SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I5P109

Volume 11 Issue 5, 84-98, May 2024

Original Article

Emerging Trends in Data Science and Big Data Analytics: A Bibliometric Analysis

Abdulaziz Yasin Nageye[1*], Abdukadir Dahir Jimale[1], Mohamed Omar Abdullahi[1], Yahye Abukar Ahmed[1]

[1]*Faculty of Computing, SIMAD University, Mogadishu, Somalia.*

[*]*Corresponding Author : nageye8@simad.edu.so*

Received: 13 March 2024 Revised: 24 April 2024 Accepted: 11 May 2024 Published: 31 May 2024

Abstract *- This bibliometric analysis explores the landscape of research in Data Science and Big Data Analytics over the period from 2010 to March 2024. Leveraging advanced bibliometric techniques, including data collection from Scopus, data screening, preprocessing, and analysis using VOSviewer, Bibliometric of R package, and Microsoft Excel, this study aims to identify key trends, patterns, and dynamics within the field. The analysis encompasses document types, publication and citation trends, contributing countries, influential authors and sources, keyword co-occurrence networks, and influential affiliations. The findings provide valuable insights into the scholarly discourse, collaboration networks, and emerging research directions in Data Science and Big Data Analytics, facilitating evidence-based decision-making and fostering innovation in the field.*

Keywords *- Data science, Big data analytics, Bibliometric analysis, Collaboration networks, Analytics.*

1. Introduction

Data Science and Big Data Analytics are emerging fields that involve the analysis of large amounts of data to gain insights and make informed decisions. These fields are being applied in various industries, including healthcare, education, and business, to leverage the power of Big Data and improve processes and outcomes [1]. Big Data refers to the large volume, velocity, and variety of data that traditional processing tools are unable to handle. Cloud Computing provides the infrastructure to store, process, and analyze Big Data cost-effectively and efficiently [2]. Machine learning techniques, such as linear regression, decision trees, random forests, and gradient boosting tree algorithms, are used in Big Data Analytics to perform accurate analysis and make predictions. Overall, Data Science and Big Data Analytics offer opportunities for organizations to gain knowledge, make data-driven decisions and improve outcomes in various sectors [3].

Data science and big data analytics are interdisciplinary fields that involve the systematic processing and analysis of large volumes of data to uncover hidden patterns and extract actionable insights for decision-making [4]. Key concepts in data science and big data analytics include statistical and visual analysis of data, the use of machine learning algorithms and models, the importance of infrastructure and tooling, the concept of big data warehouses for efficient storage and processing, and the application of data science in various domains such as business, finance, healthcare, and agriculture [5].

Big data analytics and data science rely on various popular tools and techniques. Apache Hadoop, Spark, Storm, and Flink are widely used big data processing frameworks that support big data analysis. Educational Data Mining and Learning Analytics are important concepts in the education industry, with data mining tools being used for working with big data [6]. Python-based libraries like Pandas, Numpy, STATA, and SPSS are commonly used for analyzing and mining big data sets. Apache Spark, with its innovative processing framework and high-level programming libraries, is another popular tool for analyzing big data. In the field of proteomics and metabolomics, software tools like BioContainers, Galaxy, and Nextflow are used for large-scale data analysis [7].

Machine learning algorithms commonly used in data science include supervised learning algorithms such as Naive Bayes (NB), Logistic Regression (LR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests (RF). Unsupervised learning algorithms like k-means and Gaussian Mixture Models (GMM) are also popular. Additionally, deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have gained prominence in recent years. These algorithms are used for various applications, including classification problems, fraud detection, email spam filtering, image recognition, and predicting soil parameters [8, 9].

The following languages provide the necessary tools and frameworks for processing and analyzing large volumes of

data in various data science and big data applications. Python is a popular programming language used in data science and big data analytics due to its access to powerful mathematical and statistical tools. Another widely used language is Scala, which is used in programming environments like MapReduce, Spark, and Flink for large-scale data analysis. Additionally, Java is also used in big data analytics, particularly in Apache Spark's MLlib for analyzing big datasets [10].

Big data analytics and data science face several challenges. These challenges include issues in data collection, storage, analysis, and presentation. Security is a major concern, with the need to protect big data during analytics and ensure confidentiality, integrity, availability, non-repudiation, and access control [11]. In the context of agriculture, challenges arise in handling and analyzing large amounts of data, incorporating technologies like wireless sensor networks, machine learning, the Internet of Things, drones, and robotics. The challenges in big data analytics and cloud computing include managing and analyzing large volumes of data, as well as the need for new industrial models and a creative approach to data mining. Working with big spatial data presents challenges such as dealing with bias, emphasizing practical significance, accepting messiness in data, and understanding the causes of bias and complexity [12].

A bibliometric analysis serves as a systematic and quantitative examination of scholarly publications within a specific field or topic, providing insights into the trends, patterns, and dynamics of scientific literature. By employing bibliometric techniques, researchers can evaluate the dissemination of knowledge, identify influential authors and institutions, trace the evolution of research themes over time, and uncover emerging areas of interest [13]. Bibliometric analyses often involve the collection and analysis of metadata from academic databases, including citation data, publication counts, author affiliations, and journal impact factors. Through rigorous bibliometric methodologies, researchers can gain a comprehensive understanding of the scholarly landscape, facilitating evidence-based decision-making, strategic planning, and the assessment of research impact and productivity [14].

In the realm of Data Science and Big Data Analytics, conducting a bibliometric analysis holds considerable significance for several compelling reasons. Firstly, given the rapid proliferation and interdisciplinary nature of these domains, a meticulous bibliometric study provides a structured framework for synthesizing the vast and multifaceted literature corpus [15]. This systematic approach aids researchers in identifying overarching trends and pivotal contributions, thereby facilitating a deeper understanding of the field's evolution. Secondly, in an era where evidence-based decision-making is paramount, bibliometric analyses offer invaluable insights into the dissemination and impact of research outputs. By discerning citation patterns and assessing publication metrics, stakeholders can make informed decisions regarding resource allocation, policy formulation, and strategic planning [16]. Additionally, by pinpointing knowledge gaps and emerging research frontiers, bibliometric analyses serve as catalysts for innovation and collaboration, driving advancements in Data Science and Big Data Analytics. Consequently, a robust bibliometric analysis not only enriches scholarly discourse but also fosters progress and innovation within these dynamic and rapidly evolving domains [17].

The aims and objectives of this study are multifaceted and designed to provide a comprehensive understanding of the landscape of Data Science and Big Data Analytics through the lens of bibliometric analysis. Firstly, the study seeks to map the intellectual structure of the field by identifying key research themes, influential authors, and prolific institutions. By elucidating the network of collaborations and knowledge dissemination pathways, the study aims to uncover underlying patterns and trends that have shaped the development of Data Science and Big Data Analytics. Secondly, the study aims to assess the impact and visibility of research outputs within the field, gauging citation metrics, publication trends, and journal distributions. Through meticulous data collection and analysis, the study endeavours to offer insights into the dissemination and reception of scholarly work, facilitating evidence-based decision-making and strategic planning for researchers, policymakers, and practitioners alike. Moreover, the study aims to identify emerging research areas and knowledge gaps, providing valuable guidance for future research directions and interdisciplinary collaborations. Ultimately, by elucidating the dynamics of scholarly communication and knowledge production within Data Science and Big Data Analytics, this study aims to contribute to the advancement and consolidation of these vital domains.

