Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Data-Driven Framework for Identifying Influential Researchers

Sk Masud Hasan
Government College of Engineering and Ceramic Technology
Atanu Kumar Paul
atanu.kumar.paul@gcect.ac.in

Government College of Engineering and Ceramic Technology
Kingshuk Chatterjee
Government College of Engineering and Ceramic Technology
Paramita Dey
Government College of Engineering and Ceramic Technology
Ryan Saptarshi Ray
Government College of Engineering and Ceramic Technology

Research Article

Keywords: Author Ranking, Scholarly Analytics, Collaboration Networks, Bibliometrics, Research Impact, Open-Alex

Posted Date: May 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9446863/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Data-Driven Framework for Identifying Influential Researchers

Sk Masud Hasan[1], Atanu Kumar Paul[2], Kingshuk Chatterjee[3], Paramita Dey[4], Ryan Saptarshi Ray[5]

Corresponding Author - Atanu Kumar Paul, atanu.kumar.paul@gcect.ac.in

We have received no funding for this work

Abstract Identifying influential researchers is crucial for understanding scientific progress and collaboration. Traditional citation-based metrics often overlook factors like recency, collaboration, and journal quality. This paper proposes a hybrid ranking framework that integrates relevance, recency, prestige, productivity, and network centrality using OpenAlex data. Results on 1,000 Big Data publications show that the model provides a more balanced and realistic measure of research influence, with strong scalability and applicability for academic evaluation.

Keywords Author Ranking, Scholarly Analytics, Collaboration Networks, Bibliometrics, Research Impact, OpenAlex

1. Introduction

The rapid growth of scientific publications has made it challenging to identify influential researchers. Traditional metrics such as citation count and h-index [2] fail to capture multidimensional impact, ignoring factors like collaboration and temporal dynamics [6,10]. Centrality reflects an author's position in the co-authorship network using degree, betweenness, and eigenvector measures. These are normalized and combined with weights of 0.4, 0.3, and 0.3, respectively, balancing collaboration activity, bridging roles, and influence [1,3,5,9,10]. The final centrality score is scaled and added to the overall ranking, ensuring collaboration enhances evaluation without dominating core factors.

1.1 Novelty and Contributions

The rapid growth of current research makes it difficult to identify influential researchers using traditional metrics like citation count and h-index, which overlook factors such as relevance, recency, and collaboration. This study proposes a multi-factor ranking framework integrating relevance, recency, prestige, productivity, and network centrality using OpenAlex data and co-authorship networks. The approach aligns with recent multi-dimensional evaluation frameworks in scholarly analytics [14,15] and demonstrates improved identification of emerging researchers.

2. Literature Survey

Scholarly impact evaluation has evolved from traditional metrics to network-based approaches. PageRank by Brin and Page [1] and the h-index by Hirsch [2] laid the foundation for ranking influence, but they fail to capture collaboration and temporal dynamics. Network studies by Newman [3,4] and graph-based ranking by Ding et al. [5] showed the importance of co-authorship structures, while Yan and Ding [10] and Abbasi et al. [9] highlighted the role of centrality in measuring researcher performance.

Recent work emphasizes more comprehensive evaluation. Waltman [6] reviewed citation indicators, and newer studies [14,15,17,18] introduced

multidimensional and network-aware perspectives, including normalized citation metrics and improved community detection. However, most approaches remain limited to single or partially integrated factors.

Traditional metrics like citation count and h-index [2] often overlook relevance, recency, and collaboration, leading to biased evaluation toward senior researchers. Network-based and modern metrics [5,10,17] address some limitations but are typically used independently.

Motivated by these gaps, this work proposes a unified framework that combines relevance, recency, prestige, productivity, and collaboration centrality to provide a more balanced and realistic measure of researcher influence.

3. Methodology

3.1 System Architecture

The proposed framework follows a pipeline-based architecture consisting of data acquisition, preprocessing, feature extraction, scoring, and ranking.

