West Science Interdisciplinary Studies
Vol. 01, No. 07, July and 2023: pp. 387-397

A Bibliometric Analysis of Data Science: Trends, Contributions, and Research Developments

Muhamad Juliardi[1], Ibnu Malik[2]
[1] Universitas Budi Luhur: muhamadjuliardi07@gmail.com
[2] Universitas Budi Luhur: ibnu_malik@outlook.co.id

ABSTRACT

This research paper presents an analysis of the collected articles using the Publish or Perish software, consisting of the top 290 scientific articles listed on Google Scholar from 1947 to 2023. The study aims to explore the current state of the data science field in relation to technological advancements and the prevalence of big data. The findings reveal a significant and rapid development in the field, highlighting the importance of topics such as data science, big data, data analysis, and machine learning for further investigation.

Keywords: *Data Science, Big Data, Data Analysis, Bibliometric Analysis, Publish or Perish, Mendeley, Vosviewer*

1. INTRODUCTION

In the era of the Internet of Things and Big Data, data scientists are required to extract valuable knowledge from the given data[1]. Data is becoming the most valuable asset for any organization and might be its only truly inimitable asset[2]. In the current era, Data Science has emerged as a crucial field. During this era, a significant amount of data is generated and collected across various domains. Data Science aids researchers and practitioners in extracting valuable insights and making data-driven decisions. Therefore, comprehensive literature analysis is necessary to understand trends, contributions, and research developments in Data Science.

The growth in the quantity and diversity of data has led to data sets larger than is manageable by the conventional, hands-on management tools[3]. The accessibility to large datasets enables the application of complex algorithms and data science (DS) tools[4]. In this sense, DS tools, such as machine learning (ML), have the potential to support several fields of research, such as biomedicine, neuro- science or robotics, by the automation or resolution of complex tasks in time series prediction, classification, regression, diagnostics, monitoring, and so on[5].

Data are widely considered to be a driver of better decision making and improved profitability, and this perception has some data to back it up[3]. The common theme in descriptions of the job of the data scientist is a kind of beginning-to-end narrative, whereby data scientists have a hand in many, if not all, aspects of a process that involves data. The only aspects in which they are not involved are the choice of the question itself and the decision that is ultimately made upon seeing the results. In fact, based on our experience, many real-world situations draw the data scientist into participating in those activities as well[6].

This study not only provides an overview of the current state of Data Science but also serves as a guide for future research directions. By identifying emerging research areas and existing literature gaps, this analysis can inform researchers and decision-makers about potential opportunities and challenges in the field. Ultimately, the findings of this study will contribute to the advancement of Data Science by revealing key trends, influential contributions, and research developments that shape its evolution.

The aim of this paper is to address the following questions: (1) How are articles on data science classified? (2) What are the research trends in the field of data science? (3) Which research

topics have received more publications? (4) Which data science topics offer opportunities for further research in the future? The paper begins with a literature review on the term "data science" based on previous research findings. Additionally, the research objectives are presented in Section 1. Section 2 explains the definition of data science and provides an overview of related terms. The methodology used to conduct bibliometric analysis, including the steps and methods associated with the use of databases, is described in Section 3. Section 4 presents the results of the analysis using the VOSviewer tool. Research recommendations, conclusions, and study limitations are discussed in Section 5.

1. Data Science

Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information. Although the name Data Science seems to connect most strongly with areas such as databases and computer science, many different kinds of skills - including non-mathematical skills - are needed.

The complexity of data gives rise to new perspectives that were previously impossible or difficult to achieve. For example, large-scale traditional sensor data surveys have proven to be less effective due to associated challenges such as irrelevant participants, low response rates, and unanswered questions. However, data-driven discoveries can aid in determining whom to survey, what questions need to be answered, designing implementable survey operational models, and measuring the cost-effectiveness of conducting surveys[7].

Table 1. 1Several bibliometric analyses that have been done by previous researchers on the topic of data science

Authors(s) & Year	Number of Documens Anayzed	Source	Findings
[8]	1727	Scopus database	The study aims to fill the research gap on big data analytics in enterprises. However, the findings are limited by the sample and analysis software used. The sample consisted of English journal articles from the Scopus database. Future research can overcome these limitations by exploring different databases and including documents in multiple languages. It's important to note that the VOSviewer software doesn't differentiate author contributions. Despite these limitations, the study provides structure to the fragmented literature on big data analytics by conducting an early bibliometric study. It analyzes thematic areas and proposes future research agendas. The study's protocol for bibliometric studies can be a valuable resource for future researchers.

