

Analysis of Research Trends in Indian Computer Science Scholarly Communication: A Methodological Perspective with Python

Ashoka R.¹; N. S. Harinarayana²

Officer (Learning Centre), Indian Institute of Management Indore, Indore, India¹; Senior Professor, DOS in Library and Information Science, University of Mysore, Mysuru, India²

ashokar@iimidr.ac.in; harinarayana@lisc.uni-mysore.ac.in

ABSTRACT

This study investigates research trends in the Indian computer science journal over a ten-year period (2011- 2020) using keyword-based analysis across title, abstract, and author-supplied metadata. Since the selected journals were not uniformly indexed in an international database, the study employed a combined approach of manual review and automated data extraction to ensure comprehensive coverage. Keywords were lemmatized and filtered utilizing YAKE score and frequency threshold, thereafter analysed through TF-IDF vectorization, cosine similarity, and hierarchical clustering. The top 25 semantically relevant keywords were extracted from each source. Thematic dendrograms indicated a constant tendency in wireless sensor networks, neural networks, machine learning, and cybersecurity. Hypothesis testing revealed no substantial disparity in term distribution within the metadata layer, suggesting theme congruence. The research advances metadata standards and provides a reproducible framework for trend analysis and vocabulary enhancement in scholarly communication.

KEYWORDS: Keyword Analysis, Research Trends, Computer Science Journal, YAKE, Semantic Clustering.

1. INTRODUCTION

In the evolving landscape of scholarly communication, the strategic use and analysis of keywords have become central to understanding research trend, improving information retrieval, and enhancing metadata quality. Keywords not only represent the thematic core of an academic document but also serve as a vital access point in a digital repository, indexing system, and bibliographic database. Their standardization and semantic clustering enable improved subject indexing and vocabulary control, which are essential for ensuring consistency across the bibliographic system AlShammari (2023). The ability to extract, analyse, and normalize these terms offers valuable insight into the development of the research domain and the progression of scholarly discourse R & Harinarayana, 2022). A recent study has demonstrated that the keyword-based approach is effective in detecting emerging research topics, evaluating publication dynamics, and identifying collaborative and interdisciplinary research patterns

(Kathiria & Arolkar, 2022; Lu et al., 2021). In particular, the field of computer science is characterized by rapid technological advancement and multidisciplinary convergence require robust and adaptive methodology to trace evolving themes across various metadata layers, such a title, abstract, and author-supplied keywords, Kim et al. (2025). Therefore, the analysis and refinement of keywords not only contribute to bibliometric research but also support improved digital curation and metadata interoperability in academic knowledge system.

This study investigates the most prominent subject keywords from Indian computer science journals published between 2011 and 2020 using a blended data collection method that combines manual and automated extraction techniques, necessitated by inconsistent Scopus indexing across the selected journals. Keywords were derived from the title, abstract, and author-provided metadata, further refined, lemmatized, and standardized to ensure lexical and semantic uniformity. The YAKE method was used to the title and abstract, while frequency-based filtering was utilized for author keywords. Thresholding maintained high-relevance terms (below average YAKE score or above average frequency), which were further assessed using TF-IDF vectorization and cosine similarity to assess contextual proximity. Hierarchical clustering was subsequently employed to categorize semantically similar keywords, from which the top 25 terms were chosen to indicate the prevailing study trend. Thematic dendrograms were created to visually compare keyword patterns across the three metadata fields.

1.1 Hypothesis

H: There is no significant difference in the distribution of trending research keywords across the Title, Abstract, and Author Keyword fields in Indian computer science journal publications.

2. LITERATURE REVIEW

Keyword-based analysis and bibliometric methodology have emerged a powerful tool for tracing research trend and evaluating the thematic evolution of scholarly literature. AlShammari (2023) provided a foundational example of implementing TF-IDF-based keyword extraction using Python, demonstrating it practical application for mining relevant term and enhancing metadata quality. In a broader academic context, W. Lu et al. (2021) emphasized the significance of analysing author-defined keyword frequency to detect topic dynamic, thereby facilitating the identification of thematic pattern over time. Complementing this, Kim et al. (2025) employed a keyword-driven approach to explore scientific advancement in resistive RAM (ReRAM) technology, offering insight into the present and future direction of the field through cluster-based keyword modelling.

