IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 June 2023, accepted 26 June 2023, date of publication 28 June 2023, date of current version 13 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3290488

RESEARCH ARTICLE

Analyzing the Scholarly Literature of Digital Twin Research: Trends, Topics and Structure

FRANK EMMERT-STREIB[1], SHAILESH TRIPATHI[1,2], AND MATTHIAS DEHMER[1,3,4,5]

[1]Predictive Society and Data Analytics Laboratory, Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland
[2]Production and Operations Management, University of Applied Sciences Upper Austria, 4400 Steyr, Austria
[3]Department for Biomedical Computer Science and Mechatronics, UMIT–Private University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, 6060 Tyrol, Austria
[4]Department of Computer Science, Swiss Distance University of Applied Sciences, 3900 Brig, Switzerland
[5]College of Artificial Intelligence, Nankai University, Tianjin 300071, China

Corresponding author: Frank Emmert-Streib (v@bio-complexity.com)

ABSTRACT Currently, studies involving a digital twin are gaining widespread interest. While the first fields adopting such a concept were in manufacturing and engineering, lately, interest extends also beyond these fields across all academic disciplines. Given the inviting idea behind a digital twin which allows the efficient exploitation and utilization of simulations such a trend is understandable. The purpose of this paper is to use a scientometrics approach to study the early publication history of the digital twin across academia. Our analysis is based on large-scale bibliographic and citation data from Scopus that provides authoritative information about high-quality publications in essentially all fields of science, engineering and humanities. This paper has four major objectives. First, we obtain a global overview of all publications related to a digital twin across all major subject areas. This analysis provides insights into the structure of the entire publication corpus. Second, we investigate the co-occurrence of subject areas appearing together on publications. This reveals interdisciplinary relations of the publications and identifies the most collaborative fields. Third, we conduct a trend and keyword analysis to gain insights into the evolution of the concept and the importance of keywords. Fourth, based on results from topic modeling using a Latent Dirichlet Allocation (LDA) model we introduce the definition of a scientometric dimension (SD) of digital twin research that allows to summarize an important aspect of the bound diversity of the academic literature.

INDEX TERMS Data science, digital twin, scientometrics, natural language processing.

I. INTRODUCTION

In recent years, there is a tremendous interest in the concept of a digital twin as exemplified by the ever increasing number of publications. The idea of a digital twin can be stated simply as follows: A digital twin is a digital representation of a real-world object that is essentially indistinguishable of its real-world counterpart. Here digital representation means that a digital twin is a software implementation and a real-world object could be either a system or process that has a physical representation, e.g., an engine, a biological cell or a manufactural process. For a collection of further but similar definitions see [11] and for a specific and detailed definition of a biological digital twin see [15].

David Gelernter's book Mirror Worlds [16], published in 1991, is widely credited with introducing the broad idea behind a digital twin. However, it was Michael Grieves who outlined the concept and model of a digital twin in more technical terms in the early 2000s [18]. One of the first practical applications of a digital twin can be found in a study of aircraft structure conducted by [41]. Overall, it is widely believed that digital twins will play a pivotal role in the fourth industrial revolution, as they have the potential to significantly improve the operational efficiency in manufacturing and production [1], [42].

Since its beginnings, the idea of a digital twin inspired many studies, especially in engineering and manufacturing.

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

As a consequence, in recent years more and more papers are published seemingly across all academic disciplines. In order to obtain a better understanding of this diverse corpus of publications and to get insights into trends and used methodologies, we conduct a scientometric analysis. In general, scientometrics studies the scientific activity of a field or topic using quantitative methods, with the aim of providing insights into the structure, dynamics, and impact of scientific research [22], [26]. Scientometrics involves analyzing large-scale bibliographic and citation data to identify patterns and trends in scientific communication and collaboration, e.g., as documented by publications. Sometimes, one distinguishes between scientometrics and infometrics, however, the differences are subtle and there are no generally accepted distinctions [19]. For this reason, in this paper, we will refer to our study as scientometric analysis of publications focusing on a digital twin.

For our analysis, we use data from Scopus - a citation database provided by Elsevier - that provides authoritative information about high-quality publications in essentially all fields of science, engineering and humanities. In total, we obtain over 6000 publications from Scopus allowing a comprehensive scientometric analysis. The objective of this paper is threefold. First, we gain a global overview of all publications related to the digital twin across all main subject areas of academic disciplines. This includes information about the number of citations, citations per publication, evolution of publications and citations per publication type. Importantly, these analyses will be based on either subject areas or publication types - we distinguish between 27 subject areas spanning all academic disciplines and 8 paper types - allowing insights into the structure of the entire publication corpus. Second, we investigate the co-occurrence of subject areas on publications. This allows us to gain insights into the interdisciplinary relations of the publications and identify dominating fields. Third, we conduct a keyword analysis that allows us to derive subject area-specific constituents of a ''digital twin'' and to identify trends. Finally, we perform a Latent Dirichlet Allocation (LDA) [5] analysis for topic modeling. By analyzing the frequency and co-occurrence of keywords across subject areas, LDA allows us to extract topics and provides insights about the most important themes used in the field of digital twin research. Based on this, we will introduce the definition of a scientometric dimension of digital twin research.

So far, a number of papers appeared providing also a scientometric or bibliometric analysis of a digital twin. For instance, the paper by [46] conducted a bibliometric analysis of 514 articles related to a digital twin as found in the Web of Science (WoS) database. They studied, e.g., core journals, institutions, countries and a theme map. In [47], 1158 publications from Web of Science and 745 records from the Derwent Patent Database were analyzed with a focus on the construction industry. Similarly narrow studies were provided by [29] analyzing 197 journal articles in architectural, engineering, construction, operation, and facility management (AECO-FM) industry, [20] analyzed 77 publications about architecture, engineering, construction, and facility management (AEC-FM) industry, [35] studied 817 journal papers from the Smart City, Engineering and Construction (SCEC) sectors and [27] investigated 276 publications about smart manufactoring. In contrast to these scientometric studies, our paper is different with respect to the following aspects. First, we provide an extensive analysis by using over 6000 publications from the Scopus database. Second, we do not narrowly focus on selected subject areas but study 27 main subject areas across all academic disciplines. Third, our analysis is not limited to a descriptive analysis but we conduct also an inferential analysis, e.g., by utilizing a Latent Dirichlet Allocation (LDA) model. As a consequence this allows us to derive a subject-specific, data-driven interpretations of digital twin research as present in the literature.

This paper is organized as follows. In the next section, we introduce our data and statistical methods we use for our analysis. Then we present the results of our descriptive and scientometric analysis of publication data from Scopus. Based on these findings, we provide a discussion of their meaning and finish with a summary and a conclusion.

II. METHODS

In this section, we describe the data and methods we need for our analysis. We start with a description of the data and then we provide information about the used analysis methods.

A. DEFINITION OF MAIN SUBJECT AREAS

For our analysis, we use a categorization of academic disciplines. Specifically, we use the authorative definition of such a categorization provided by Scopus. In Table 1, we show an overview of these categories consisting of 27 main fields spanning all areas of science, engineering and humanities. This allows us to get a comprehensive overview of the entire literature utilizing a digital twin.

We would like to note that every listed publication in Scopus is assigned to at least one of the these 27 categories. That means publications that are narrowly focused on only one topic will be uniquely labeled by just one category but interdisciplinary studies or broad research work can have more than one category label. In the results section, this will be studied in detail to learn about the interdisciplinarity of the subject.