Despite the rapid growth and interdisciplinary nature of Data Science and Big Data Analytics, there remains a significant gap in understanding the evolving scholarly landscape and identifying key trends. While these fields have garnered substantial attention and investment, there is a lack of comprehensive studies that systematically analyze the breadth and depth of research contributions, particularly through a bibliometric lens. Existing literature offers insights into specific aspects of data science and big data analytics but often lacks a holistic perspective that encompasses the entire scholarly discourse. Furthermore, as these fields continue to evolve with technological advancements and emerging research areas, there is a pressing need for updated analyses that reflect the current state of research and identify emerging trends and challenges. Thus, this study aims to fill this research gap by conducting a rigorous bibliometric analysis to provide a comprehensive understanding of the

landscape of Data Science and Big Data Analytics. Through systematic data collection, analysis, and interpretation, the study seeks to identify key research themes, influential authors, publication trends, and citation patterns, thereby offering valuable insights for researchers, practitioners, and policymakers engaged in these dynamic fields.

The remaining parts of the paper are structured into five main sections. The first section, Related Work, provides a comprehensive review of existing literature and prior research relevant to the field of Data Science and Big Data Analytics. The second section outlines the methodology employed in conducting the bibliometric analysis, including data collection sources, inclusion criteria, and analytical techniques. Following this, the findings of the analysis are presented in the third section, encompassing insights into research themes, influential authors, publication trends, and citation patterns within Data Science and Big Data Analytics. Subsequently, the fourth section offers a comprehensive discussion of the findings, providing contextualization, interpretation, and implications for research, practice, and policy. Lastly, the paper concludes by synthesizing the key findings, highlighting the significance of the study, and suggesting avenues for future research in the field.

2. Related Work

The Related Work section offers a comprehensive overview of prior research endeavours and seminal studies within the realm of Data Science and Big Data Analytics. Through an examination of diverse scholarly contributions, this section provides valuable insights into the evolution, trends, and key themes shaping the field. From exploring the application of machine learning models in specific domains to examining the transformative impact of data analytics across sectors, the Related Work section delves into a rich tapestry of literature that informs and contextualizes the current study. By synthesizing and critically evaluating prior research, this section sets the stage for the subsequent methodological approach adopted in our bibliometric analysis.

Ankit Verma and Hansraj proposed in their study Big Data and Deep Learning to enhance intrusion detection systems, evaluating machine learning models like Deep Feed-Forward Neural Networks, Random Forests, and Gradient Boosting Tree. The study also explores the use of machine learning for flight delay prediction, improving prediction accuracy with a random forest-based model. Additionally, the study delves into the application of machine learning algorithms in various sectors like healthcare, production, sales, IoT devices, and organizations, focusing on data patterns and decision-making. The research also compares the performance of different machine learning techniques on datasets like Cardiotocography (CTG) using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [18].

Rahul Rn focuses his study on the evolution of data analytics, starting from the era of data warehousing and business intelligence to the current era of big data and data science. It highlights how businesses have transitioned from using basic analytics to uncover insights to leveraging big data analytics for immediate decision-making. The study also emphasizes the impact of data analytics across various sectors like finance, healthcare, energy management, and retail. Additionally, it discusses the role of data analytics in personalized recommendations, gaming, and energy management [19].

Matthew A. Waller and Stanley E. Fawcett emphasize their study on the importance of domain knowledge in data science, particularly in the context of logistics and supply chain management. Researchers highlight the need for a combination of analytical skills and business understanding to address challenges in predictive analytics and data-driven decision-making effectively. Additionally, the study encourages research submissions on data science, predictive analytics, and big data that are relevant to logistics and supply chain management, highlighting the transformative potential of these tools in the field [20].

Radwa Elshawi and Sherif Sakr investigate their study of the convergence of big data systems and machine learning challenges, aiming to establish Big Data Science as a Service. It highlights the era of Big Data, characterized by the 3Vs: Volume, Velocity, and Variety, emphasizing the need for new technical architectures and analytics to unlock business value. The research explores the evolution of big data processing platforms, led by Hadoop and transitioning to newer platforms like Hive, Impala, and Spark for specialized data processing tasks. Additionally, it discusses the emergence of big data science as a discipline combining statistics, artificial intelligence, and computer science to extract value from abundant data sources. The study also addresses the role of cloud computing in providing scalable and cost-effective infrastructure for big data analytics, emphasizing the shift towards centralized data centers and the pay-as-you-go pricing model. Furthermore, it touches on the challenges of data availability, collaboration, and interoperability in the realm of big data analytics, highlighting the importance of open data initiatives and online data markets for fostering innovation and insights [21].

Daphne R. Raban and Avishag Gordon proposed the evolution of data science and big data research and conducted a bibliometric analysis to explore the trends and dynamics in these fields. The research focused on various aspects such as the growth of publications, research areas covered, author commitment, research funding, subject dispersion, publication dynamics, and limitations of the study. Key findings include a significant increase in the number of publications on Big Data (BD) and Data Science

(DS) from 2010 to 2019, with BD experiencing a more rapid growth rate. BD literature tends to focus on computer science, management, medical sciences, and engineering, while DS literature is more disciplinary dispersed. The study also highlighted differences in author commitment, research funding sources, publication dynamics, and the core journals covering BD and DS. Additionally, it discussed the limitations of relying on Web of Science (WoS) data and the implications of the study's findings [22].

Hamed Jahani et al. focused their study on conducting a systematic review of methodologies used in supply chain and logistics research related to Data Science and Big Data Analytics (DS&BDA). The research aims to develop a conceptual framework by addressing specific research questions and objectives. These questions included strategies for identifying relevant studies, understanding research processes and guidelines, investigating research topics and methodologies in DS&BDA within the context of supply chain and logistics, and identifying gaps in the literature for using DS&BDA techniques in this field. The study outlined a detailed research process involving research planning, conducting reviews, and reporting results. It also discussed the review protocol used, the search strategy employed in academic databases, the selection criteria for articles, and the analysis of search results from various databases. Additionally, the study analyzed recent relevant review studies, identified gaps in research topics, and categorized lessons learned from the content analysis of selected review articles. The research also delved into the methodologies used in review studies, gaps in research topics, and the distribution of publications per year and venue, providing insights into the trends and approaches in DS&BDA research within the supply chain and logistics domain [23].

Justin Zuopeng Zhang et al. proposed to conduct a comprehensive bibliometric analysis focusing on big data analytics and machine learning research over a span of fifteen years (2006-2020). The research employed bibliometric methods to analyze patterns in published articles, including keyword topics and content analysis of journals, universities, and countries. The study was structured into three phases: planning the review process, implementing the bibliometric protocol, and reporting the findings.

Methodologies such as bibliographic coupling, citation analysis, co-citation analysis, co-word analysis, and prestige analysis were utilized to extract insights from the data. The results included descriptive and bibliometric analyses, revealing trends in publication and citation, author productivity, country collaboration networks, and organization-wise analysis. The study identified top authors, universities, and journals in the field, providing valuable insights into the landscape of big data analytics and machine learning research [24].

The existing literature in the field of Data Science and Big Data Analytics reflects a burgeoning interest in leveraging advanced analytical techniques to address a myriad of challenges across various domains. However, amidst this extensive body of research, several notable gaps emerge that our study aims to bridge. While previous studies have explored the application of machine learning models, such as Deep Feed-Forward Neural Networks and Random Forests, in specific domains like intrusion detection systems and flight delay prediction, there remains a need for a comprehensive examination of the broader landscape of Data Science and Big Data Analytics. Additionally, while some research has emphasized the evolution and impact of data analytics across sectors like finance, healthcare, and logistics, there is a lack of comprehensive analyses that integrate insights from diverse domains to provide a holistic understanding of the field. Furthermore, existing studies often focus on specific methodologies or applications within Data Science and Big Data Analytics, overlooking the interdisciplinary nature of the field and its implications for research and practice. Our study seeks to address these gaps by conducting a rigorous bibliometric analysis that encompasses a wide range of research themes, methodologies, and applications within Data Science and Big Data Analytics, thereby contributing to a more nuanced understanding of this rapidly evolving field.