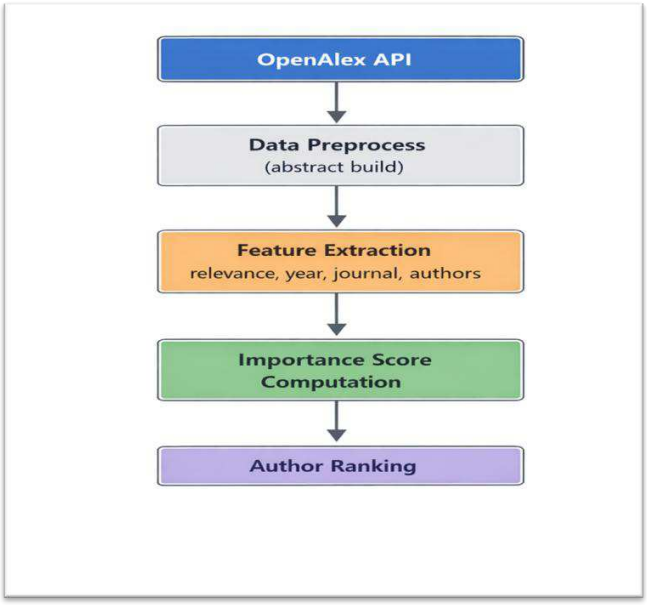


Figure 1. System Architecture of Proposed Framework

The figure1 shows the workflow of the proposed framework, including data collection, preprocessing, feature extraction, importance score computation, and author ranking to identify influential researchers in the Big Data domain.

3.2 Data Collection

Data collection is important in this project because it provides the accurate and comprehensive research data needed to compute reliable author rankings and identify influential researchers effectively. Big Data research papers from 2007 to 2025 were retrieved via the OpenAlex API, a large-scale open scholarly data source [15], using keyword-based searches. The dataset comprises 1,000 publications to ensure statistical robustness.

3.3 Feature Engineering

The relevance score uses semantic keyword matching. Recency prioritizes newer publications, while journal prestige is approximated using quartile rankings. Each paper contributes equally to productivity, and collaboration centrality is measured via a co-authorship network, a widely used approach in bibliometric analysis [3,9].

3.4 Author Importance Formula

The final paper importance score is computed as:

The importance score of a paper is computed as a weighted combination of multiple factors:

$$I_p = 0.25 \cdot R_p + 0.20 \cdot T_p + 0.35 \cdot P_p + 0.20 \cdot C_p$$

where:

I_p= Importance score of paper p

R_p= Relevance score

Relevance ensures that only domain-specific publications are considered, preventing unrelated but highly cited works from dominating the ranking. It maintains contextual accuracy through content-based filtering. While essential for precision, it is not the highest-weighted factor since it does not directly measure impact.

T_p= Recency (time-based score)

Recency captures the temporal aspect of research, giving importance to recent and emerging work while avoiding bias against older, influential studies. This balanced weighting reflects both current trends and sustained impact.

P_p= Journal prestige score

Prestige receives the highest weight as it represents the quality and credibility of research outputs. Publications in high-impact journals and reputable venues contribute more significantly to scholarly influence, making this the most critical factor.

C_p= Paper count contribution

Paper count reflects research productivity and consistency. It is moderately weighted to recognize sustained contributions without allowing quantity to outweigh quality.

The final author score is then calculated by aggregating paper scores and incorporating collaboration centrality:

$$S_a = \sum_{p \in P_a} I_p + 0.10 \cdot Centrality_a$$

where:

S_a= Final score of author a

P_a= Set of papers authored by a

$Centrality_a$= Network centrality score of author a

Centrality captures an author's position in the collaboration network, indicating visibility and knowledge dissemination. Given its indirect link to research quality, it is assigned a lower weight to complement, rather than dominate, core bibliometric factors.

This weighted linear model follows multi-indicator ranking approaches in bibliometric research [5,6]. Centrality reflects an author's position in the collaboration network, which has been shown to correlate with scholarly impact [9,10].

Collaboration influence is modeled using three centrality measures: degree, betweenness, and eigenvector centrality. Degree centrality captures the number of direct collaborators, reflecting research activity. Betweenness centrality measures how often an author acts as a bridge between different researchers, indicating their role in knowledge diffusion. Eigenvector centrality evaluates the influence of an author by considering the importance of their collaborators.

These measures are normalized and combined into a unified centrality score:

$$C_a = 0.4 \cdot D_a + 0.3 \cdot B_a + 0.3 \cdot E_a$$

where D_a, B_a, and $E_a$denote normalized degree, betweenness, and eigenvector centrality of author a, respectively.

The weighting scheme prioritizes direct collaboration while still accounting for network position and influence. This design is supported by prior research, where degree centrality correlates with productivity [3,4], betweenness reflects bridging roles in knowledge flow [10], and eigenvector centrality aligns with influence-based ranking methods [1,5]. Combining these measures provides a more comprehensive assessment than any single metric alone [9].