B. SEARCH AND SELECTION OF RELEVANT PAPERS

In order to include only publications that are closely related to studies about the digital twin, we use the following filtering for querying the Scopus database. Specifically, we limit our search to publications that use the term ''digital twin'' in the title. This is very restrictive and has the advantage that only publications with a dedicated focus on a digital twin are considered. That means publications using the term ''digital twin'' in a broader context of a study will not be considered.

As a result from this query, we find 6314 publications in the Scopus database. These publications can assume one

TABLE 1. Overview of 27 main subject areas provided by the scopus database used for our study.

1.	AGRI	Agricultural and Biological Sciences
2.	ARTS	Arts and Humanities
3.	BIOC	Biochemistry, Genetics and Molecular Biology
4.	BUSI	Business, Management and Accounting
5.	CENG	Chemical Engineering
6.	CHEM	Chemistry
7.	COMP	Computer Science
8.	DECI	Decision Sciences
9.	DENT	Dentistry
10.	EART	Earth and Planetary Sciences
11.	ECON	Economics, Econometrics and Finance
12.	ENER	Energy
13.	ENGI	Engineering
14.	ENVI	Environmental Science
15.	HEAL	Health Professions
16.	IMMU	Immunology and Microbiology
17.	MATE	Materials Science
18.	MATH	Mathematics
19.	MEDI	Medicine
20.	MULT	Multidisciplinary
21.	NEUR	Neuroscience
22.	NURS	Nursing
23.	PHAR	Pharmacology, Toxicology and Pharmaceutics
24.	PHYS	Physics and Astronomy
25.	PSYC	Psychology
26.	SOCI	Social Sciences
27.	VETE	Veterinary

of the following eight paper types: ''Article'', ''Book'', ''Book Chapter'', ''Conference Paper'', ''Editorial'', ''Letter'', ''Review'' and ''Short Survey''. Additional information we gather for each publication are publication year, keywords and number of citations. All of these features will be analyzed in the results section.

For finding the publications, we utilized the ''rscopus'' package in R to retrieve papers relevant to our search keyword ''digital twin''. Our search gives a total of 6314 papers, from which we extracted additional attributes such as subject area, citations, year of publication, subject area label, and keywords using the ''eid'' identifier. The subject area labels cover a broad range of sub-categories across various fields, from which we use the major 27 categories listed in Table 1.

C. TEXT ANALYSIS AND NATURAL LANGUAGE PROCESSING

For analyzing keywords provided by the publications, we conduct a text analysis based on natural language processing (NLP). For this, we use the following steps outlined below.

1) PREPROCESSING

For the preprocessing of the keywords, we follow [44]. Specifically, for preparing the keywords for the analysis we filter them in the following way:

- Remove the term ''digital twin''.
- Remove all keyword containing the term ''twin''.
- Select only keywords with a frequency >= 5.
- Remove keywords with length <= 5.

2) ANALYSIS OF KEYWORDS

In order to identify important keywords we perform two types of analysis. The first studies the frequency of keywords and the second tests their statistical significance.

For the first analysis we select the top 35 keywords of each subject area and exame their proportional representation in different years for digital-twin-related concepts and implementations. Importantly, from a top keyword selection we calculate the probablity of each keyword by

$$p_i(j) = \frac{\text{frequeny of the } i^{th} \text{ keyword in subject area } j}{\text{total frequency of all keywords in subject area } j}, \tag{1}$$

for every subject area, and select the keywords with $p_i(j) > 0.01$.

The second analysis involves identifying significant keywords within a particular subject area. For this analysis, we used the following hypothesis: Suppose we have a set of all subject area $SA = \{sa_1, sa_2, \ldots sa_n\}$ where $|SA|$ gives the total number of subject areas. If the probability of success of a keyword $kw_i \in sa_j$ is greater than $\frac{1}{|SA|}$, then kw_i is significantly represented in sa_j compared to a random appearance. We conducted a hypothesis test to select significant keywords, with the null hypothesis (H_0) stating that the true probability of success is less than or equal to $\frac{1}{|SA|}$, and the alternative hypothesis (H_1) stating that the true probability of success is greater than $\frac{1}{|SA|}$.

$$H_0 : \text{True probablity of success} \leq \frac{1}{|SA|}$$
$$H_1 : \text{True probablity of success} > \frac{1}{|SA|}$$

To test these hypothesis, we apply a Binomial test where the probablity of k successes is estimated from a Binomial distribution. The z-score is calculated as follows:

$$z(kw_i, sa_j) = \frac{(k - np)}{\sqrt{np(1-p)}} \tag{2}$$

where the n is the total occurrence of a keyword (kw_i) irrespective of a subject area and k is the number of successes if it occurs in a particular subject area sa_j, and $p = \frac{1}{|SA|}$. For controling the false discovery rate (FDR), we used the Benjamini-Hochberg (BH) procedure [13] and select keywords for which the FDR is less than a significance level of $\alpha = 0.05$.

3) SUB-SUBJECT AREAS ANALYSIS WITH A BIPARTITE GRAPH

For a publication it is possible to possible to have in addition to subject areas also sub-subject areas. For example subject area engineering (''ENGI'') can be subdivided into *Control and Systems Engineering, Aerospace Engineering, Ocean Engineering, Mechanical Engineering, Electrical and Electronic Engineering*. Simlarly this applies to other subject areas as well. Using such sub-subject areas allows to see

which keywords are related to the fine structure of subject areas.

For this analysis we first construct a bipartite graph as follows: Let document D_i have a set of sub-subject areas $sa^{sub} = \{sa_1^{sub}, sa_2^{sub}, \ldots sa_p^{sub}\}$, and keywords, $kw = \{kw_1, kw_2, \ldots kw_n\}$. We construct a bipartite graph, $G_i = (sa^{sub}, kw, E_{sa^{sub},kw})$, where $E_{sa^{sub},kw} = (sa^{sub}, kw)$ is the connection matrix between sub-subject areas and keywords. The final graph G is constructed by the taking union of all $\{G_1, G_2, \ldots G_m\}$ subgraphs constructed from m documents, $G = G_1 \cup G_2 \cup \cdots \cup G_m$. We then apply *louvian* [6] to the bipartite graph to detect its modules Dugué [12] modularity.

4) STATISTICAL ANALYSIS AND TOPIC MODELING

For the analysis of the data from Scopus, we use various methods. Specifically, for identifying significant correlations between the rank order of different publication statistics, we use Spearman's rank correlation test [36] for a significance level of $\alpha = 0.05$.

For topic modeling, we use a LDA (Latent Dirichlet Allocation) model to discover underlying topics or themes in different subject areas [5], [21]. LDA utilizes a Gibbs sampling technique to estimate the parameters of the model. For the LDA analysis, we assume that each subject area in the collection is a mixture of a small number of topics, and each topic is a probability distribution over keywords in the whole keywords set. The goal of LDA is to discover the underlying topics and their associated probabilities in the collection as well as the distribution of topics in each subject area. This is obtained by maximizing the likelihood of the observed data, given the topic assignments for each word in each subject area, and the topic distributions across the entire matrix of subject areas (rows) and keywords (columns). For the topic modelling using LDA we used Gibbs sampling to estimate the distribution of topics in each subject area, as well as the probability distribution of keywords in each topic.

In our analysis, we first apply LDA to optimize the number of topics for all subject areas. For this, we run a LDA analysis for 2 to 27 topics and calculate density-based metrics for adaptive LDA model selection proposed by Cao et al. [8] and Deveaud et al. [10] which maximize the intra topic and minimize inter topic similarity, metrices for optimal number of topics.

