systematic analysis of FakeScholar's 668 misclassified test instances reveals two primary failure modes with distinct structural characteristics. False negatives—325 cases in which fraudulent papers are incorrectly classified as genuine—are heavily concentrated among Retraction Watch papers in the data manipulation category. These papers were authored by genuine researchers who fabricated specific experimental results while the remainder of the manuscript reflects authentic scientific work: the prose is natural, the citation neighbourhood is legitimate, the author network reflects real collaborative relationships, and the reported numerical values are plausible within their declared ranges. All four detection signals are collectively defeated, representing the genuinely hard cases that motivate future work on finer-grained semantic anomaly detection.

False positives—343 cases in which genuine papers are incorrectly flagged—are disproportionately concentrated among two subpopulations. The first comprises papers from research communities with legitimately high self-citation norms, such as narrow applied subfields in which a small number of research groups produce the dominant literature, causing citation graph topology to superficially resemble paper mill patterns. The second comprises early-career researchers from institutions with limited international collaboration networks, whose co-authorship graphs are sparse and geographically clustered in ways that the author network encoder incorrectly

associates with ghost authorship. Both failure modes represent important bias concerns addressed in Section VIII.

D. Limitations

FakeScholar carries several limitations that constrain its current applicability and scope. First, the system requires citation graph data that may be entirely unavailable for papers submitted as preprints with no citation history, degrading the system to a three-module variant at reduced performance. Second, the author network encoder depends on third-party metadata infrastructure—specifically the Semantic Scholar Author API and OpenAlex—whose availability, completeness, and accuracy are outside the authors' control. Third, the synthetic AI-generated training data is derived from a single model family (GPT-4) and may not generalise optimally to papers generated by architecturally distinct models with different stylistic signatures. Fourth, the 9.1% false positive rate, while substantially lower than any evaluated baseline, may remain unacceptably high for fully automated decision-making in editorial contexts; responsible deployment therefore requires human reviewer oversight of all flagged papers rather than autonomous rejection or retraction.

VIII. Ethical Considerations

The deployment of an automated academic fraud detection system carries substantive ethical responsibilities that must be addressed explicitly before any production integration into editorial or institutional workflows.

A. False Positives and Harm to Legitimate Researchers

The most immediate ethical concern is the professional harm that false positive classifications may inflict upon legitimate researchers. An incorrectly flagged paper can damage an author's reputation, generate unwarranted editorial suspicion, trigger institutional investigations, and produce career consequences wholly disproportionate to the available evidence. This risk is not uniformly distributed across the research population. As identified in Section VII, early-career researchers and authors from institutions with limited international collaboration networks are disproportionately represented among false positives, meaning the system's errors fall most heavily on researchers who are already structurally disadvantaged within the global academic system. Rigorous demographic and institutional bias auditing—across career stage, country of origin, institution type, and language background—is a non-negotiable prerequisite for responsible deployment and is acknowledged as a limitation requiring dedicated future investigation.

B. Bias Against Non-Native English Speakers

Scientific manuscripts authored by researchers whose first language is not English may exhibit linguistic regularities—reduced lexical diversity, higher grammatical uniformity, simpler subordinate clause structures—that superficially resemble statistical signatures of LLM-generated text. The text encoder must be rigorously evaluated and actively debiased with respect to author language background before deployment. Inclusion of diverse multilingual scientific corpora spanning papers authored in translated or non-native English across all training and evaluation splits is recommended as a primary mitigation strategy.

C. Adversarial Robustness and the Arms Race Problem

The public availability of FakeScholar's detection architecture and methodology creates a legitimate adversarial concern. Actors aware of the system's detection criteria—particularly the structural graph signals—can in principle engineer submissions to evade them. A well-resourced paper mill could attempt to construct synthetic citation neighbourhoods that satisfy GCN-learnable structural criteria. This arms race dynamic is inherent to any published detection system and mandates a commitment to continuous model updating, ensemble strategies incorporating unpublished signal components, and the integration of features that are intrinsically costly to manipulate. FakeScholar should therefore be treated as one layer within a broader institutional integrity framework rather than a complete solution in isolation.