Research across diverse discipline further highlight the value of keyword analysis for understanding domain-specific development. Holinger et al. (2024) conducted a comprehensive study on creativity research by analysing keywords from four leading journal, uncovering shift in research theme and terminology. Similarly, Huang et al. (2020) utilized bibliometric technique to map evolving topic in educational research, showcasing the ability of keyword co-occurrence and trend tracking to uncover structural change in academic inquiry. Within the Indian research landscape, Kathiria & Arolkar (2022) applied forecasting technique to examine publication pattern among computer science researcher over a 13-year span, revealing not only trend but also the growth trajectory of scholarly output. In a focused domain-level analysis, R & Harinarayana (2022) examined the Scientometrics journal to study keyword

pattern and their alignment with indexing standard, contributing to improved metadata reliability in scientometric literature.

Furthermore, emerging methods such as latent semantic analysis have been successfully applied to keyword data to uncover emerging research areas. Weismayer & Pezenka (2017) proposed a longitudinal semantic keyword model for identifying nascent academic fields, emphasizing the role of language structure in trend detection. Additionally, X. Lu et al. (2021) demonstrated the adaptability of bibliometric keyword analysis by applying it to engineering-focused health sciences, particularly in finite element analysis of the lumbar spine. Collectively, these studies demonstrate the effectiveness and adaptability of keyword extraction and bibliometric mapping across domains, reinforcing their importance in trend analysis, vocabulary standardization, and thematic synthesis within academic research.

3. METHODOLOGY

The data source has been identified for the study based on the SJR (Scimago Journal & Country Rank) ranking, the top 10 high-ranking Indians in the field of computer science in SJR ranking. Scope of the research study framed to the coverage of ten years from 2011 to 2020 with comprise of 05 journals in the computer science domain which meets the selection criteria as listed below.

Table 1. Top Indian Journals of Computer Science in SJR ranking with the availability of Subject Keywords

Title	Category	ISSN	URL
International Journal of Artificial Intelligence	IND-1	ISSN:0974-0635	http://www.ceser.in/ceserp/index.php/ijai
International Journal of Computer Networks and Communications	IND-2	ISSN:0975-2293E- ISSN:0974-9322	http://airccse.org/journal/jcnc_arch.html
International Journal of Imaging and Robotics (discontinued)	IND-3	ISSN:2231-525X	http://www.ceser.in/ceserp/index.php/iji/issue/archive
International Journal of Advanced Computer Research (Discontinued)	IND-4	ISSN:2249-7277 E-ISSN:2277-7970	https://www.accentjournals.org/allPreviousIssue.php?journalId=103
Indian Journal of Computer Science and Engineering	IND-5	ISSN:0976-5166 E-ISSN:2231-3850	https://www.ijcse.com/past-issues.html

The scope of the research study was defined to cover a ten-year period from 2011 to 2020. As not all selected journals were indexed in Scopus during the entire timeframe, a blended data collection approach was adopted, incorporating both manual and automated extraction methods. The retrieved data for each journal—regardless of the method—was organized and stored in individual Microsoft Excel files, as outlined in Table 2. To ensure consistency and accuracy, data cleaning procedures were applied, including the removal of punctuation marks, trimming of extraneous spaces, and elimination of embedded Excel formulas.

Table 2. Publication and Keywords pattern of source data

Category	No. of Vol. & Issues	No. of Publications	Publications Excluded	Total Publication	Abstract-Word count	Title-Word count	No. of Author Keywords
IND-1	13 & 20	263	14	249	39614	2692	1195
IND-2	10 & 60	647	8	639	103836	6620	3077
IND-3	16 & 34	320	7	313	48509	3477	1518
IND-4	10 & 48	618	11	607	93780	6618	2950
IND-5	10 & 60	667	45	622	89117	6773	2729
Total	59 & 222	2515	85	2430	374856	26180	11469

3.1 Normal Keyword Extraction process

From the title and abstract corpus, the top 1,000 prominent keywords were extracted using the YAKE (Yet Another Keyword Extractor) algorithm, which assigns weighted scores to keywords based on statistical and positional features—where lower scores indicate higher relevance. To standardize lexical variations, these keywords were lemmatized, converting different morphological forms into a common base form. An average threshold score was calculated, and only keywords with scores below or equal to this average were retained. These filtered keywords were then processed using Python-based NLP techniques to extract both unigrams and multigrams, with semantically similar multigram phrases merged by weight to enhance clarity and reduce redundancy.