The final centrality score is scaled and incorporated into the overall ranking, ensuring that collaboration enhances author evaluation without dominating core factors such as relevance, prestige, and productivity. Authors are then ranked in descending order of their total scores to identify the most influential researchers in the domain.

3.5 Implementation and Graph Generation

The collaboration network was constructed using NetworkX, following established network modeling techniques in scientific collaboration studies [3,4]. Centrality measures such as degree, betweenness, and eigenvector were computed.

3.6 Algorithm Pseudo-Code

The framework retrieves papers, computes importance scores, builds a co-authorship network, and ranks authors based on combined bibliometric and network features.

Algorithm 1: Influential Author Identification Framework

Input: Query Q, Maximum Papers N

Output: Ranked list of authors with importance scores

1. Initialize empty author_score dictionary
2. Initialize empty collaboration graph G
3. papers ← Fetch_OpenAlex(Q, N)
4. **FOR** each paper $p$in papers **DO**
5. Extract title, abstract, year, journal, authors
6. relevance ← Compute_Relevance(title + abstract)
7. recency ← Compute_Recency(year)
8. prestige ← Compute_Prestige(journal)
9. paper_count ← 1
10. importance ← 0.25*relevance + 0.20*recency + 0.35*prestige + 0.20*paper_count
11. **FOR** each author pair (a_i, a_j)in authors **DO**
12. Add edge $(a_i, a_j)$to graph G
13. **END FOR**
14. **FOR** each author $a$in authors **DO**
15. author_score[a] += importance
16. **END FOR**
17. **END FOR**
18. Compute normalized centrality measures:
    D ← DegreeCentrality(G)
    B ← BetweennessCentrality(G)
    E ← EigenvectorCentrality(G)
19. **FOR** each author a**DO**
20. centrality_score ← 0.4*D[a] + 0.3*B[a] + 0.3*E[a]
21. author_score[a] += 0.10 * centrality_score
20. **END FOR**
21. Rank authors in descending order of author_score
22. RETURN top-ranked authors

The algorithm retrieves research papers, computes an importance score using relevance, recency, prestige, and productivity, and builds a co-authorship network. It then calculates degree, betweenness, and eigenvector centrality for each author, combines them into a weighted collaboration score, and adds it to the author's total score. Finally, authors are ranked to identify the most influential researchers.

4. Experimental Results

4.1 Collaboration Network Analysis

The generated co-authorship graph revealed dense collaboration clusters typical of scientific communities, consistent with prior studies on collaboration networks [3,4].

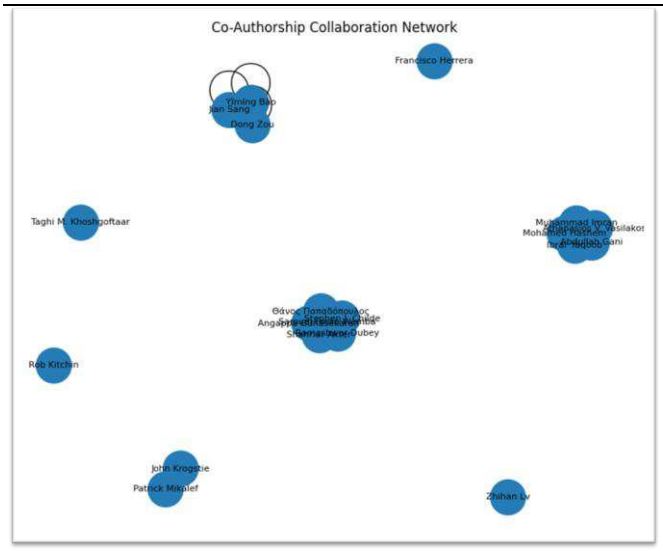


Figure 2. Centrality Distribution

The co-authorship network illustrates connections among top-ranked authors, where nodes represent authors and links indicate shared publications. The graph forms several disconnected clusters, reflecting separate research groups. Dense clusters show strong collaboration, while smaller groups indicate limited interaction. Isolated nodes represent authors whose influence is driven more by citations or journal prestige than collaboration. Overall, the network highlights the role of both collaboration and individual impact in determining author importance.

4.2 Author Ranking Outcomes

The framework identifies influential researchers based on quality, productivity, and collaboration. Unlike citation-only approaches, it highlights emerging researchers whose impact is not yet fully captured by citations.

