Khorshidi *et al. Journal of Big Data* (2026) 13:80
https://doi.org/10.1186/s40537-026-01412-w

Journal of Big Data

RESEARCH Open Access

A bibliometric analysis of the first decade of the Journal of Big Data

Mohammad Sadegh Khorshidi[1], José M. Merigó[1*], Ghassan Beydoun[1], Borko Furht[2] and Taghi M. Khoshgoftaar[2]

*Correspondence:
José M. Merigó
Jose.Merigo@uts.edu.au
[1]School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway, Ultimo, NSW 2007, Australia
[2]Department of Electrical Engineering and Computer Science, College of Engineering and Computer Science, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA

Abstract

This study provides a bibliometric overview of the first decade of the *Journal of Big Data* (JBD) since its launch in 2014. Using bibliographic records for 1,023 articles and reviews indexed in Scopus (2014–2024) and 917 records in Web of Science, the analysis combines performance indicators with science-mapping techniques, including co-citation, bibliographic coupling, and keyword co-occurrence implemented in VOSviewer and bibliometrix. The results document rapid growth in output, a Scopus/SCImago *h*-index of 91 by 2024, and strong international contributions led by the USA, China, India, and European countries. Citation and topical structures reveal three main thematic cores centred on big data infrastructures (e.g., Hadoop, MapReduce, Apache Spark), machine and deep learning (e.g., convolutional and recurrent neural networks), and application domains such as cybersecurity, healthcare analytics, and sentiment analysis. The findings characterize JBD's position within the wider data science ecosystem, showing how its publications concentrate influence in a small number of highly cited surveys and frameworks while supporting an increasingly diverse long tail of topics and emerging research fronts.

Keywords Big data, Bibliometrics, Scientometric mapping, Journal retrospectives, Scopus, Web of Science

Introduction

The growth of data-intensive research across computational, social, and organizational domains has established big data analytics as a central epistemological and technological force. This growth, driven by advancements in data generation, storage, and processing, has produced a corresponding surge in scholarly output, institutional investment, and applied development. Within this context, the Journal of Big Data (JBD), launched in 2014 by SpringerOpen, emerged to serve the expanding needs of the research community working at the intersection of large-scale data systems, scalable algorithms, and real-world applications [35]. The journal has developed as an open-access venue for contributions spanning analytics algorithms, infrastructure design, machine learning, natural language processing, cloud and edge computing, and applied domains such as healthcare, finance, cybersecurity, and smart urban systems [58]. These themes all align

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

with the growing complexity and ubiquity of big data as a defining feature of contemporary science and society [9, 20, 78].

JBD's editorial scope spans a wide methodological and disciplinary range. It typically welcomes contributions from fields as varied as computer science, applied mathematics, engineering, information systems, and increasingly, business analytics and the social sciences. The editorial board reinforces this diversity: Editors-in-Chief Professors Borko Furht and Taghi Koshgoftaar (Florida Atlantic University) are supported by a geographically and topically diverse group of editorial members with expertise in data mining, parallel computing, deep learning, high-performance systems, and data ethics [59]. Such breadth is relevant for a domain characterized by heterogeneity in data types, modeling techniques, and application contexts.

During its first decade, JBD has expanded substantially in volume and visibility. From publishing six articles in 2014, the journal grew to over 180 articles yearly by 2023 and accumulated more than 1,000 publications within a decade [88]. Several of its papers have become widely cited within the global data science community. For example, Shorten and Khoshgoftaar's [93] survey on image data augmentation has accumulated over 9,000 citations, while Alzubaidi et al.'s [6] comprehensive review of deep learning has accumulated over 5,000 citations and more than 600,000 downloads. These citation and download figures, together with Altmetric Attention Scores and dissemination on Twitter/X, blogs, and policy portals, indicate substantial academic and public uptake [101]. JBD also scores highly on key bibliometric benchmarks. According to Clarivate [23], the journal reached a 2023 Journal Impact Factor of 8.6, with a five-year impact factor of 12.4. In Scopus-based indicators, JBD's 2023 CiteScore is 22.3, placing it above benchmark outlets such as *Expert Systems with Applications* (CiteScore 15.0; Journal Impact Factor 7.5) and *Information Sciences* (CiteScore 14.4; Journal Impact Factor 6.8). In Scopus and SCImago, its *h*-index reached 91 by 2024, and its SJR value of 1.979 confirmed its Q1 ranking in Information Systems and Computer Science Applications [88]. These metrics indicate that JBD currently outperforms several comparable journals such as *Big Data* (Mary Ann Liebert; Journal Impact Factor 2.0; CiteScore 9.0), *ACM Journal of Data and Information Quality* (*h*-index 29), and *Data Science Journal*, which, while respected, lacks JCR indexing. The full open-access model enhances discoverability and allows for global readership without subscription barriers.

As big data continues to shape research across disciplines [70], journals like JBD provide a venue for disseminating new methods, theories, and use cases. Its performance in both traditional and alternative metrics, including over 2.7 million downloads in 2024 and coverage on media platforms, indicates that it operates both as a scholarly venue and as a conduit for real-world engagement [101]. This dual role is particularly visible in domains concerned with ethical AI, public-sector analytics, and digital transformation [30, 113].

While performance and altmetric indicators reflect JBD's prominence, they do not provide visibility of its thematic and internal evolution. Bibliometric analysis offers a structured methodology for evaluating the intellectual trajectory, thematic concentration, and citation dynamics of academic journals [38, 76]. In recent years, journal-level bibliometric retrospectives have been adopted by major publication outlets across disciplines as a means of reflecting on institutional influence and research direction. Examples include retrospectives of *Omega* [55], *Information Processing & Management*

[62], *Environmental and Resource Economics* (Polyakov et al., 2018), and *Scandinavian Journal of Economics* [33]. Such retrospectives are not new and have been carried out for a long time, for example, in the *Journal of Financial Economics* (Schwert, 1993), *The Accounting Review* [49], and the *Journal of Urban Economics* [4]. Such studies benchmark the scientific contribution of journals through citation indicators (e.g., *h*-index, citations per paper, total outputs) and map the structural relationships among articles using co-citation, bibliographic coupling, and keyword clustering techniques [18, 61, 97, 110, 111]. They enable identification of dominant research clusters, influential authors and institutions, temporal trends in knowledge production, and evolving topical foci.

Given JBD's interdisciplinary scope and rapid ascent, a bibliometric retrospective can illuminate how the journal's internal structures align with the broader evolution of big data research, while offering evidence of its impact on both foundational and applied domains. In contrast to earlier journal-level retrospectives, this study offers a decade-long examination of JBD that highlights how its publication growth, citation patterns, and thematic clusters differ from the trajectories observed in comparable information systems and data science outlets. By contextualizing JBD's internal evolution alongside broader shifts in big data research, the study provides a holistic view of how the journal has developed as a focal point for emerging analytical and computational themes. In light of these important considerations, the present study provides a comprehensive bibliometric overview of the JBD from 2014 to 2024. The analysis is guided by six research questions: (1) What is the current position of JBD in the scientific community? (2) Which are the most cited articles of JBD? (3) Which documents does JBD cite more frequently? (4) Who are the most productive authors, institutions, and countries of JBD? (5) How is the co-citation and bibliographic coupling structure of JBD? (6) What is the keyword and topical structure of JBD?

To address these questions, we adopt a dual-method strategy combining performance analysis and structural mapping. Performance metrics include total publications, citations, average citations per paper, and *h*-index. Structural analysis is performed using bibliographic coupling, co-citation analysis, and keyword co-occurrence techniques implemented in VOSviewer [110, 116]. This integrated approach has been validated in earlier retrospectives [7, 30] and is especially relevant for journals like JBD that operate across multiple disciplinary boundaries.

The remainder of this article is organized as follows. Sect. "Methods" describes the data collection process and bibliometric techniques used in the study. Sect. "Results" presents the results, beginning with an overview of publication and citation patterns (3.1), influential papers (3.2), and citing articles (3.3)Sect. "Mapping JBD with VOSviewer software" maps JBD's structural dynamics using VOSviewer, including bibliographic coupling and co-citation analysis (4.1) and topical clustering based on keywords (4.2). Sect. "Conclusions" concludes the study by summarizing the key findings (5.1), discussing practical implications (5.2), and identifying limitations and avenues for future research (5.3). In addition, detailed results on the most productive authors, institutions, and countries are reported in Supplementary Material S1, and the corresponding bibliographic-coupling analyses for these units are presented in Supplementary Material S2.

Methods

As outlined in the introduction, the increasing prominence of JBD within the global scholarly landscape warrants a structured bibliometric examination to trace its intellectual development, citation performance, and thematic evolution. Bibliometrics, as originally conceptualized by Alan Pritchard [76], refers to the quantitative analysis of written publications and has since evolved to encompass a broad range of citation-based and network-oriented indicators [17, 82]. Early work on citation indexing by Eugene Garfield [37] led to the institutionalization of performance indicators such as the impact factor [12], providing much of the conceptual and technical foundation subsequently used in bibliometric analysis [10, 39]. In this study, a combination of performance analysis and science mapping is used to examine the growth trajectory and intellectual landscape of JBD.

Following established protocols in bibliometric research [30, 74], the methodological design is guided by the SPAR-4-SLR framework, which structures scientific reviews into the stages of identification, acquisition, purification, evaluation, and reporting. Table 1 summarizes how each phase is implemented in this study. The objective is to provide a comprehensive bibliometric retrospective of JBD's publication history from 2014 to 2024, aligned with the six research questions introduced in the previous section on journal positioning, influential works, citation patterns, author productivity, intellectual structure, and topical development. To operationalize these questions, Elsevier's Scopus database is used as the primary source for JBD publications and their citing documents, because of its broad coverage of peer-reviewed literature in science and technology and its integration with citation and subject classification tools [89].

The Scopus search was conducted in June 2025 using the "Source Title" filter restricted to *Journal of Big Data* and the publication years 2014–2024. The initial retrieval produced 1,220 records. After refining the results by enforcing an exact source-title match and excluding non-journal content (e.g., editorials, notes, corrections, conference papers, and early-access records without final bibliographic information), 1,171 documents remained. A further restriction to the 2014–2024 window, excluding records dated 2025 to avoid incomplete citation windows, reduced the dataset to 1,031 items. Limiting the corpus to documents classified as "Article" or "Review" yielded a final set of 1,023 JBD publications. This Scopus corpus underpins all article-level performance indicators and rankings reported in the results. The full search strings, field tags, and filter settings used in Scopus are provided in the Supplementary Material to facilitate replication.

To complement the Scopus dataset, a parallel search was carried out in Clarivate's Web of Science (WoS) Core Collection. WoS is widely recognized for its curated indexing and compatibility with Journal Citation Reports (JCR) [23, 113]. Because JBD was indexed in WoS at a later stage than in Scopus, the WoS search was constrained to the 2018–2024 period and conducted in the "Publication Titles" field using *Journal of Big Data* as the exact journal name. The initial WoS retrieval returned 1,070 records. After excluding documents dated 2025, removing retracted items, and restricting the corpus to articles and reviews, the final WoS dataset comprised 917 documents. WoS data are used for journal-level indicators and science mapping.

Table 1 Summary of the bibliometric review protocol for the JBD (2014–2024) based on SPAR-4-SLR protocol

	Activity
	Identification
Assembling	Domain: Journal of Big Data (JBD)
	Research objective: Provide a bibliometric retrospective of JBD
	Research questions: 1) What is the current position of JBD in the scientific community? 2) Which are the most cited articles of JBD? 3) Which documents does JBD cite more frequently? 4) Who are the most productive authors, institutions and countries of JBD? 5) How is the co-citation and bibliographic coupling structure of JBD? 6) How is the keyword and topical structure of JBD?
	Acquisition
	Search platform: Scopus
	Search period: 2014–2024. The search was performed in June 2025
	Source Title: Journal of Big Data. Results = 1,220
	Filter Source Title: Journal of Big Data: 1,171 documents
	Search in Web of Science Core Collection (Publication Titles) (2018–2024): 1,070 documents
	Purification
Arranging	Filter 1. Exclude 2025 (2014–2024): 1,031 documents Filter 2: Article & Review. Final number: 1,023 documents
	WoS Core Collection (with equivalent filters): 917 documents
	Computer software: VOSviewer (version 1.6.20), bibliometrix (version 4.1), biblioshiny and Microsoft Excel
	Additional results: JCR (WoS), Sources (Scopus) and SciVal (Scopus)
	Evaluation
Assessing	Performance analysis: 1) Rankings with bibliometric indicators according to the number of articles and citations 2) Bibliometric indicators: Number of documents and citations, cites per paper, the *h*-index, cites per year, papers and cites per capita, and temporal classification
	Graphical analysis: 1) Mapping with VOSviewer and bibliometrix software 2) Analysis with co-citation, bibliographic coupling, and co-occurrence of author keywords
	Agenda proposal method: 1) Present a bibliometric review of JBD during the last decade (2014–2024) 2) Analyze the leading and current trends of JBD 3) Study open questions for future research
	Reporting
	Reporting conventions: Explanations with numbers, tables and figures
	Limitations: 1) The bibliographic data is from June 2025 but changes through time 2) The drawbacks of the Scopus and WoS databases affect this study 3) It is not easy to compare the publication and citation patterns of different subfields

Data cleaning and normalization

Before analysis, the Scopus and WoS datasets were independently cleaned and harmonized. Within each database, duplicate records were identified and removed by jointly matching article titles, DOIs, author lists, and publication years. Document types were reverified so that only peer-reviewed research articles and reviews were retained, and early-access records were assigned to their final publication year or excluded when final metadata were not yet available. Journal titles, ISSNs, publisher names, and country and institutional affiliations were standardized across records using the identifiers and thesauri supplied by each database, supplemented by manual inspection for high-frequency entities and obvious variants.