[9]	4081	WoS Database	It has been observed that significant works in this field are being conducted worldwide, including both developed and developing countries. China is emerging as the leading contributor in terms of the total number of publications, while Singapore has the highest per-capita publication rate. Furthermore, there is a growing linkage between different types of publications, particularly between engineering journals and industrial journals. This indicates that these techniques are gaining increasing industrial importance. In conclusion, data-based process monitoring is rapidly developing and being implemented in process industries. However, the pace of application in process industries is not keeping up with the pace of theoretical development.
[10]	478	WoS Database	The analysis of document co-citation and its evaluation reveals the presence of four distinct clusters that connect Big Data analytics with various management phenomena. These clusters include the theoretical development of Big Data analytics, the transition of management towards Big Data analytics, the relationship between Big Data analytics and firm resources, capabilities, and performance, and the application of Big Data analytics in supply chain management.
[11]	7868	Scopus database	Utilizing bibliometric and network analysis, we conducted a comprehensive literature review on the topic of Big Data and supply chain management (SCM) spanning a 10-year period from 2006 to 2016. Our study aimed to shed light on the role of scientific journals in advancing research on Big Data and the contributions of individual researchers to this emerging field. As far as we are aware, this is the first study to identify the top contributing authors, countries, and key research topics in this area. Despite its limitations, we believe that our study offers valuable insights and encourages further exploration of the field of Big Data and SCM by researchers.

2. METHODOLOGY: A BIBLIOMETRIC ANALYSIS

The purpose of this paper is to answer the questions of how articles on data science classified, what are the research trends in the field of data science, what research topics are the subject of more publications, what data science topics will provide opportunities for further research in the future.

2.1. Search for specific journals on the topic of data science

A Bibliometric review, commonly used in scientific fields, involves a quantitative analysis of journal articles, books, or other written communications[12]. The first step of the process is to conduct a Google database search for articles that particularly discuss data science.

2.2. Journal metrics information

This section explicitly describes the profiles and metrics of the selected journals. Table 2 presents important information obtained from the one chosen journal. These metric details were extracted from metadata using the Publish or Perish (PoP) application on June 11, 2023.

Table 2. Metrics information of selected journals

Metrics data	
Publication years	1947-2023
Citation years	76
Papers	290
Citations	997526
Cites/year	13125.34
Cites/paper	3439.74
Authors/paper	2.78
h-index	270
g-index	290
hI,norm	248
hI,annual	3.26

2.3. Reference management

After the articles have been downloaded from the journal website, the next step is to tidy up the references using the Mendeley application. References are necessary to ensure that the metadata for each article is complete, such as information about the author, keywords, abstracts, and other details.

2.4. Bibliometric analysis

The subsequent step involves performing a bibliometric analysis. The software employed for conducting bibliometric analysis in this study is VosViewer.

3. RESULTS AND DISCUSSION

To address the initial objective of this paper regarding the classification of data science articles, the VosViewer software was employed to create a map based on textual data extracted from the title and abstract fields. By utilizing the full counting method, a total of 5216 terms were identified. With a minimum number of occurrences of a term of 10 times, 95 thresholds were found. However, for each of the 95 terms, a relevance score will be calculated. Based on this score, the most relevant terms will be selected automatically by default as much as 60%, so we get the 57 most appropriate words. However, the verification process still has to be done manually by eliminating unrelated words, such as editorial, sample, abstract, and others. Thus, the total words that can be included in making a map are 100 words.