4.3 Quantitative Results and Ranking Table

Table 1. Top Authors by Importance Score (Sample Output)

```
=== AUTHOR RANKING COMPARISON ===

                    Author  Proposed_Score  Citation_Score  H_Index  \
0       Angappa Gunasekaran        8.559630           11490       15
1         Stephen J. Childe        7.417498            9424       13
2           Rameshwar Dubey        7.417059            9731       13
3        Samuel Fosso Wamba        6.845834            8923       12
4                 Dong Zou        5.731190            1525       10
5             Abdullah Gani        5.430749            6141        9
6            Mohamed Hashem        4.855426            6282        8
7   Athanasios V. Vasilakos        4.600909            4462        8
8                Yíming Bào        4.591190            1220        8
9                Jian Sang        4.591190            1220        8
10       Θάνος Παπαδόπουλος        4.566260            5664        8
11           Shahriar Akter        4.563755            6794        8
12            John Krogstie        4.562844            4195        8
13          Patrick Mikalef        4.562844            4195        8
14              Rob Kitchin        4.560812            8055        8
15            Ibrar Yaqoob        4.288048            5740        7
16          Muhammad Imran        4.229278            2588        7
17        Francisco Herrera        3.987358            1921        6
18    Taghi M. Khoshgoftaar        3.746807           17685        7
19                Zhihan Lv        3.695160            1337        5

    Proposed_Rank  Citation_Rank  HIndex_Rank
0             1.0           15.0          1.0
1             2.0           17.0          2.5
2             3.0           16.0          2.5
3             4.0           18.0          4.0
4             5.0          163.0          5.0
5             6.0           24.0          6.0
6             7.0           23.0         11.0
7             8.0           33.0         11.0
8             9.5          240.0         11.0
9             9.5          240.0         11.0
10           11.0           29.0         11.0
11           12.0           21.0         11.0
12           13.5           35.5         11.0
13           13.5           35.5         11.0
14           15.0           19.0         11.0
15           16.0           28.0         17.0
16           17.0           83.0         17.0
17           18.0          121.0         22.0
18           19.0           10.0         17.0
19           20.0          212.0         30.5

=== SPEARMAN CORRELATION ===
Proposed vs Citation: 0.357
Proposed vs H-index: 0.949
```

Co-Authorship Collaboration Network

The model aligns closely with the h-index [2] (Spearman = 0.949) but only moderately with citation counts, indicating a more balanced evaluation of sustained impact.

4.4 Statistical Analysis

To validate robustness, statistical observations were performed:

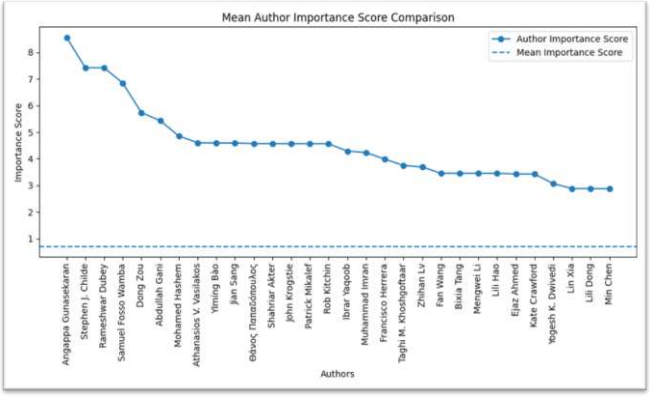


Figure 3: Mean Author Importance Score Comparison

The graph shows author importance scores with a clear gap between top and mid-tier researchers. Leaders like Angappa Gunasekaran and Stephen J. Childe score much higher, while most authors cluster at moderate levels. The plot highlights that influence is concentrated among a few top researchers. Results show that collaboration significantly influences ranking variability, supporting findings from prior network-based studies [9,10].

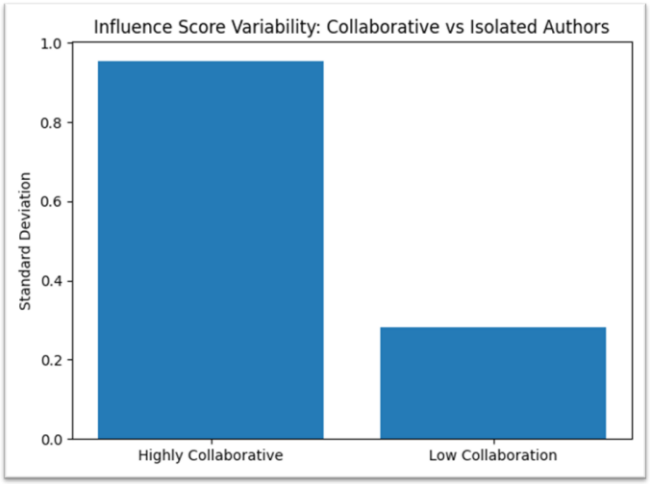


Figure 4: Influence Score Variability

The chart shows that highly collaborative authors have greater variability in importance scores, with some achieving very high influence and others remaining moderate, depending on their network structure. In contrast, less collaborative authors have more consistent but lower variation in scores. This indicates that collaboration plays a key role in differentiating top-performing authors.