For identifying the optimal number of clusters in a dendrogram, we use the *Ball & Hall* (BH) index [2], [50] that estimates the average distance of instances to their respective cluster centroids. The BH-index is obtained by

$$BH = \frac{1}{C}\sum_{k=1}^{C}\sum_{i\in C_k} \|x_i - c_k\|^2 \tag{3}$$

where C is the number of clusters, c_k are centroids of cluster C_k and x_i are vectors of observations of the ith instance in cluster C_k. The optimal value of BH is the maximum value of the second successive differences.

All parts of the analysis are conducted by using the statistical programming language R [31].

III. RESULTS

In the following, we study the publication data from Scopus. This analysis is subdivided into four parts. First, we provide a global overview of all publications related to the digital twin across all main subject areas of academic disciplines. This includes information about the number of citations, citations per publication, evolution of publications and citations per publication type. Second, we investigate the co-occurrence of subject areas on publications. This allows us to learn about the interdisciplinary relations of the publications. Third, we conduct a trend and keyword analysis for obtaining insights into domain-specific keywords and their usage over time. Fourth, we perform a topic modeling analysis to gain a detailed understanding of the diversity of keywords.

A. GLOBAL OVERVIEW

We start our analysis by providing an overview of publication statistics. In Figure 1 A, we show the total number of publications for each subject area that have more than 10 publications. That means the categories dentistry (DENT), nursing (NURS), psychology (PSYC) and veterinary (VETE) have been removed from the 27 available categories (see Table 1).

From Figure 1 A one can see that engineering (ENGI) and computer science (COMP) have by far the most publications followed by mathematics (MATH). In the next group, we find 12 subject areas which all still have more than 100 publications where the largest four are decision sciences (DECI), energy (ENER), materials science (MATE) and physics and astronomy (PHYS). Finally, there are 8 subject areas with less than 100 publications.

Figure 1 B shows the number of citations for the subject areas. The subject areas with the highest number of citations are the same as for the total number of publications (Figure 1 A) and also the order of the remaining subject areas seems similar. In order to confirm this, we perform Spearman's rank correlation test, r, for the number of publications and the number of citations.

$$r(\text{number of publications, number of citations}) = 0.974$$
$$\text{with p-value} = 4.496e-15$$

As a result, we obtain a correlation of $r = 0.974$ with a p-value of $4.496e-15$ which is significant for a significance level of $\alpha = 0.05$. It is also interesting to note that the subject area with the smallest number of citations (pharmacology, toxicology and pharmaceutics (PHAR)) is also the subject area with the smallest number of publications.

As a measure of importance for publications, we show in Figure 1 C the number of citations per publications which is given by

$$\text{citations per publication} = \frac{\text{number of citations}}{\text{number of publications}}. \tag{4}$$

for each subject area. Here the subject area with the highest number of citations per publication is economics, econometrics and finance (ECON) and the subject area with the smallest number is again pharmacology, toxicology and pharmaceutics (PHAR). Overall, the order of the subject areas in this figure looks considerably different to Figure 1 A and B. To confirm this, we perform Spearman's rank correlation tests for correlation, r, and obtain the following results:

$$r(\text{number of publications, citations per publication}) = 0.204 \quad \text{with p-value} = 0.350$$

and

$$r(\text{number of citations, citations per publication}) = 0.357 \quad \text{with p-value} = 0.094$$

As one can see for a significance level of $\alpha = 0.05$ non of the tests is significant indicating that the observed orders of the subject areas are different from each other.

Next, we take a look at the time course of publications. In Figure 2, we show the number of publications for the subject areas over the years. It is interesting to note that dedicated publications about digital twin started only recently in 2005, however, it took a few more years until 2016 to reach a noticeable number of publications per year. Since then the number of publications grows continuously across all subject areas.

For our next analysis, we focus on the article types of publications. In Figure 3 A and B, we show the number of publications and the number of citations per publication type respectively. The paper types ''conference paper'' and ''article'' are by far the most common publication types. Interestingly, the highest number of citations per publication type is obtained for ''book'' followed by ''review'' despite the fact that the number of publications in these categories is low to moderate.

Figure 3 C shows a heatmap giving the number of publications per subject area and publication type. The highest numbers are observed for engineering (ENGI) and computer science (COMP) for the publication types article and conference paper. It is interesting to note that 6 out of the 7 published books are also in these two subject areas. To identify significant entries in Figure 3 C, we estimate the mean values in the cells expected by chance via a Binomial distribution with $p = 1/23$ and n corresponding to the sum of the columns. From this and a significance level of $\alpha = 0.05$ we estimate the following threshold values for the 8 article types: 279.9 (article), 1.1 (book), 22.4 (book chapter), 317.3 (conference paper), 6.6 (editorial), 1.0 (letter), 26.3 (review) and 2.3 (short survey). In Figure 3 C, we highlight all significant cell entries in yellow.

In Table 3, we show a list of the 10 most cited publications across all subject areas. Most of these publications (6) are articles and 3 conference papers. It is important to highlight that only three publications are categorized by a single subject area (engineering (ENGI)) while all other publications are assigned to two or more subject areas. This will be studied in more detail in the next section. The most diverse publication is a book chapter by [18] assigned to four subject categories (ECON, BUSI, MATH and ENGI).

TABLE 2. Subject specific summary statistics of the number of citations for the top 8 subject areas (ENGI, COMP, MATH, MATE, DECI, ENER, PHYS, SOCI) in Figure 1 B. The columns correspond to the mean value of citations, top 20%, top 5% and top 1% of citations per subject area.

subject areas	mean number of ctiations	top 20%	top 5%	top 1%
ENGI	13.5	11.0	49.9	223.4
COMP	11.6	10.0	45.0	168.6
MATH	7.4	6.6	25.9	95.0
MATE	16.8	11.0	59.0	392.9
DECI	8.3	7.0	33.0	138.9
ENER	7.2	8.0	31.0	88.5
PHYS	8.7	9.0	35.8	83.6
SOCI	9.6	11.2	59.8	99.4

Regarding the distribution of citations it is interesting to note that the mean number of citations of all publications is 11.3. Furthermore, 20% of all publications have more than 10 citations, 5% of all publications have more than 42 citations and 1% of all publications have more than 167 citations. When looking at subject specific numbers for the top 8 subject areas (ENGI, COMP, MATH, MATE, DECI, ENER, PHYS, SOCI) in Figure 1 B, we find the results shown in Table 2. These results are similar to the subject area independent results indicating overall a rapid decay in the number of citations per paper.

In order obtain more detailed insights about the top cited publications in particular application domains, we show in Table 4 two examples. Specifically, the top part shows the five most cited publications in the earth and environmental sciences (EART, EVNI) and the bottom part lists publications in the life sciences (BIOC, HEAL, IMMU, MEDI, PHAR, NEUR, NURS). As one can see from this table, only three publications in the life sciences are within the top 1% of the most cited papers but all of the shown publication are within the top 5%.

B. INTERDISCIPLINARY RELATIONS

For the next analysis part, we focus on interdisciplinary relations between publications. To gain insights into such interdisciplinary relations of publications we conduct two types of analysis. First, we study the distribution of subject areas per publication and, second, we investigate the co-occurrence of such fields.