For performance analysis, citation indicators derived from Scopus (e.g., total citations and citations per paper) are reported directly for the JBD corpus. Journal-level indicators from WoS (Impact Factor, five-year Impact Factor, and subject-category percentiles) and from Scopus (CiteScore and subject percentiles, including SciVal-based metrics) are reported separately and compared using normalized rankings rather than combined raw counts, to avoid double-counting and cross-database inconsistencies.

For science-mapping analyses, the cleaned Scopus JBD corpus (2014–2024) and the WoS-based list of external references (2018–2024) are used as separate inputs. Co-citation and bibliographic-coupling networks are built either on JBD publications (to analyze JBD's internal intellectual structure) or on the external reference set (to analyze its citation backbone).

Microsoft Excel is used for preliminary data inspection, descriptive statistics, and ranking of publications, authors, institutions, and countries. Network-based visualizations are generated using VOSviewer [110], which supports the construction and visualization of bibliographic-coupling [61], co-citation [97], and keyword co-occurrence networks [18]. Additional bibliometric indicators and structural metrics are computed using the *bibliometrix* R package and its web-based interface biblioshiny [7]. These tools enable the graphical exploration of influential documents, collaboration and citation structures, and thematic clusters.

The evaluation phase combines performance analysis with graphical structural mapping. Performance indicators include the total number of documents, total citations, citations per paper, annual publication growth, the h-index [52], and temporal classification of output. These metrics are used to assess JBD's relative standing among peer journals and to quantify the scholarly impact of its contributions [42, 62].

Graphical analyses focus on three primary dimensions: (1) co-citation analysis to identify shared intellectual roots, (2) bibliographic coupling to map topical proximity among publications, and (3) co-occurrence analysis of author keywords to uncover thematic structures. Networks in VOSviewer are generated using fractional counting and association-strength normalization, which limit the influence of highly prolific units and provide comparable similarity metrics across maps [7, 110, 111]. For each network, conservative inclusion thresholds (e.g., minimum citation or occurrence counts and a minimum number of links for co-citation and bibliographic-coupling graphs) are chosen after inspecting the size and readability of candidate maps,the exact thresholds used for each figure are reported in the corresponding captions. Analyses are performed twice, once on the full JBD corpus and once on the subset of top-cited papers, to compare the core and peripheral knowledge structures. Sensitivity checks with adjacent threshold values and cross-verification of aggregate citation counts against the underlying Scopus and WoS exports indicate that the main clusters and link patterns are robust to reasonable variations in parameter settings.

Finally, the reporting phase follows transparent documentation practices consistent with SPAR-4-SLR. Descriptive statistics, tables, and visualizations are used jointly to explain the findings, while acknowledging three main limitations: (1) bibliographic data retrieved in June 2025 are subject to future updates; (2) the coverage and metadata standards of Scopus and WoS differ, which may constrain cross-database comparisons; and (3) heterogeneity in citation and publication patterns across subfields within big data research complicates the interpretation of influence in strictly comparable terms.

Results

Publication and citation structure of JBD

Tracing the publication volume and cumulative citation yearly provides a foundational perspective on the journal's scholarly positioning over time. From 2014 to 2024, JBD published 1,023 documents, collectively receiving 64,985 citations. Table 2 presents the annual citation trend of JBD from 2014 to 2024, offering a comprehensive view of the journal's productivity, citation accumulation, and influence stratification across its first decade. The data reveal a consistent upward trend in annual output, from only 6 papers in 2014 to 181 in 2024, with the most rapid growth phase beginning in 2019. This quantitative rise reflects both increased submission volumes and the journal's consolidation as a visible outlet within the data science ecosystem. However, the growth in output has not been accompanied by a proportional increase in citation yield, as the temporal distribution of total citations is heavily skewed toward earlier cohorts. For instance, while the journal's output increased nearly 17-fold between 2014 and 2024, total annual citations peaked in 2019 (20,115) and have since declined, with 2024 attracting only 1,655 citations despite being the most prolific publication year. This pattern is partly explained by shorter citation windows for recent cohorts but also suggests a shift from a few highly visible synthesis papers in the journal's formative years toward a broader base of more specialized contributions, which is typical as a field moves from emergence to consolidation.

The citation concentration in 2019 is particularly notable. With only 112 publications, this year accounts for 30.9% of all citations received by the journal over the decade, driven primarily by the publication of several highly cited survey articles, as confirmed in Table 3. By contrast, years with higher output such as 2021 (160 papers) and 2023 (179 papers) received significantly fewer citations—11,417 and 3,159, respectively—illustrating the temporal lag inherent in citation accumulation but also suggesting a saturation of mid-tier contributions in later years. The concentration of citations around the 2018–2019 cohorts is largely driven by a small set of survey and methodological overview articles on deep learning and big data infrastructures that rapidly became 'go-to' references for subsequent work. By contrast, many later publications focus on more specialized or domain-specific applications, which have shorter citation windows and

Table 2 Annual publication output and citation stratification of JBD (2014–2024)

Year	TP	TC	≥ 500	≥ 100	≥ 50	≥ 20	≥ 10	≥ 5	≥ 1	T50
2014	6	349	0	1	2	3	3	4	6	1
2015	26	5,829	3	13	14	20	22	24	26	9
2016	26	5,656	1	2	5	15	22	25	26	2
2017	49	2,221	0	8	14	25	35	44	48	1
2018	53	3,416	2	8	15	26	39	46	51	4
2019	112	20,115	5	22	36	76	93	102	112	12
2020	107	7,630	2	18	30	64	80	95	107	11
2021	160	11,417	2	13	38	79	116	138	157	6
2022	124	3,538	0	8	14	41	70	96	122	3
2023	179	3,159	0	3	13	38	82	132	169	1
2024	181	1,655	0	2	8	21	37	66	144	0
Total	1,023	64,985	15	98	189	408	599	772	968	50
%	100	-	1.5	9.6	18.5	39.9	58.6	75.5	94.6	4.9

TP and TC = Total papers and citations; ≥ 500, ≥ 100, ≥ 50, ≥ 20, ≥ 10, ≥ 5, ≥ 1 = Number of papers with equal or more than 500, 100, 50, 20, 10, 5 and 1 citations; T50 = Number of documents in the Top 50 of Table 3

Table 3 Top 50 most cited articles in JBD, ranked by total citations [1–3, 5, 8, 13, 21, 24, 26, 32, 46, 51, 53, 56, 60, 65, 66, 72, 73, 85, 86, 91, 92, 98–100, 102, 103, 105–107, 115, 119] and citation rate (2014–2024)

R	Title	Author/s	Year	TC	C/Y
1	A survey on image data augmentation for deep learning	Shorten C.; Khoshgoftaar T.M	2019	8,833	1,472.17
2	Review of deep learning: concepts, CNN architectures, challenges, applications, future directions	Alzubaidi L.; Zhang J.; Humaidi A.J.; Al-Dujaili A.; Duan Y.; Al-Shamma O.; Santamaría J.; Fadhel M.A.; Al-Amidie M.; Farhan L	2021	4,874	1,218.50
3	A survey of transfer learning	Weiss K.; Khoshgoftaar T.M.; Wang D.D	2016	4,721	524.56
4	Survey on deep learning with class imbalance	Johnson J.M.; Khoshgoftaar T.M	2019	2,024	337.33
5	Deep learning applications and challenges in big data analytics	Najafabadi M.M.; Villanustre F.; Khoshgoftaar T.M.; Seliya N.; Wald R.; Muharemagic E	2015	1,960	196.00
6	Big data in healthcare: management, analysis and future prospects	Dash S.; Shakyawar S.K.; Sharma M.; Kaushik S	2019	1,163	193.83
7	CatBoost for big data: an interdisciplinary review	Hancock J.T.; Khoshgoftaar T.M	2020	880	176.00
8	Internet of Things is a revolutionary approach for future technology enhancement: a review	Kumar S.; Tiwari P.; Zymbler M	2019	869	144.83
9	Deep convolutional neural network based medical image classification for disease diagnosis	Yadav S.S.; Jadhav S.M	2019	838	139.67
10	Big data analytics: a survey	Tsai C.-W.; Lai C.-F.; Chao H.-C.; Vasilakos A.V	2015	668	66.80
11	Selecting critical features for data classification based on machine learning methods	Chen R.-C.; Dewi C.; Huang S.-W.; Caraka R.E	2020	644	128.80
12	Sentiment analysis using product review data	Fang X.; Zhan J	2015	584	58.40
13	A survey on addressing high-class imbalance in big data	Leevy J.L.; Khoshgoftaar T.M.; Bauder R.A.; Seliya N	2018	582	83.14
14	A survey on missing data in machine learning	Emmanuel T.; Maupong T.; Mpoeleng D.; Semong T.; Mphago B.; Tabona O	2021	577	144.25
15	Big healthcare data: preserving security and privacy	Abouelmehdi K.; Beni-Hessane A.; Khaloufi H	2018	523	74.71
16	Cybersecurity data science: an overview from machine learning perspective	Sarker I.H.; Kayes A.S.M.; Badsha S.; Alqahtani H.; Watters P.; Ng A	2020	480	96.00
17	Uncertainty in big data analytics: survey, opportunities, and challenges	Hariri R.H.; Fredericks E.M.; Bowers K.M	2019	453	75.50
18	Survey on categorical data for neural networks	Hancock J.T.; Khoshgoftaar T.M	2020	420	84.00
19	Text data augmentation for deep learning	Shorten C.; Khoshgoftaar T.M.; Furht B	2021	418	104.50
20	A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications	Alzubaidi L.; Bai J.; Al-Sabaawi A.; Santamaría J.; Albahri A.S.; Al-dabbagh B.S.N.; Fadhel M.A.; Manoufali M.; Zhang J.; Al-Timemy A.H.; Duan Y.; Abdullah A.; Farhan L.; Lu Y.; Gupta A.; Albu F.; Abbosh A.; Gu Y	2023	393	196.50
21	Step away from stepwise	Smith G	2018	392	56.00
22	Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset	Kasongo S.M.; Sun Y	2020	386	77.20
23	A survey of open source tools for machine learning with big data in the Hadoop ecosystem	Landset S.; Khoshgoftaar T.M.; Richter A.N.; Hasanin T	2015	369	36.90

Table 3 (continued)

R	Title	Author/s	Year	TC	C/Y
24	Survey of review spam detection using machine learning techniques	Crawford M.; Khoshgoftaar T.M.; Prusa J.D.; Richter A.N.; Al Najada H	2015	363	36.30
25	Intelligent video surveillance: a review through deep learning techniques for crowd analysis	Sreenu G.; Saleem Durai M.A	2019	342	57.00
26	A survey on platforms for big data analytics	Singh D.; Reddy C.K	2015	337	33.70
27	A systematic review and research perspective on recommender systems	Roy D.; Dutta M	2022	318	106.00
28	Customer churn prediction in telecom using machine learning in big data platform	Ahmad A.K.; Jafar A.; Aljoumaa K	2019	316	52.67
29	The use of big data analytics in healthcare	Batko K.; Ślęzak A	2022	314	104.67
30	Smart literature review: a practical topic modelling approach to exploratory literature review	Asmussen C.B.; Møller C	2019	314	52.33
31	A comprehensive survey of anomaly detection techniques for high dimensional big data	Thudumu S.; Branch P.; Jin J.; Singh J.J	2020	310	62.00
32	Boosting methods for multi-class imbalanced data classification: an experimental review	Tanha J.; Abdi Y.; Samadi N.; Razzaghi N.; Asadpour M	2020	304	60.80
33	Big data privacy: a technological perspective and review	Jain P.; Gyanchandani M.; Khare N	2016	303	33.67
34	A survey on heterogeneous transfer learning	Day O.; Khoshgoftaar T.M	2017	302	37.75
35	Analysis of diabetes mellitus for early prediction using optimal features selection	Sneha N.; Gangil T	2019	296	49.33
36	Big Data: Deep Learning for financial sentiment analysis	Sohangir S.; Wang D.; Pomeranets A.; Khoshgoftaar T.M	2018	286	40.86
37	Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities	Seyedan M.; Mafakheri F	2020	276	55.20
38	Comparative analysis of deep learning image detection algorithms	Srivastava S.; Divekar A.V.; Anilkumar C.; Naik I.; Kulkarni V.; Pattabiraman V	2021	269	67.25
39	Deep Learning applications for COVID-19	Shorten C.; Khoshgoftaar T.M.; Furht B	2021	266	66.50
40	Short-term stock market price trend prediction using a comprehensive deep learning system	Shen J.; Shafiq M.O	2020	252	50.40
41	Visualizing big data with augmented and virtual reality: challenges and research agenda	Olshannikova E.; Ometov A.; Koucheryavy Y.; Olsson T	2015	251	25.10
42	Intrusion detection and big heterogeneous data: a Survey	Zuech R.; Khoshgoftaar T.M.; Wald R	2015	251	25.10
43	A review of data mining using big data in health informatics	Herland M.; Khoshgoftaar T.M.; Wald R	2014	251	22.82
44	Transfer learning: a friendly introduction	Hosna A.; Merry E.; Gyalmo J.; Alom Z.; Aung Z.; Azim M.A	2022	245	81.67
45	An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities	O'Donovan P.; Leahy K.; Bruton K.; O'Sullivan D.T.J	2015	239	23.90
46	Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage	Sarker I.H.; Kayes A.S.M.; Watters P	2019	217	36.17
47	Sensor data quality: a systematic review	Teh H.Y.; Kempa-Liehr A.W.; Wang K.I.-K	2020	214	42.80
48	On the sustainability of smart and smarter cities in the era of big data: an interdisciplinary and transdisciplinary literature review	Bibri S.E	2019	212	35.33
49	A comprehensive evaluation of ensemble learning for stock-market prediction	Nti I.K.; Adekoya A.F.; Weyori B.A	2020	211	42.20
50	A survey on data-efficient algorithms in big data era	Adadi A	2021	210	52.50

R = Rank; TC = Total citations; C/Y = Cites per year

naturally accumulate fewer citations at this stage. This skewed pattern is consistent with the broader literature on citation concentration in fast-growing fields and reflects differences in topical breadth, perceived foundational value, and publication timing rather than simple differences in quality across JBD contributions.

