The Development in Data Science due to the Integration of Blockchain Technology

AKSHAT GAURAV[1],DOMENICO SANTANIELLO[2]
[1]1Ronin Institute, Montclair, USA (e-mail: akshat.gaurav@ieee.org)
[2]University of Salerno, Italy. (e-mail: dsantaniello@unisa.it)

ABSTRACT The exponential rise of data is a consequence of the evolution of digital technology and the use of smart devices. Also as a result of this, the data science field has grown rapidly in recent years. There are, however, several problems that need fixing. The privacy, security, and decentralized nature of blockchain technology have earned it widespread acclaim. Due to blockchain technology, the way we get and distribute information will shift radically. Researchers think that blockchain technology may help alleviate some of the challenges faced by data scientists and boost the growth of the field. In this context, we analyze the development in the field of data science due to the integration of blockchain technology. We used the Scopus database to collect the relevant database.

KEYWORDS Data Science; Big data; Blockchain

I. INTRODUCTION

Information is becoming more diverse. As the volume of data continues to grow, it becomes more difficult to parse the many components. This calls for novel approaches to data analysis. Because of this increase in data's diversity and volume, the field of data science has advanced rapidly. Moreover, information security and privacy are major issues that must be addressed [1], [2]. Information security and protection throughout the Data Science industry's many value chains—from data collection to data production to data analysis to data sharing—faces a serious threat. Due to its decentralized nature and high level of security, Blockchain is uniquely suited to this position. Many obstacles in the field of data science may be overcome with the help of blockchain technology. data science allows us to extract critical data and snip out unnecessary product details [3]. However, issues related to privacy, security, data sharing, uneven distribution, and unequal demand are creating friction in the development of Data Science. The simplicity, safety, adaptability, and protection offered by Blockchain have direct analogs in the field of data science.

II. LITERATURE SURVEY

Author in [4] presents a Browser-Side Context-Aware Sanitization of Suspicious HTML5 Code for Halting the DOM-Based XSS Vulnerabilities in Cloud. Author in [5] proposed a lightweight mutual authentication protocol based on elliptic curve cryptography for IoT devices. Author in [6] proposed a secure Timestamp-Based Mutual Authentication Protocol for IoT Devices Using RFID Tags. Author in [7] proposed a novel framework for risk assessment and resilience of critical infrastructure towards climate change. Author in [8] proposed a secure and energy efficient-based E-health care framework for green internet of things. Author in [9] proposed a novel coverless information hiding method based on the average pixel value of the sub-images. Author in [10] proposed a secure Machine Learning scenario from Big Data in Cloud Computing via Internet of Things network. Author in [11] presets a review on advances in security and privacy of multimedia big data in mobile and cloud computing. Author in [12] proposed myocardial infarction detection based on deep neural network on imbalanced data. Author in [13] proposed a reputation score policy and Bayesian game theory based incentivized mechanism for DDoS attacks mitigation and cyber defense. Author in [14] proposed a DDoS Attacks and Defense Mechanisms in Various Web-Enabled Computing Platforms: Issues, Challenges, and Future Research Directions. Author in [15] proposed a context Aware Recommender Systems. Author in [16] proposed a cross-lingual transfer method and distributed MinIE algorithm on apache spark. Author in [17] proposed a smart defense against distributed Denial of service attack in IoT networks using supervised learning classifiers. Author in [18] proposed an adaptive Feature Selection and Construction for Day-Ahead Load Forecasting using Deep Learning Method. Author in [19] proposed and analysis of artificial intelligence-based technologies and approaches on sustainable entrepreneurship. Author in [20] proposed a digital Watermarking-Based Cryptosystem for Cloud Resource Provisioning.