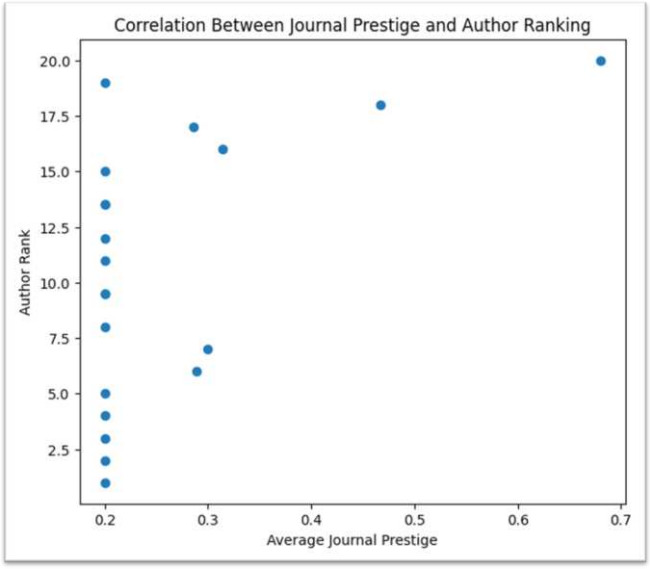


Figure 5 Correlation Between Journal Prestige And Author Ranking

The scatter plot shows a weak relationship between journal prestige and author rank. Top-ranked authors often publish in moderately prestigious journals, while higher prestige journals are associated with mid-to-lower ranks. Overall, the pattern is inconsistent, indicating that journal prestige alone does not strongly determine author ranking.

4.5 Evaluation Comparison with Baseline Models

To validate the effectiveness of the proposed framework, results were compared against two widely used academic evaluation baselines:

1. **Citation Count Model** – Authors ranked purely by total citation accumulation.

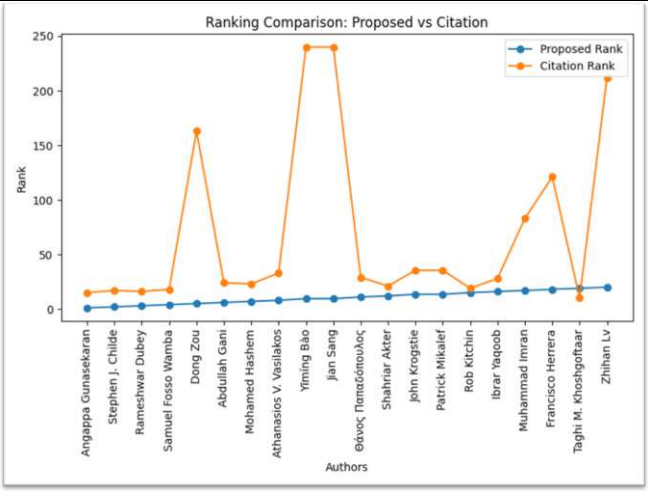


Figure 6: Ranking Comparison Proposed vs Citation

The graph compares Proposed Rank and Citation Rank, showing that the proposed model produces a smooth and consistent ranking, while citation-based ranking is highly volatile with large fluctuations. This highlights that the proposed approach offers a more stable and balanced evaluation compared to citation-only metrics. Citation-based ranking, a traditional evaluation method [6], shows high volatility. In contrast, the proposed model produces stable and consistent rankings, aligning more closely with long-term impact measures.