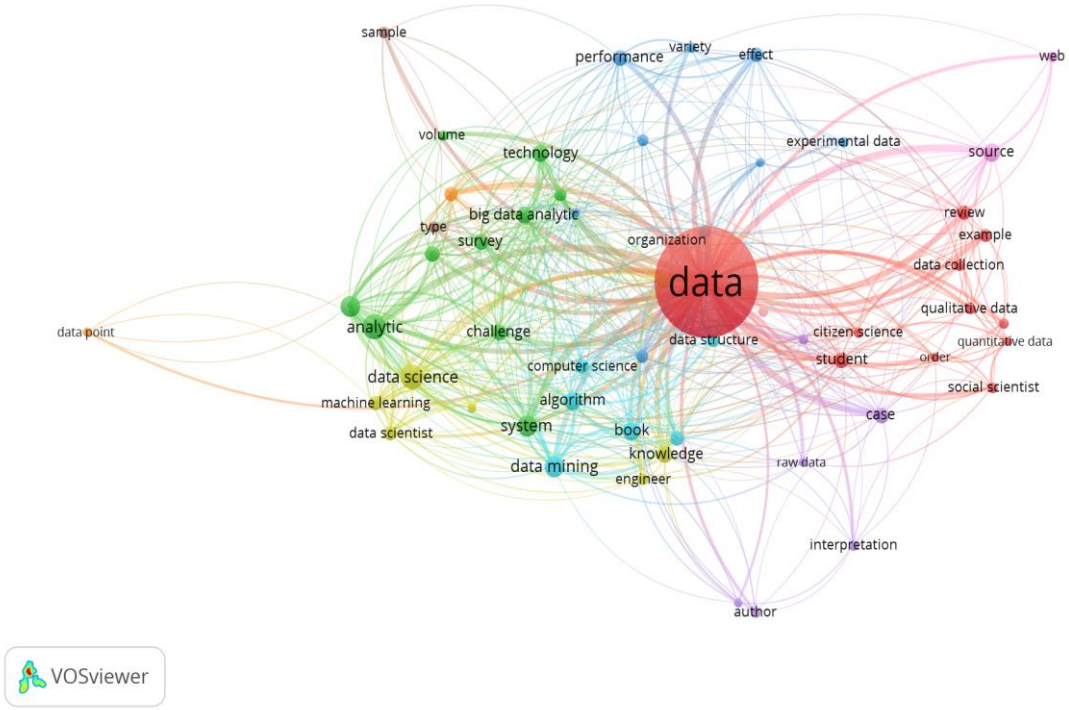

Figure 1. Network visualization map of keywords

Figure 1 illustrates multiple clusters, each represented by a distinct color. The clusters consist of frequently occurring words within the articles, indicating the presence of eleven distinct classifications of published articles. For further information, refer to Table 3.

Table *3.* Clusters and keywords therein

Cluster	Total items	Most frequent keyword (occurrences)	Keyword
1	11	data (1370), student (30), review (24)	citizen science, data, data collection, example, order, political science, qualitative data, quantitative data, review, social scientist, student

2	10	analytic (72), big data (49), big data analytic (35)	analytic, big data, big data analytic, challenge, data analytic, framework, survey, system, technology, volume
3	9	effect (26), engineering (20), text (20)	Business, data development analysis, effect, engineering, experimental data, life science, performance, text, variety
4	7	data science (69), data scientist (26), machine learning (24)	data science, data scientist, engineer, information, knowledge, machine learning, predictive analytic
5	6	case (30), form (15), interpretation (14)	author, behavioral science, case, form, interpretation, raw data
6	6	data mining (54), book (42), algorithm (38)	algorithm, area, book, computer science, data mining, data structure
7	2	data point (12), value (27)	data point, value
8	2	sample (16), type (16)	sample, type
9	2	source (42), web (14)	source, web
10	2	nature (13), organization (22)	nature, organization

To answer the trend of data science research, we can actually see the answer from the cluster itself. Figure 2 shows the density visualization of published articles. Cluster 1, with the word "data" appearing most frequently, indicates the research trend in the field of data, particularly data science.