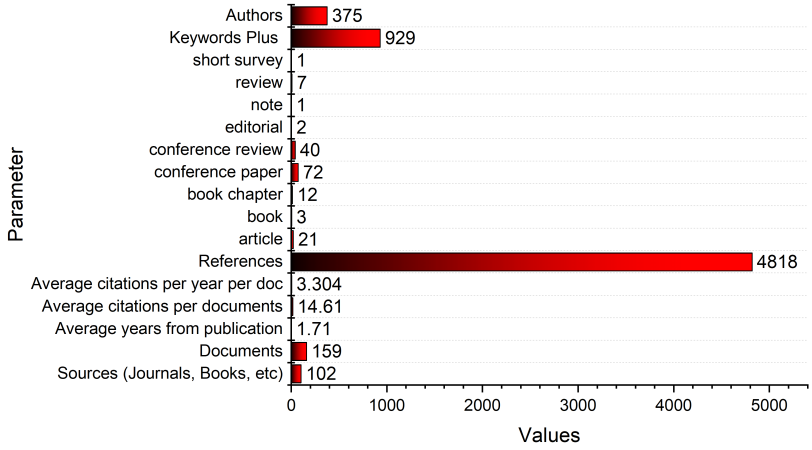

FIGURE 1: General Information

III. RESEARCH METHODOLOGY

In this article, we analyze the development in the field of data science due to the integration of blockchain technology. We search the Scopus database using the following query:

TITLE-ABS-KEY (blockchain AND "data science")

The above-defined query extracts all the articles that have "blockchain" and "data science" in their *title, abstract, or keywords*.

IV. RESULTS AND DISCUSSION

In this subsection, we analyze all the papers that are extracted from the Scopus database. The collected articles are from different sources, as represented in Figure 2. However, the majority of papers are published at international conferences. Also, as represented in Figure 3, out of the total researchers that are working in to develop new theories of blockchain for the development of the data science domain are from computer science. The collected papers have 375 unique authors and 4818 authors.

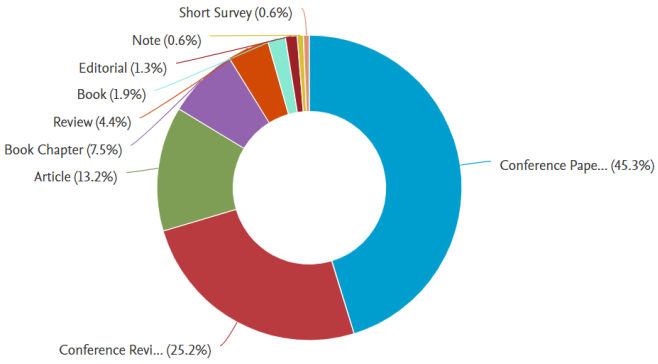

FIGURE 2: different Data Types

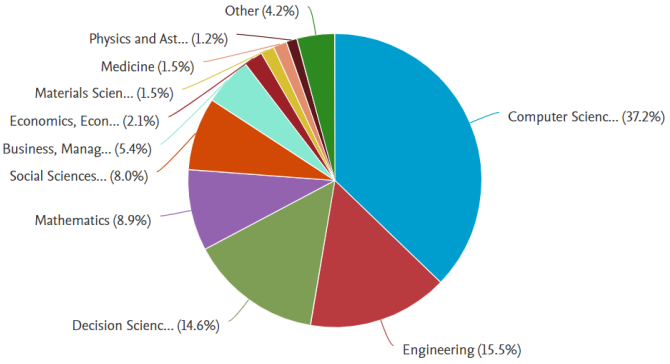

FIGURE 3: Data Types

A. ANALYSIS DISTRIBUTION OF COUNTRIES

in this subsection, we analyze the distribution of published articles according to their geographical location. Figure 4 presents the distribution of papers according to the countries. From Figure 4 it is clear that most productive countries are :

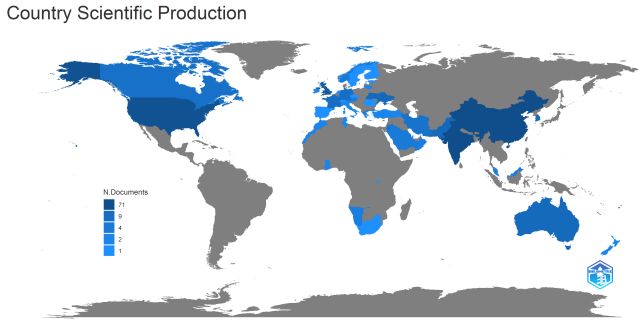

FIGURE 4: Country Scientific Production

- INDIA (71)
- CHINA (70)
- USA (55)
- UK (31)