The citation stratification metrics provide further insight into the distribution of influence. Of the 1,023 documents published, only 15 (1.5%) exceeded 500 citations and 98 (9.6%) exceeded 100 citations, highlighting the dominance of a small elite subset of publications. Nonetheless, the middle tiers exhibit robust performance: 408 papers (39.9%) attracted at least 20 citations and 599 (58.6%) received 10 or more citations, indicating sustained scholarly attention beyond the top-cited fringe. Furthermore, 94.6% of papers received at least one citation, suggesting a high retrieval rate across the corpus with minimal uncited output. This pattern reflects both the journal's high visibility and the applicability of its published works across a wide spectrum of domains.

Temporal trends in the number of highly cited articles also reveal shifting citation dynamics. For example, 2015 produced 13 papers with 100 + citations and three papers exceeding 500 citations, while 2022 and 2023, with larger outputs, generated fewer papers in the highest citation bands. This reflects a dilution effect, where a growing share of annual publications does not necessarily translate to more highly cited work. It also underscores the outsized impact of early contributions during formative years when fewer publications received broader attention within the emerging field.

The rightmost column of Table 2, showing the number of documents appearing in the journal's Top 50 cited articles, adds another layer of insight. Years 2015, 2019, and 2020 clearly dominate, contributing 9, 12, and 11 papers respectively to the Top 50 list. Later years, including 2023 and 2024, contribute only one or zero documents, underscoring the time-dependent nature of citation accumulation. While some lag is expected, the near absence of recent publications in the upper echelon indicates that, despite increased productivity, later-year outputs have not yet reached comparable visibility or influence. For JBD, this implies that visibility in the big data community has so far been driven by a small set of broadly scoped surveys and reviews, while recent years are characterized by a larger volume of more targeted work with yet unrealized citation potential. From an editorial perspective, these dynamics highlight the need to balance high-level synthesizing articles with focused methodological and application-oriented studies as the journal matures.

The box-whisker plot in Fig. 1 provides an overall overview of annual citation performance, revealing strong inter-year heterogeneity and the presence of high-leverage outliers that shape overall metrics. While Table 2 highlighted temporal peaks in total citations, Fig. 1 disaggregates those totals to show variability within each year. For instance, 2015, 2016, 2019, 2020, and 2021 stand out not only due to their total citations (5,829; 5,656; 20,115; 7,630; and 11,417, respectively) but also for their skewed distributions with extreme outliers exceeding 700 citations. In 2015, despite a modest number of publications ($n = 26$), one paper surpassed 1,900 citations, inflating the mean far above the median and indicating the influence of a small set of dominant articles. The whiskers and interquartile ranges have shrunk markedly in recent years, particularly from 2022 to 2024, suggesting a more compressed citation distribution with fewer standout contributions. However, this pattern should be interpreted with caution: the apparent absence of highly cited papers in these years is largely due to citation latency, as recent publications

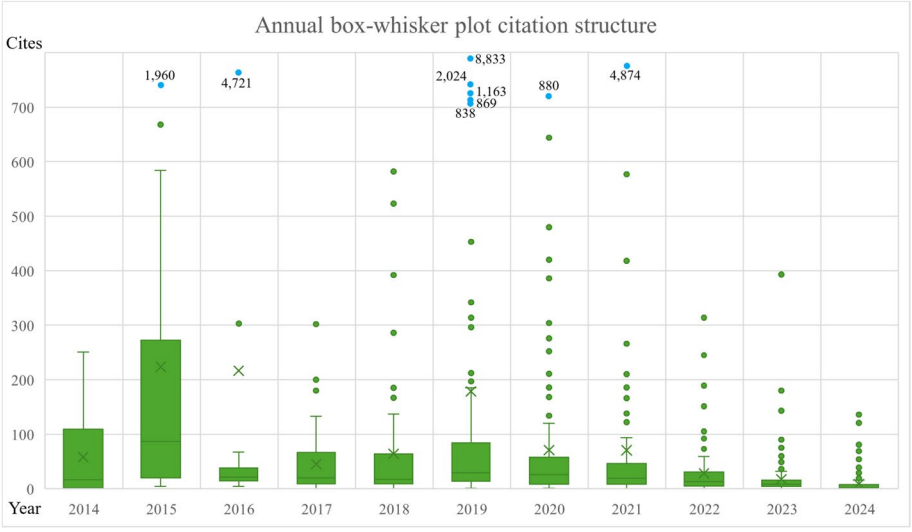


Fig. 1 Citation distribution of JBD publications by year, highlighting interquartile ranges and extreme outliers (2014–2024)

have yet to accumulate sufficient citations. This is consistent with the underrepresentation of recent papers in the Top 50 cited list (Table 2) and reflects a structural delay rather than a decline in scholarly quality.

The elevated presence of single-paper outliers in 2016, 2019, and 2021 reinforces the role of influential surveys and reviews in driving annual citation spikes. Meanwhile, the relative consistency of median values across recent years indicates a stabilization of baseline impact. Figure 1 confirms that JBD's citation landscape is right-skewed, with long-tailed distributions typical of interdisciplinary journals that attract both foundational syntheses and more specialized empirical contributions. The visual stratification illustrates a shift from earlier citation concentration toward broader, though less extreme, scholarly uptake in later phases—albeit shaped by the citation window of each cohort.

Figure 2 depicts the annual performance of JBD across two leading bibliometric databases—Clarivate's WoS [23] and Elsevier's Scopus [89]—capturing the evolution of its Impact Factor, CiteScore, and disciplinary percentiles from 2016 to 2024. The journal's Scopus-based CiteScore has demonstrated a marked upward trend, rising from 4.5 in 2016 to a peak of 23.9 in 2022, before stabilizing at 22.3 in 2024. This trajectory reflects a sustained increase in article influence across broader citation windows and reaffirms the journal's growing recognition among data science researchers and applied technologists.

In contrast, the WoS Impact Factor, available from 2021 onward due to delayed indexing, evolved from 10.83 in its inaugural year to 6.4 in 2024, indicating a strong entry but subsequent moderation in two-year citation density. Despite this numerical decrease, the journal retained its WoS Percentile ranking above the 90th percentile, signaling a consistent placement in the top decile within its categories. The Best Percentile in Scopus, reflecting JBD's highest placement across all assigned subject areas, remained above 80% throughout 2016–2024, with a slight decline in 2019 but full recovery thereafter.

The dual-axis trendlines collectively reinforce JBD's position as a Q1-ranked journal across both platforms, characterized by rapid ascent and sustained competitiveness in citation-based evaluations. The discrepancy between CiteScore and Impact Factor trends reflects differences in citation windows and source coverage, but both metrics

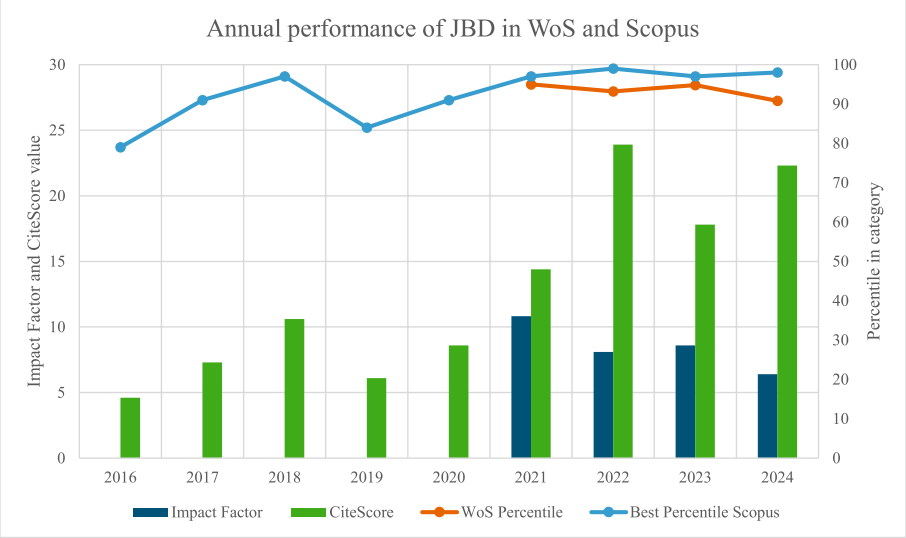


Fig. 2 Evolution of JBD's CiteScore, Impact Factor, and percentile rankings in Scopus and WoS (2016–2024)

converge in portraying JBD as an increasingly influential venue at the intersection of computer science, engineering, and data-intensive disciplines.

Influential papers in JBD

Identifying highly cited papers provides insight into coalescing thematic anchors and intellectual impact of a journal's corpus. Within JBD, citation leadership is dominated by comprehensive surveys and interdisciplinary reviews that consolidate knowledge and shape emerging research agendas. These papers drive visibility through high citation volumes and act as intellectual scaffolding for subsequent contributions across fields such as deep learning, data augmentation, transfer learning, and healthcare analytics.

As shown in Table 3, the top 50 most cited articles in JBD reveal an asymmetric citation structure led by a few landmark papers. The most cited article, "*A survey on image data augmentation for deep learning*" by Shorten and Khoshgoftaar [93], accrued 8,833 citations, with an exceptional annual average of 1,472.2. This paper exemplifies the journal's ability to host reference works that rapidly gain widespread adoption. The second-ranked paper, "Review of deep learning" by Alzubaidi et al. [6], received 4,874 citations in just four years, averaging 1,218.5 citations annually—an indication of its foundational role in modern deep learning discourse [6]. Note that the first paper is currently the 13th most cited paper of 2019 of all sciences (13th of 3,319,050 documents published in English and available in Scopus), and the fourth most cited in computer science (4th of 513,035 documents). The second paper is the 28th most cited document of 2021 of all sciences (28th of 3,766,877 documents) and the sixth most cited in computer science (6th of 558,736 documents).

Further demonstrating the journal's contribution to methodologically central themes, Weiss et al. [114] on transfer learning ranks third with 4,721 citations, followed by Johnson and Khoshgoftaar [57] on class imbalance in deep learning with 2,024 citations, and Najafabadi et al. [71] addressing deep learning challenges in big data with 1,960 citations [57, 71, 114]. Notably, 14 papers have surpassed 1,000 citations, and 25 exceeded 500 citations, a rare outcome in a journal of relatively recent origin.

Thematically, the top-cited documents are clustered around machine learning (e.g., [31, 94, 95]), healthcare analytics (e.g., [11, 25]), and data infrastructure (e.g., [108],Singh & Reddy, 2015). Contributions such as Hancock and Khoshgoftaar's ([44, 45]) review on CatBoost and Emmanuel et al.'s [31] paper on missing data offer more technical granularity while maintaining broad relevance through applied scenarios [31, 44, 45].

The majority of the top-cited works are authored by researchers based in the United States and Middle East, with several recurring authors—including Khoshgoftaar, Alzubaidi, and Shorten—appearing multiple times. This recurrence suggests the emergence of citation networks within JBD where a few author clusters contribute disproportionately to the journal's visibility.

These findings underscore the dual role of JBD as both a repository for high-impact surveys and a platform for applied, domain-specific advances. The influence of these papers extends across disciplines such as computer vision, cybersecurity, smart healthcare, and recommender systems, confirming the journal's broad epistemic footprint. The dominance of survey and review articles in these areas indicates that JBD has largely supported the consolidation and standardization of methods that underpin many data-intensive applications in society. As big data research increasingly targets issues such as healthcare delivery, critical infrastructure, and digital services, these anchoring contributions provide a methodological base on which more specialized and domain-specific studies can build.