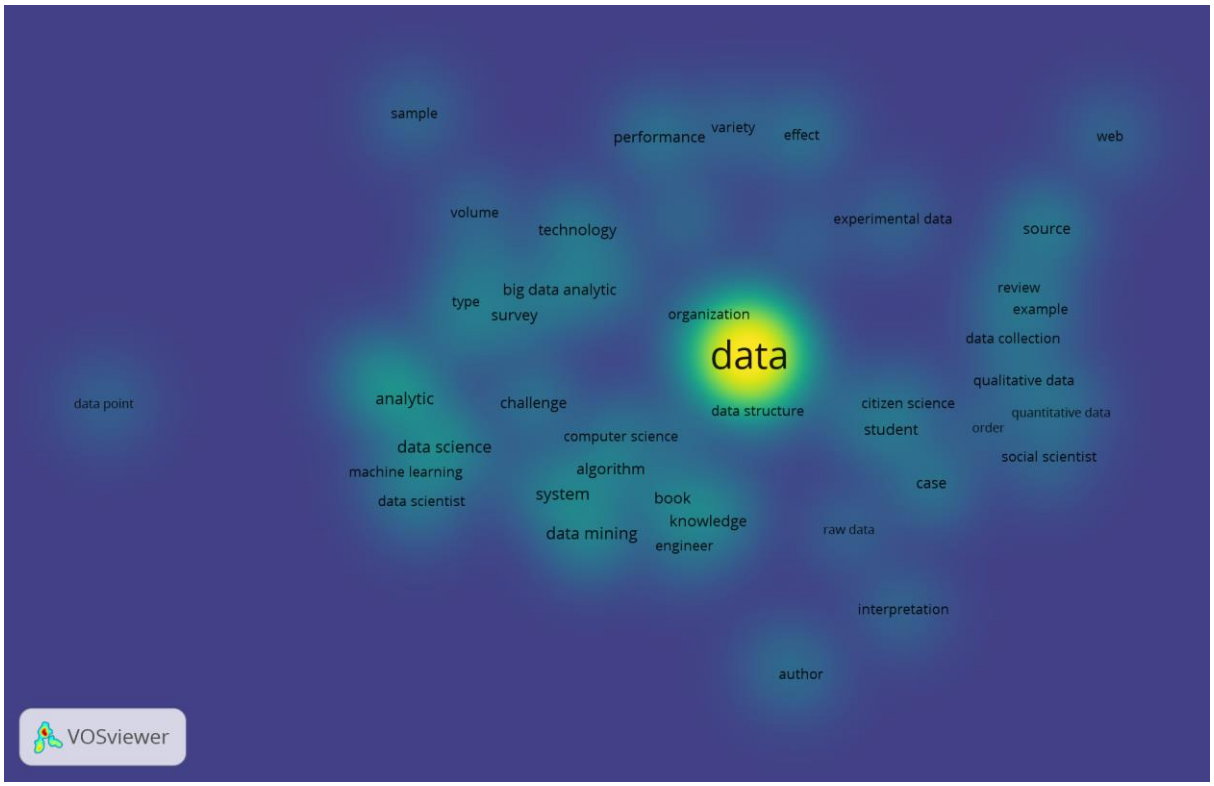

Figure 2. Density visualization map of keywords

There is one cluster that appears in the mapping results, at least in the keywords, namely cluster 4. This cluster covers topics related to data science and machine learning. Additionally, in each cluster, there are some keywords that rarely appear, such as web and experimental data. This indicates that there are still research gaps that are highly likely to become trends in the future, adapted to current and future world conditions. From the researcher's perspective, there are also ten clusters, as presented in Figure 3.