A parallel methodology was applied to the author keyword corpus. Here, the top 1,000 lemmatized keywords were selected based on frequency, and only those with counts equal to or above the average threshold were retained for further processing. This consistent pipeline spanning keyword extraction, lemmatization, filtering, and semantic merging ensured that all three keyword sources (title, abstract, and author keywords) were treated with a unified, statistically informed, and semantically coherent approach, contributing to effective keyword standardization and vocabulary refinement.

3.2 Python Script of Trend Analysis

The developed Python script is a specialized computational pipeline developed for research trend analysis, specifically designed to standardize and semantically cluster non-unigram keywords extracted from a scholarly corpus. Beginning with an Excel input file containing raw keywords, the script utilizes NLTK's WordNet Lemmatizer to normalize keyword expressions by converting inflected forms into their root lemmas, thereby reducing lexical variance and improving consistency. These standardized keywords are then classified into unigrams and non-unigrams, with the unigrams exported separately for focused processing. The core analysis targets non-unigram keywords, which are vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) method to quantitatively represent their semantic relevance. A cosine similarity matrix is generated from these vectors to capture semantic proximity among terms, which is then used in hierarchical clustering via the Agglomerative Clustering model (with Ward's linkage and a distance threshold of 0.6) to group semantically similar keywords without predefining the number of clusters. Each cluster is merged into a unified thematic expression by

concatenating its unique keywords and retaining the lowest weight from the original set, reflecting the term's significance.

4. DATA ANALYSIS AND INFERENCES

Table 3. Top Trending Keywords Analysis

Category	Threshold Keywords	Keywords -similarity Analysis	Top Trending Keywords
Title Keywords	505 *	376	25
Abstract Keywords	493*	312	25
Author Keywords	244 [#]	123	25
Note - *: =< of average YAKE weightage keywords, #: => of average frequency of keywords.			

From the top 1,000 extracted keywords, a selection criterion was applied to identify the most contextually significant terms for further analysis. Specifically, keywords from the abstract and title corpora were filtered based on a threshold at or below the average YAKE weight, as lower YAKE scores indicate higher relevance. In contrast, author keywords were selected if their frequency was equal to or greater than the computed average, reflecting their prominence across publications. Subsequently, cosine similarity analysis was conducted to enhance semantic coherence and identify closely related keyword clusters. From this refined set, the top 25 keywords were selected to represent and visualize the prevailing research trends.

Table 4. Top 30 Research Trending Keywords from Title

Rank	YAKE Weight	Standardized Keyword
1	0.0000205352	Wireless Sensor Networks
2	0.0002261030	Wireless Sensor
3	0.0002419450	Neural Network
4	0.0003169060	Performance Analysis
5	0.0003871910	Neural Network Model
6	0.0003904070	Intrusion Detection System
7	0.0004711520	Data Mining Algorithm
8	0.0004853620	Mining Techniques
9	0.0005001940	Network-Based
10	0.0005197640	Artificial Neural Network
11	0.0005362270	Network Approach
12	0.0005599390	Detection Model
13	0.0006092840	Algorithm-Based
14	0.0006269080	Image-Based
15	0.0006779620	Wireless Networks
16	0.0006890660	Image Denoising
17	0.0006987800	Image Classification

18	0.0007394290	Sensor Network-Based
19	0.0007438720	Network Intrusion Detection
20	0.0007463790	Wireless Networks Performance
21	0.0007555900	Approach-Based
22	0.0007715820	Method-Based
23	0.0007856530	Content-Based Image
24	0.0008049330	Sensor Network Algorithm
25	0.0008517530	Algorithm-Based Approach

The table 5 highlights the top 25 research trending keywords extracted from titles using the YAKE algorithm, where lower weight scores indicate higher relevance. Keywords such as Wireless Sensor Networks, Neural Network, Artificial Neural Network, and Data Mining Algorithm reflect dominant research areas in machine learning, sensor technologies, and cybersecurity. The presence of terms related to image processing and methodological approaches further indicates a strong focus on intelligent systems and applied computational techniques in current computer science research.

