Early Detection of Retractable Papers with Scientometrics & Machine Learning

Nantheesan Raveenthiran

BSc Computer Science — University of Southampton

16 May 2025

Motivation — scale of the problem

- **Publication boom**: 1.9 M papers (2016) $\rightarrow$ 2.8 M (2022).
- **Retraction surge**: 1.8 k (2013) $\rightarrow$ 10 k (2023).
- **Time-to-retract**: mean 34 $\pm$ 36 months $\Rightarrow$ ample time for harm.

Cascading effects of late retractions

- Fraudulent findings cited $\sim$400 $\times$ before withdrawal.
- Author's prior work loses 6.9 % citations (Lu *et al.* 2024).
- Spill-over up to 4 citation hops; 25 % of authors leave academia.
- Public trust eroded — e.g. hydroxychloroquine during COVID-19.

Research objectives

1. Build a supervised model to **rank** papers by retraction risk.
2. Compare *metadata*, *citation metrics*, and network centrality.
3. Prototype a metadata-only *Meta-Net* graph for low-resource cases.
4. Achieve precision ≥ 0.9 for editorial triage.

Hypotheses

- **H1**: Network centrality lifts ROC-AUC $\geq +5$ pp over metadata baseline.
- **H2**: SBERT-based Meta-Net adds predictive signal beyond text.
- **H3**: Temporal citation trajectories (CTPIR) will further improve recall (future work).

Data sources

Source	Coverage	Volume	Key fields
PubMed	1996–2024	1.8 M docs	title, abstract, year, authors
Retraction Watch	1980–2025	12 220 docs	flag, date, reason
OpenCitations	1950–2024	122 M edges	citing $\leftrightarrow$ cited DOIs
arXiv physics	1991–2024	700 docs	cross-domain validation

Final balanced set: **3 558** papers

(1 779 retracted / 1 779 control).

ETL pipeline

1. Merge PubMed + Retraction Watch by PMID.
2. Clean text (lowercase, stop-words, lemmatisation with spaCy).
3. Transform heavy-tail counts: $\log(1 + x)$, then z-score.
4. 2-hop OpenCitations crawl $\rightarrow$ directed graph (Graph-Tool, 8 threads).
5. Stratified split 70/30 by retraction year.

Pre-processing stats

- Clean vocabulary: 48 k tokens; TF-IDF truncated to 5 k dims.
- Subject one-hot (9 classes) $\rightarrow$ SVD 100 for LR; tree models use raw.
- Graph density: 1.85×10^{-6}; giant component covers 98 % of nodes.

Feature block A: Traditional metadata

- TF-IDF 1–3 grams (5 000 dims).
- Author count, publication year, paper age.
- Subject SVD vectors (100 dims).

Feature block B: Citation metrics

- In-degree, out-degree counts.
- Age-normalised citation rate.
- PageRank ($d = 0.85$), betweenness centrality.

Feature block C: Citation network

- Graph size: 9.84 M nodes, 18.13 M edges.
- Centrality vectors computed per paper (degree, PageRank, betweenness).
- Reveals hierarchical, self-citation clusters typical of paper mills.

Feature block D: Meta-Net (SBERT)

- Sentence embeddings: SBERT all-MiniLM-L6 (384-d).
- Mutual k-NN ($k = 10$) on cosine similarity $\rightarrow$ 18 927 edges.
- Edge weight $= 0.7\,\text{cosine} + 0.2\,\text{subject match} + 0.1\,\text{author bin}$.
- Planned upgrade: SPECTER 2.0 embeddings + MeSH Jaccard.

Modelling & validation

- Baseline: Logistic Regression (interpretable).
- Ensembles: Random Forest (400 trees, depth 10) and XGBoost (500 trees, $\eta = 0.05$).
- Hyper-parameter search: Optuna (40 trials, log-loss).
- Metrics: Accuracy, Precision @0.7, Recall, ROC-AUC.

Performance on hold-out (30 % / 1 067 papers)

Features	Best model	Acc.	Prec.@0.7	AUC
Metadata	RF	0.81	0.84	0.81
Citation metrics	RF	0.80	0.83	0.79
Citation network	RF	0.81	0.85	0.81
Meta-Net	RF	0.59	0.63	0.60
Fusion (no bias feats)	RF	0.83	0.86	0.83

Error analysis (fusion model)

- Threshold 0.7 $\Rightarrow$ TP 918, FP 140, FN 185, TN 824.
- Precision 0.87 Recall 0.83 F1 0.85.
- FP cluster: high PageRank but low citation rate — possible early fraud signal.

Limitations

- **Domain skew**: biomedical focus; physics subset AUC drops 0.03.
- **Static graph**: ignores temporal trajectories after publication.
- **Meta-Net sparsity**: SBERT cosine ≥ 0.7 yields weak semantics.
- Ethical risk: geographic/journal bias removed but still monitored.

Future work

1. Integrate SPECTER 2.0 and full-text embeddings.
2. Temporal GNN with CTPIR slopes to model citation decay.
3. Expand dataset via Crossref Dimensions ($+5\times$ size).
4. Deploy high-precision pilot with journal editors (human-in-the-loop).

Take-home message

Adding citation-network centrality to classic metadata delivers a high-precision early-warning score, cutting editors' screening workload five-fold and shrinking the 34-month retraction lag.

Questions?