- PAKISTAN (12)
- UKRAINE (12)
- AUSTRALIA (11)
- SOUTH KOREA (10)

B. KEYWORD ANALYSIS

Apart from the distribution of papers according to their geographical locations, the distribution of keywords is also a good factor to analysis the variation of the research topic. Figure 5 presents the distribution of keywors. In the Figure 5, as the frequency of the occurrence of the keyword increases, its size increases in the Figure 5. From Figure 5 it is clear that the following are the frequently occurred keywords:

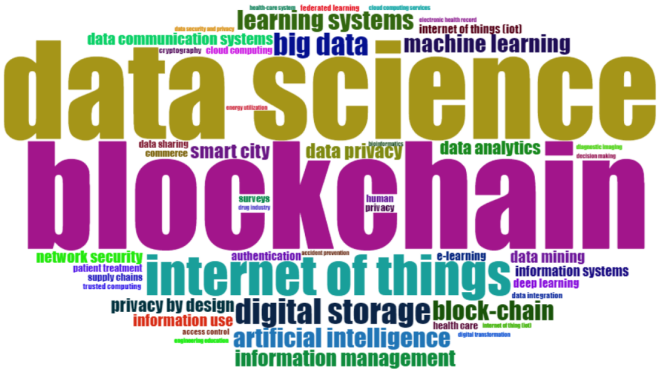

FIGURE 5: Keyword Analysis

- blockchain (71)
- data science (65)
- internet of things (27)
- digital storage (17)
- big data (15)
- artificial intelligence (13)
- block-chain (13)
- learning systems (13)
- machine learning (12)
- information management (11)

C. DOCUMENT DISTRIBUTION

This subsection describes the scientific distribution of the papers. We compiled a total of 159 works, including articles from journals, conference proceedings, book reviews, and book/chapter publications that are indexed in Scopus. The most-cited publications in the subject of data science provide a broad overview of the major themes and theoretical frameworks that have been developed by the academic community. Table 1 presents the distribution of the paper according to the total number of citations.

V. CONCLUSION

The Internet of Things, digital transaction records, and supply chain management are all areas where blockchain technology might be useful. Businesses are increasingly focused on the pioneering efforts of blockchain technology. As a result of blockchain technology, consumers may conduct transactions directly with one another, cutting out the middleman. There is no need for middlemen when transferring, storing, executing, supervising, or trading anything of value, from hard currency to Hollywood blockbusters. In its current form, data science is the study of how to use both organized and unstructured data to answer important questions and address pressing issues. Blockchain technology with data science may make for a powerful combination, with many opportunities and risks, to efficiently manage large amounts

TABLE 1: Highly Cited Papers

Paper	Total Citations
AZARIA A, 2016, PROC - INT CONF OPEN BIG DATA, OBD [21]	1243
KLERKX L, 2019, NJAS WAGENINGEN J LIFE SCI [22]	321
YU B, 2018, IEEE CLOUD COMPUT [23]	94
MIN S, 2019, J BUS LOGIST [24]	91
ENGIN Z, 2019, COMPUT J [25]	67
SWAN M, 2018, ADV COMPUT [26]	40
FREITAG C, 2021, PATTERNS [27]	38
SHUBINA V, 2020, DATA [28]	35
CHUNG S-H, 2021, TRANSP RES PART E LOGIST TRANSP REV [29]	28
GUPTA H, 2021, IND MANAGE DATA SYS [30]	28
DOKU R, 2019, PROC - IEEE INT CONF INF REUSE INTEGR DATA SCI, IRI [31]	21
KOMNINOS N, 2019, SMART CITIES IN THE POST-ALGORITHMIC ERA: INTEGRATING TECHNOLOGIES, PLATFORMS AND GOV-a [32]	17
LAWRENZ S, 2019, ACM INT CONF PROC SER	16
CAO L, 2021, INT J DATA SCI ANAL [33]	12
JOHNSON N, 2020, PROC - ADV COMPUT COMMUN TECHNOL HIGH PERFORM APPL, ACCTHPA [34]	11
KOMNINOS N, 2019, SMART CITIES IN THE POST-ALGORITHMIC ERA: INTEGRATING TECHNOLOGIES, PLATFORMS AND GOV [32]	11
BELHADI A, 2021, AD HOC NETW [35]	11
MIKROYANNIDIS A, 2019, PROC FRONT EDUC CONF FIE [36]	10
CAMINO R, 2020, IEEE INT CONF BLOCKCHAIN CRYPTOCURRENCY, ICBC [37]	10
JESSE N, 2018, IFAC-PAPERSONLINE [38]	10

of data without sacrificing quality. in this context, we analyze the development in the field of data