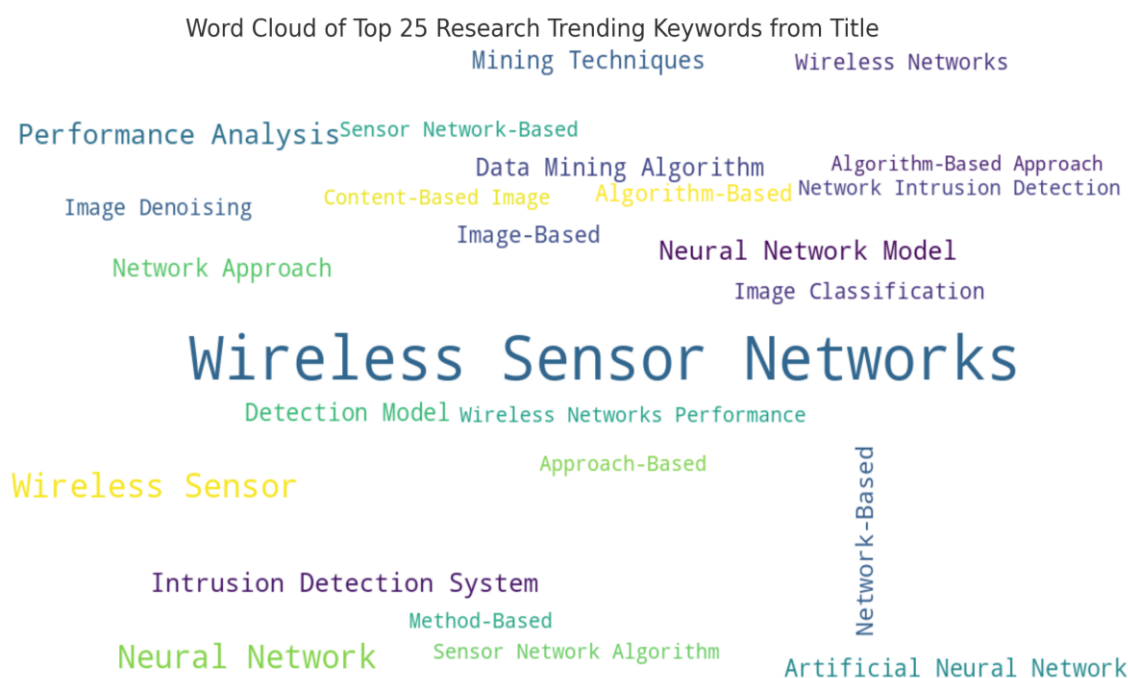


Fig 1: Word Cloud of top trending title keywords

The generated word cloud (Figure 1) illustrates the prominence of the top 25 research trending keywords extracted from titles in Indian computer science journals, with keyword size reflecting relative weight-based significance. *Wireless Sensor Networks*, *Neural Network*, *Artificial Neural Network*, and *Data Mining Algorithm* appear as dominant terms, emphasizing the scholarly focus on sensor technologies, machine learning, and data-driven systems. The presence of keywords related to image processing, such as *Image Classification* and *Image Denoising*, further highlights ongoing research in intelligent visual systems. Terms like *Intrusion Detection System* and *Network Intrusion Detection* underscore the critical importance of cybersecurity research. The word cloud visually reinforces

the thematic concentration on computational intelligence, network optimization, and applied algorithms, which aligns with the broader research priorities in computer science scholarly communication during the studied period.