The citation landscape of JBD publications is shaped not only by articles published within the journal but also by foundational works that are frequently cited by JBD authors. Table 4 lists the top 40 most cited sources cited within JBD papers between 2018 and 2024, highlighting the journal's intellectual dependencies and the broader epistemic infrastructure on which its scholarship is built. The most cited reference is the seminal work by Krizhevsky et al. [64] on convolutional neural networks, published in the *Communications of the ACM*, with 41 citations, closely followed by the ResNet paper by He et al. [48] in the *IEEE Proceedings of the CVPR* (40 citations), and Pedregosa et al.'s [75] foundational scikit-learn library paper in the *Journal of Machine Learning Research* (39 citations). These entries signal JBD's strong methodological alignment with core deep learning and machine learning libraries and architectures. The frequent citations to Dean and Ghemawat's [27] MapReduce framework and Breiman's [16] Random Forest algorithm further illustrate the influence of scalable computing and ensemble learning paradigms on big data research.

Transformer-based architectures have also made an imprint, with Vaswani et al. [112] appearing in sixth place with 33 citations. This reflects the integration of natural language processing innovations into the journal's discourse. Several references also pertain to foundational statistical and machine learning theory, such as Friedman [34], Chawla et al. [19] on SMOTE, and Tukey [109] on exploratory data analysis—highlighting the methodological rigor underpinning many JBD contributions. The inclusion of working papers and conference proceedings—such as Devlin et al. [29] on BERT and Simonyan & Zisserman's [96] VGGNet—underscores the fast-moving, preprint-friendly culture of AI and big data communities, and JBD's receptivity to early, high-impact research. Additionally, multiple articles from JBD itself appear in the list, including those by Leevy et al. [69], Shorten and Khoshgoftaar [93], and Johnson and Khoshgoftaar [57], reinforcing the internal citation ecosystem within the journal's own body of work.

Table 4 Most frequently cited references in JBD [14, 15, 22, 28, 36, 40, 41, 43, 47, 50, 54, 67, 68, 80, 81, 83, 84, 87, 90, 104, 117, 118] publications (WoS, 2018–2024)

Rank	Year	First author	Reference	Vol	Page	Type	TC
1	2017	Krizhevsky A	Commun ACM	v60	p84	A	41
2	2016	He KM	Proc CVPR IEEE		p770	C	40
3	2011	Pedregosa F	J Mach Learn Res	v12	p2825	A	39
4	2004	Dean J	OSDE '04		p137	C	36
5	2001	Breiman L	Mach Learn	v45	p5	A	35
6	2017	Vaswani A	Adv Neur In	v30		C	33
7	2012	Graves A	Stud Comput Intell	v385	p1	BC	32
8	2016	Chen TQ	KDD'16		p785	C	29
9	2015	Simonyan K	Arxiv. DOI: 1409.1556			WP	29
10	2018	Leevy JL	J Big Data	v5		A	28
11	2002	Chawla NV	J Artif Intell Res	v16	p321	A	28
12	2016	Goodfellow I	Adapt Comput Mach Le		p1	B	21
13	2015	Tsai CW	J Big Data	v2	p1	A	21
14	2009	Deng J	Proc CVPR IEEE		p248	C	21
15	1949	Tukey JW	Biometrics	v5	p99	A	20
16	2019	Shorten C	J Big Data	v6		A	18
17	2015	Najafabadi MM	J Big Data	v2	p1	A	18
18	2010	Zaharia M	2nd U W Hot Top Cl Comp	v2010	p95	C	18
19	2003	Blei DM	J Mach Learn Res	v3	p993	A	18
20	2001	Friedman JH	Ann Stat	v29	p1189	A	18
21	2017	Huang G	Proc CVPR IEEE		p2261	C	17
22	2015	Lecun Y	Nature	v521	p436	A	17
23	2015	Russakovsky O	Int J Comput Vision	v115	p211	A	17
24	1998	Lecun Y	P IEEE	v86	p2278	A	17
25	2017	Ren SQ	IEEE T Pattern Anal	v39	p1137	A	16
26	2016	Zaharia M	Commun ACM	v59	p56	A	16
27	2015	Schmidhuber J	Neural Networks	v61	p85	A	16
28	2014	Chen M	Mobile Netw Appl	v19	p171	A	16
29	2009	Seliya N	Proc Int C Tools Art		p59	C	16
30	2019	Devlin J	NAACL HLT 2019	vol. 1	p4171	C	15
31	2016	Szegedy C	Proc CVPR IEEE		p2818	C	15
32	2015	Ronneberger O	Lect Notes Comput Sc	v9351	p234	C	15
33	2012	Han J	Mor Kauf D		p1	B	15
34	2009	He HB	IEEE T Knowl Data En	v21	p1263	A	15
35	1996	Breiman L	Mach Learn	v24	p123	A	15
36	2019	Devlin J	Arxiv. DOI: 1810.04805			WP	14
37	2019	Johnson JM	J Big Data	v6		A	14
38	2018	Herland M	J Big Data	v5		A	14
39	2015	Gandomi A	Int J Inform Manage	v35	p137	A	14
40	2015	King DB	ACS Sym Ser	v1214	p1	B	14

TC = Total citations; A = Article; B = Book; BC = Book chapter; C = Conference proceedings; WP = Working paper

These references delineate a hybrid foundation of theoretical robustness, algorithmic innovation, and applied relevance. They capture the breadth of domains intersecting within JBD—from core computer science to healthcare, finance, and privacy—while affirming its methodological reliance on canonical models, widely adopted toolkits, and neural architecture.

Citing articles of JBD

To assess the scholarly reach and disciplinary penetration of JBD, an analysis of the citing documents is conducted using Scopus. The dataset is extracted by selecting the

"Cited by" records associated with JBD's 1,023 articles published between 2014 and 2024. Filters are applied to include only final-stage journal articles and reviews (excluding early access and conference papers), yielding a refined corpus of 30,057 citing documents from the 2014–2024 period. This filtered view offers a more stable indicator of JBD's formal impact within peer-reviewed scholarly communication. Table 5 presents a detailed breakdown of the citing articles, categorized by year of publication, most frequent citing authors, affiliated institutions, and contributing countries. The volume of citations has grown rapidly, with 2024 (10,375 citations), 2023 (7,629), and 2022 (5,512) being the most prolific years—highlighting the journal's recent acceleration in influence.

The most active individual citing JBD publications is Taghi M. Khoshgoftaar (65 papers), followed by Ibrahim H. Sarker (37), Laith Alzubaidi (28), and both Simon Elias Bibri and Kalpesh Kotecha (28 and 26, respectively). These names overlap with some of the journal's most prolific contributors, pointing to internal citation reinforcement and author-driven thematic ecosystems. This also suggests that clusters of authors are both contributing to and building upon JBD's literature, particularly in areas like deep learning, healthcare analytics, and smart cities. Institutionally, the Ministry of Education of the People's Republic of China leads with 473 citing articles, followed by the Chinese Academy of Sciences (395), King Saud University (232), and Vellore Institute of

Table 5 Institutional, national, and author-level profile of JBD citing articles

R	Year	TP	Author	TP	University	TP	Country	TP
1	2024	10,375	Khoshgoftaar, T.M	65	Min Educ PR China	473	China	6,838
2	2023	7,629	Sarker, I.H	37	Chinese Academy of Sciences	395	India	4,413
3	2022	5,512	Alzubaidi, L	28	King Saud U	232	USA	4,072
4	2021	3,296	Bibri, S.E	28	Vellore Inst Tech	208	UK	1,787
5	2020	1,698	Kotecha, K	26	King Abdulaziz U	197	South Korea	1,646
6	2019	829	Khan, M.A	25	Zhejiang U	186	Saudi Arabia	1,461
7	2018	430	Ashraf, I	23	CNRS France	173	Australia	1,184
8	2017	187	Fadhel, M.A	22	Tsinghua U	165	Germany	1,143
9	2016	78	Dewi, C	21	Wuhan U	154	Italy	1,019
10	2015	22	Acharya, U.R	20	Pr Nourah Bint Abdulrahman U	151	Canada	1,005
11	2014	1	Ahmad, T	20	Shanghai Jiao Tong U	146	Malaysia	955
12			Chen, R.C	20	U Electronic Sci Tech China	132	Spain	947
13			Pardamean, B	20	Southeast U	131	Indonesia	850
14			Kadry, S	19	Central South U	130	Pakistan	744
15			Damaševičius, R	17	Harbin Inst Tech	129	Iran	743
16			Ding, W	17	SRM Inst Sci Tech	127	Turkey	636
17			Khan, H.U	17	Pr Sattam Bin Abdulaziz U	125	France	617
18			Leevy, J.L	17	U Toronto	123	Egypt	588
19			Mridha, M.F	17	National U Singapore	120	Japan	573
20			Park, E	17	Norwegian U Sci Tech	118	Taiwan	572
21			Abualigah, L	16	U Tech Sydney	116	Brazil	505
22			Attallah, O	16	Sichuan U	112	UAE	440
23			Ghadi, Y.Y	16	K L Deemed to be U	111	Iraq	417
24			Gu, Y	16	U Sains Malaysia	110	Netherlands	405
25			Nam, Y	16	U College London	110	Poland	380
26			Rahmani, A.M	16	Beihang U	107	Bangladesh	357
27			Sarno, R	16	UNSW Sydney	107	Russia	345
28			Kumar, N	15	Bina Nusantara U	107	Morocco	339
29			Moni, M.A	15	Shenzhen U	104	Portugal	324
30			3 authors	15	2 institutions	104	2 countries	316

R = Rank; TP = Total papers

Technology (208). These institutions are geographically diverse yet consistently strong in applied AI research, data mining, and smart system development.

At the country level, China dominates the citation landscape with 6,838 citing documents, followed by India (4,413), the United States (4,072), and the United Kingdom (1,787). A significant number of citations also originate from South Korea, Saudi Arabia, Germany, and Australia, demonstrating the journal's resonance in both Global North and Global South contexts. Several emergent research economies—such as Pakistan (744), Iran (743), and Egypt (588)—also contribute substantially to the citing body, reflecting JBD's accessibility and relevance to diverse academic communities. These results highlight a key characteristic of JBD's growth: its citing network is both geographically broad and authorially recursive, with strong representation from institutions and researchers who are active in adjacent technical domains. This structural profile underscores JBD's function as both a methodological resource and a thematic hub for big data and machine learning discourse.

Table 6 provides a dual-layered view of JBD's citation network, identifying the most frequent journals and subject areas in which citing articles appear. This insight not only reflects the breadth of JBD's scholarly influence but also its functional integration across disciplines that consume and build upon its published research. Among journals, *IEEE Access* ranks first with 1,352 citing articles, a pattern consistent with its broad scope and high publication volume in applied computing and engineering. Following closely are *Applied Sciences* (658), *Sensors* (582), and *Scientific Reports* (433), all of which maintain multidisciplinary remits with strong emphasis on data-centric innovation. Notably, JBD is also self-cited 422 times, indicating a sustained ecosystem of internally coherent research, a phenomenon observed in fast-growing, thematically focused journals.

The spread of citing sources is not limited to technical domains. Alongside core AI and data science journals like *Expert Systems with Applications*, *Neural Computing and Applications*, and *Applied Soft Computing*, citation activity extends into biomedical (*PLOS One*, *Diagnostics*), sustainability (*Sustainability*, *Remote Sensing*), and even health and humanities-related outlets (*Heliyon*, *Journal of Theoretical and Applied Information Technology*). These patterns point to JBD's cross-sectoral utility and emphasize its role in bridging traditional disciplinary silos.

The subject area distribution reinforces this observation. Computer Science leads with a dominant 17,171 citing documents, followed by Engineering (11,895), and Mathematics (3,902). The long tail includes Medicine (3,261), Materials Science (3,123), and Physics and Astronomy (2,877), signaling interdisciplinary adoption. Meanwhile, citations from Social Sciences (2,499), Decision Sciences (1,774), and Business, Management & Accounting (1,701), highlight the journal's relevance for socio-technical research and data-driven decision-making. This wide topical dispersion, extending even to Psychology, Economics, Pharmacology, and Dentistry, confirms that JBD's outputs are not only technically impactful but are also increasingly instrumental across knowledge systems where big data methodologies underpin discovery, prediction, and optimization. The journal's ability to seed research across scientific, engineering, medical, and managerial domains demonstrates both its epistemic versatility and its integrative value in the current data-intensive research paradigm. The strong representation of citing articles from rapidly developing research economies, together with uptake in fields such as medicine, sustainability, and management, suggests that JBD's open-access model is facilitating the

Table 6 Top citing journals and subject areas for JBD articles

R	Journal	TP	PN	Subject area	TP
1	IEEE Access	1,352	89,242	Computer Science	17,171
2	Applied Sciences	658	69,869	Engineering	11,895
3	Sensors	582	34,775	Mathematics	3,902
4	Scientific Reports	433	225,978	Medicine	3,261
5	J Big Data	422	1,023	Materials Science	3,123
6	Electronics	358	21,962	Physics and Astronomy	2,877
7	Multimedia Tools and Applications	339	17,651	Social Sciences	2,499
8	Expert Systems with Applications	325	14,349	Biochemistry, Genetics & Molecular Biology	2,286
9	Int J Advanced Computer Sci Applic	284	8,399	Environmental Science	1,823
10	Sustainability	273	86,906	Decision Sciences	1,774
11	Remote Sensing	254	35,096	Business, Management & Accounting	1,701
12	PLOS One	229	223,456	Earth and Planetary Sciences	1,425
13	Neural Computing and Applications	222	8,929	Chemistry	1,392
14	Mathematics	194	21,267	Chemical Engineering	1,351
15	Computers Materials and Continua	167	5,445	Agricultural and Biological Sciences	1,326
16	Diagnostics	154	13,774	Multidisciplinary	1,112
17	Computers in Biology and Medicine	150	6,702	Energy	1,082
18	SN Computer Science	137	3,403	Neuroscience	723
19	Heliyon	135	40,674	Health Professions	550
20	Engineering Applic Artificial Intelligence	131	6,065	Economics, Econometrics & Finance	480
21	J Theoretical Applied Information Tech	123	7,379	Arts and Humanities	355
22	Knowledge Based Systems	123	7,360	Psychology	298
23	Applied Soft Computing	120	4,500	Pharmacology, Toxicology & Pharmaceutics	222
24	Biomedical Signal Processing and Control	120	5,675	Immunology and Microbiology	204
25	Indonesian J Electrical Engin Computer Sci	117	6,060	Nursing	167
26	Int J Intelligent Syst Applic Engin	114	2,743	Dentistry	59
27	Information	113	3,008	Veterinary	59
28	Computational Intelligence Neuroscience	107	4,881		
29	Energies	102	52,375		
30	Artificial Intelligence Review	100	1,652		

R = Rank; TP = Total papers; PN = Total publications between 2014 and 2024

diffusion of big data methods into regions and sectors where data-driven infrastructures are still consolidating. This pattern aligns with broader shifts in big data scholarship toward supporting digital transformation in public services, industrial production, and health systems.