3. Method

The methodology employed in this study aims to rigorously analyze the landscape of research in Data Science and Big Data Analytics through a comprehensive bibliometric approach. The overarching objective is to gain insights into the trends, patterns, and dynamics shaping these fields over the past decade, from 2010 to March 2024. By systematically collecting, filtering, and analyzing scholarly literature, this study seeks to identify key Publications and Citations Trends, Contributing Countries, Influential Authors, Influential Sources, Keyword Co-occurrence Networks, and Influential Affiliations. Through the application of advanced bibliometric techniques, the study aims to provide valuable insights for researchers, practitioners, and policymakers engaged in Data Science and Big Data Analytics. The methodology is structured to ensure transparency, reproducibility, and reliability of the findings, aligning with the objectives of generating actionable knowledge and facilitating evidence-based decision-making in the field.

3.1. Data Collection and Search Strategy

For our study, we utilized a targeted search strategy to gather relevant scholarly literature from the Scopus database. We employed a search query formulated to capture publications specifically related to Data Science and Big Data Analytics. The search strategy included keywords such as "data science" and "big data analytics," ensuring a

comprehensive coverage of relevant literature. Additionally, we restricted the search to publications between 2010 and March 2024 to encompass recent developments in the field. To refine the dataset, we focused on specific document types, including Articles, books, conference papers, Reviews, Book chapters, and Conference reviews. Furthermore, to maintain consistency and facilitate analysis, we limited our search to English-language publications. This approach ensured that our dataset comprised scholarly literature directly pertinent to Data Science and Big Data Analytics, providing a solid foundation for our bibliometric analysis. The following is the search string that was used to fetch the data from Scopus: Search String = TITLE-ABS-KEY (data AND science AND big AND data AND analytics) AND PUBYEAR > 2009 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE , “cp”) OR LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “ch”) OR LIMIT-TO (DOCTYPE , “re”) OR LIMIT-TO (DOCTYPE , “cr”) OR LIMIT-TO (DOCTYPE , “bk”)) AND (LIMIT-TO (LANGUAGE , “English”))

3.2. Data Filtering and Selection

Following the initial data collection process, we meticulously filtered and selected relevant scholarly literature based on predefined inclusion and exclusion criteria. Inclusion criteria comprised various document types such as articles, conference papers, book chapters, reviews, and books, ensuring a comprehensive representation of scholarly outputs. Moreover, publications from the period between 2010 and March 2024 were included to capture recent advancements in the field. Additionally, to maintain consistency and facilitate analysis, only publications in English were considered. Conversely, exclusion criteria were employed to eliminate non-scholarly materials like editorials, letters, and conference abstracts.

Publications outside the specified timeframe, before 2010 or after March 2024, were excluded to focus on contemporary research. Moreover, materials in languages other than English were omitted to ensure uniformity in language and comprehension. Furthermore, publications lacking peer review or academic rigour were excluded to maintain the credibility and reliability of the dataset. The application of these rigorous criteria ensured that the final dataset comprised high-quality scholarly literature directly relevant to our study of Data Science and Big Data Analytics. Table 1 below shows the inclusion and exclusion criteria.

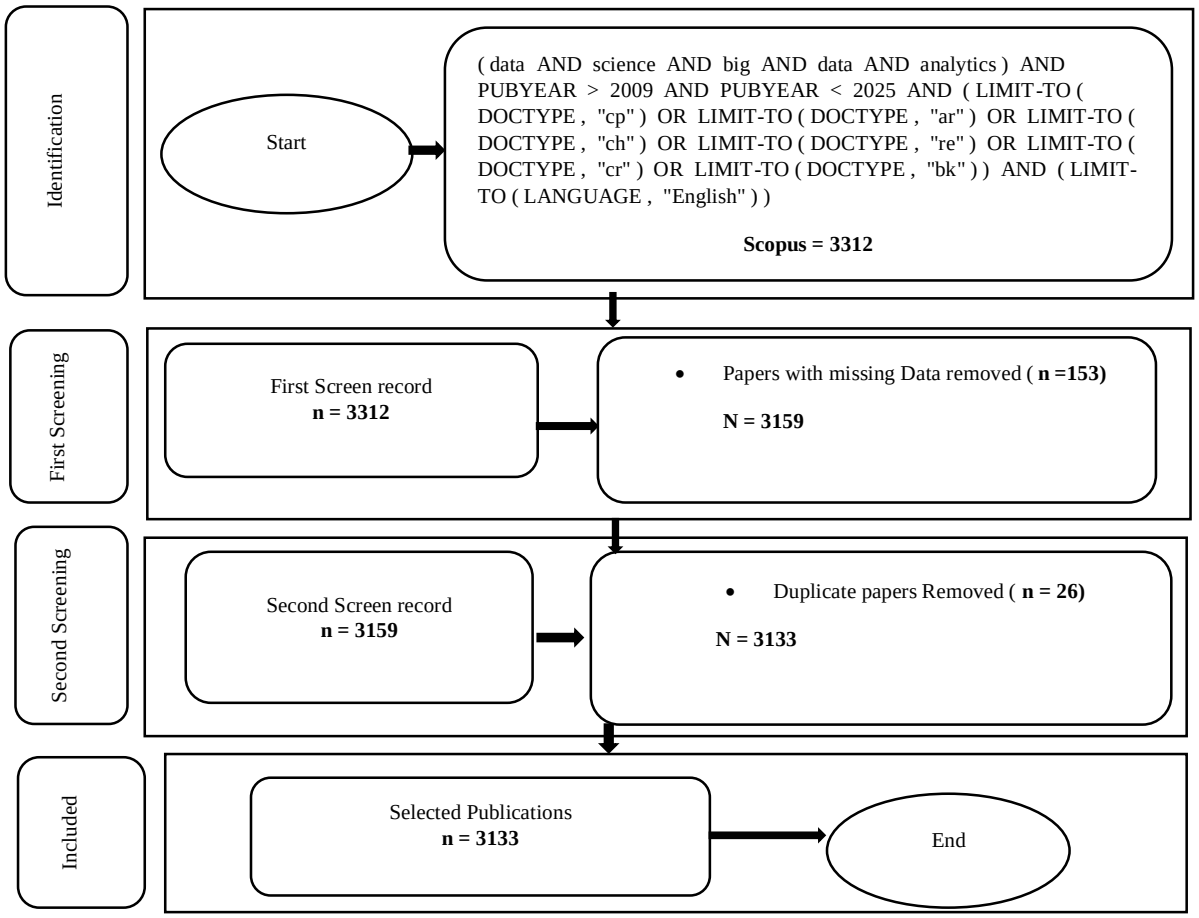

Fig. 1 Search and screening strategy

Table 1. Inclusion and exclusion criteria

Inclusion	Exclusion
Publication Related to Data Science and Big Data Analytics	Publication not related to Data Science and Big Data Analytics
Limited to Article, Book, Conference Paper, Review, Book Chapter, Conference Review	Excluded Non-Article, Non-Book, Non-Conference Paper, Non-Review, Non-Book Chapter, Non-Conference Review
Publication between 2010 and March 2024	Publication before 2010, after March 2024
English Language Publications	Languages other than English

3.3. Data Screening and Preprocessing

The data screening process involved several stages aimed at refining the dataset to ensure its relevance and reliability for analysis. Initially, all retrieved records were screened based on their titles and abstracts to assess their alignment with the research topic. Subsequently, papers with missing values were removed, and duplicates were also removed to eliminate redundancy and ensure data integrity. The remaining records underwent full-text screening to assess further their eligibility based on the predefined inclusion and exclusion criteria. Any discrepancies or uncertainties during the screening process were resolved through discussion and consensus among the research team members.

Following data screening, the selected publications underwent preprocessing to enhance the quality and usability of the dataset for analysis. This involved standardizing data formats, resolving inconsistencies, and encoding categorical variables for uniformity. Additionally, data-cleaning techniques were applied to address missing values, typographical errors, and outliers that could potentially skew the analysis results.