REFERENCES

[1] A. Gaurav, V. Arya, and D. Santaniello, "Analysis of machine learning based ddos attack detection techniques in software defined network," Cyber Security Insights Magazine (CSIM), vol. 1, no. 1, pp. 1–6, 2022.

[2] R. K. S. Rajput, D. Goyal, A. Pant, G. Sharma, V. Arya, and M. K. Rafsanjani, "Cloud data centre energy utilization estimation: Simulation and modelling with idr," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–16, 2022.

[3] K. Pathoee, D. Rawat, A. Mishra, V. Arya, M. K. Rafsanjani, and A. K. Gupta, "A cloud-based predictive model for the detection of breast cancer," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–12, 2022.

[4] B. B. Gupta, S. Gupta, and P. Chaudhary, "Enhancing the browser-side context-aware sanitization of suspicious html5 code for halting the dom-based xss vulnerabilities in cloud," International Journal of Cloud Applications and Computing (IJCAC), vol. 7, no. 1, pp. 1–31, 2017.

[5] A. Tewari and et al., "A lightweight mutual authentication protocol based on elliptic curve cryptography for iot devices," International Journal of Advanced Intelligence Paradigms, vol. 9, no. 2-3, pp. 111–121, 2017.

[6] A. Tewari and et al., "Secure timestamp-based mutual authentication protocol for iot devices using rfid tags," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 16, no. 3, pp. 20–34, 2020.

[7] N. Kumar and et al., "A novel framework for risk assessment and resilience of critical infrastructure towards climate change," Technological Forecasting and Social Change, vol. 165, p. 120532, 2021.

[8] M. Kaur and et al., "Secure and energy efficient-based e-health care framework for green internet of things," IEEE Transactions on Green Communications and Networking, vol. 5, no. 3, pp. 1223–1231, 2021.

[9] L. Zou and et al., "A novel coverless information hiding method based on the average pixel value of the sub-images," Multimedia tools and applications, vol. 78, no. 7, pp. 7965–7980, 2019.

[10] C. L. Stergiou and et al., "Secure machine learning scenario from big data in cloud computing via internet of things network," in Handbook of computer networks and cyber security. Springer, 2020, pp. 525–554.

[11] B. B. Gupta, S. Yamaguchi, and D. P. Agrawal, "Advances in security and privacy of multimedia big data in mobile and cloud computing," Multimedia Tools and Applications, vol. 77, no. 7, pp. 9203–9208, 2018.

[12] M. Hammad and et al., "Myocardial infarction detection based on deep neural network on imbalanced data," Multimedia Systems, vol. 28, no. 4, pp. 1373–1385, 2022.

[13] A. Dahiya and et al., "A reputation score policy and bayesian game theory based incentivized mechanism for ddos attacks mitigation and cyber defense," Future Generation Computer Systems, vol. 117, pp. 193–204, 2021.

[14] A. Singh and et al., "Distributed denial-of-service (ddos) attacks and defense mechanisms in various web-enabled computing platforms: Issues, challenges, and future research directions," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 18, no. 1, pp. 1–43, 2022.

[15] M. Casillo and et al., "Context aware recommender systems: A novel approach based on matrix factorization and contextual bias," Electronics, vol. 11, no. 7, p. 1003, 2022.

[16] P. Do and et al., "Building a knowledge graph by using cross-lingual transfer method and distributed minie algorithm on apache spark," Neural Computing and Applications, pp. 1–17, 2020.