These patterns indicate that JBD has evolved from a primarily computer science–oriented outlet into a cross-sector platform where applied AI and data-analytics research are disseminated to a wide range of disciplines. The concentration of citations in high-throughput applied journals such as *IEEE Access* and *Applied Sciences* suggests that JBD's articles are frequently used as methodological and application-oriented references in the broader big data ecosystem, which is consistent with the citing profiles of related venues such as *Big Data* and *Information Sciences*.

.

Mapping JBD with VOSviewer software

General overview of JBD based on bibliographic coupling and co-citation analysis

To characterize the intellectual foundation and thematic orientation of JBD, we conducted a co-citation analysis of journals using a threshold of 30 minimum citations and 100 co-citation links. This network, illustrated in Fig. 3, visualizes the disciplinary influence of journals most frequently cited alongside JBD articles between 2018 and 2024, as indexed in Web of Science.

The map reveals six major clusters, each representing a disciplinary or topical community that frequently co-occurs in the citation context of JBD publications. The red cluster, which centers on *IEEE Access* and *Lecture Notes in Computer Science*, is the most densely connected region of the map. It represents an applied computing and systems engineering cluster with strong ties to conference proceedings, procedural literature, and practical implementations in computer science. Journals such as *Procedia Computer Science*, *Expert Systems with Applications*, and *Knowledge-Based Systems*, fall within or overlap with this sphere but bridge into the green cluster through shared citations.

The purple cluster, anchored by JBD itself and *Information Processing & Management*, signals the core domain of big data, information systems, and informetrics. This cluster integrates both technical and managerial aspects of data-intensive research, with strong citation ties to *Decision Support Systems*, *Future Generation Computer Systems*, and *Information Sciences*. This group reflects the journal's foundational literature base and interdisciplinary scope. The blue cluster, featuring *arXiv*, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *Pattern Recognition*, and *Advances in Neural Information Processing Systems*, represents the computer vision and deep learning community. These journals are methodologically rigorous and strongly theoretical,

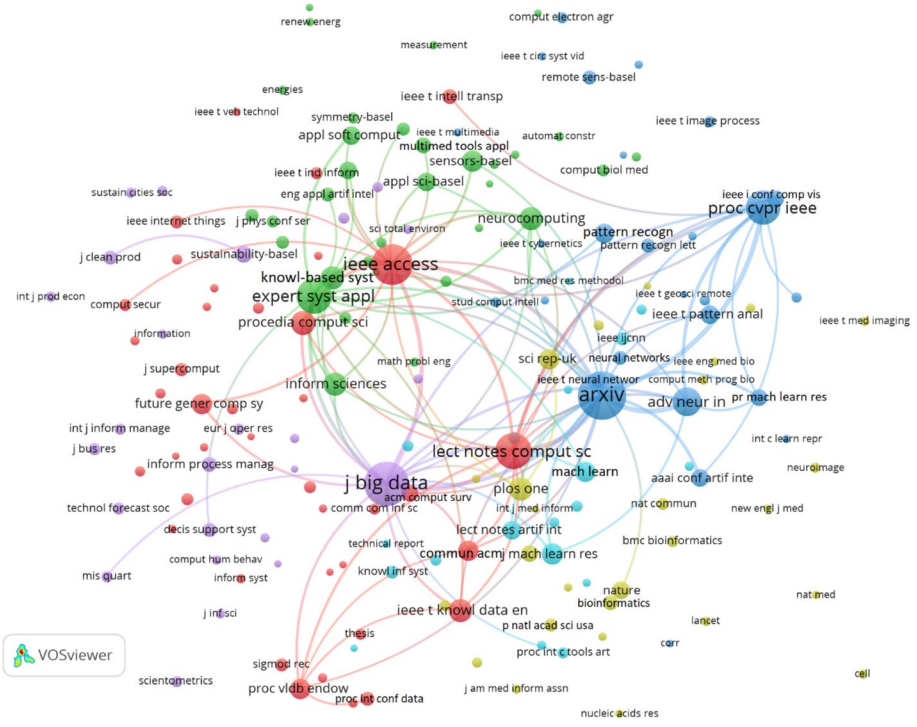


Fig. 3 Co-citation network of journals cited in JBD articles (WoS 2018–2024), with a minimum citation threshold of 30 and at least 100 co-citation links

indicating that JBD is often situated near cutting-edge machine learning research, particularly in image analysis, neural networks, and algorithmic foundations.

The green cluster, linked to *Expert Systems with Applications*, *Knowledge-Based Systems*, and *Information Sciences*, emphasizes AI-driven decision systems, knowledge representation, and applied machine intelligence. It demonstrates how JBD is connected to expert systems literature, drawing from robust foundations in algorithmic decision support and intelligent applications. The yellow cluster includes *PLOS One*, *Scientific Reports* (UK), *Bioinformatics*, and *Nature*. These journals indicate a cross-domain interest in open science, reproducibility, and interdisciplinary problem-solving, with emphasis on data-centric research in biomedical and natural sciences. Their co-citation signals that JBD is frequently used in broader applications beyond computing, especially in health, environment, and life sciences.

Finally, the cyan cluster, which includes *Machine Learning*, *Journal of Machine Learning Research*, and *Lecture Notes in Artificial Intelligence*, highlights theoretical ML research with ties to computational learning theory and algorithm development. The proximity of this group to both the blue and red clusters demonstrates the algorithmic depth within JBD-cited literature. This structure illustrates that JBD is intellectually positioned at the intersection of applied data science, systems engineering, artificial intelligence, and interdisciplinary research domains. Its bibliographic coupling network reveals strong methodological ties to machine learning and systems design, while its co-citation landscape confirms ongoing relevance to data-driven research in the life sciences, engineering, and computational social sciences. Viewed longitudinally, the increasing weight of clusters anchored in deep learning and computer vision, alongside interdisciplinary outlets such as Scientific Reports and PLOS One, points to a gradual shift in JBD's intellectual centre of gravity from infrastructure- and systems-focused work toward model-centric and application-driven research. At the same time, the persistent presence of information-systems and decision-support journals in the co-citation neighbourhood indicates that JBD continues to bridge algorithmic innovation with managerial and socio-technical perspectives.

Figure 4 presents a bibliographic coupling map of highly cited documents published in JBD between 2014 and 2024. Nodes represent individual papers with at least 50 citations and the figure visualizes the 100 strongest bibliographic links. Note that the edge thickness indicates the strength of coupling (shared references). The size of each node reflects its total citation count, enabling identification of central and influential works. The most prominent cluster, shown in green, is centered on Shorten et al. [93] and Alzubaidi et al. [6], both of which focus on deep learning and image data augmentation in big data contexts. This green cluster encompasses foundational works that define JBD's intersection with deep learning and computer vision, including Yadav & Jadhav [115], and Najafabadi et al. [71]. The density and connectivity of this group highlight its thematic coherence and centrality within the JBD corpus.

The red cluster is led by Dash et al. [25] and Kumar et al. [65], reflecting a strong orientation toward big data analytics frameworks, decision support systems, and optimization strategies. These papers share methodological commonality, particularly in hybrid intelligent systems and data-driven decision-making. A blue cluster is structured around Kasongo & sun [60], which emphasizes cybersecurity and intrusion detection using deep

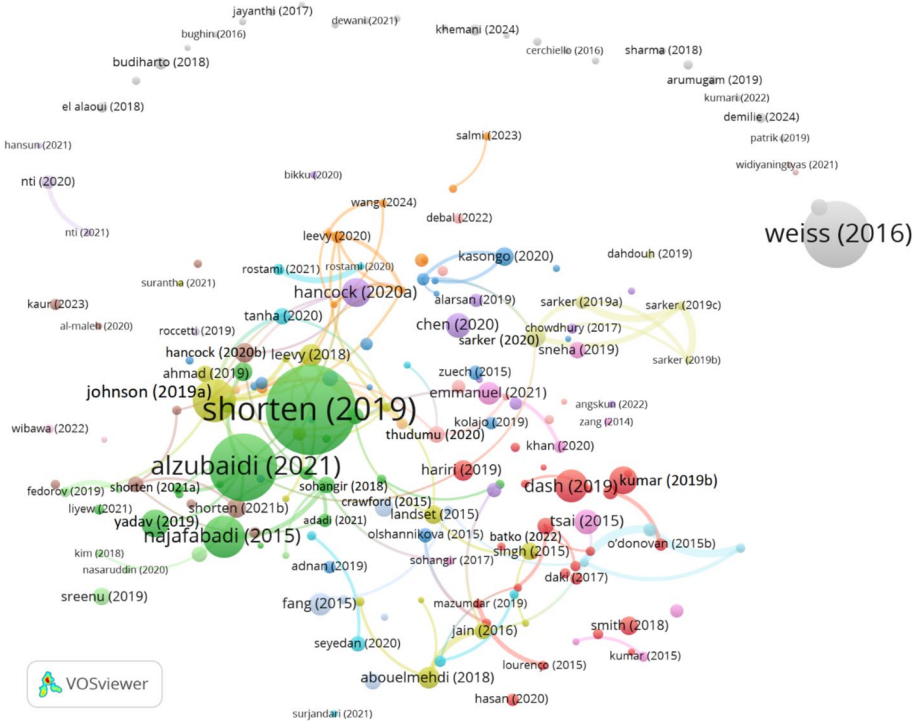


Fig. 4 Bibliographic coupling of documents published in JBD: minimum threshold of 50 citations and 100 links

learning. Though smaller, this cluster signifies an emerging subfield in JBD literature linking big data with network security.

Other smaller clusters include the purple node built around Hancock et al. [44]. The yellow node is built around Johnson & Khoshgoftaar [57] and Leevy et al. [69], reflecting literature on privacy in big data and health informatics applications. The isolated gray node represents Weiss et al. [114], a key paper on class imbalance, which has limited bibliographic coupling with other JBD publications despite its high citation count. This suggests thematic divergence, likely because its influence extends beyond the immediate scope of JBD's core thematic clusters. This map demonstrates how a small set of seminal publications scaffolds the conceptual foundation of JBD, forming tightly connected thematic cores that reflect the journal's primary areas of influence. The dominance of a few survey and framework papers within these cores also reveals a marked citation dependency on deep-learning and big data overview contributions, whereas topics such as privacy, graph-based learning, and causal inference remain more peripheral despite their growing importance in the wider data science literature. This asymmetry suggests that JBD's historical strengths lie in consolidating high-impact methodological overviews and application-driven work, with opportunities for future volumes to cultivate a more diversified foundational base.

Keyword and topic analysis of JBD

The co-occurrence analysis of author keywords provides insight into the thematic composition, the conceptual structure, and the evolving research priorities within JBD. By examining the frequency and co-linkage of keywords, this method identifies dominant research topics, emerging areas, and inter-topic relationships. The network not only reveals the prominence of foundational fields such as machine learning and big data

analytics but also highlights the emergence of novel methodological and domain-specific focuses. To explore these dynamics, Fig. 5 visualizes the co-occurrence network of author keywords in JBD from 2018 to 2024, considering a minimum occurrence threshold of five and 100 bibliographic links. The co-occurrence map in Fig. 5 identifies four major conceptual cores. The largest and most interconnected node is "big data", located centrally and colored predominantly in blue, indicating sustained relevance since 2019. This term forms the anchor of a broader computational cluster incorporating "big data analytics," "cloud computing," "Hadoop," "MapReduce," and "Apache Spark", terms associated with infrastructure-level and large-scale data processing frameworks. These keywords occupy the purple-blue zone, reflecting earlier-stage foundational research.