Preprocessing also included the extraction of relevant metadata such as publication year, author affiliations, keywords, and citation counts for further analysis. By meticulously preprocessing the dataset, we ensured its readiness for comprehensive bibliometric analysis and interpretation. The above Figure 1 shows the search and screening strategy.

Figure 1 above illustrates the iterative nature of the data screening process, starting from the initial retrieval of records to the final selection of relevant publications. Preprocessing activities are depicted as integral steps in refining the dataset for subsequent analysis, highlighting their importance in ensuring data quality and reliability.

3.4. Data Analysis

Upon assembling the dataset, we conducted comprehensive bibliometric analyses to extract valuable insights and trends in the field of Data Science and Big Data Analytics. We employed a combination of quantitative and qualitative methods to analyze publication patterns, citation networks, keyword co-occurrences, and author affiliations.

For bibliometric visualization and network analysis, we facilitated the visualization of co-authorship networks, keyword co-occurrence networks, and citation networks, allowing us to identify key clusters and trends within the literature. We utilized VOSviewer, a widely used software tool for constructing and visualizing bibliometric networks.

Additionally, we employed Bibliometrix, an R package specifically designed for bibliometric analysis, to perform statistical analyses and generate visualizations of publication trends, author productivity, and citation patterns [25]. This tool enabled us to conduct more advanced analyses by providing deeper insights into the scholarly landscape. Furthermore, Microsoft Excel was utilized for data management, preprocessing, and basic statistical analyses. Excel's versatility allowed us to organize and manipulate the dataset efficiently, facilitating data cleaning, filtering, and preparation for further analysis.

By leveraging these tools in tandem, we were able to conduct a comprehensive and robust bibliometric analysis, uncovering significant trends, influential authors, and emerging research topics within the realm of Data Science and Big Data Analytics.

4. Results and Discussions

4.1. Document Types

In this subsection, we explore the various document types prevalent in research related to data science and big data analytics. By analyzing the distribution of document types, we gain insights into the diverse forms of scholarly output within the field.

Our analysis reveals a range of document types contributing to the scholarly discourse in data science and big data analytics. Among these, articles represent a significant portion of scholarly output, with 1,086 documents. These articles serve as primary avenues for disseminating original research findings and empirical studies, reflecting the depth and breadth of research conducted within the field.

Additionally, conference papers play a crucial role in knowledge dissemination, with 1,345 documents identified in our analysis. Conference papers serve as vital platforms for sharing cutting-edge research, facilitating collaboration among researchers, and advancing the field's frontiers through the exchange of ideas and findings.

The presence of book chapters, reviews, and books further enriches the scholarly landscape in data science and big data analytics. Our analysis identified 305 book chapters, 254 reviews, and 143 books, highlighting the diversity of scholarly contributions and the varied formats through which knowledge is synthesized and disseminated within the field.

Understanding the distribution of document types is essential for researchers, practitioners, and policymakers to navigate the scholarly landscape effectively. By recognizing the prevalence of different document types and their respective roles, stakeholders can identify relevant resources, engage with diverse forms of research output, and develop strategies for disseminating research findings effectively. Below, Figure 2 shows the different document types that the data consists.

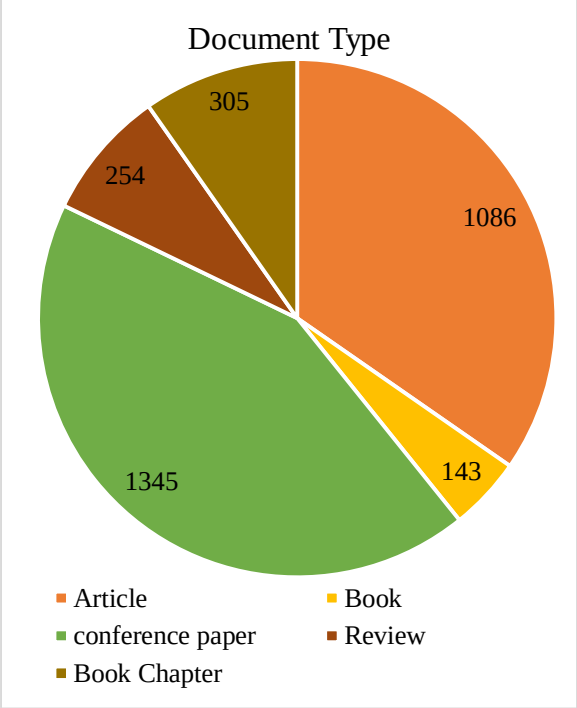

Fig. 2 Document types

4.2. Publications and Citations Trends

In this section, we delve into the publication and citation trends in the field of data science and big data analytics, analyzing the number of documents published and citations received annually from 2010 to 2024. This analysis provides valuable insights into the evolving scholarly activity and impact within the field.

Over the past decade, there has been a notable growth trajectory in publication output, indicative of the increasing interest and engagement in data science and big data analytics research. The scholarly output in 2010 was negligible, with no documents published. However, subsequent years witnessed a steady rise in publication output. Notable increases were observed from 2014 onwards, with peak years in 2020, with 406 documents published, and 2021, with 434 documents published.

Concurrently, citation trends have shown a corresponding increase over the years, reflecting the growing impact and influence of research within the field. Notable peaks in citation counts coincide with peak publication years, indicating the resonance and impact of research contributions during these periods. For instance, in 2012, publications received 5,776 citations, showcasing the significance of research outputs during that year. Similarly, 2016 and 2018 also stood out as influential years in terms of citation counts, with 10,257 and 8,080 citations, respectively.

However, while the overall trend exhibits growth in both publication output and citations, there are variations and fluctuations in specific years. Recent years, specifically 2023 and 2024 (up to March), have seen a stabilization or slight decline in publication output compared to the preceding years. This trend may reflect various factors, such as saturation of research topics or shifting priorities within the field.

Understanding these publication and citation trends is crucial for stakeholders to identify emerging research directions, allocate resources effectively, and assess the scholarly impact of research outputs. The observed trends underscore the dynamic nature of research in data science and big data analytics, characterized by growth, innovation, and evolving scholarly discourse. Figure 3 shows the publications and citation trends.

4.3. Contributing Countries

In this subsection, we examine the countries contributing to research in the fields of data science and big data analytics. By analyzing the distribution of research output across different countries, we gain insights into the global landscape of scholarly activity and collaboration within the field.

Our analysis reveals a diverse range of countries actively engaged in research related to data science and big data analytics. The United States emerges as a leading contributor with 987 documents and an impressive citation count of 35,322. Similarly, the United Kingdom demonstrates substantial involvement, with 253 documents and 6,821 citations. China, Germany, and Canada also feature prominently, each making significant contributions to the scholarly discourse with 316 documents and 5,607 citations, 167 documents and 4,286 citations, and 180 documents and 4,114 citations, respectively. Other countries such as the Netherlands, Ireland, and India exhibit notable research activity, underscoring the global nature of research in data science and big data analytics.

The Netherlands contributed 71 documents with 3,693 citations, Ireland contributed 19 documents with 3,521 citations, and India contributed 504 documents with 3,500 citations.

The distribution of contributions by country highlights the collaborative and interconnected nature of research within the field. Collaborations across borders enable the exchange of knowledge, expertise, and resources, driving innovation and advancement in data science and big data analytics.

Furthermore, recognizing the varied contributions of different countries enhances the inclusivity and diversity of research efforts, ensuring that insights and perspectives from diverse cultural and geographical contexts are incorporated into the scholarly discourse. By fostering international collaboration and cooperation, stakeholders can harness the collective expertise and resources of the global research community to address complex challenges and drive innovation in data science and big data analytics. The below Figure 4 shows the contributing countries' network.