The first analysis is shown in Figure 4 A. The histogram shows the frequency of publications in dependence on the number of subject areas found per publication. It is interesting to note that the vast majority of publications is connected to more than one subject area, however, the number of publications with only one subject area is still 2042 papers. Specifically, the probability to observe two or more subject areas per publication is given by

$$\Pr(\text{two or more subject areas}) = 0.65 = \frac{4147}{6294}. \quad (5)$$

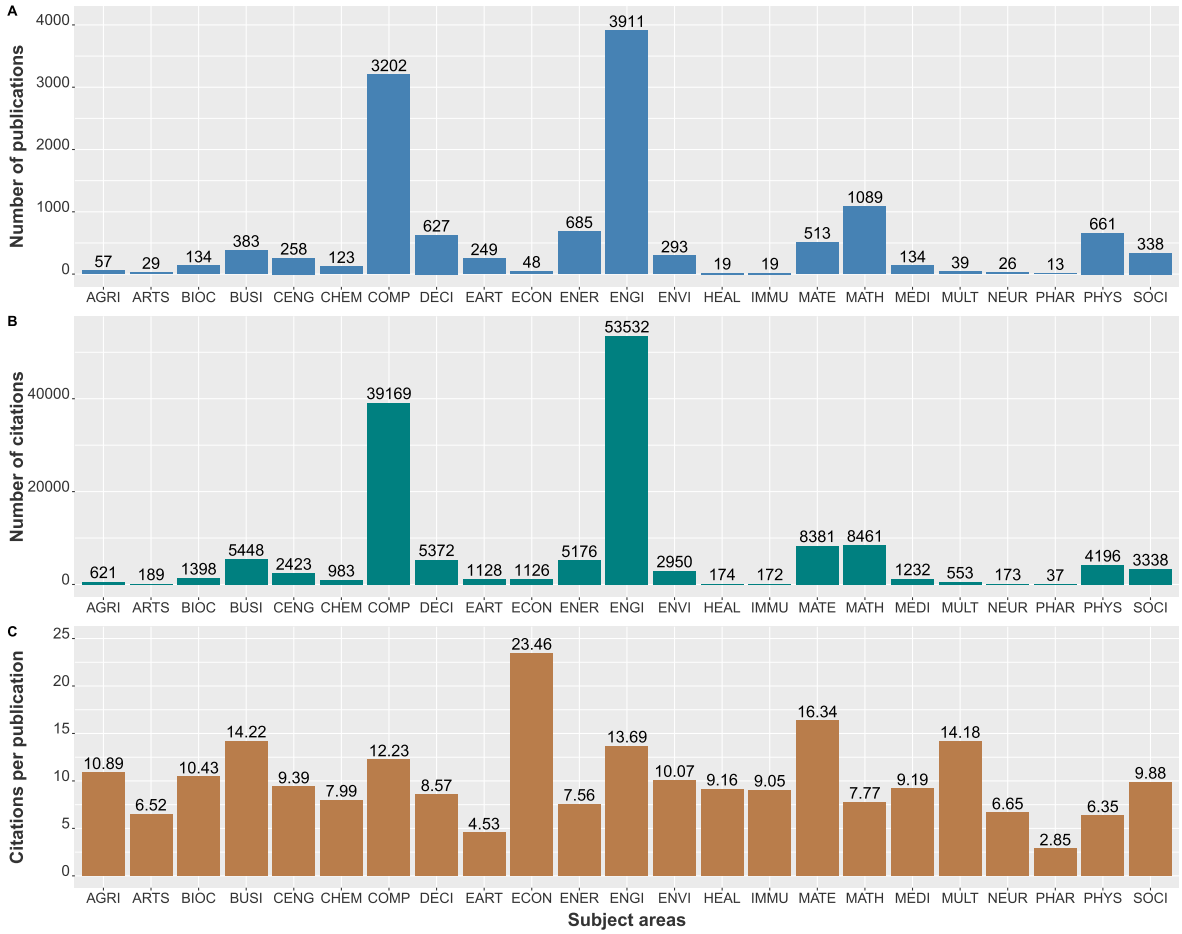


FIGURE 1. Overview of publications in Scopus about digital twin. A: Total number of publications in the 23 subject areas of Scopus that have more than 10 publications. B: Number of citations of these publications. C: Number of citations per publication.

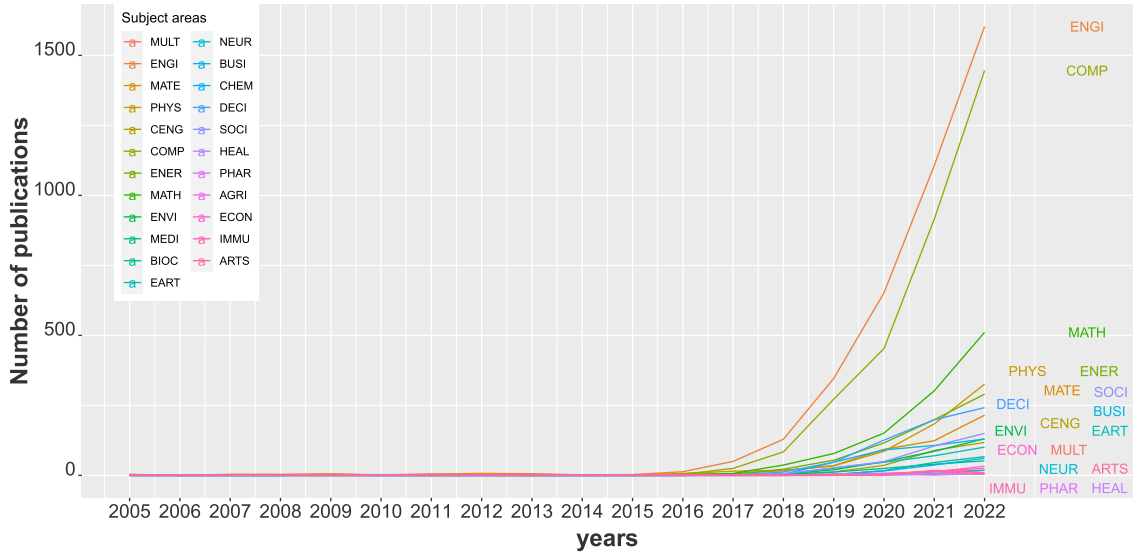


FIGURE 2. Evolution of the number of publications over time for the 23 subject areas shown in Figure 1 A-C.

Overall, the distribution is rapidly decaying and only the paper by [3] is connected to seven subject areas, namely, COMP, DECI, ENER, ENGI, MATH, MEDI, PHYS.

The pairwise co-occurrence of subject areas per publication is shown in Figure 4 B. This heatmap shows all possible pairwise co-occurrences regardless of the total