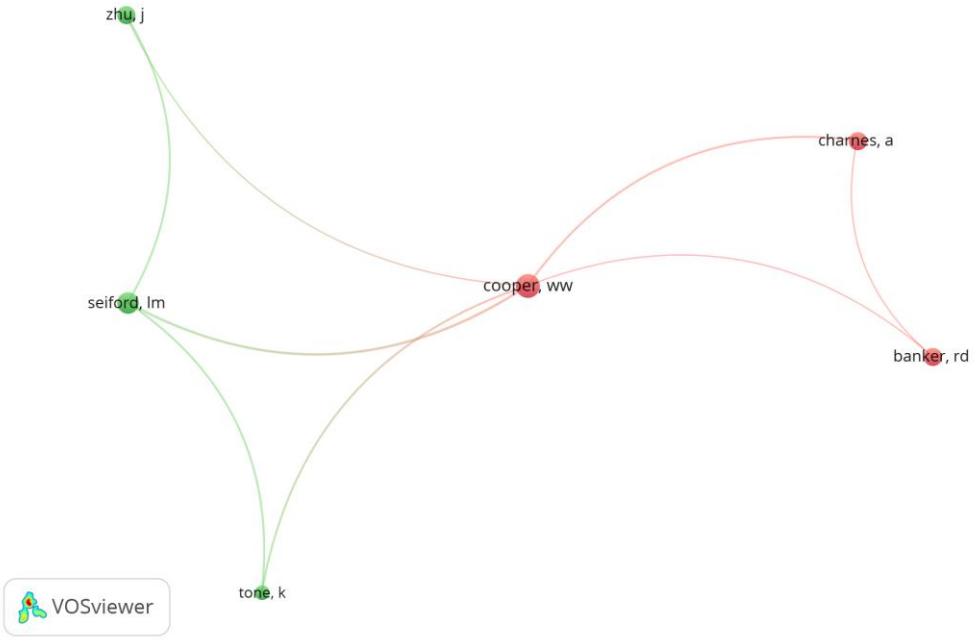

Figure 3. Network visualization map of authors

From Figure 3, it can be observed that there are five prominent figures from each cluster marked with large dots in each cluster. Only authors who have a connection in their publications are shown in the picture. In Table 4, information is presented regarding the most cited documents, along with other detailed elements, which were calculated on June 11, 2023.

Table 4. The top ten cited documents

Citations	Authors and year	Title
484	[13]	Data science: A comprehensive overview
210	[14]	Big data and data science: what should we teach?
169	[7]	Data science: Challenges and directions
85	[15]	Data science: supporting decision-making
55	[16]	Data science vs. statistics: two cultures?
54	[17]	Data science for business: benefits, challenges and opportunities
51	[18]	Agile big data analytics: AnalyticsOps for data science
47	[19]	A review of data science in business and industry and a future view
45	[20]	Skills Requirements of Business Data Analytics and Data Science Jobs: A Comparative Analysis

40	[21]	A review of artificial intelligence methods for data science and data analytics: Applications and research challenges

There is a noticeable trend of significant increase in the number of citations in many previous papers for documents on big data and data science. However, more recent materials tend to receive fewer citations, except from authors who have previously conducted research in this field and have a good reputation. Furthermore, to identify the research topics that have been the subject of more publications, we can refer to Table 5.

Table 5. The 15 Most and Fewer Occurrences Terms

Most occurrences		Fewer occurrences	
Occurrences	Term	occurrences	Term
1370	Data	10	Predictive Analysis
72	Analytic	11	Life Science
69	Data science	11	Business
54	Data Minning	12	Quantitative data
52	System	12	Order
49	Big Data	12	Data point
42	Source	12	Behavioral science
42	Book	12	Raw data
38	Algorithm	13	Author
35	Big data analytic	13	Political science
35	Knowledge	13	Nature
34	Technology	14	Web
30	Case	14	Interpretation
30	Student	14	Experimental data
29	Information	15	Social scientist

4. CONCLUSIONS

The articles were collected using the Publish or Perish software, which includes the top 290 scientific articles listed on Google Scholar from 1947 to 2023. In the context of this research, we conclude that the field of data science is experiencing rapid development in parallel with technological advancements and the growth of big data in the present era. Based on the findings of this study, there are also specific issues that can be explored in more detail. Topics such as data science, big data, data analysis, and machine learning provide opportunities for further research.

This study has at least two limitations. Firstly, it is based on articles listed on Google Scholar. Secondly, despite the use of formal tools such as PoP software, VOSviewer, and Mendeley, subjective assessments by the authors still exist and may introduce errors. Future studies should utilize a larger

sample size by involving other journals. It is recommended for future research to focus on more specific and reliable sources, such as the Scopus index, to generate more diverse bibliometric maps.

REFERENCES

[1] J. M. Alonso, C. Castiello, and C. Mencar, *A bibliometric analysis of the explainable artificial intelligence research field*, vol. 853. Springer International Publishing, 2018. doi: 10.1007/978-3-319-91473-2_1.

[2] O. B. Nielsen, "A Comprehensive Review of Data Governance Literature," *Issue Nr*, vol. 8, no. 8, p. 3, 2017, [Online]. Available: http://aisel.aisnet.org/iris2017http://aisel.aisnet.org/iris2017/3

[3] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management," *J. Bus. Logist.*, vol. 34, no. 2, pp. 77–84, 2013, doi: 10.1111/jbl.12010.