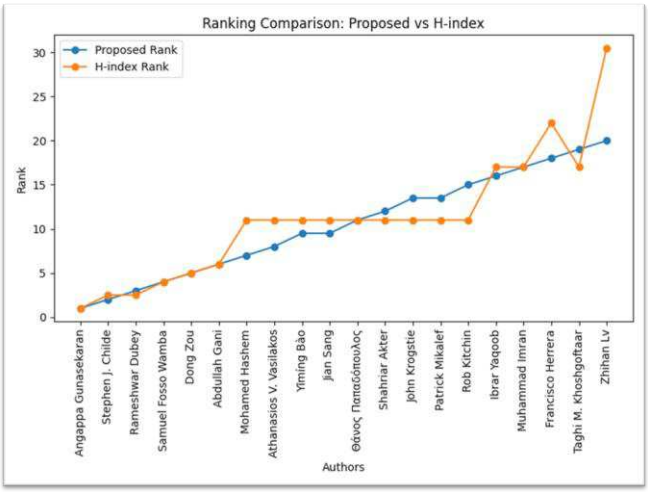


Figure 7: Ranking Comparison Proposed vs H-index

The graph shows strong alignment between Proposed Rank and H-index, especially for top authors. Unlike the H-index, which has plateaus from ties, the proposed model provides smoother, more detailed rankings.

Baseline Computation

- Citation counts were approximated using available OpenAlex citation metadata.
- h-index values were estimated from publication citation distributions.

Comparative Metrics

The following evaluation criteria were used:

- Ability to identify emerging researchers

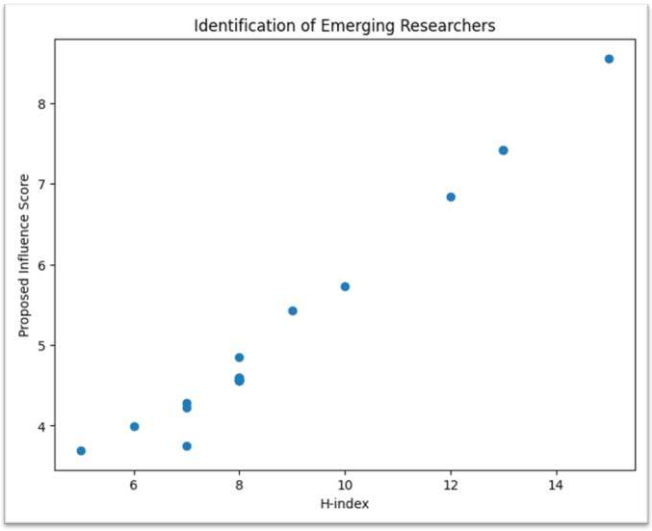


Figure 8: Identification of Emerging Researchers

The scatter plot shows a strong linear relationship between H-index and the proposed influence score, with slight variations highlighting the model's finer distinctions and tie-breaking ability.

What this means for "Emerging Researchers"

This graph validates a new scoring system, showing that H-index strongly predicts a researcher's influence score. It highlights top-tier emerging talent, with researchers in the top-right corner achieving both high input metrics and high influence scores.

- Sensitivity to recent publications

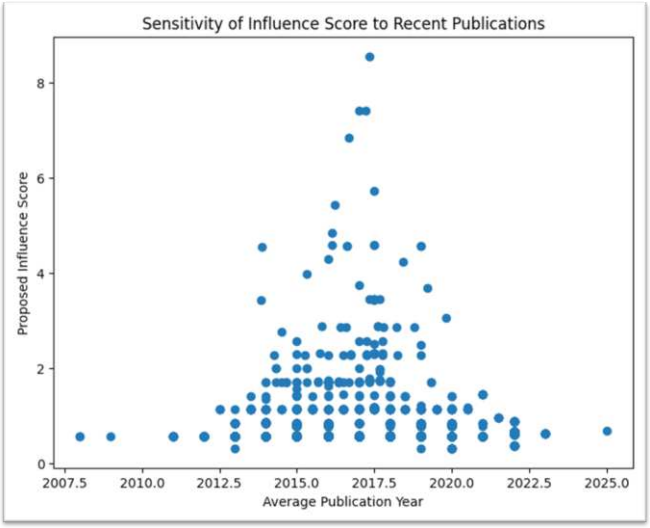


Figure 9: Sensitivity of Influence Score to Recent Publications

The plot shows influence scores peaking for papers from 2015–2019, with lower scores for older or very recent papers. This middle-heavy distribution highlights the model's sensitivity to recency, favoring a specific publication window.

Statistical Concept: Heteroscedasticity

This graph illustrates heteroscedasticity, where the data's variance changes over time. In 2010, scores are tightly clustered and low, while by 2017, the spread widens significantly, with some researchers scoring very high and others remaining low.