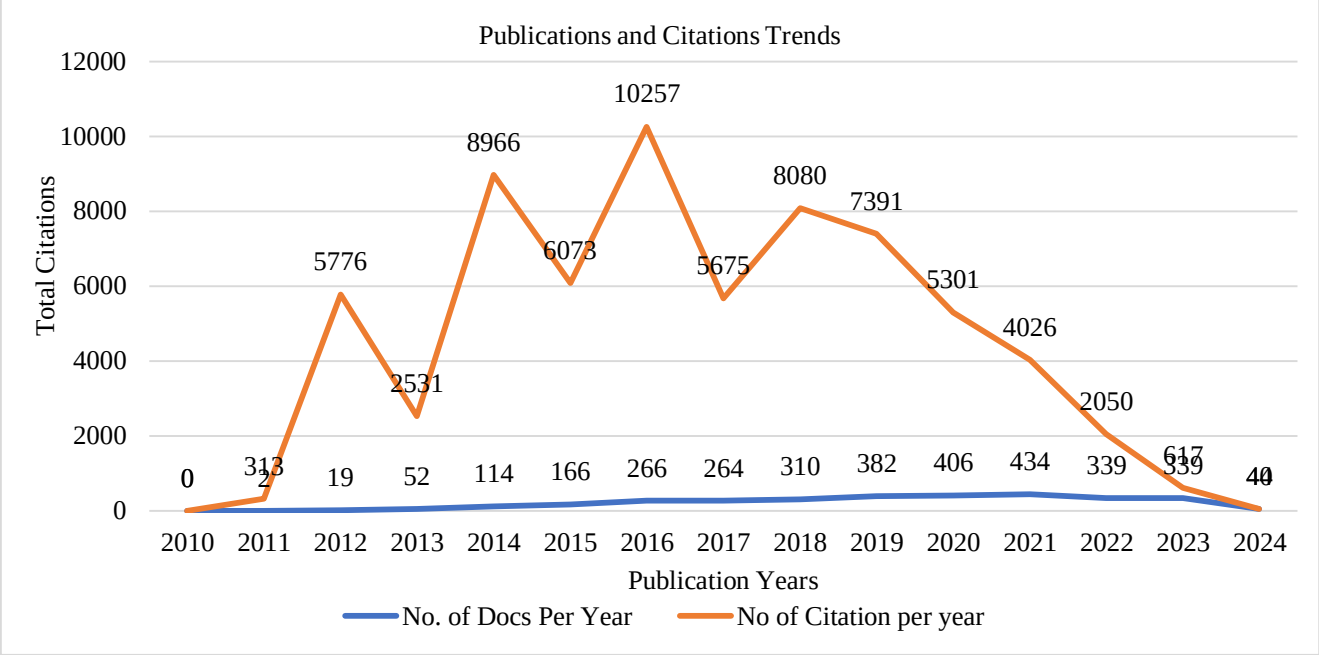

Fig. 3 Publications and citations trends

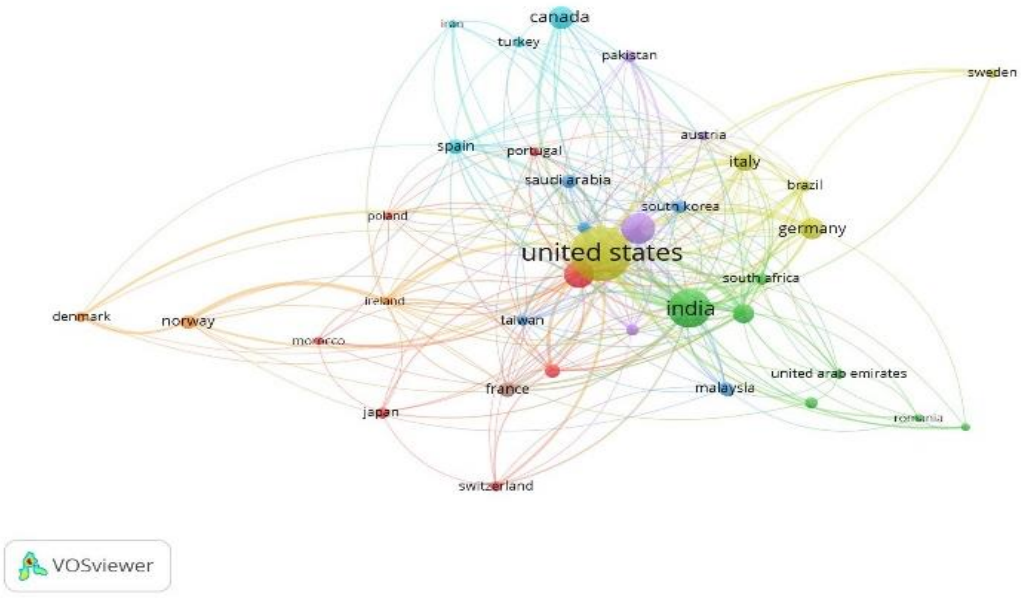

Fig. 4 Contributing countries network

Table 2. Top ten most influential authors

Author	h_index	g_index	m_index	TC	NP	PY_start
Wang J	8	15	0.727	1275	15	2014
Leung Ck	16	26	1.6	848	59	2015
Cuzzocrea A	11	21	1	561	21	2014
Chen Y	7	18	0.636	477	18	2014
Liu Y	9	18	0.818	356	24	2014
Bibri Se	9	18	1.286	328	20	2018
Dinov Id	7	10	0.7	317	10	2015
Zhang Y	8	17	0.8	308	18	2015
Vatrapu R	9	10	0.818	298	10	2014
Chen H	7	15	0.538	297	15	2012

4.4. Influential Authors

In this subsection, we identify and analyze the top ten most influential authors in the field of data science and big data analytics. By examining metrics such as h-index, g-index, and m-index, we gain insights into the impact and significance of their scholarly contributions.

Our analysis reveals a diverse array of authors who have made significant contributions to the field. Among them, Wang J stands out with an h-index of 8 and a total citation count (TC) of 1,275 since 2014, indicating substantial impact and influence within the research community. Similarly, Leung CK possesses an h-index of 16 and a TC of 848 since 2015, showcasing notable contributions to the field. Cuzzocrea A, with an h-index of 11 and a TC of 561 since 2014, and Chen Y, with an h-index of 7 and a TC of 477 since 2014, have also made significant contributions to the field. Additionally, Liu Y, Bibri SE, and Dinov ID exhibit notable research output, each contributing substantially since their respective start years. Authors such as Zhang Y, Vatrapu R, and Chen H have also made notable contributions to the field, further enriching the scholarly discourse in data science and big data analytics.

The identification of influential authors provides valuable insights into the key contributors shaping the field. Understanding the contributions of these scholars is essential for identifying emerging trends, assessing the quality of research output, and fostering collaboration within the field. By recognizing the impact of individual scholars, stakeholders can identify potential collaborators, mentors, and research partners, thus facilitating knowledge exchange and advancement within the field. The Table 3. shows the top ten most influential authors.

4.5. Influential Sources

In this subsection, we identify and analyze the top ten most influential sources in the field of data science and big data analytics.

These sources serve as key platforms for disseminating knowledge, shaping research trends, and fostering scholarly discourse within the field.

IEEE Access emerges as a significant platform, boasting an h-index of 12 and a total citation count (TC) of 909 since 2016. Following closely, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) holds an h-index of 11 and a TC of 479 since 2012, underscoring its importance in various subfields. Sustainability (Switzerland) has garnered attention with an h-index of 11 and a TC of 564 since 2018, reflecting the increasing focus on sustainability within data science and big data analytics research. The ACM International Conference Proceeding Series, with an h-index of 9 and a TC of 392 since 2012, serves as a crucial venue for presenting research findings and innovations. Similarly, Advances in Intelligent Systems and Computing, with an h-index of 8 and a TC of 265 since 2015, have contributed significantly to the advancement of intelligent systems and computing methodologies. Big Data, with an h-index of 8 and a TC of 199 since 2013, addresses critical challenges and opportunities associated with large-scale data processing and analysis. Furthermore, Big Data and Society provides a platform for exploring the societal implications of big data technologies and practices, boasting an h-index of 8 and a TC of 1649 since 2014.