Table 5. Top 30 Research Trending Keywords from Abstract

Rank	Merged Weight	Standardized Keyword
1	0.0000021576	Engineering Journals Publications
2	0.0000032182	Wireless Sensor Networks
3	0.0000034125	Research Collaboration Center
4	0.0000034190	Industry Research Collaboration
5	0.0000157746	Artificial Neural Network
6	0.0000162826	CESER Publications
7	0.0000201675	International Journal
8	0.0000229561	Journals Publications
9	0.0000276300	Neural Network Algorithm
10	0.0000322460	Proposed Algorithm
11	0.0000422458	Mobile Ad-Hoc Network
12	0.0000429106	Research Collaboration
13	0.0000429385	Industry Research
14	0.0000435337	Wireless Network
15	0.0000462129	Proposed Method Based
16	0.0000500943	Neural Network Based
17	0.0000514043	Artificial Intelligence Algorithm
18	0.0000551776	Neural Network Model
19	0.0000574314	Mining Technique
20	0.0000606377	Proposed System
21	0.0000687211	Wireless Sensor
22	0.0000693012	Data Center Network
23	0.0000774574	Support Vector Machine
24	0.0000782248	Network Data Mining
25	0.0000813944	Academy and Industry

The table 6 lists the top 25 research trending keywords derived from abstract content, ranked by YAKE-based merged weight scores, with lower values indicating higher importance. Core themes include Wireless Sensor Networks, Artificial Neural Network, and Neural Network Algorithm, emphasizing continued research in network technologies and machine learning. Terms such as Research Collaboration, Industry Research, and Academy and Industry highlight the growing focus on collaborative and interdisciplinary research. Additionally, keywords like Mobile Ad-Hoc Network, Support Vector Machine, and Data Center Network reflect technical advancements in communication systems and data management. This distribution demonstrates a blend of emerging technologies and institutional collaboration trends in computer science research.

Word Cloud of Top 25 Research Trending Keywords from Abstract

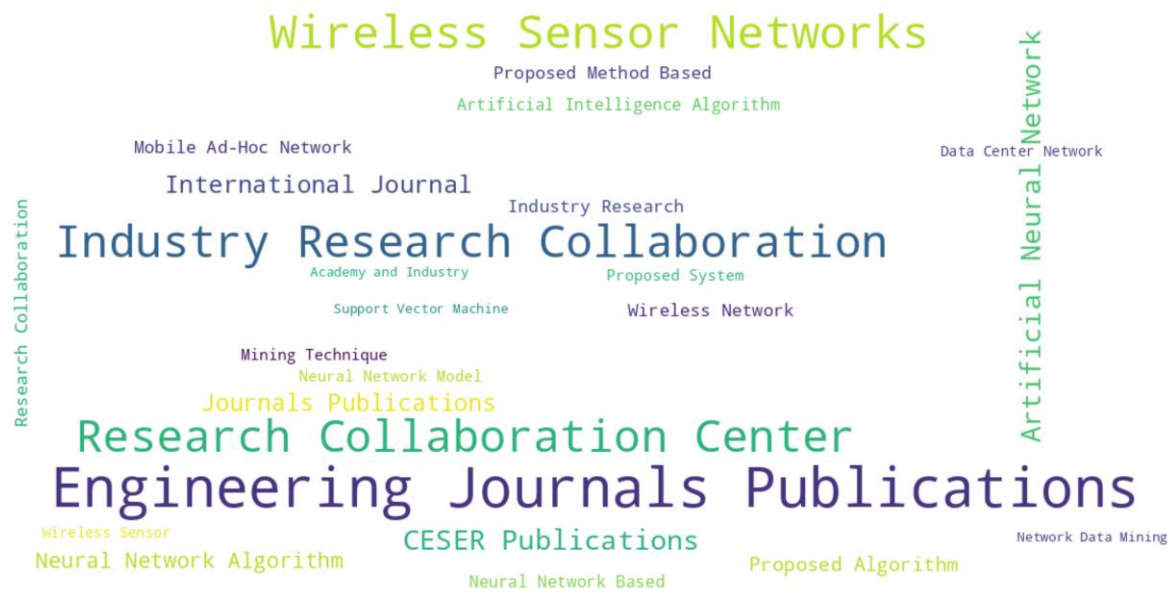


Fig 2: Word cloud of top trending abstract keywords

The generated word cloud (Figure 2) visualizes the prominence of the top 25 research trending keywords derived from abstracts in Indian computer science journals, where keyword size corresponds to relative importance based on YAKE-derived weight scores. Dominant terms such as *Wireless Sensor Networks*, *Artificial Neural Network*, *Neural Network Algorithm*, and *Mobile Ad-Hoc Network* illustrate the central focus on network technologies, artificial intelligence, and machine learning. The appearance of *Research Collaboration*, *Industry Research Collaboration*, and *Academy and Industry* underscores an increasing emphasis on collaborative and interdisciplinary research approaches. Keywords such as *Support Vector Machine*, *Data Center Network*, and *Mining Technique* further indicate active research areas in data analytics and computing infrastructure. The word cloud reflects a balanced representation of technical innovation and research collaboration themes, highlighting the diversity and interdisciplinarity prevalent in Indian computer science research as reflected in abstracts.