Adjacent to this, the "machine learning" and "deep learning" nodes (yellow cluster) represent a second prominent axis, closely linked to application-focused terms like "neural network," "convolutional neural network," "transfer learning," "classification," and "image classification." These keywords are temporally situated between 2021 and 2022, highlighting a transition toward model-centric exploration and algorithmic innovation. A third notable focus involves "artificial intelligence," "intrusion detection," "IoT," "healthcare," and "bioinformatics", illustrating JBD's increasing interdisciplinary applications. Terms like "prognosis," "hepatocellular carcinoma," and "immunotherapy" (colored in orange and red) represent more recent biomedical entries, suggesting a growing cross-domain integration from 2022 onward.

Lastly, emerging methodological innovations are evident in keywords such as "large language models," "SMOTE," "undersampling," and "high-performance computing," each represented by small orange nodes indicative of post-2022 uptake. This cluster reflects methodological refinement and adaptation in response to imbalanced datasets, computational scaling, and generative model paradigms. The keyword landscape in JBD shows a temporal and conceptual evolution from infrastructural development (2019–2020) to algorithmic sophistication and real-world applications (2021–2023), culminating in recent advancements in generative models and biomedical integration (2023–2024).

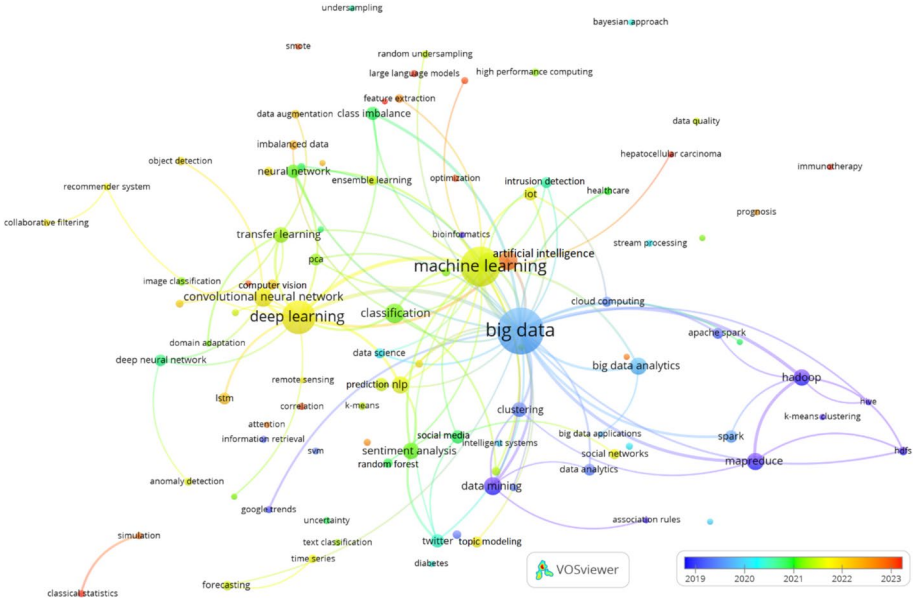


Fig. 5 Co-occurrence of author keywords in JBD: minimum occurrence threshold of 5 and 100 links

Figure 6 presents a strategic diagram generated using bibliometrix [7], employing the Spinglass clustering algorithm [79] to visualize the structural positioning of research themes in JBD based on author keyword co-occurrence. The two axes represent Callon's centrality (x-axis) and Callon's density (y-axis), indicating, respectively, how thematically connected a cluster is to the rest of the network, and how internally cohesive or developed it is. In the upper-right quadrant (Motor Themes) lie highly central and well-developed topics such as "machine learning," "artificial intelligence," "feature selection," "deep learning," "convolutional neural network," and "natural language processing." These themes are mature, extensively developed, and serve as technical engines of the field.

The lower-right quadrant (Basic Themes) is dominated by "big data," "data mining," and "MapReduce." These themes exhibit high centrality but low density, indicating foundational importance with wide interdisciplinary relevance, yet relatively lower current research expansion—suggesting stabilization or potential saturation. In the upper-left quadrant (Niche Themes), "classical statistics," "simulation," and "Bayesian approach" emerge as methodologically deep yet thematically peripheral. Their development suggests specialized use cases or legacy systems within the domain.

The lower-left quadrant (Emerging or Declining Themes) includes "prediction," "Google Trends," and "recurrent neural network," which possess limited thematic relevance and internal development. These may represent either nascent experimental directions or declining research focus in the JBD corpus. Centrally positioned concepts such as "classification," "data science," "clustering," "class imbalance," and "intrusion detection" act as bridging topics with moderate centrality and density. Their positions suggest evolving significance as they integrate into broader methodological or applied paradigms. This thematic mapping confirms that JBD research is anchored in high-impact, machine learning-driven infrastructures, while selectively expanding into emerging challenges like data imbalance and application-specific anomaly detection.

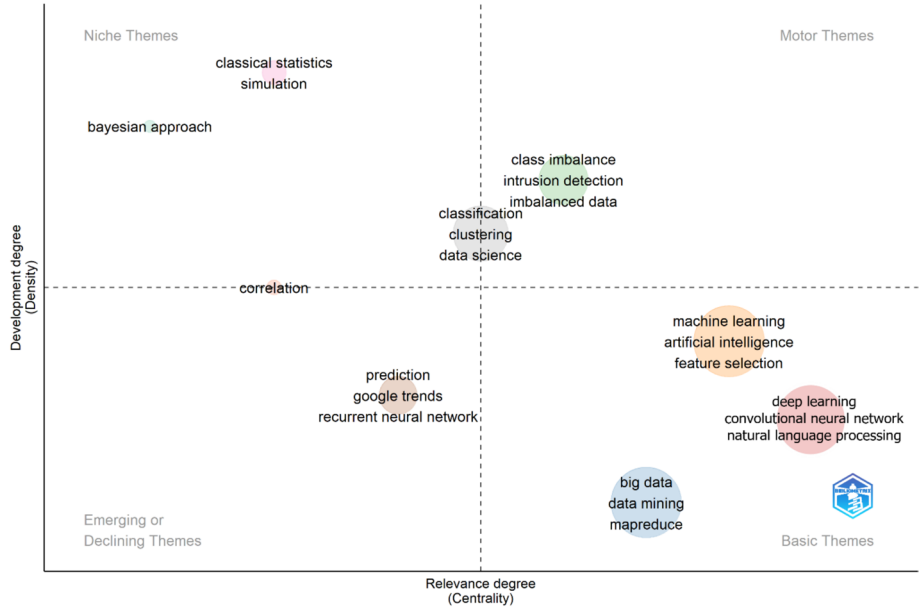


Fig. 6 Thematic map with the 250 most frequent author keywords and minimum cluster frequency of 5 (clustering algorithm: Spinglass)

Figure 7 presents a WordCloud visualization of the most frequently occurring author keywords in JBD publications, generated using bibliometrix [7]. The font size of each keyword corresponds to the square root of its raw frequency, a transformation applied to minimize the visual dominance of the top three terms, "big data" (224 occurrences), "machine learning" (164), and "deep learning" (118), relative to the remaining terms. For instance, with the square root scaling, their respective sizes are approximately 15, 13, and 11, while the fourth most frequent term, "classification" (39 occurrences), scales to approximately 6.2, enabling more equitable visual distribution.

The WordCloud clearly reinforces the thematic pillars of the journal: "big data" remains the most prominent, anchoring the field's focus on massive-scale data processing and management. This is followed by "machine learning" and "deep learning," confirming the methodological centrality of automated predictive modeling. Other dominant terms include "artificial intelligence," "feature selection," "classification," "sentiment analysis," "mapreduce," and "convolutional neural network," which highlight both technical strategies and specific analytical tasks relevant to big data environments.

The presence of keywords such as "transfer learning," "iot," "anomaly detection," "hadoop," and "social networks" illustrates the journal's coverage of contemporary applications and infrastructure. Meanwhile, domain-specific and methodological topics like "covid-19," "forecasting," "genetic algorithm," and "data augmentation" reflect responsiveness to recent global challenges and evolving research paradigms. This visualization offers a concise, yet rich, thematic snapshot of the most salient concepts in JBD literature, complementing the structural keyword co-occurrence and thematic evolution analyses. Taken together with the co-citation and topic-cluster analyses, the keyword evidence suggests a thematic evolution in JBD from infrastructure-oriented big data and data mining toward mature machine-learning pipelines, with more recent but still less consolidated growth in areas such as privacy, healthcare analytics, and cyber-security. This trajectory mirrors broader developments in big data research while highlighting specific niches—such as class imbalance, intrusion detection, and sentiment analysis—where JBD has become a recurrent reference point.



Fig. 7 WordCloud of the most frequent author keywords

Table 7 presents the top 30 most productive and influential index keywords associated with publications in JBD, ranked by total publication count (TP). The metrics include total citations (TC), *h*-index (H), average citations per publication (C/P), the number of documents cited at least 100 and 10 times ($\geq 100, \geq 10$), and the number of documents among the 50 most cited in the dataset (T50). These indicators collectively reflect the thematic prominence, impact, and durability of each keyword in the journal's discourse. "Big data" dominates with 317 publications and 35,692 citations, appearing in 47 papers with over 100 citations and 209 papers with over 10 citations, reinforcing its foundational role in the journal. Despite being ranked fourth in TP, "deep learning" demonstrates the highest average citation rate (160.3), supported by a strong *h*-index of 49 and a considerable T50 count of 13, indicating its high-impact contributions. Similarly, "deep neural networks" and "convolution" show exceptional C/P ratios (343.5 and 336.9, respectively), although with fewer appearances, reflecting concentrated high-impact usage rather than thematic breadth.

Notably, "learning systems," "classification (of information)," and "learning algorithms" also rank high in productivity and citation performance, suggesting consistent research engagement with foundational computational approaches. Meanwhile, terms such as

Table 7 Most productive and influential index keywords in JBD based on Scopus-indexed documents

R	Keyword	TP	TC	H	C/P	≥ 100	≥ 10	T50
1	Big Data	317	35,692	65	112.6	47	209	28
2	Machine Learning	207	10,002	42	48.3	15	118	8
3	Learning Systems	172	18,868	48	109.7	26	125	14
4	Deep Learning	161	25,813	49	160.3	24	112	13
5	Classification (of Information)	151	16,624	48	110.1	24	130	13
6	Data Mining	129	13,530	39	104.9	18	95	10
7	Learning Algorithms	115	13,590	38	118.2	17	95	10
8	Forecasting	114	4,128	34	36.2	9	76	4
9	Data Analytics	98	9,400	39	95.9	20	77	9
10	Performance	89	8,964	37	100.7	11	88	7
11	Data Handling	81	5,061	28	62.5	12	52	5
12	Large Datasets	81	3,425	25	42.3	8	55	3
13	Social Networking (online)	78	3,438	30	44.1	8	60	3
14	Convolutional Neural Networks	64	7,600	21	118.8	6	33	4
15	Feature Selection	61	3,557	26	58.3	8	43	4
16	Deep Neural Networks	56	19,238	28	343.5	8	36	7
17	Artificial Intelligence	56	6,583	14	117.6	5	19	4
18	Support Vector Machines	56	4,478	29	80	8	36	7
19	Convolution	53	17,854	31	336.9	11	54	6
20	Clustering Algorithms	53	4,563	19	86.1	8	30	5
21	Semantics	53	3,004	20	56.7	2	30	1
22	Digital Storage	52	1,780	23	34.2	4	31	1
23	Sentiment Analysis	49	3,093	25	63.1	9	38	4
24	Decision Making	49	2,763	25	56.4	5	37	3
25	Information Management	47	1,686	18	35.9	6	29	1
26	Map-reduce	47	1,269	18	27	1	31	1
27	Diseases	47	1,178	16	25.1	3	25	0
28	Decision Trees	45	4,120	25	91.6	11	39	6
29	Long Short-Term Memory	45	1,775	23	39.4	4	31	0
30	2 Keywords	44	-	-	-	-	-	-

R = Rank; TP = Total papers; TC = Total citations; H = h-index; C/P = Cites per paper; ≥ 100, ≥ 10 = Number of papers with equal or more than 100 and 10 citations; T50 = Number of documents in the Top 50 of Table 3

"forecasting," "sentiment analysis," and "decision trees" contribute modestly in volume but show respectable influence metrics, suggesting niche but relevant subfields. Lower-frequency terms like "map-reduce" and "long short-term memory" exhibit modest citation counts despite their technical importance, which may indicate limited but specific application scopes. Collectively, the table illustrates how JBD's indexed keyword landscape balances foundational constructs in data science (e.g., big data, machine learning), evolving algorithmic strategies (e.g., CNNs, LSTMs), and emerging domain applications (e.g., sentiment analysis, diseases), aligning with earlier co-occurrence patterns identified in Fig. 5. From an interpretive standpoint, the concentration of very high C/P values in a narrow set of deep-learning-related keywords, contrasted with more moderate but stable impact for topics such as data mining and decision trees, implies that JBD's influence has become increasingly tied to neural-network-based approaches, which may make the journal's citation profile sensitive to shifts in the deep learning research cycle.