[17] B. Gupta, P. Chaudhary, X. Chang, and N. Nedjah, "Smart defense against distributed denial of service attack in iot networks using supervised learning classifiers," Computers & Electrical Engineering, vol. 98, p. 107726, 2022.

[18] R. Jiao and et al., "Adaptive feature selection and construction for day-ahead load forecasting use deep learning method," IEEE Transactions on Network and Service Management, vol. 18, no. 4, pp. 4019–4029, 2021.

[19] B. B. Gupta, A. Gaurav, P. K. Panigrahi, and V. Arya, "Analysis of artificial intelligence-based technologies and approaches on sustainable entrepreneurship," Technological Forecasting and Social Change, vol. 186, p. 122152, 2023.

[20] S. Kumar, S. Kumar, N. Ranjan, S. Tiwari, T. R. Kumar, D. Goyal, G. Sharma, V. Arya, and M. K. Rafsanjani, "Digital watermarking-based cryptosystem for cloud resource provisioning," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–20, 2022.

[21] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," 2016, pp. 25–30.

[22] L. Klerkx, E. Jakku, and P. Labarthe, "A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda," NJAS - Wageningen Journal of Life Sciences, vol. 90-91, 2019.

[23] B. Yu, J. Wright, S. Nepal, L. Zhu, J. Liu, and R. Ranjan, "Trust chain: Establishing trust in the iot-based applications ecosystem using blockchain," IEEE Cloud Computing, vol. 5, no. 4, pp. 12–23, 2018.

[24] S. Min, Z. Zacharia, and C. Smith, "Defining supply chain management: In the past, present, and future," Journal of Business Logistics, vol. 40, no. 1, pp. 44–55, 2019.

[25] Z. Engin and P. Treleaven, "Algorithmic government: Automating public services and supporting civil servants in using data science technologies," Computer Journal, vol. 62, no. 3, pp. 448–460, 2019.

[26] M. Swan, "Blockchain for business: Next-generation enterprise artificial intelligence systems," Advances in Computers, vol. 111, pp. 121–162, 2018.

[27] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. Blair, and A. Friday, "The real climate and transformative impact of ict: A critique of estimates, trends, and regulations," Patterns, vol. 2, no. 9, 2021.

[28] V. Shubina, S. Holcer, M. Gould, and E. Lohan, "Survey of decentralized solutions with mobile devices for user location tracking, proximity detection, and contact tracing in the covid-19 era," Data, vol. 5, no. 4, pp. 1–40, 2020.

[29] S.-H. Chung, "Applications of smart technologies in logistics and transport: A review," Transportation Research Part E: Logistics and Transportation Review, vol. 153, 2021.

[30] H. Gupta, S. Kumar, S. Kusi-Sarpong, C. Jabbour, and M. Agyemang, "Enablers to supply chain performance on the basis of digitization technologies," Industrial Management and Data Systems, vol. 121, no. 9, pp. 1915–1938, 2021.

[31] R. Doku, D. Rawat, and C. Liu, "Towards federated learning approach to determine data relevance in big data," 2019, pp. 184–192.

[32] N. Komninos, A. Panori, and C. Kakderi, Smart cities beyond algorithmic logic: digital platforms, user engagement and data science, 2019.

[33] L. Cao, Q. Yang, and P. Yu, "Data science and ai in fintech: an overview," International Journal of Data Science and Analytics, vol. 12, no. 2, pp. 81–99, 2021.

[34] N. Johnson, M. Santosh Kumar, and T. Dhannia, "A study on the significance of smart iot sensors and data science in digital agriculture," 2020, pp. 80–88.

[35] A. Belhadi, Y. Djenouri, G. Srivastava, A. Jolfaei, and J.-W. Lin, "Privacy reinforcement learning for faults detection in the smart grid," Ad Hoc Networks, vol. 119, 2021.

[36] A. Mikroyannidis, J. Domingue, M. Bachler, and K. Quick, "Smart blockchain badges for data science education," vol. 2018-October, 2019.

[37] R. Camino, C. Torres, M. Baden, and R. State, "A data science approach for detecting honeypots in ethereum," 2020.

[38] N. Jesse, "Organizational evolution - how digital disruption enforces organizational agility," vol. 51, no. 30, 2018, pp. 486–491.