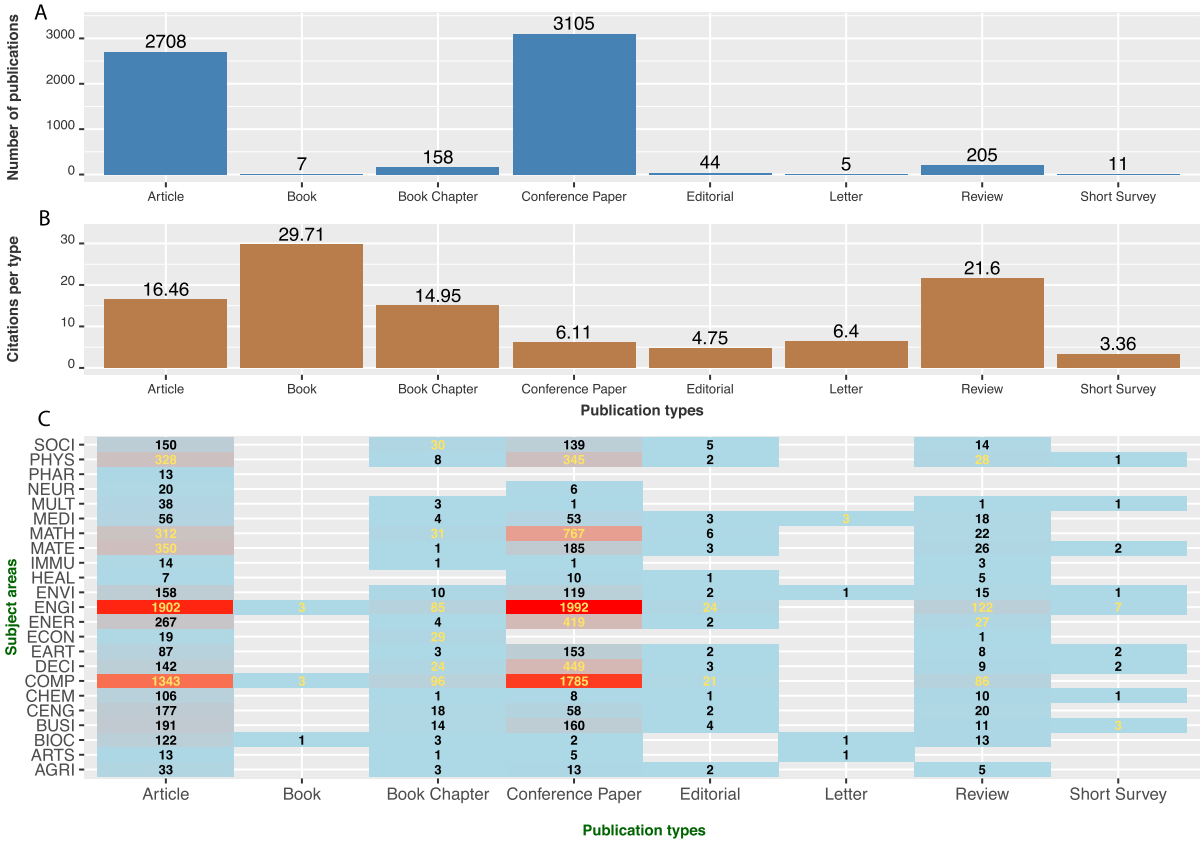


FIGURE 3. Overview of publications per article type. A: Number of publications per article type. B: Citations per publication. C: Heatmap of the pairwise co-occurrence of the 23 subject areas and 8 article types. Significant values are highlighted in yellow.

TABLE 3. Most cited publications across all subject areas and publication types.

Title	Subject area	Citations	Ref
Article: Digital twin-driven product design, manufacturing and service with big data	ENGI, COMP	1296	[39]
Article: Digital Twin in Industry: State-of-the-Art	ENGI, COMP	1014	[39]
Conference paper: Digital Twin in manufacturing: A categorical literature review and classification	ENGI	938	[25]
Book chapter: Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems	ECON, BUSI, MATH, ENGI	915	[18]
Article: A Review of the Roles of Digital Twin in CPS-based Production Systems	ENGI, COMP	751	[28]
Article: About the importance of autonomy and digital twins for the future of manufacturing	ENGI	750	[32]
Conference paper: Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison	COMP, MATE, ENGI	715	[30]
Article: Shaping the digital twin for design and production engineering	ENGI	681	[33]
Conference paper: The digital twin paradigm for future NASA and U.S. air force vehicles	ENGI, MATE, PHYS	648	[17]
Article: Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing	COMP, MATE, ENGI	638	[40]

number of subject areas of a publication. Given the results from the preceding analysis it is not surprising to see that the highest numbers are obtained for engineering (ENGI) and computer science (COMP) followed by MATH-COMP and MATH-ENGI. Significant values are again highlighted in yellow and the threshold for this test is 94. Interestingly, the common significant subject areas for engineering (ENGI) and computer science (COMP) are the eight fields SOCI, PHYS, MATH, MATE, ENER, DECI, CENG and BUSI. Hence, these 10 fields engage in significantly more collaborations with each other than other fields.

To gain insights into higher-dimensional dependencies, we repeat a similar analysis as for Figure 4 B but for three subject areas instead of two. The problem with such an analysis is the visualization because it requires a three-dimensional representation. In order to avoid such visualization issues we

TABLE 4. Most cited publications in earth and environmental sciences (EART, EVNI) (top) and the life sciences (bottom: BIOC, HEAL, IMMU, MEDI, PHAR, NEUR, NURS).

Title	Subject area	Citations	Ref
Article: The Digital Twin of the City of Zurich for Urban Planing	SOCI, PHYS, EART	84	[34]
Article: Development of a bridge maintenance system for pre-stressed concrete bridges using 3D digital twin model	ENGI, EART	79	[37]
Review: Digital twin and CyberGIS for improving connectivity and measuring the impact of infrastructure construction planning in smart cities	SOCI, EART	61	[38]
Article: From LiDAR point cloud towards digital twin city: Clustering city objects based on Gestalt principles	PHYS, ENGI, COMP, EART	42	[49]
Article: Virtual monitoring method for hydraulic supports based on digital twin theory	EART	48	[48]
Article: Triboelectric nanogenerator sensors for soft robotics aiming at digital twin applications	CHEM, BIOC, PHYS	182	[23]
Article: Digital Twins in health care: Ethical implications of an emerging engineering paradigm	BIOC, MEDI	174	[7]
Review: The 'Digital Twin' to enable the vision of precision cardiology	MEDI	167	[9]
Article: Semi-Supervised Support Vector Machine for Digital Twins Based Brain Image Fusion	NEUR	118	[45]
Review: Digital twins to personalize medicine	BIOC, MEDI	80	[4]

show two-dimensional projections for selected subject areas. The results of two projects are shown in Figure 4 C and D. Specifically, Figure 4 C is a projection on engineering (ENGI) and Figure 4 D on computer science (COMP) because these two fields are most frequently involved in pairwise collaborations, corresponding to the highest column (or row) sums in the heatmap in Figure 4 B. The cells in the two heatmaps in Figure 4 C and D show

$$\text{counts(i, j, ENGI)} \tag{6}$$

$$\text{counts(i, j, COMP)} \tag{7}$$

where i and j correspond to the 19 subject areas in Figure 4 C and D (subject areas with no publications were removed).

For the three-dimensional co-occurrence of subject areas in Figure 4 C and D, we observe that DECI, MATH, MATE and PHYS are the most dominating combinations excluding COMP and ENGI respectively.

C. TREND ANALYSIS AND IMPORTANCE OF KEYWORDS

1) FREQUENCY ANALYSIS OF KEYWORDS

In order to identify trends, we analysis the top 35 high-frequency keywords for different subject areas and study their prevalence over the years. To obtain the year-wise proportion of each selected keyword we condition on the subject area and the year. Based on this, we calculate the proportion value for a given keyword as the fraction of its frequency count and the total number of all frequency counts. Hence, the resulting proportion value for each keyword is a normalized number between zero and one. We visualize these results using stacked histograms where one histgram is obtained for a year and subject area. The results of this analysis are shown in Figure 5.