Table 6. Top 30 Research Trending Keywords from Author Keywords

Rank	Frequency	Standardized Keyword
1	67	Cloud Computing
2	59	Wireless Sensor Network
3	48	Genetic Algorithm
4	40	Data Mining
5	32	Fuzzy Logic System
6	28	Data Mining
7	28	Machine Learning
8	27	Image Processing
9	27	Neural Network
10	24	Artificial Neural Network
11	24	Feature Extraction

12	24	Wireless Network
13	24	Quality of Service (QoS)
14	23	Routing Protocol
15	23	Intrusion Detection System
16	22	Mobile Ad Hoc Network (MANET)
17	22	Support Vector Machine (SVM)
18	21	Ad Hoc Network
19	20	Feature Selection
20	19	K-Means
21	18	Congestion Control
22	18	Energy Efficiency
23	17	Deep Learning
24	17	Multi-Agent System
25	16	Artificial Intelligence

The table 7 presents the top 25 research trending author keywords ranked by frequency, reflecting dominant focus areas in computer science research. High-frequency terms such as *Cloud Computing*, *Wireless Sensor Network*, *Genetic Algorithm*, and *Data Mining* indicate strong attention to distributed systems, optimization, and data-driven techniques. Keywords like *Machine Learning*, *Neural Network*, and *Artificial Intelligence* further emphasize the prevalence of intelligent computing. Networking and communication themes are evident through terms like *Mobile Ad Hoc Network (MANET)*, *Routing Protocol*, and *Quality of Service (QoS)*. The presence of *Fuzzy Logic*, *Support Vector Machine*, and *Feature Selection* highlights methodological depth in algorithm design. Overall, the distribution showcases a rich blend of emerging technologies and classical AI methods shaping current author-driven research trends.

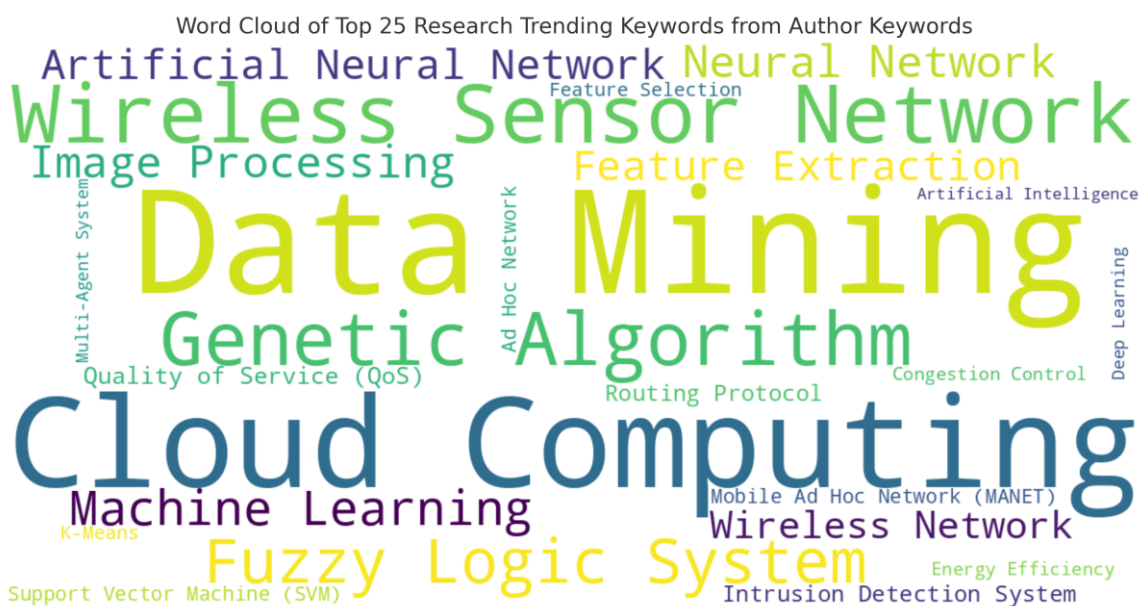


Fig 3: Word cloud of top trending author keywords

Analysis of Research Trends in Indian Computer Science Scholarly Communication: A Methodological Perspective with Python