Future Generation Computer Systems contributes significantly to the advancement of computer systems and technologies, with an h-index of 8 and a TC of 327 since 2015. The Journal of Big Data serves as a prominent outlet for research in the field, with an h-index of 8 and a TC of 2048 since 2015. Lastly, the Journal of Medical Internet Research focuses on the intersection of medical research and internet technologies within the realm of big data analytics, holding an h-index of 8 and a TC of 235 since 2019.

Table 3. Top ten most influential sources

Source	h_index	g_index	m_index	TC	NP	PY_start
IEEE Access	12	26	1.333	909	26	2016
Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	11	19	0.846	479	76	2012
Sustainability (Switzerland)	11	23	1.571	564	23	2018
ACM International Conference Proceeding Series	9	18	0.692	392	63	2012
Advances in Intelligent Systems and Computing	8	14	0.8	265	119	2015
Big Data	8	14	0.667	199	17	2013
Big Data and Society	8	11	0.727	1649	11	2014
Future Generation Computer Systems	8	9	0.8	327	9	2015
Journal of Big Data	8	12	0.8	2048	12	2015
Journal of Medical Internet Research	8	10	1.333	235	10	2019

Recognizing the influence of these top ten sources provides valuable insights into the landscape of research dissemination and scholarly discourse within the field. Researchers can utilize these platforms to access high-impact publications, stay abreast of emerging trends, and contribute meaningfully to the advancement of knowledge in data science and big data analytics.

Understanding the prominence of these sources enables researchers to identify suitable venues for publishing their work, thereby enhancing visibility and impact within the academic community. By acknowledging the significance of influential sources, stakeholders can make informed decisions regarding resource allocation, collaboration opportunities, and research dissemination strategies. Table 3 shows the top 10 most influential sources.

4.6. Keyword Co-Occurrence Network

The Keyword Co-occurrence Network provides valuable insights into the interconnectedness of key terms in data science and big data analytics, with each term residing within a specific cluster and exhibiting varying degrees of link weight and link strength.

Among the top 10 influential keywords within this network, several pivotal concepts stand out. “Big Data,” occupying Cluster 6, emerges as a cornerstone of research, boasting a substantial link weight (125) and link strength (2107), indicative of its widespread citation and prevalence in scholarly discourse.

Similarly, “Data Science,” found in Cluster 7, commands attention with a significant number of links (120) and a high link weight (1137), suggesting its central role in academic literature and its increasing prominence in the field.

“Machine Learning,” clustered within Cluster 5, is another key focal point, with a substantial link weight (111) and link strength (956). This term has garnered significant attention across scholarly works and remains a central theme in data science research.

“Big Data Analytics” and “Data Analytics,” both residing within Cluster 1, underscore the importance of analytical techniques in extracting insights from vast datasets. These terms exhibit high link weights (112, 101) and link strengths (595, 652), indicating their pivotal roles in scholarly discourse.

“Data Mining,” clustered within Cluster 7, and “Analytics,” found in the same cluster, further exemplify the breadth and depth of research in data science. These terms have been pivotal in scholarly discourse and are highly cited, as reflected in their significant link weights (88, 101) and link strengths (504, 652).

Moreover, “Artificial Intelligence,” “Cloud Computing,” and “Deep Learning,” clustered within Clusters 5, 4, and 5, respectively, each contribute significantly to the network. These terms highlight the intersection of cutting-edge technologies and data science research, with each term exhibiting high link weights (76, 60, 71) and link strengths (414, 226, 273), indicative of their influence and relevance in scholarly literature.

In summary, the top 10 influential keywords in the Keyword Co-occurrence Network provide valuable insights into the prevailing themes and trends in data science and big data analytics. Through a visual representation of this network, researchers can glean a deeper understanding of the intricate relationships and interdependencies among these critical concepts. Figure 5 below shows the keyword co-occurrence network.

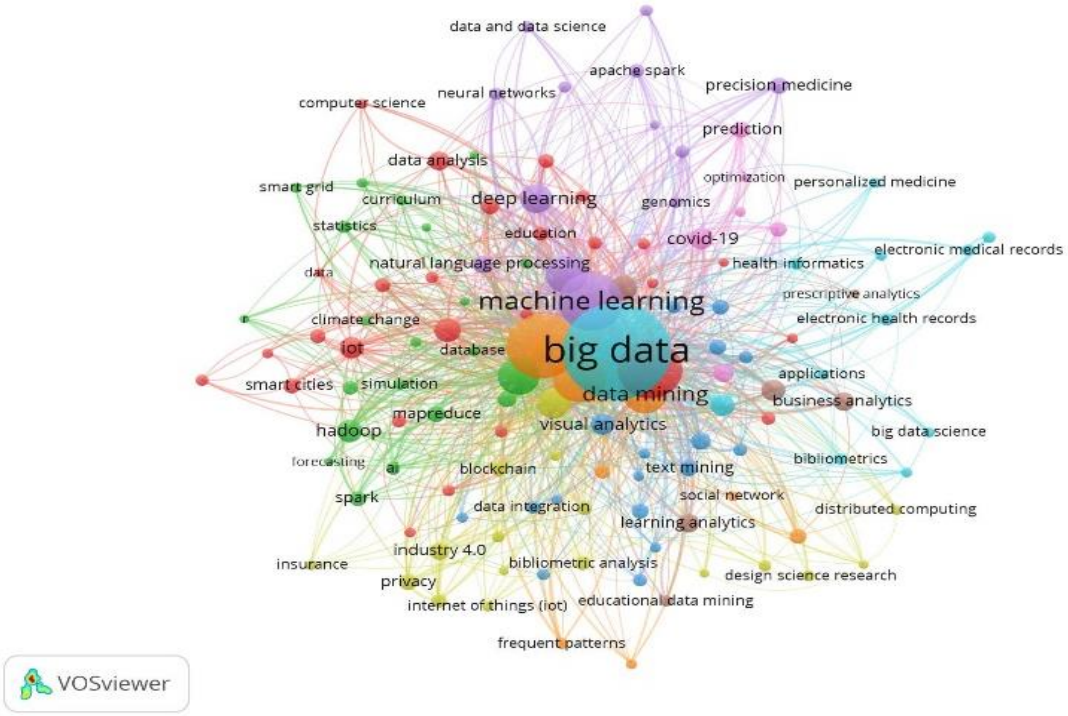

Fig. 5 Keyword co-occurrence network

4.7. Influential Affiliations

The influential affiliations in the field of data science and big data analytics are essential indicators of academic productivity and research impact. The top 10 affiliations, based on the number of publications, offer valuable insights into the institutions driving innovation and scholarly discourse in this domain.

Leading the list is the University of Manitoba, with a significant number of publications (162), reflecting its prominent role in advancing research in data science and related fields. Following closely is the L V Prasad Eye Institute, with 51 publications, underscoring its contributions to interdisciplinary research at the intersection of healthcare and data analytics.

The University of Michigan and Stanford University occupy prominent positions on the list, with 44 and 39 publications, respectively, highlighting their commitment to cutting-edge research and education in data science. These institutions serve as hubs of innovation, fostering collaboration and driving progress in the field.

Notably, several affiliations, such as the University of Toronto, the University of California, and the University of Minnesota, demonstrate a strong presence in the academic landscape, with a considerable number of publications (27, 26, and 26, respectively). These institutions play pivotal roles in shaping the direction of research and training the next generation of data scientists and analysts.