Table 8 presents the most prominent research topics and topic clusters within JBD from 2014 to 2023, as extracted from Scopus via the SciVal platform. Each topic is defined as a co-citation-based topical cluster [63], and the table lists its total publications (TP), Field-Weighted Citation Impact (FWCI), and Prominence Percentile (PP), which jointly assess topic-specific productivity, field-normalized impact, and global visibility [77]. The top-ranked topic, "Intrusion Detection System,Deep Learning; Network Security", comprises 23 documents, with an FWCI of 5.39 and a PP of 99.89, signifying intense scholarly attention and high field-normalized impact. The second-ranked topic, "Big Data; Decision-Making; Data Analytics", exhibits both high productivity (20 papers and a slightly stronger FWCI (5.63, confirming its sustained relevance and citation impact. "Imbalanced Data; Support Vector Machine; Class Imbalance" stands out with the highest FWCI (7.51, reinforcing the technical and applied value of methodological advances in handling skewed datasets.

Table 8 Top thematic topics in JBD (2014–2023) based on SciVal clusters and field-normalized citation impact

R	Topic	TP	FWCI	PP
1	Intrusion Detection System; Deep Learning; Network Security	23	5.39	99.89
2	Big Data; Decision-Making; Data Analytics	20	5.63	99.74
3	Big Data; Map-Reduce; Cloud Computing	19	1.83	95.16
4	Imbalanced Data; Support Vector Machine; Class Imbalance	18	7.51	99.49
5	Market Forecasting; Neural Network; Commerce	11	3.8	99.66
6	Deep Learning; Aspect-Based Sentiment Analysis; Long Short-Term Memory Network	9	6.71	99.56
7	Association Rules; Frequent Itemsets; Data Mining	9	1.23	95.88
8	Social Networking (Online); Sentiment Classification; Data Mining	8	3.15	99.32
9	Big Data; Health Care; Data Analytics	8	8.81	94.77
10	Fraud Detection; Health Care; Machine Learning	8	4.2	85.08
11	Feature Selection; Support Vector Machine; Data Mining	7	2.81	99.17
12	Recommender Systems; Collaborative Filtering; E-Commerce	7	4.59	98.53
13	Anonymization; Data Privacy; Cloud Computing	7	3.73	93.52
-	9 Topics	6	-	-
-	5 Topics	5	-	-
-	13 Topics	4	-	-
-	31 Topics	3	-	-
-	63 Topics	2	-	-
-	339 Topics	1	-	-

R = Rank; TP = Total papers; FWCI = Field-weighted citation impact (data from Scopus); PP = Worldwide prominent percentile (according to Scopus and FWCI)

Other high-performing topics include "Deep Learning; Aspect-Based Sentiment Analysis; LSTM" (FWCI = 6.71) and "Health Care; Big Data; Data Analytics" (FWCI = 8.81), the latter of which, despite a lower PP (94.77), reflects the multidisciplinary utility of JBD in medical informatics. Emerging themes such as "Recommender Systems" and "Anonymization; Data Privacy" also exhibit high PP values (> 93), suggesting growing global attention. Notably, the long tail of topic diversity is reflected in the 339 additional topics represented by only a single paper each, indicating that while JBD maintains a strong core in big data, machine learning, and security, it also supports a wide and evolving thematic landscape. This table complements the previous keyword analyses (Table 7) by revealing co-citation-driven topic-level clusters with high centrality in global research, grounded in Scopus' algorithmic classification and citation normalization via FWCI.

Table 8 presents the most prominent research topics and topic clusters within JBD from 2014 to 2023, as extracted from Scopus via the SciVal platform. Each topic is defined as a co-citation-based topical cluster [63], and the table lists its TP, FWCI, and PP, which jointly assess topic-specific productivity, field-normalized impact, and global visibility [77].

The top-ranked topic, "Intrusion Detection System; Deep Learning; Network Security", comprises 23 documents, with an FWCI of 5.39 and a PP of 99.89, signifying intense scholarly attention and high field-normalized impact. The second-ranked topic, "Big Data; Decision-Making; Data Analytics", exhibits both high productivity (20 papers) and a slightly stronger FWCI (5.63), confirming its sustained relevance and citation impact. "Imbalanced Data; Support Vector Machine; Class Imbalance" stands out with the highest FWCI (7.51), reinforcing the technical and applied value of methodological advances in handling skewed datasets.

Other high-performing topics include "Deep Learning; Aspect-Based Sentiment Analysis; LSTM" (FWCI = 6.71) and "Health Care; Big Data; Data Analytics" (FWCI = 8.81), the latter of which, despite a lower PP (94.77), reflects the multidisciplinary utility of JBD in medical informatics. Emerging themes such as "Recommender Systems" and "Anonymization; Data Privacy" also exhibit high PP values (> 93), suggesting growing global attention. Notably, the long tail of topic diversity is reflected in the 339 additional topics represented by only a single paper each, indicating that while JBD maintains a strong core in big data, machine learning, and security, it also supports a wide and evolving thematic landscape. This table complements the previous keyword analyses (Tables 11 and 12) by revealing co-citation-driven topic-level clusters with high centrality in global research, grounded in Scopus' algorithmic classification and citation normalization via FWCI.

Table 9 presents the leading topic clusters in JBD between 2014 and 2023, derived from Scopus through the SciVal analytics platform. Each cluster aggregates conceptually linked topics based on co-citation and bibliographic coupling, enabling the identification of broader thematic areas across the journal's literature. For each cluster, total publications, FWCI, and PP are reported, following the metrics established in prior bibliometric literature [63, 77]. The top-ranked cluster, "Machine Learning,Data Mining; Artificial Intelligence", accumulates 67 publications and demonstrates a strong citation influence (FWCI = 3.91) with a PP of 93.39, confirming its foundational role in JBD's scholarly ecosystem. The second cluster, "Sentiment Analysis; Natural Language Processing; Machine

Table 9 Top co-citation-based topic clusters in JBD (2014–2023) based on SciVal data

R	Topic Cluster	TP	FWCI	PP
1	Machine Learning; Data Mining; Artificial Intelligence	67	3.91	93.39
2	Sentiment Analysis; Natural Language Processing; Machine Learning	66	2.81	99.67
3	Image Segmentation; Deep Neural Network; Object Detection	63	8.03	100
4	Data Mining; Artificial Intelligence; Information System	49	4.18	79.13
5	Data Mining; Graph Neural Network; Social Network Analysis	39	1.26	95.16
6	Query Processing; Cloud Computing; Input/Output	36	1.33	26.35
7	Network Security; Intrusion Detection; Machine Learning	30	4.65	89.27
8	Data Mining; Artificial Intelligence; Machine Learning	26	1.48	57.55
9	Query Processing; Cloud Computing; Internet of Things	22	1.17	23.41
10	Commerce; Machine Learning; Transport	20	3.22	90.71
11	Data Mining; Artificial Intelligence; e-Commerce	17	2.48	85.41
12	Data Mining; Image Segmentation; Artificial Intelligence	16	1.12	38.91
13	Ontology; Semantic Web; Linked Data	13	0.67	49.24
14	Causal Inference; Regression Analysis; Randomized Controlled Trial	12	5.04	81.62
15	Transport; Data Mining; Information System	10	0.71	94.76
16	Social Media; Health Literacy; COVID-19	9	1.19	88.48
17	Robot; Action Recognition; Gait Analysis	8	1.47	93.26
18	Machine Learning; Artificial Intelligence; Knowledge Acquisition	7	4.52	72.66
-	8 Topic Clusters	6	-	-
-	7 Topic Clusters	5	-	-
-	7 Topic Clusters	4	-	-
-	21 Topic Clusters	3	-	-
-	34 Topic Clusters	2	-	-
-	91 Topic Clusters	1	-	-

Abbreviations are available in Table 8

Learning", has 66 papers with an exceptional PP of 99.67, indicating global prominence and relevance in text-based analytics.

The third cluster, "Image Segmentation; Deep Neural Network; Object Detection," exhibits the highest citation normalization (FWCI = 8.03) and a perfect PP of 100, suggesting it is both cutting-edge and globally recognized. Other impactful themes include "Network Security and Intrusion Detection" (FWCI = 4.65), "Commerce and Transport via Machine Learning" (FWCI = 3.22), and "Causal Inference with RCTs" (FWCI = 5.04), indicating methodological diversification. While most clusters maintain high relevance, some such as "Query Processing; IoT; Cloud Computing" and "Ontology and Linked Data" report lower PP and FWCI, highlighting areas of niche or emerging development. Overall, this table illustrates that JBD's core strength lies in combining machine learning foundations with application-driven domains such as NLP, healthcare, security, and computer vision. These findings complement Table 8 by emphasizing not just individual topics but interconnected thematic constellations with measurable influence and visibility across the global research landscape.

Comparing these clusters with topic portfolios reported for other leading big-data and AI outlets suggests that JBD occupies a relatively application-oriented position in the ecosystem, with a strong emphasis on security, health, and commerce alongside its methodological core. This configuration reinforces the journal's role as a conduit through which advances in machine learning and data analytics diffuse into diverse applied communities, while also indicating underexplored space for future contributions in areas such as causal modelling, explainability, and responsible AI.

Conclusions

General findings

This bibliometric investigation provides a comprehensive overview of the JBD's scientific output between 2014 and 2024. The journal has demonstrated substantial growth in publication volume and citation impact, particularly since 2019, marking a rapid ascent in both volume of scholarly output and global recognition. Its thematic trajectory has been shaped predominantly by the intersection of machine learning, deep learning, and big data. These three areas have consistently emerged as pervasive themes across multiple analyses—including keyword co-occurrence, topic clustering, and highly cited contributions. Authorship analysis revealed a high degree of international collaboration. The United States, China, India, and the United Kingdom have been core contributors to the journal's knowledge base. Notwithstanding the diversity of contributing nations, collaboration networks suggest regional clustering and centrality effects, with a few institutions accounting for a disproportionate share of high-impact output.

On the thematic front, keyword evolution and thematic mapping (via bibliometrix and VOSviewer) identified four principal knowledge structures: foundational topics such as "big data" and "machine learning" with high centrality but moderate density; central themes like "deep learning" and "natural language processing"; niche topics including "simulation" and "Bayesian approaches"; and emerging or less developed themes such as "Google Trends" and "causal inference." The convergence of application-specific studies—e.g., in healthcare, cybersecurity, and sentiment analysis—with methodological advances reflects the journal's interdisciplinary character and the growing complexity of research frontiers it hosts. Furthermore, topic-level analytics using SciVal metrics revealed that several clusters, such as "image segmentation with deep neural networks" and "intrusion detection using deep learning," exhibit high field-weighted citation impacts (FWCI > 5.0), indicating not only academic visibility but also the journal's relevance to pressing real-world challenges. Overall, JBD has evolved into a globally influential platform for methodological innovation and domain-specific exploration in the big data ecosystem.

Practical implications

The findings from this bibliometric review of JBD yield important practical insights for scholars, practitioners, and policy-makers operating within the broader big data ecosystem. First, the dominance of topics such as machine learning, deep learning, classification, and data mining suggests that JBD functions as a methodological hub for the development and dissemination of scalable algorithms and data-driven solutions. Researchers aiming to contribute to these core areas can leverage the identified citation patterns and keyword clusters to position their work within high-impact thematic niches. Second, the presence of topically diverse applications, ranging from intrusion detection and fraud analytics to sentiment classification and healthcare forecasting, illustrates the journal's practical orientation and timely responsiveness to societal needs. Authors focusing on applied research will find JBD a receptive venue for interdisciplinary studies that bridge technical depth with domain relevance.

Institutionally, the analysis of bibliographic coupling at author, organization, and country levels offers a roadmap for strategic collaborations. Institutions in emerging research regions such as Southeast Asia, North Africa, and the Middle East, which have

shown increasing publication trends, could benefit from forming consortia with high-performing centers in North America and Western Europe to boost research visibility and citation impact. The identification of high field-weighted citation impact (FWCI) in specific thematic clusters—especially in areas such as imbalanced data learning, neural network-based image segmentation, and healthcare analytics—can guide funding agencies and policy planners in prioritizing investment in fast-evolving and high-impact research domains. For industry stakeholders, these trends signal opportunities for technological adoption, talent acquisition, and research partnerships rooted in the evolving big data landscape curated by JBD.

From an editorial perspective, these dominance patterns suggest scope for further diversifying both geographical participation and methodological contributions through targeted calls for papers and special issues, particularly in underrepresented regions and application domains.

Limitations and future research

The scope of this bibliometric analysis was comprehensive, but several limitations must be acknowledged. First, the study primarily relied on metadata extracted from Scopus and WoS. They are both widely recognized for their coverage and reliability, but they may not encompass all relevant contributions, especially preprints, regional publications, or emerging venues indexed elsewhere. This selective coverage may introduce publication and citation biases, particularly affecting early-career researchers or institutions in developing regions. Second, while advanced bibliometric techniques such as co-citation, bibliographic coupling, and keyword co-occurrence were employed, these methods offer retrospective and structural insights rather than causal inferences. Interpretations of influence, collaboration, or thematic evolution must therefore be contextualized as descriptive rather than predictive. Similarly, citation-based metrics such as the *h*-index or FWCI, although widely adopted, do not capture the nuanced quality or societal impact of individual publications.