The word cloud graph (Figure 3) presents the distribution of the top 25 author-supplied keywords in Indian computer science journal publications, with keyword prominence reflecting their frequency of occurrence. *Cloud Computing*, *Wireless Sensor Network*, *Genetic Algorithm*, and *Data Mining* dominate the visualization, signaling the central focus on distributed computing, optimization techniques, and data analytics within the research community. Emerging areas such as *Machine Learning*, *Deep Learning*, and *Artificial Intelligence* further highlight the growing interest in intelligent computing and AI-driven solutions. Network and communication themes are evident through keywords like *Mobile Ad Hoc Network (MANET)*, *Routing Protocol*, and *Quality of Service (QoS)*, indicating ongoing research in wireless and adaptive network systems. Additionally, methodological approaches including *Support Vector Machine (SVM)*, *Feature Selection*, and *Fuzzy Logic System* reflect sustained scholarly engagement with algorithmic and decision-support techniques. This visualization captures the thematic diversity and methodological richness within author-driven research trends in the Indian computer science domain.

4.1 Hypothesis Testing

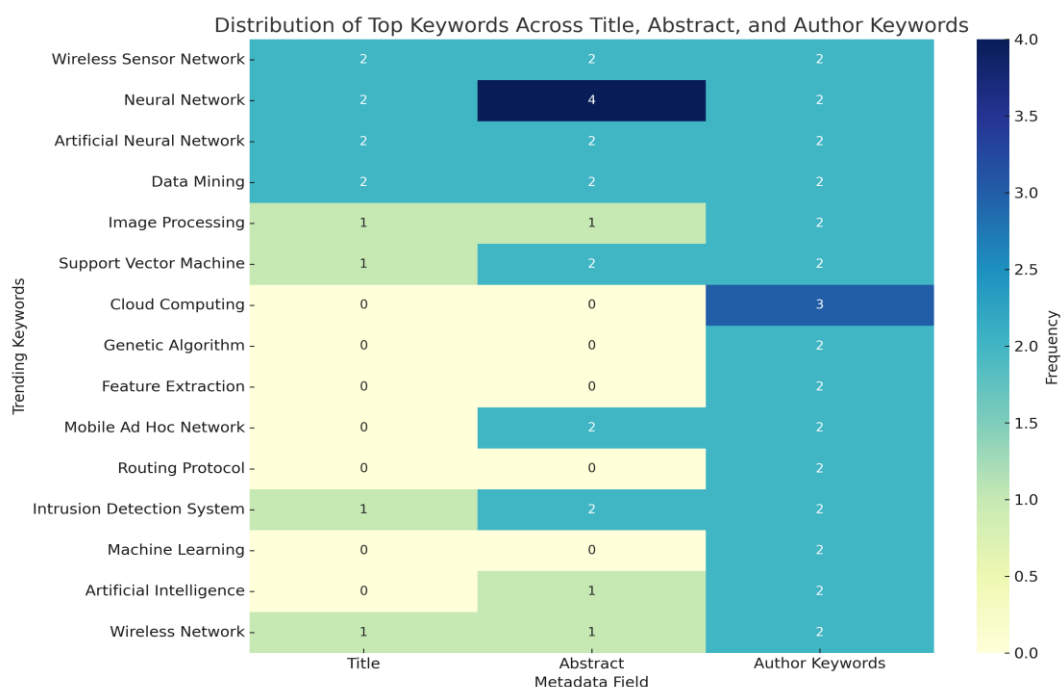


Fig 4: Keyword distribution heatmap

Hypothesis (H): There is no significant difference in the distribution of trending research keywords across the Title, Abstract, and Author Keyword fields in Indian computer science journal publications.