Other notable affiliations include George Mason University and the University of South Carolina, each with 25 and 22 publications, respectively. These institutions contribute significantly to the body of knowledge in data science, enriching scholarly discourse and advancing the frontiers of research.

In summary, the influential affiliations highlighted in this section underscore the diverse and collaborative nature of research in data science and big data analytics. Through their collective efforts, these institutions drive innovation, foster interdisciplinary collaboration, and contribute to the advancement of knowledge in this rapidly evolving field. The below Figure 6 shows the top ten most influential affiliations.

The bibliometric analysis conducted in this study sheds light on the landscape of research in data science and big data analytics, providing valuable insights into trends, patterns, and dynamics shaping the field. By synthesizing findings from various aspects such as document types, publication and citation trends, contributing countries, influential authors, sources, keyword co-occurrence networks, and influential affiliations, we offer actionable insights for researchers, practitioners, and policymakers.

Our analysis reveals a diverse array of document types contributing to the scholarly discourse, with articles and conference papers emerging as primary avenues for disseminating research findings.

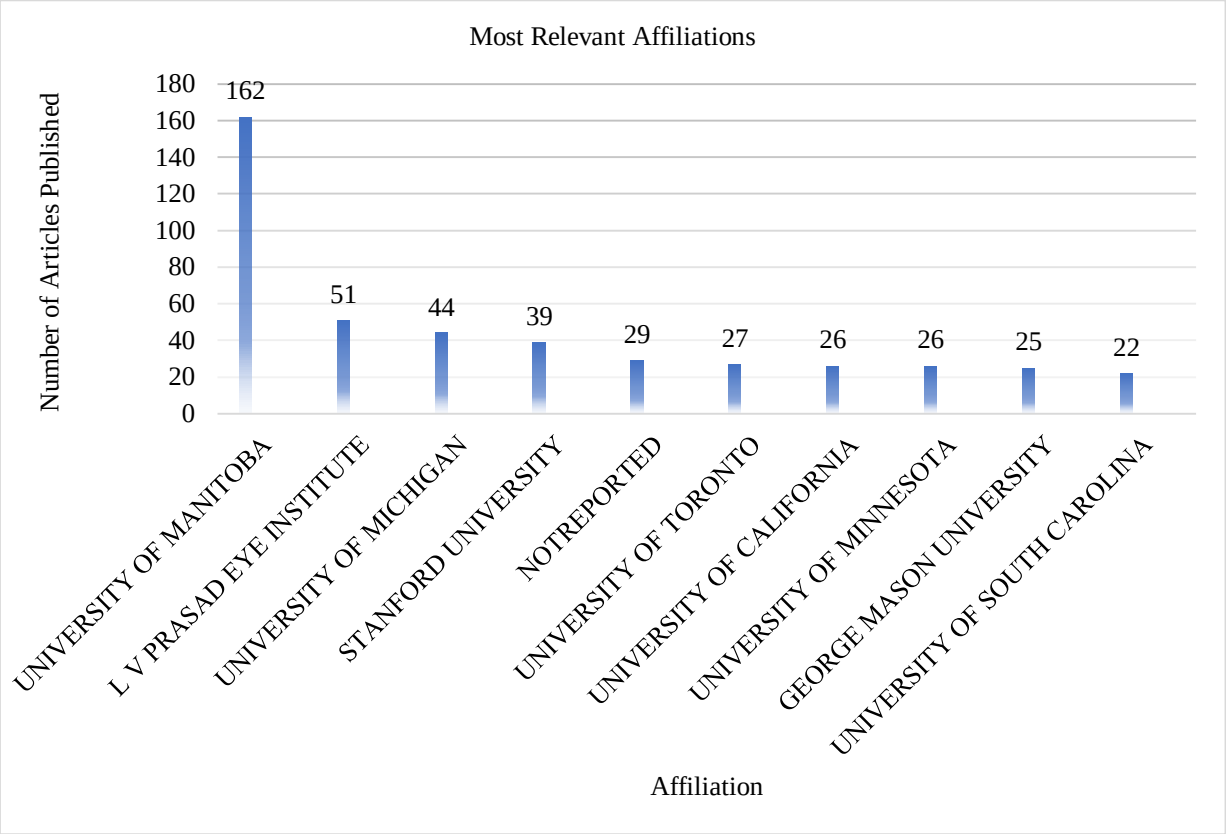

Fig. 6 Top ten influential affiliations

The notable growth trajectory in publication output over the past decade reflects increasing interest and engagement in the field, driven by advancements in technology, methodologies, and applications. Concurrently, the corresponding increase in citation counts underscores the growing impact and influence of research contributions.

The analysis of contributing countries highlights a global landscape of research activity and collaboration, with leading contributors such as the United States, the United Kingdom, China, Germany, and Canada demonstrating substantial involvement in advancing the field. Collaborations across borders enable the exchange of knowledge, expertise, and resources, driving innovation and advancement in data science and big data analytics.

Identification of influential authors provides valuable insights into key contributors shaping the field, with scholars such as Wang J, Leung CK, and Cuzzocrea A making significant contributions. Recognizing the impact of individual scholars facilitates knowledge exchange, collaboration, and advancement within the field. Analysis of influential sources highlights key platforms for disseminating knowledge and shaping research trends. Platforms such as IEEE Access, Lecture Notes in Computer Science, and Sustainability play crucial roles in advancing research agendas and fostering scholarly discourse.

The Keyword Co-occurrence Network provides insights into prevailing themes and trends, with concepts such as "Big Data," "Data Science," and "Machine Learning" emerging as pivotal themes. By visualizing the interconnectedness of key terms, researchers can gain a deeper understanding of critical concepts, facilitating knowledge synthesis and trend identification.

Identification of influential affiliations highlights institutions driving innovation and scholarly discourse in data science and big data analytics. Institutions such as the University of Manitoba, L V Prasad Eye Institute, and the University of Michigan demonstrate significant contributions to research output, fostering interdisciplinary collaboration and contributing to knowledge advancement.

In conclusion, our bibliometric analysis offers valuable insights for navigating the landscape of research in data science and big data analytics. By leveraging these insights, stakeholders can make informed decisions, identify emerging research directions, and contribute meaningfully to the advancement of knowledge and practice in the field.

5. Limitations of the Study

While our bibliometric analysis provides valuable insights into the landscape of research within the field of

Data Science and Big Data Analytics, it is important to acknowledge several limitations inherent to the study.

Firstly, the study solely relies on data extracted from the Scopus database. While Scopus is a comprehensive source covering a wide range of scholarly literature, it does not encompass all publications in the field. Other databases, such as Web of Science and Google Scholar, may offer different coverage and may include additional relevant publications that were not captured in our analysis. Secondly, the study is limited by its timeframe, covering publications from 2010 to March 2024. While this timeframe provides a decade-long perspective on research trends, it may not capture the most recent developments in the field, especially considering the rapidly evolving nature of Data Science and Big Data Analytics.

Additionally, our analysis focuses on quantitative metrics such as publication counts, citation counts, and author metrics. While these metrics offer valuable insights into the scholarly impact and visibility of research outputs, they do not capture qualitative aspects such as the novelty, significance, or methodological rigor of individual studies. Furthermore, the study's analysis is based on predefined keywords and search criteria, which may introduce bias and overlook relevant publications that do not align precisely with the selected terms. The choice of keywords and search strategy may influence the scope and breadth of the analysis, potentially omitting niche or emerging research areas within Data Science and Big Data Analytics. Lastly, our analysis predominantly focuses on English-language publications, potentially excluding non-English research outputs and limiting the representation of global perspectives within the field. This language bias may overlook valuable contributions from non-English-speaking regions and researchers, thereby limiting the comprehensiveness and inclusivity of the study.