For this figure, we selected the six subject areas ENGI, COMP, SOCI, MEDI, ENER and MATH to obtain a good overview of different disciplines. As one can from Figure 5, in the initial years, only a few common keywords are popular, such as modeling and simulation-related keywords. However, in recent years, the concept of digital twin has diversified and is no longer limited to just modeling or simulations, as evident from the increasing number of associated keywords. For instance, Figure 5 reveals that the digital twin concept is widely used in Industry 4.0, smart manufacturing, digitalization, and machine learning, with AI-related techniques being commonly employed in various subject areas. Nevertheless, the core concepts, such as simulation, modeling, real-time, and others, continue to show a significant presence, emphasizing that the fundamental understanding of the digital twin is crucial to recognize its wider potential. Furthermore, the digital twin concept has expanded to other fields of research, including smart city, smart grid, precision medicine, healthcare delivery, sustainability, and virtual reality, indicating its evolution towards a process-oriented direction, encompassing detailed characteristics and requirements of various applications thereby diversifying the scope of research. Another observation from Figure 5 is that regardless of the subject area there is a consolidation process over the years with respect to the keywords. This is sensible because it indicates the maturing of the fields in a way that there is general agreement what a digital twin means.

To give a succinct summary of this consolidation, we show in Table 5 the top 5 keywords for 2022-2023 of different subject areas ENGI, COMP, ENER, MEDI, MATH, and SOCI (as used in Figure 5). All of these keywords can be also found in the stacked histograms in Figure 5 which are clearly conserved over time after the initial phase. For instance, for ENGI, the top 5 most frequent keywords are modeling, simulations, optimization, performance, and real-time, while for COMP they are modeling, real-time, simulations, performance, and cyberphysical systems; see Table 5.

2) SIGNIFICANCE ANALYSIS OF KEYWORDS

The next analysis aims to identify important keywords in comparison to other subject areas. For this analysis, we are using statistical hypothesis testing based on a Binomial test [14]. On a technical note, we would like to remark that

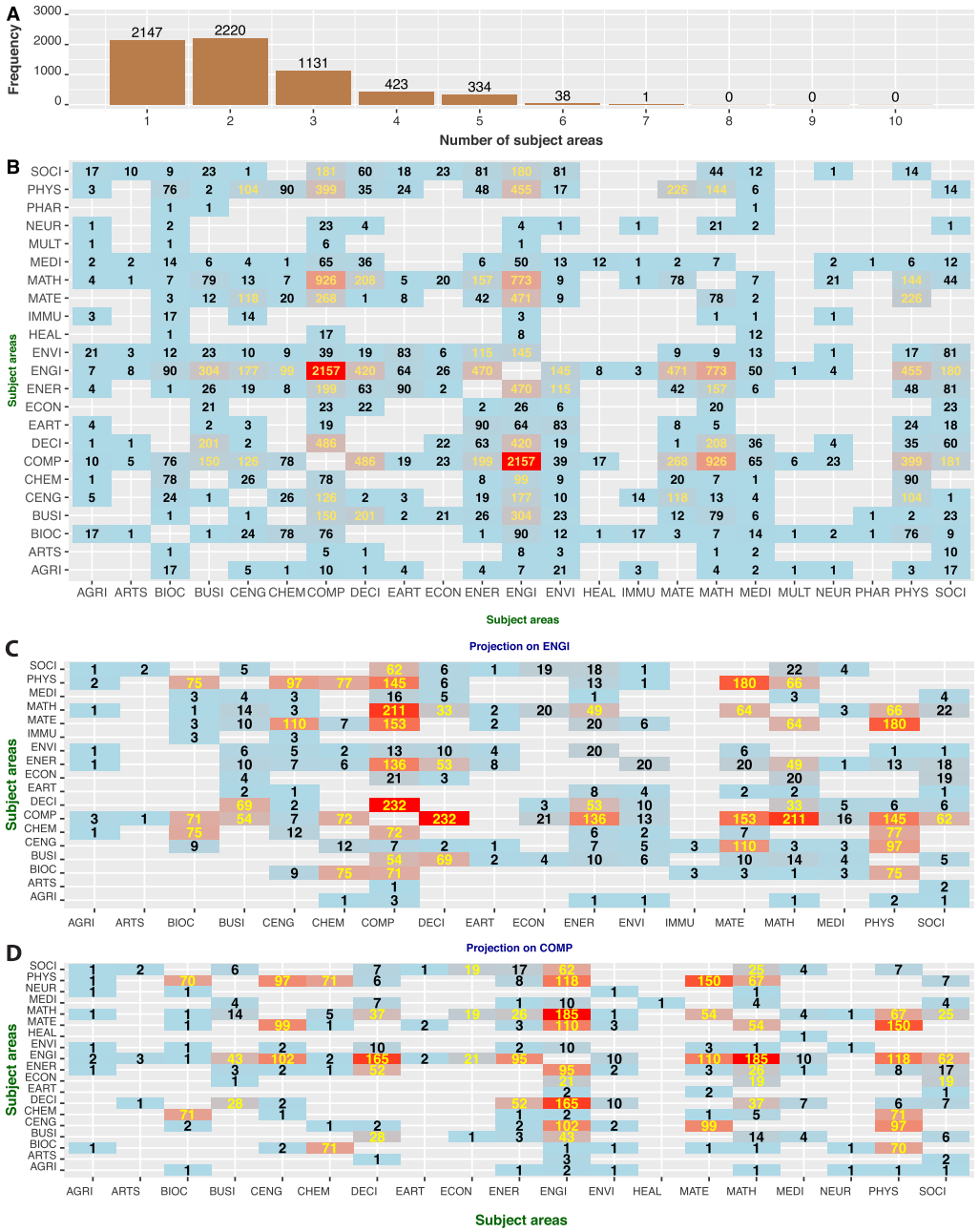


FIGURE 4. Common subject areas per publication. A: Histogram of the number of subject areas per publication. B: Heatmap of the pairwise co-occurrence of subject areas on publications. Significant values are highlighted in yellow whereas the threshold is 94. C: Projection on engineering (ENGI) with a significance threshold of 23. D: Projection on computer science (COMP) with a significance threshold of 19.

TABLE 5. Top 5 most frequent keywords of the six subject areas in Figure 5 for the years in 2022-2023.

subject area	Top 5 keywords
ENGI	modeling, simulations, optimization, real time, performance
COMP	modeling, simulations, real time, performance, cyberphysical systems
SOCI	digitalization, modeling, simulations, real time, smart city
MEDI	digitalization, humans, artificial intelligence, technology, modeling
ENER	modeling, energy, simulations, optimization, real time
MATH	modeling, simulations, real time, cybephysical systems, machine learning

we use a FDR (false discovery rate) control as a multiple testing correction for this analysis. This analysis differentiates keywords for subject areas by comparing them with all other subject areas. For example, it would exclude high-frequency