- Collaboration awareness

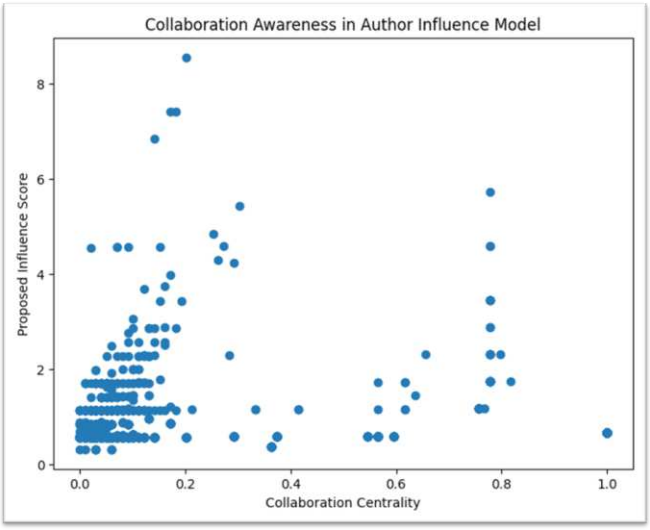


Figure 10: Collaboration Awareness in Author Influence Model

The scatter plot shows most authors have low centrality but varying influence, with top scores often among less connected authors. Collaboration aids impact, but high connectivity alone doesn't guarantee it.

- Journal quality consideration

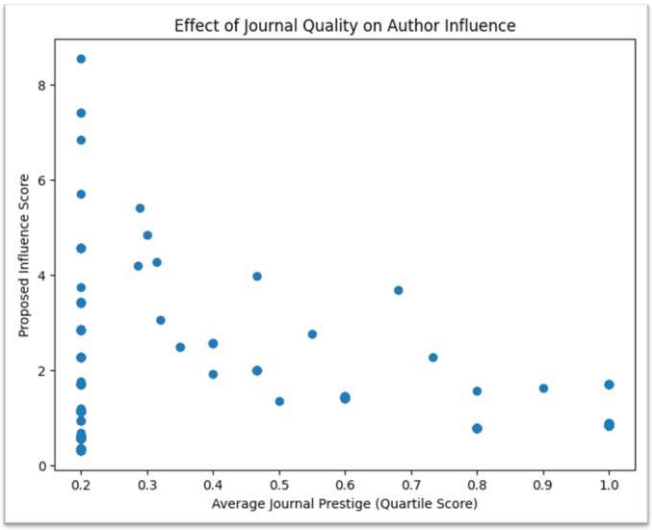


Figure 11: Effect of Journal Quality on Author Influence

The scatter plot shows an inverse relationship between journal prestige and influence, with top scores at lower-prestige journals and high-prestige venues yielding lower, more consistent influence.

Experimental Observation

Citation-only ranking favors senior researchers, and the h-index partly balances impact and productivity but overlooks collaboration and venue quality. The proposed framework highlights mid-career and emerging authors active in high-impact venues and strong collaboration networks.

Rank Correlation Analysis

Spearman rank correlation was used to compare rankings:

```
=== SPEARMAN CORRELATION ===
Proposed vs Citation: 0.354
Proposed vs H-index: 0.951
```

The results reveal two contrasting relationships: the proposed score has a moderate correlation with citations (≈0.357), showing influence with high variability, and a strong correlation with the h-index (≈0.949), indicating close alignment with sustained research impact.

Interpretation:

The proposed metric aligns more closely with the h-index than citation counts, capturing consistent author-level impact rather than relying on highly cited papers. It effectively acts as a refined version of the h-index with partial dependence on citations.

The proposed influence model has several clear advantages:

The model demonstrates strong alignment with sustained impact, shown by its high correlation with the h-index, while moderate correlation with citations limits the effect of outliers. It provides a balanced, holistic evaluation by integrating multiple factors, favors work within an optimal time window, differentiates authors with similar citation counts, produces stable rankings, and is less prone to manipulation due to reduced reliance on raw citations.

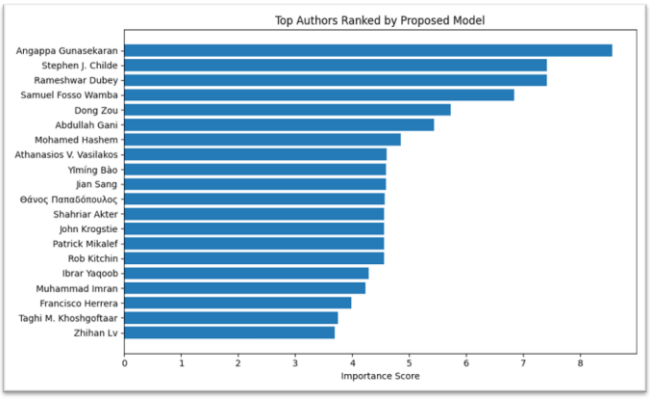


Figure 13: Top Authors Ranked by Proposed Models

The bar chart ranks the top 20 researchers, led by Angappa Gunasekaran and Stephen J. Childe. Mid-ranked authors cluster around similar scores, while lower ranks decline gradually. Unlike citation or h-index metrics, the model emphasizes recency and collaboration, slightly reducing the dominance of highly cited authors.