Despite these limitations, our bibliometric analysis offers a comprehensive overview of research trends, influential authors, publication outlets, and keyword co-occurrence networks within the field of Data Science and Big Data Analytics. By acknowledging these limitations, future studies can build upon our findings and address these gaps to provide a more nuanced understanding of this dynamic and interdisciplinary field.

6. Conclusion and Future Research Directions

In summary, our bibliometric analysis provides valuable insights into the landscape of research within the field of Data Science and Big Data Analytics. Through an examination of document types, publication and citation trends, contributing countries, influential authors, influential sources, keyword co-occurrence networks, and influential affiliations, we have identified key themes, trends, and contributors shaping the scholarly discourse in this rapidly evolving field.

Our analysis reveals a diverse array of document types contributing to the scholarly output within Data Science and Big Data Analytics, including articles, conference papers, book chapters, reviews, and books. Furthermore, we observe a notable growth trajectory in publication output and citation counts over the past decade, underscoring the increasing interest and impact of research within the field. Moreover, our analysis highlights the global nature of research in Data Science and Big Data Analytics, with contributions from a wide range of countries and institutions. The collaborative and interconnected nature of research is evident as researchers from different regions collaborate to address complex challenges and advance knowledge within the field. Additionally, we identify influential authors, sources, and affiliations driving innovation and scholarly discourse within Data Science and Big Data Analytics. These key contributors play pivotal roles in shaping research trends, fostering interdisciplinary collaboration, and driving progress in the field.

Looking ahead, several avenues for future research emerge from our analysis. Firstly, there is a need for continued exploration of emerging research topics and methodologies within Data Science and Big Data Analytics, particularly in response to advancements in technology and changes in societal needs. Secondly, future research could focus on addressing the limitations identified in our study, such as expanding the scope of databases used for analysis, extending the timeframe of the study to capture more recent developments, and incorporating qualitative assessments of research quality and impact. Furthermore, there is a need for research that explores the ethical, social, and legal implications of data-driven technologies, as well as research that examines strategies for addressing challenges related to data privacy, security, and bias. Moreover, future research could explore the application of novel methodologies, such as machine learning algorithms and natural language processing techniques, to analyze and extract insights from large-scale scholarly literature datasets.

In conclusion, our study provides a comprehensive overview of research trends and contributors within the field of Data Science and Big Data Analytics. By identifying key themes, trends, and future research directions, our analysis serves as a foundation for advancing knowledge and driving innovation in this dynamic and interdisciplinary field.

Acknowledgement

We are deeply grateful to SIMAD University for their generous financial support towards our research paper. Their contribution has played a pivotal role in enabling the realization of this project, providing essential resources and opportunities for its successful completion. The funding

provided by SIMAD University has not only alleviated financial constraints but has also empowered me to pursue rigorous academic inquiry and produce high-quality scholarly work. I extend my heartfelt appreciation to SIMAD University for its unwavering commitment to fostering research excellence and for its instrumental role in advancing knowledge and innovation in the academic community.

References

[1] Allison S. Theobold, "Human Centered Data Science: Ungrading in an Introductory Data Science Course," *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education*, Turku, Finland, pp. 327-333, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Blend Berisha, Endrit Mëziu, and Isak Shabani, "Big Data Analytics in Cloud Computing: An Overview," *Journal of Cloud Computing*, vol. 11, no. 1, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Saba Ameer et al., "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," *IEEE Access*, vol. 7, pp. 128325-128338, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Iqbal H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Computer Science*, vol. 2, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ananta Charan Ojha, and Subhendu Kumar Pani, *Data Science and Big Data Analytics*, 1st ed., Big Data Analytics and Computing for Digital Forensic Investigations, CRC Press, pp. 1-22, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Vahid Ashrafimoghari, "Big Data and Education: Using Big Data Analytics in Language Learning," *arXiv*, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yasset Perez-Riverol, and Pablo Moreno, "Scalable Data Analysis in Proteomics and Metabolomics Using BioContainers and Workflows Engines," *Proteomics and Systems Biology*, vol. 20, no. 9, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shraddha Surana, Yogesh Wadadekar, and Divya Oberoi, "Machine Learning for Scientific Discovery," *arXiv*, pp. 1-4, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] S. Pasari, P. Asudani, and A. Mehta, "Machine Learning Algorithms in Geostatistical Data Analysis: Formulation and Observation," *International Conference on Advances in Earth and Environmental Studies*, NIT Raipur, India, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Leonidas Fegaras, "Compile-Time Query Optimization for Big Data Analytics," *Open Journal of Big Data*, vol. 5, no. 1, pp. 35-61, 2019. [Google Scholar] [Publisher Link]

[11] Tewodrose Tilahun, and Solomon Tsegaye, "The Security Challenges of Big Data Analytics: A Systematic Literature Review," *Asian Journal of Research in Computer Science*, vol. 14, no. 4, pp. 184-197, 2022. [CrossRef] [Publisher Link]

[12] Chris Brunsdon, and Alexis Comber, "Big Issues for Big Data: Challenges for Critical Spatial Data Analytics," *Journal of Spatial Information Science*, no. 21, pp. 89-98, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Shutao Wang et al., "Hot Topics and Frontier Evolution of Science Education Research: A Bibliometric Mapping from 2001 to 2020," *Science & Education*, vol. 32, no. 3, pp. 845-869, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Chien-Heng Chou, Sa Ly Ngo, and Phung Phi Tran, "Renewable Energy Integration for Sustainable Economic Growth: Insights and Challenges via Bibliometric Analysis," *Sustainability*, vol. 15, no. 20, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Maria Drogkoula, Konstantinos Kokkinos, and Nicholas Samaras, "A Comprehensive Survey of Machine Learning Methodologies with Emphasis in Water Resources Management," *Applied Sciences*, vol. 13, no. 22, pp. 1-44, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Robin Chin Roemer, and Rachel Borchardt, *Meaningful Metrics, A 21st Century Librarian's Guide to Bibliometrics, Altmetrics, and Research Impact*, Association of College and Research Libraries, a Division of the American Library Association, pp. 1-241, 2015. [Google Scholar] [Publisher Link]

[17] Kaihua Chen, Yi Zhang, and Xiaolan Fu, "International Research Collaboration: An Emerging Domain of Innovation Studies?," *Research Policy*, vol. 48, no. 1, pp. 149-168, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ankit Verma, and Hansraj, "Big Data Analytics Using Machine Learning Techniques for Prediction on Datasets," *Computational Intelligence and Machine Learning*, vol. 4, no. 1, pp. 6-10, 2023. [CrossRef] [Publisher Link]

[19] Rahul Reddy Nadikattu, "Research on Data Science, Data Analytics and Big Data," *International Journal of Engineering, Science and Mathematics*, vol. 9, no. 5, pp. 99-105, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Matthew A. Waller, and Stanley E. Fawcett, "Data Science, Predictive Analytics, and Big Data: A Revolution that will Transform Supply Chain Design and Management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77-84, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[21] Radwa Elshawi et al., "Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service," *Big Data Research*, vol. 14, pp. 1-11, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[22] Daphne R. Raban, and Avishag Gordon, "The Evolution of Data Science and Big Data Research: A Bibliometric Analysis," *Scientometrics*, vol. 122, pp. 1563-1581, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Hamed Jahani, Richa Jain, and Dmitry Ivanov, "Data Science and Big Data Analytics: A Systematic Review of Methodologies Used in the Supply Chain and Logistics Research," *Annals of Operations Research*, pp. 1-58, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Justin Zuopeng Zhang et al., "Big Data Analytics and Machine Learning: A Retrospective Overview and Bibliometric Analysis," *Expert Systems with Applications*, vol. 184, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Massimo Aria, and Corrado Cuccurullo, "Bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis," *Journal of Informetrics*, vol. 11, no. 4, pp. 959-975, 2017. [CrossRef] [Google Scholar] [Publisher Link]