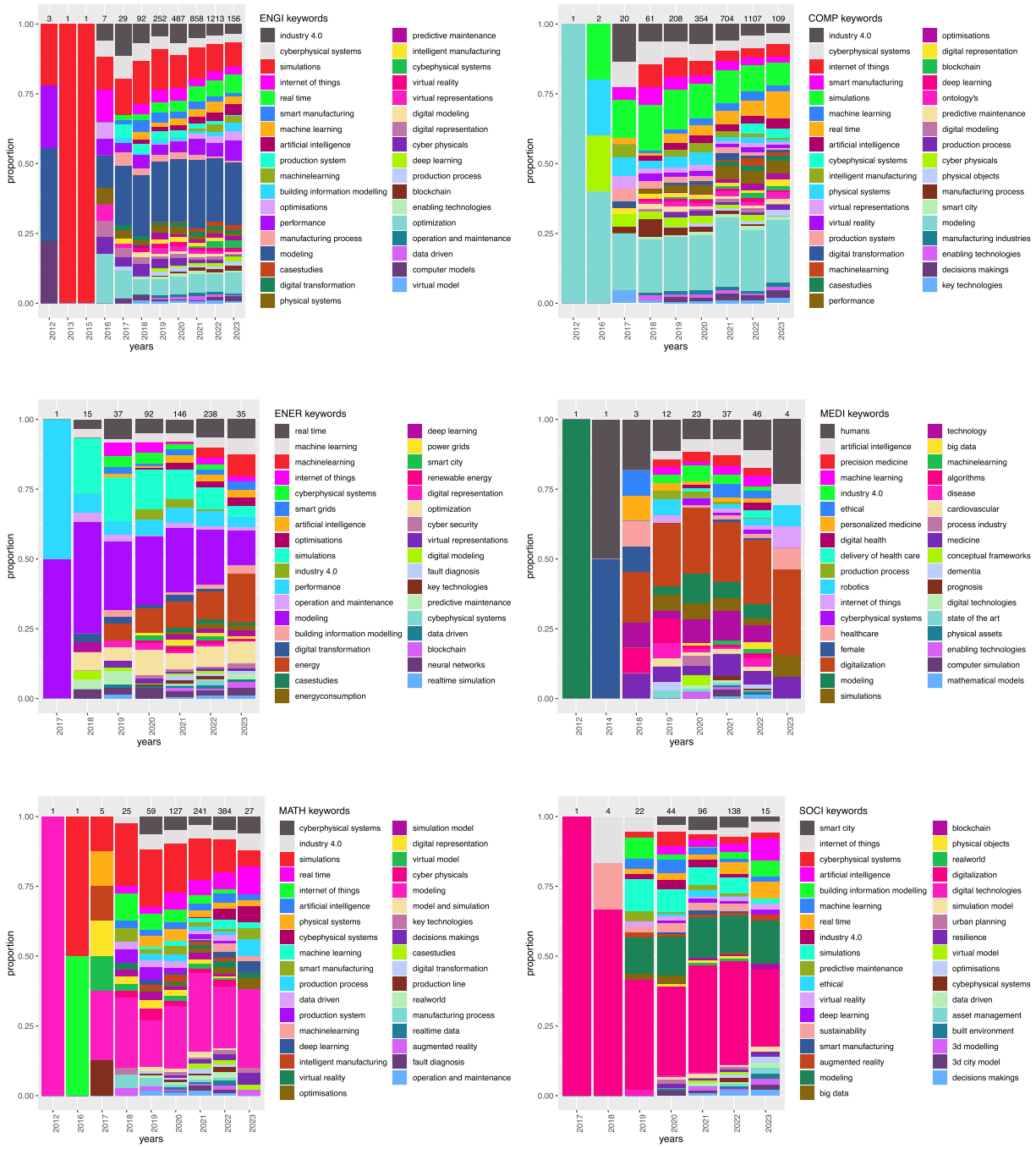


FIGURE 5. **Trend and frequency analysis of the top keywords over time for six subject areas. The color highlights different keywords.**

keywords if they are equally distributed across subject areas (see the discussion in the method section). In Table 6, we show a maximum of 20 significant keywords for each subject area ordered based on their significance (from low to high). In total, we find 2193 (COMP), 3147 (ENGI), 21 (ENER), 32 (MATH), 6 (MEDI), and 1 (SOCI) significant keywords. It is important to note that we found already that the subject areas COMP and ENGI occur together more often (see Figure 4), and therefore, it is plausible that the keywords are strongly overlapping between them. Thus, we observe similar significant keywords in these subject areas. Furthermore, one can see that there is also considerable overlap with the most frequent keywords shown in Table 5.

3) MODULE ANALYSIS OF A BIPARTITE GRAPH

Finally, we conduct an analysis of sub-subject areas and keyword associations using a bipartite module detection algorithm [6]. For this we construct a bipartite graph and

then apply a module detection algorithm. The results of this analysis are shown in Table 7. The applied Louvain algorithm finds 11 modules, with a modularity score of 0.343. This high score indicates that sub-subject areas can be grouped under different keywords that are specific to those research areas. From each module, we select the top 10 subject areas and the top 20 keywords. These results are useful for understanding the common themes of different sub-subjects along with the common keywords in those respective sub-subject areas. For example, the sub-subject areas of module two (see Table 7) are mainly related to energy and sustainability. The common research theme of these sub-subject areas can be understood based on the co-occurring keywords, which show that energy and sustainability-related areas mainly focus on real-time modeling leveraging new technologies (IoT, digitalization) using AI or data-driven models for various operational and maintenance tasks that can save energy and result in sustainable outcomes.

D. TOPIC MODELING

Next, we study topic modelling with the Latent Dirichlet Allocation (LDA) [5] and use metrics from [8] and [10] for the optimization. Basically, the LDA model allows to generate two types of probability distributions one is a set of topic distributions over subject areas and the other is a set of keyword distributions within each topic. From this one can gain insights into the underlying themes in digital twin research and the relationships between topics and subject areas. In our case, keywords will be organized as topics.

The results from the optimization proceedure to find the optimal number of topics is shown in Figure 6 A. Specifically, Deveaud's metric [10] shown on the left-hand side determines the optimal number of topics that maximizes the calculated information divergence between topics. This approach finds an optimal value for eight topics. Second, Cao's metric [8] shown on the right-hand side is based on the convergence of the cosine distance and the cardinality between topics in the LDA. Figure 6 A (right) shows that the optimal topics converge when the number of topics reaches eight. That means for our data both metrics suggest the same number of optimal topics.

Using this cutoff for eight topics, the resulting distribution of these topics for different subject areas is shown in Figure 6 B. This figure illustrates the subject areas where the topics are ranked consecutively. Regarding the top ranks of the topics one finds the following: Topic 7 ranks first in *COMP, DECI ENGI, and MATH*, Topic 5 ranks first in *AGRI, ARTS, HEAL, IMMU, MEDI, PHAR*, Topic 3 ranks first in *EART, ENER, ENVI, SOCI*, Topic 4 ranks first in *BIOC, CHEM, MULT, PHYS*, Topic 2 ranks first in *NEUR*, Topic 6 is in *BUSI, ECON*, and Topic 8 is in *CENG, MATE*. That means 7 of the 8 topics rank first for at least one subject area. Only Topic 1 provides never the largest proportion.

In order to obtain overall proportions for all subject areas, we average over all individual distributions in Figure 6 B. This gives the subject area independent results shown in Table 8. Aside from the overall proportions this table includes also the top keywords for each topic. The topics with the highest overall probabilities are topics 7, 5, 3, and 4 respectively. When discussing the ranks of topics for subject areas, we found that Topic 1 provided never the largest proportion of a subject area. From Table 8, one can see that a reason for this is due to the low proportion of Topic 1 which is 0.080.

Regarding the meaning of the topics one can recognize a structure with respect to specific domains. For example, Topic 1 pertains to the operation and maintenance of smart grids, incorporating key terms such as physical modeling, real-time monitoring, and data-driven approaches. It encompasses various aspects of digital twin technology and aims to facilitate efficient energy management within smart grids. Topic 2 mainly focuses on cyber physical sytem, along with edge computing and key technologies, AI, and simulation for intelligent manufacturing and production, these technologies enable the creation of a digital twin that can be used to monitor and optimize the performance of a physical system or product through simulation. Topic 6 relates to the digital twin research with the incorporation of blockchain technology for supply chain management that can enhance the security and reliability of supply chain management and enable new business models through smart contracts, autonomous vehicles, and trekkings. Topic 8 relates to the digital twin and research for additive manufacturing and related technologies. Other topics include a mix of AI and ML applications for smart energy grids, industrial manufacturing, transportation, and other industrial applications.