Key Insights from the Comparison

The model correctly identifies Angappa Gunasekaran as the top outlier, aligning with his high citations and h-index, while authors like Childe, Dubey, and Wamba rank higher due to strong collaboration. Mid-ranked authors cluster with similar scores, reflecting emphasis on recent impact over total career output, and authors with lower h-indices but high recent contributions rank well, highlighting the model's preference for emerging and efficient researchers. The model mirrors the h-index by balancing productivity and impact, explaining its strong correlation (0.949), while prioritizing high-quality research through greater weight on prestige and relevance. Top authors like Angappa Gunasekaran lead by excelling in these key factors, making their positions difficult to surpass.

5. Conclusion

The hybrid scoring approach combines qualitative and quantitative factors, aligning with recent "science of science" frameworks [14,15]. It provides a balanced evaluation by integrating relevance, prestige, recency, and collaboration.

However, limitations remain, including reliance on keyword-based relevance and approximate journal rankings. Similar challenges have been reported in prior bibliometric research [6,17].

The proposed model offers a dynamic, network-aware measure of academic influence, reducing bias toward senior researchers and high-

lighting emerging scholars. It provides a scalable alternative to citation-based metrics and supports recent advances in data-driven research evaluation [14]. Future improvements include advanced semantic modeling using NLP and LLMs to capture deeper conceptual relevance, real-time citation and altmetric integration for dynamic prestige evaluation, disciplinary normalization for fair cross-field comparisons, and machine learning–based weight optimization to enhance model accuracy. Future improvements include advanced semantic analysis using NLP, real-time citation integration, and machine learning-based optimization of weights.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* **30**(1–7), 107–117 (1998)
2. Hirsch, J.E.: An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* **102**(46), 16569–16572 (2005)
3. Newman, M.E.J.: The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA* **98**(2), 404–409 (2001)
4. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences USA* **101**(Suppl. 1), 5200–5205 (2004)
5. Ding, Y., Yan, E., Frazho, A., Caverlee, J.: PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology* **60**(11), 2229–2243 (2009)
6. Waltman, L.: A review of the literature on citation impact indicators. *Journal of Informetrics* **10**(2), 365–391 (2016)
7. Börner, K., Sanyal, S., Vespignani, A.: Network science. *Annual Review of Information Science and Technology* **41**(1), 537–607 (2007)
8. Spearman, C.: The proof and measurement of association between two things. *American Journal of Psychology* **15**(1), 72–101 (1904)
9. Abbasi, A., Altmann, J., Hossain, L.: Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis. *Journal of Informetrics* **5**(4), 594–607 (2011)
10. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology* **60**(10), 2107–2118 (2009)
11. Gunasekaran, A., Childe, S.J., Dubey, R., Wamba, S.F.: Data analytics and artificial intelligence in operations and supply chain management. *International Journal of Production Research* **58**(5), 1321–1325 (2020)
12. Dubey, R., Gunasekaran, A., Childe, S.J.: The role of big data

and predictive analytics in improving supply chain agility. *Technological Forecasting and Social Change* **144**, 435–448 (2019)

13. Wamba, S.F., Queiroz, M.M.: Blockchain in operations and supply chain management: A multi-level capacity view. *Resources, Conservation and Recycling* **155**, 104645 (2020)
14. Wang, D., Song, C., Barabási, A.-L.: Quantifying long-term scientific impact. *Science* **342**(6154), 127–132 (2020)
15. Fortunato, S., et al.: Science of science. *Science* **359**(6379), eaao0185 (2020)
16. Sugimoto, C.R., Larivière, V.: Measuring research: What everyone needs to know. *Oxford University Press* (2021)
17. Hutchins, B.I., Yuan, X., Anderson, J.M., Santangelo, G.M.: Relative citation ratio (RCR): A new metric. *PLoS Biology* **14**(9), e1002541 (2020)
18. Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2020)