[4] S. Barik, M. Moharana, S. Bhutia, N. Tripathy, and A. MOHAN, "Advances in data science, trends, and applications of artificial intelligence within the interaction between natural and artificial computation," *Neurocomputing*, vol. 12, no. 06, pp. 314–343, 2022.

[5] J. M. Górriz *et al.*, "Artificial intelligence within the interplay between natural and artificial Computation: advances in data science, trends and applications," *Neurocomputing*, 2020, doi: 10.1016/j.neucom.2020.05.078.

[6] R. D. Peng and H. S. Parker, "Perspective on Data Science," pp. 1–20, 2022.

[7] L. Cao, "Data science: Challenges and directions," *Commun. ACM*, vol. 60, no. 8, pp. 59–68, 2017, doi: 10.1145/3015456.

[8] S. Khanra, A. Dhir, and M. Mäntymäki, "Big data analytics and enterprises : a bibliometric synthesis of the literature," *Enterp. Inf. Syst.*, vol. 00, no. 00, pp. 1–32, 2020, doi: 10.1080/17517575.2020.1734241.

[9] F. Khan, S. A. Imtiaz, and S. Ahmed, "Review A bibliometric review and analysis of data-driven fault detection and diagnosis methods for process systems," 2018, doi: 10.1021/acs.iecr.8b00936.

[10] L. Ardito, V. Scuotto, M. Del Giudice, A. Messeni, and L. Ardito, "A bibliometric analysis of research on Big Data analytics for business and management analytics," 2018, doi: 10.1108/MD-07-2018-0754.

[11] D. Mishra, A. Gunasekaran, T. Papadopoulos, and S. J. Childe, "Big Data and supply chain management : a review and bibliometric analysis," *Ann. Oper. Res.*, 2016, doi: 10.1007/s10479-016-2236-y.

[12] R. Heersmink and J. Van Den Hoven, "Bibliometric mapping of computer and information ethics," pp. 241–249, 2011, doi: 10.1007/s10676-011-9273-7.

[13] L. Cao, "Data science: A comprehensive overview," *ACM Comput. Surv.*, vol. 50, no. 3, 2017, doi: 10.1145/3076253.

[14] I. Y. Song and Y. Zhu, "Big data and data science: what should we teach?," *Expert Syst.*, vol. 33, no. 4, pp. 364–373, 2016, doi: 10.1111/exsy.12130.

[15] D. J. Power, "Data science: supporting decision-making," *J. Decis. Syst.*, vol. 25, no. 4, pp. 345–356, 2016, doi: 10.1080/12460125.2016.1171610.

[16] I. Carmichael and J. S. Marron, "Data science vs. statistics: two cultures?," *Japanese J. Stat. Data Sci.*, vol. 1, no. 1, pp. 117–138, 2018, doi: 10.1007/s42081-018-0009-3.

[17] M. M. de Medeiros, N. Hoppen, and A. C. G. Maçada, "Data science for business: benefits, challenges and opportunities," *Bottom Line*, vol. 33, no. 2, pp. 149–163, 2020, doi: 10.1108/BL-12-2019-0132.

[18] N. W. Grady, J. A. Payne, and H. Parker, "Agile big data analytics: AnalyticsOps for data science," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-January, pp. 2331–2339,

2017, doi: 10.1109/BigData.2017.8258187.

[19] G. Vicario and S. Coleman, "A review of data science in business and industry and a future view," *Appl. Stoch. Model. Bus. Ind.*, vol. 36, no. 1, pp. 6–18, 2020, doi: 10.1002/asmb.2488.

[20] Z. Radovilsky, V. Hegde, A. Acharya, and U. Uma, "Skills Requirements of Business Data Analytics and Data Science Jobs: A Comparative Analysis," *Comp. Anal. J. Supply Chain Oper. Manag.*, vol. 16, no. 1, pp. 1–20, 2018, [Online]. Available: https://www.csupom.com/uploads/1/1/4/8/114895679/v16n1p5.pdf

[21] C. V. Krishna, H. R. Rohit, and Mohana, "A review of artificial intelligence methods for data science and data analytics: Applications and research challenges," *Proc. Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2018*, pp. 591–594, 2019, doi: 10.1109/I-SMAC.2018.8653670.