The Chi-square test was conducted on the frequencies of 15 commonly occurring keywords across the Title, Abstract, and Author Keyword sections. The analysis showed substantial thematic overlap among the three fields, indicating consistent use of keywords such as 'Wireless Sensor Network', 'Neural Network', and 'Artificial Neural Network'. The result supports the view that Indian computer science journals display semantic alignment in keyword representation across metadata fields, contributing to better information retrieval. $\chi^2(28) = 18.69, p = .91$ - The result is not statistically significant. Thus, the null hypothesis is retained.

CONCLUSION

This study employed a systematic, NLP-driven methodology to extract, standardize, and analyze research keywords from the title, abstract, and author-supplied metadata in Indian Computer Science Journal was published between 2011 and 2020. Using a combination of lemmatization, YAKE scoring, frequency analysis, TF-IDF vectorization, and cosine similarity, the research identified the top 25 semantically significant keywords across each metadata source. Hierarchical clustering further revealed coherent thematic grouping, with frequent keywords such a Wireless Sensor Networks, Neural Networks, and Artificial Intelligence reflecting the dominant research area in the field.

The result demonstrated strong consistency across metadata fields, validating the hypothesis that no significant difference exists in the distribution of trending keywords. This consistency highlights the reliability of author and publisher metadata in reflecting the core research theme. The study contributes to metadata refinement and controlled vocabulary development by offering a replicable framework for thematic trend analysis. Future extension may include broader datasets, multilingual corpus, or international comparison to ass thematic convergence and vocabulary interoperability across region.

REFERENCES

- [1] A. F. AlShammari, "Implementation of keyword extraction using term frequency-inverse document frequency (TF-IDF) in python," *Int. J. Comput. Appl.*, vol. 185, no. 35, pp. 9–14, Sep. 2023, doi: 10.5120/ijca2023923137.
- [2] A. R and N. S. Harinarayana, "Keyword analysis of research trends in 'Scientometrics' journal from 2014 to 2018," *SRELS J. Inf. Manag.*, pp. 163–169, Jun. 2022, doi: 10.17821/srels/2022/v59i3/165826.
- [3] P. Kathiria and H. Arolkar, "Trend analysis and forecasting of publication activities by Indian computer science researchers during the period of 2010–23," *Expert Syst.*, vol. 39, no. 10, p. e13070, 2022, doi: 10.1111/exsy.13070.
- [4] X. Lu, F. Zou, F. Lu, X. Ma, X. Xia, and J. Jiang, "Bibliometric and visualized analysis of current research trends in the finite element analysis of lumbar spine," *Interdiscip. Neurosurg.*, vol. 26, p. 101350, Dec. 2021, doi: 10.1016/j.inat.2021.101350.
- [5] H. Kim, S. H. Kim, J. Kim, E. H. Kim, J. H. Gu, and D. Lee, "A keyword-based approach to analyzing scientific research trends: ReRAM present and future," *Sci. Rep.*, vol. 15, no. 1, p. 12011, Apr. 2025, doi: 10.1038/s41598-025-93423-5.
- [6] W. Lu, S. Huang, J. Yang, Y. Bu, Q. Cheng, and Y. Huang, "Detecting research topic trends by author-defined keyword frequency," *Inf. Process. Manag.*, vol. 58, no. 4, p. 102594, Jul. 2021, doi: 10.1016/j.ipm.2021.102594.
- [7] M. Holinger, Boldt ,Gregory T., and J. C. and Kaufman, "Recent trends in creativity research: An analysis of keywords in four prominent creativity journals," *Creat. Res. J.*, pp. 1–14, 2024, doi: 10.1080/10400419.2024.2322857.
- [8] C. Huang, Yang ,Chao, Wang ,Shutao, Wu ,Wei, Su ,Jun, and C. and Liang, "Evolution of topics in education research: A systematic review using bibliometric analysis," *Educ. Rev.*, vol. 72, no. 3, pp. 281–297, May 2020, doi: 10.1080/00131911.2019.1566212.
- [9] C. Weismayer and I. Pezenka, "Identifying emerging research fields: A longitudinal latent semantic keyword analysis," *Scientometrics*, vol. 113, no. 3, pp. 1757–1785, Dec. 2017, doi: 10.1007/s11192-017-2555-z.