D. Responsible Deployment Guidelines

FakeScholar must be deployed exclusively as a decision-support tool providing a ranked risk score for human editorial review, never as an autonomous system for manuscript rejection or retraction initiation. All deployment contexts must include transparent communication to authors about the scoring methodology, the factors that influence the fraud score, and clear documented appeal mechanisms through which authors may contest flagged assessments. Score thresholds should be calibrated by individual publishers to reflect their specific tolerance for false positives relative to false negatives, with conservative high-recall settings appropriate only when human review capacity is sufficient to handle the resulting flag volume. No author-level identifiers should be surfaced in any aggregated reporting derived from the system's outputs.

IX. Conclusion and Future Work

A. Summary of Contributions

This paper has introduced FakeScholar, the first multimodal deep learning framework for holistic academic fraud detection, jointly exploiting textual, citation-graph-structural, author-network, and statistical anomaly signals within a single end-to-end trainable architecture. Evaluated on a rigorously constructed benchmark dataset of 50,000 papers balanced between genuine and fraudulent samples, FakeScholar achieves an F1-score of 0.912 and an AUROC of 0.961, outperforming the strongest single-modality baseline by 12.3 percentage points in F1 and demonstrating substantially greater robustness to paraphrasing-based evasion than any evaluated text-only system. Ablation analysis confirms that all four modalities contribute positively and independently to system performance, with the citation graph encoder providing the single largest marginal contribution—a finding that carries important practical implications for the design of future fraud

detection systems. Cross-domain experiments further demonstrate that structural fraud signatures generalise meaningfully across scientific disciplines, with FakeScholar achieving an F1 of 0.821 on biomedical papers despite exclusive training on computer science literature. The benchmark dataset and pre-trained model weights will be made publicly available to support the research community's ongoing efforts to defend the integrity of the scientific record.

B. Future Work

Several directions for future investigation are identified. The most immediate priority is the extension of FakeScholar to additional input modalities: automated analysis of embedded figures and tables for image duplication and manipulation artifacts, inspection of supplementary data files for distributional anomalies inconsistent with claimed experimental conditions, and extraction of code availability and reproducibility signals from linked repositories. Second, integration of FakeScholar into a real-time manuscript screening API designed for direct deployment within journal submission management systems is planned, with system latency and throughput optimised to satisfy the practical requirements of high-volume publishers. Third, the present work treats the fraud detection problem as a static classification task; future work will investigate adversarial training approaches and reinforcement learning-based strategies to maintain detection efficacy as adversarial evasion techniques evolve—directly addressing the arms race concern raised in Section VIII. Fourth, the current system does not model temporal dynamics in citation and authorship networks; incorporating temporal graph neural networks to detect characteristic growth patterns of coordinated citation manipulation campaigns represents a promising avenue for improving recall on the hardest fraudulent cases identified in the failure analysis. Finally, extending the author network encoder to incorporate reviewer identity signals—where available through transparent review platforms—may substantially improve detection of fake peer review, which currently represents one of the most difficult fraud categories for the system to identify reliably.

REFERENCES

[1] I. Oransky and A. Marcus, "Retraction Watch," 2010. [Online]. Available: https://retractionwatch.com

[2] J. P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine*, vol. 2, no. 8, p. e124, Aug. 2005.

[3] E. M. Bik, A. Casadevall, and F. C. Fang, "The prevalence of inappropriate image duplication in biomedical research publications," *mBio*, vol. 7, no. 3, pp. e00809-16, May/Jun. 2016.

[4] C. Else, "How this paper mill detected 400 potentially fake cancer papers," *Nature*, vol. 591, pp. 19–20, Mar. 2021.

[5] J. A. Byrne and J. Christopher, "Digital magic or sleight of hand? Lessons learned from the paper mill phenomenon and potential futures for the academic publishing system," *Research Integrity and Peer Review*, vol. 5, no. 1, p. 22, Nov. 2020.

[6] R. Van Noorden, "Publishers: Stem the tide of fake academic papers," *Nature*, vol. 603, pp. 206–207, Mar. 2022.

[7] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, Mar. 2023.

[8] P. Demir, "Hallucinated but authoritative: The epistemic risks of AI-authored academic manuscripts," *Accountability in Research*, vol. 31, no. 2, pp. 134–151, 2024.

[9] E. Tian, "GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods," GPTZero Technical Report, 2023. [Online]. Available: https://gptzero.me

[10] Turnitin LLC, "Turnitin AI writing detection," Turnitin White Paper, 2023. [Online]. Available: https://www.turnitin.com/solutions/ai-writing

[11] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics: System Demonstrations*, Florence, Italy, Jul. 2019, pp. 111–116.