Third, temporal limitations should be noted. Given the journal's relatively recent establishment and the accelerated evolution of the big data domain, some trends—particularly in fast-moving AI subfields like Large Language models, foundation models, and generative AI—may not yet be fully represented in the available datasets. This temporal lag constrains the ability to evaluate the long-term influence of newer publications or emerging topics. Future research could address these limitations by integrating altmetric indicators, full-text content analysis, and machine learning-based topic modeling methods such as dynamic topic modeling, temporal citation-burst analysis, and predictive mapping of emerging themes to yield richer and more dynamic insights. Longitudinal studies comparing JBD with peer journals in overlapping domains (e.g., *Information Sciences*, *Expert Systems with Applications*, *IEEE Access*) can illuminate a wider field view to identify subtle shifts and JBD's role. Furthermore, incorporating author-level demographics, such as gender or academic age, could help explore diversity, equity, and inclusion patterns within the journal's scholarly community. Also, a more granular explanation of why specific articles, countries or authors become highly cited relative to others would require a complementary qualitative or survey-based research design (for example, examining country research systems, reading practices, syllabus adoption, or practitioner uptake), and is therefore beyond the scope of this descriptive

bibliometric study; we explicitly point to this as a promising direction for future work on JBD's influence.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-026-01412-w.

Additional file 1.

Acknowledgements
We would like to thank the editor and the reviewers for valuable comments that have significantly improved the quality of the paper.

Author contributions
M.K: Original draft writing, and Tables 1–2,7, S1-S5. J.M: Conceptualization, supervision, proofreading, and Tables 3–6,8–9, and Figs. 1–7, S1-S4. G.B: Proofreading and supervision. B.F: Proofreading and supervision. T.K: Proofreading and supervision.

Funding
Open Access funding enabled and organized by CAUL and its Member Institutions. No external funding has been received for this submission.

Data availability
Data is available on request.

Declarations

Competing interests
Borko Furht and Taghi Khoshgoftaar are co-Editors-in-Chief of the Journal of Big Data and were not involved in the editorial review or decision-making process for this article. The authors declare no other competing interests.

Received: 16 September 2025 / Accepted: 5 March 2026
Published online: 22 March 2026

References

1. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data. 2018;5(1):1–18.
2. Adadi A. A survey on data-efficient algorithms in big data era. J Big Data. 2021;8(1):24.
3. Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data. 2019;6(1):1–24.
4. Allen MT, Kau JB. Contributing authors and institutions to the Journal of Urban Economics: 1974–1989. J Urban Econ. 1991;30(3):373–84.
5. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-Dabbagh BSN, et al. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. Journal of Big Data. 2023;10(1):46.
6. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data. 2021;8(1):1–74.
7. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959–75.
8. Asmussen CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data. 2019;6(1):1–18.
9. Assunção MD, Calheiros RN, Bianchi S, Netto MA, Buyya R. Big data computing and clouds: trends and future directions. J Parallel Distrib Comput. 2015;79:3–15.
10. Bar-Ilan J. Informetrics at the beginning of the 21st century—a review. J Informetr. 2008;2(1):1–52.
11. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data. 2022;9(1):3.
12. Bensman SJ. Garfield and the impact factor. Annu Rev Inf Sci Technol. 2007;41(1):93–155.
13. Bibri SE. On the sustainability of smart and smarter cities in the era of big data: An interdisciplinary and transdisciplinary literature review. J Big Data. 2019;6:25.
14. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
15. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40.
16. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
17. Broadus RN. Toward a definition of "bibliometrics." Scientometrics. 1987;12(5–6):373–9.
18. Callon M, Courtial J-P, Turner WA, Bauin S. From translations to problematic networks: An introduction to co-word analysis. Soc Sci Inf. 1983;22(2):191–235.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
20. Chen M, Mao S, Liu Y. Big data: A survey. Mob Netw Appl. 2014;19:171–209.
21. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data. 2020;7(1):52.
22. Chen T, & Guestrin C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016; (pp. 785–794).

23. Clarivate. 2025; Journal Citation Reports. https://clarivate.com
24. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. J Big Data. 2015;2:1–24.
25. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: Management, analysis and future prospects. J Big Data. 2019;6(1):1–25.
26. Day O, Khoshgoftaar TM. A survey on heterogeneous transfer learning. J Big Data. 2017;4:1–42.
27. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. In OSDI. 2004;4(1):5–17.
28. Deng J, Dong W, Socher R, Li LJ, Li K, & Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 2009; 248–255 IEEE.
29. Devlin J, Chang MW, Lee K, & Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In; *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2019; 1 4171–4186.
30. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. J Bus Res. 2021;133:285–96.
31. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data. 2021;8:1–37.
32. Fang X, Zhan J. Sentiment analysis using product review data. J Big Data. 2015;2:1–14.
33. Figuerola-Wischke A, Merigó JM, Gil-Lafuente AM, Kydland FE, Amiguet L. The Scandinavian Journal of Economics at 125: A bibliometric overview. Scand J Econ. 2024;126(4):643–97.
34. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
35. Furht B, Khoshgoftaar T. Data science, big data, and machine learning are coming of age. J Big Data. 2025;12:161.
36. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage. 2015;35(2):137–44.
37. Garfield E. Citation indexes for science. Science. 1955;122(3159):108–11.
38. Garfield E. Citation analysis as a tool in journal evaluation. Science. 1972;178:471–9.
39. Glänzel W, Moed HF, Schmoch U, Thelwall M. Bibliometrics as a research field. Scientometrics. 2019;121:139–46.
40. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
41. Graves A. Supervised sequence labelling with recurrent neural networks. Springer; 2012.
42. Guan J, Yu J, Liu X. Assessing journal impact and scientific influence with advanced citation metrics. J Informet. 2025;19(2):254–73.
43. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. Waltham: Morgan Kaufmann Publishers; 2012.
44. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. Journal of Big Data. 2020a;7(1):94.
45. Hancock JT, Khoshgoftaar TM. Survey on categorical data for neural networks. Journal of Big Data. 2020b;7(1):28.
46. Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. 2019;6(1):1–16.
47. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
48. He K, Zhang X, Ren S, & Sun J. Deep residual learning for image recognition. In; *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2016; 770–778.
49. Heck JL, Bremser WG. Six decades of *The Accounting Review*: a summary of author and institutional contributors. Account Rev. 1986;61(4):735–44.
50. Herland M, Khoshgoftaar TM, Bauder RA. Big data fraud detection using multiple Medicare data sources. J Big Data. 2018;5(1):1–21.
51. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1:1–35.
52. Hirsch JE. An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A. 2005;102(46):16569–72.
53. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data. 2022;9(1):102.
54. Huang G, Liu Z, Van Der Maaten L, & Weinberger KQ. Densely connected convolutional networks. In; *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017; 4700–4708.
55. Hussain W, Merigó JM, Rahimi I, Lev B. Half a century of *Omega – The International Journal of Management Science*: a bibliometric analysis. Omega. 2024;129:103226.
56. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. J Big Data. 2016;3:1–25.
57. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6(1):1–54.
58. Journal of Big Data. 2025a; Aims and scope. https://journalofbigdata.springeropen.com/about
59. Journal of Big Data. 2025b; Editorial board. https://journalofbigdata.springeropen.com/editorial-board
60. Kasongo SM, Sun Y. Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. J Big Data. 2020;7(1):105.
61. Kessler MM. Bibliographic coupling between scientific papers. Am Doc. 1963;14(1):10–25.
62. Khorshidi MS, Merigó JM, Beydoun G. Half a century of *Information Processing & Management*: A bibliometric retrospective. Inf Process Manag. 2025;62(6):104238.
63. Klavans R, Boyack KW. Research portfolio analysis and topic prominence. J Informetr. 2017;11(4):1158–74.
64. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
65. Kumar S, Tiwari P, Zymbler M. Internet of Things is a revolutionary approach for future technology enhancement: A review. J Big Data. 2019;6(1):1–21.
66. Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data. 2015;2:1–36.
67. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
68. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
69. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data. 2018;5(1):1–30.
70. Mayer-Schönberger V, & Cukier K. *Big data: A revolution that will transform how we live, work, and think*. 2013; Houghton Mifflin Harcourt.

71. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2:1–21.
72. Nti IK, Adekoya AF, Weyori BA. A comprehensive evaluation of ensemble learning for stock-market prediction. J Big Data. 2020;7:20.
73. Olshannikova E, Ometov A, Koucheryavy Y, Olsson T. Visualizing big data with augmented and virtual reality: challenges and research agenda. J Big Data. 2015;2:1–27.
74. Paul J, Lim WM, O'Cass A, Hao AW, Bresciani S. Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). Int J Consum Stud. 2021;45(4):O1–16.
75. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, … & Duchesnay É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 2011; *12*, 2825–2830.
76. Pritchard A. Statistical bibliography or bibliometrics? J Doc. 1969;25(4):348–9.
77. Purkayastha A, Palmaro E, Falk-Krzesinski HJ, Baas J. Comparison of two article-level, field-independent citation metrics: field-weighted citation impact (FWCI) and relative citation ratio (RCR). J Informetr. 2019;13(2):635–42.
78. Rawat KS, & Sood SK. Emerging trends and global scope of big data analytics: A scientometric analysis. 2021; *Quality & Quantity*, 55(4).
79. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006;74(1):016110.
80. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49.
81. Ronneberger O, Fischer P, & Brox T. U-Net: Convolutional networks for biomedical image segmentation. In; *Medical Image Computing and Computer-Assisted Intervention – MICCAI* 2015; 9351, 234–241. Springer.
82. Rousseau R. Forgotten founder of bibliometrics. Nature. 2014;510:218.
83. Roy D, Dutta M. A systematic review and research perspective on recommender systems. J Big Data. 2022;9(1):59.
84. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, … & Fei-Fei L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision,* 2015; *115*, 211–252.
85. Sarker IH, Kayes ASM, Watters P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. J Big Data. 2019;6(1):1–28.
86. Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. J Big Data. 2020;7:1–29.
87. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
88. SCImago. 2025; Journal rankings. https://www.scimagojr.com
89. Scopus. 2025; Scopus database. https://www.scopus.com
90. Seliya N, Khoshgoftaar TM, & Van Hulse JA. study on the relationships of classifier performance metrics. In; *21st IEEE International Conference on Tools with Artificial Intelligence* 2009; 59–66. IEEE.
91. Seyedan M, Mafakheri F. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. J Big Data. 2020;7(1):53.
92. Shen J, Shafiq MO. Short-term stock market price trend prediction using a comprehensive deep learning system. J Big Data. 2020;7:1–33.
93. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48.
94. Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for COVID-19. J Big Data. 2021;8(1):1–54.
95. Shorten C, Khoshgoftaar TM, Furht B. Text data augmentation for deep learning. J Big Data. 2021;8(1):101.
96. Simonyan K, & Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014; *arXiv preprint* arXiv:1409.1556.
97. Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inf Sci. 1973;24(4):265–9.
98. Smith G. Step away from stepwise. J Big Data. 2018;5:32.
99. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019;6(1):1–19.
100. Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big data: deep learning for financial sentiment analysis. J Big Data. 2018;5(1):1–25.
101. SpringerOpen. 2025; Journal of Big Data – metrics & usage. https://journalofbigdata.springeropen.com
102. Sreenu G, Saleem Durai MA. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data. 2019;6(1):1–27.
103. Srivastava S, Divekar AV, Anilkumar C, Naik I, Kulkarni V, Pattabiraman V. Comparative analysis of deep learning image detection algorithms. J Big Data. 2021;8(1):66.
104. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, & Wojna Z. Rethinking the Inception architecture for computer vision. In; *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2016; 2818–2826.
105. Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data. 2020;7:1–47.
106. Teh HY, Kempa-Liehr AW, Wang KIK. Sensor data quality: a systematic review. J Big Data. 2020;7(1):11.
107. Thudumu S, Branch P, Jin J, Singh J. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data. 2020;7:1–30.
108. Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2:1–32.
109. Tukey JW. Comparing individual means in the analysis of variance. Biometrics. 1949;5(2):99–114.
110. Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523–38.
111. Van Eck NJ, & Waltman L. 2023; *VOSviewer manual*. https://www.vosviewer.com
112. Vaswani A, Shabeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, … & Polosukhin I. Attention is all you need. In; *Advances in Neural Information Processing Systems,* 2017; *30*.
113. Waltman L. A review of the literature on citation impact indicators. J Informetr. 2016;10(2):365–91.
114. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3:1–40.
115. Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. J Big Data. 2019;6(1):1–18.

116. Yazdanjue N, Yazdanjouei H, Gharoun H, Khorshidi MS, Rakhshaninejad M, Amiri B, et al. A comprehensive bibliometric analysis on social network anonymization: current approaches and future directions. Knowl Inf Syst. 2025;67(1):29–108.
117. Zaharia M, Chowdhury M, Franklin M. J, Shenker S, & Stoica I. Spark: Cluster computing with working sets. In; *2nd USENIX Workshop on Hot Topics in Cloud Computing* 2010; *(HotCloud 10)*.
118. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, … & Stoica I. Apache Spark: A unified engine for big data processing. *Communications of the ACM,* 2016; *59*(11), 56–65.
119. Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: a survey. J Big Data. 2015;2:1–41.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.