[12] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proc. 34th AAAI Conf. Artif. Intell.*, New York, NY, USA, Feb. 2020, pp. 549–556.

[13] Y. Lin, D. Xu, T. Jiang, H. Liu, and G. Luo, "Hierarchical graph attention networks for fake news detection with user credibility modelling," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, UAE, Dec. 2022, pp. 3083–3096.

[14] S. Fortunato *et al.*, "Science of science," *Science*, vol. 359, no. 6379, p. eaao0185, Mar. 2018.

[15] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA, Jul. 2023, pp. 24950–24962.

[16] J. Kirchenbauer, J. Geiping, Y. Wen, J. Kaddour, A. Ahmad, and T. Goldstein, "A watermark for large language models," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA, Jul. 2023, pp. 17061–17084.

[17] V. Verma, E. Fleisig, N. Tomlin, and D. Klein, "Ghostbuster: Detecting text ghostwritten by large language models," in *Proc. 2024 Conf. North American Chapter Assoc. Comput. Linguistics*, Mexico City, Mexico, Jun. 2024, pp. 1702–1717.

[18] I. Solaiman *et al.*, "Release strategies and the social impacts of language models," arXiv preprint arXiv:1908.09203, Aug. 2019.

[19] X. Yao, J. Li, and Y. Shi, "Knowledge graph-enhanced fake news detection via graph attention networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Miami, FL, USA, Nov. 2022, pp. 681–690.

[20] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, vol. 178, no. 4060, pp. 471–479, Nov. 1972.

[21] M. Seeber, G. Cattaneo, L. Meoli, and P. Mallig, "Self-citations as strategic response to the use of metrics for career decisions," *Research Policy*, vol. 48, no. 2, pp. 478–491, Mar. 2019.

[22] E. A. Fong and A. W. Wilhite, "Authorship and citation manipulation in academic research," *PLOS ONE*, vol. 12, no. 12, p. e0187394, Dec. 2017.

[23] M. Teplitskiy, D. Acuna, E. Duede, and J. A. Evans, "Network structure of retracted and corrected publications: Evidence of localised fraud and systemic error," arXiv preprint arXiv:2010.03183, Oct. 2020.

[24] J. A. Byrne, "Strengthening responsible research conduct and integrity: The role of institutional policy and oversight mechanisms," *Research Integrity and Peer Review*, vol. 7, no. 1, p. 3, Apr. 2022.

[25] E. M. Bik, "Setting the bar for image manipulation in scientific publications," *FACETS*, vol. 7, pp. 217–232, Jan. 2022.

[26] G. Cabanac and C. Labbe´, "Prevalence of nonsense in published biomedical literature: The Ike Antkare phenomenon," in *Proc. Annu. Meeting Amer. Soc. Inf. Sci. Technol.*, vol. 50, no. 1, Montreal, Canada, Nov. 2013.

[27] M. Dadkhah, G. Borchardt, and A. Maliszewski, "Getting our voices heard: Challenges and threats in scholarly communication from the perspective of academic authors," *European Science Editing*, vol. 43, no. 3, pp. 58–61, 2017.

[28] B. Gipp, N. Meuschke, and C. Breitinger, "Citation-based plagiarism detection: Detecting disguised and cross-language plagiarism using citation pattern analysis," in *Proc. ACM/IEEE Joint Conf. Digital Libraries*, Fort Worth, TX, USA, Sep. 2014, pp. 433–434.

[29] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process.*, Hong Kong, China, Nov. 2019, pp. 3615–3620.

[30] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online, Jul. 2020, pp. 2270–2282.

[31] M. Ostendorff *et al.*, "Neighbourhood contrastive learning for scientific document representations with citation embeddings," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, UAE, Dec. 2022, pp. 11670–11688.

[32] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, Apr. 2017.

[33] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process.*, Hong Kong, China, Nov. 2019, pp. 3982–3992.
[34] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1024–1034.
[35] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Pisa, Italy, Dec. 2008, pp. 413–422.
[36] M. J. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Hoboken, NJ, USA: Wiley, 2012.
[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Representations*, New Orleans, LA, USA, May 2019.
[38] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, "S2ORC: The Semantic Scholar Open Research Corpus," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online, Jul. 2020, pp. 4969–4983.
[39] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A fully-open index of the world's research outputs," arXiv preprint arXiv:2205.01833, May 2022.
[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.