Finally, we use the distributions obtained from the LDA model to perform a hierarchical clustering. Specifically, we use the set of topic distributions over subject areas (shown in Figure 6 B) to estimate a dendrogram where a topic distribution serves as profile vector for a subject area. For the distance measure we use the Euclidean distance and for the hierarchical clustering Ward's method. The result of the clustering is shown in Figure 7. This dendrogram consists of 23 branches corresponding to the same 23 subject areas as in Figure 4 A. The optimal number of clusters (three) was identified with the *Ball & Hall* index [2] giving a value of 0.4507. This index is widely used because of its consistency that has been demonstrated for a large number of different data sets [24], [43]. In our case, we study the robustness by varying the publication years up to 2023, 2022, 2021 and 2020 respectively of the considered publications for which the *Ball & Hall* index results always in three optimal cluster. For the results in Figure 7 these three optimal clusters consist of the following subject areas: *BIOC, CHEM*, and *AGRI, ARTS, EART, ECON, ENER, ENVI, HEAL, IMMU, MEDI, MULT, NEUR, PHAR, SOCI*, and *PHYS, CENG, MATE, BUSI, DECI, MATH, COMP, ENGI*. Overall, the observed clusters are sensible as can be seen from the groups (BIOC & CHEM) or (PHYS, DECI, MATH, COMP, ENGI) which makes intuitively sense.

TABLE 6. Significant keywords as a result from a Binomial test and a FDR control.

Subject area	Significant keywords
COMP	cyberphysical systems(<.005), industry 4.0(<.005), smart manufacturing(<.005), internet of things(<.005), simulations(<.005), real time(<.005), machine learning(<.005), artificial intelligence(<.005), intelligent manufacturing(<.005), ontology(<.005), virtual reality(<.005), blockchain(<.005), production system(<.005), digital transformation(<.005), manufacturing industries(<.005), predictive maintenance(<.005), deep learning(<.005), decisions makings(<.005), production process(<.005), optimisations(<.005), smart city(<.005), edge computing(<.005), reinforcement learning(<.005), cloudcomputing(<.005), physical assets(<.005)
ENGI	industry 4.0(<.0005), cyberphysical systems(<.0005), simulations(<.0005), smart manufacturing(<.0005), real time(<.0005), internet of things(<.0005), machine learning(<.0005), production system(<.0005), artificial intelligence(<.0005), building information modelling(<.0005), manufacturing process(<.0005), optimisations(<.0005), predictive maintenance(<.0005), digital modeling(<.0005), modeling(<.0005), digital transformation(<.0005), virtual reality(<.0005), operation and maintenance(<.0005), optimization(<.0005), computer models(<.0005), production process(<.0005), blockchain(<.0005), deep learning(<.0005), data driven(<.0005), virtual commissioning (<.0005)
MEDI	humans(<.0005), precision medicine(<.0005), female(0.00689), digital health(0.00719), ethical(0.01617), personalized medicine(0.03746)
ECON	No significant keywords
SOCI	smart city(<.0005)
MATH	cyberphysical systems(<.0005), real time(<.0005), model and simulation(<.0005), production process(<.0005), simulation model(<.0005), virtual model(<.0005), autonomous systems(<.0005), humancentered computing(0.0047), cyber security(0.0064), production line(0.0083), realtime data(0.0089), decisions makings(0.0115), intelligent manufacturing(0.0146), deep learning(0.0216), production system(0.0292), industry 4.0(0.0303), runtimes(0.0312), virtual reality(0.0329), dataflow(0.0329), human computer interaction(0.0329), virtual simulation(0.0427), modelbase systems engineering(0.0497), augmented reality(0.0497), real time simulation(0.0497)

From Figure 7, one can see that there is clearly further sub-structure in the dendrogram suggesting possibly more than three clusters. However, the *Ball & Hall* index is known to be more conservative by selecting only the most prominent clusters that are evidently present in a dendrogram.

Based on the dendrogram in Figure 7 B, we suggest the introduction of a new measure we call scientometric dimension (SD). Specifically, we define a scientometric dimension as the optimal number of clusters in a dendrogram found with the *Ball & Hall* index based on the output of a LDA. In our case $SD = 3$. This allows a quantitative summary of the observed diversity we found across a number of different analysis methods and indicates that keywords attributed to a digital twin are field-specific, although, there are also commonalities which bound this diversity. Overall, this underlines that the concept of a digital twin is a flexible idea and hints why a universally accepted, domain-independent definition has yet to be established.

IV. DISCUSSION

In this paper, we studied the scholarly literature of digital twin research. Abstractly, this can be seen as an exploration of the space of published literature. Based on data from Scopus, we made a number of interesting observations. On a global view, we observe that by far most articles have been published in engineering (ENGI) and computer science (COMP) and these fields receive also the highest number of citations; see Figure 1. In contrast, among the least active fields are health professions (HEAL), immunology & microbiology (IMMU) and pharmacology, toxicology & pharmaceutics (PHAR). These findings are intuitive considering that in engineering and computer science the usage of simulations has a long tradition. However, it is interesting to see that mathematics (MATH) and physics & astronomy (PHYS), which are the authoritative fields for developing simulations, fall somewhat short in this respect. A reason for the reluctant usage of the term ''digital twin'' in these fields may be related to the lack of a clear (formal) definition to set it apart from (ordinary) simulations. Regarding the efficiency of publications, it is worth highlighting that publications in economics, econometrics and finance (ECON) receive most citations per publication.

From analyzing publication types, we find that the most frequent publication types are journal articles and conference papers. Interestingly the highest citations per publication type are for books followed by reviews receive. From studying the co-occurrence of subject areas and publication types we find that by far most publications are for conference papers and journal articles in engineering (ENGI) and computer science (COMP); see Figure 3. Regarding the citations per subject area we find papers in materials science have the highest mean number of citations; see Table 2.

The second part of our analysis focuses on interdisciplinary relations among subject areas. From this we find that the mean number of subject areas per publication is 2.16 and the mode of this distribution is 2. That means most publications are only associated to two subject areas. Furthermore, it is interesting to see that the maximal number of subject areas per publication is as high as 7; see Figure 4 A. Less surprising it that the two fields with the highest number of co-occurrences are engineering (ENGI) and computer science (COMP). From Figure 4 B, we find that the most interdisciplinary field is engineering (ENGI) having significant co-occurrences with 11 other subject areas, namely,

TABLE 7. Sub-subject areas and their associated keywords found from modules in a bipartite graph connecting sub-subject areas and keywords. Shown are the top 10 sub-subject areas and the top 20 keywords ordered based on frequency (low to high). The first column corresponds to the module numbers.

	Subject area	keywords (within the module)
1	industrial and manufacturing engineering, computer science applications, software, mechanical engineering, computer graphics and computeraided design, mechanics of materials, humancomputer interaction, algebra and number theory, computational mechanics, language and linguistics	products quality, manufacturing is, object recognition, digital factories, manufacturing equipment, product manufacturing, assembly systems, collaborative assembly, automated generation, indoor object acquisition, product development, object acquisition, transdisciplinary engineering, solid modelling, process planning, tool wear, process monitoring, manufacturing service, product data, system design
2	energy engineering and power technology, sustainability and the environment, renewable energy, civil and structural engineering, building and construction, fuel technology, planning and development, geography, geotechnical engineering and engineering geology, energy	real time, internet of things, data driven modelling, smart city, simulations, digital transformation, predictive maintenance, operation and maintenance, modeling, virtual reality, built environment, energyconsumption, virtual representations, realtime simulation, energy efficiency, construction, energy systems, decisions makings, big data, realworld
3	artificial intelligence, reliability and quality, safety, control and optimization, information systems and management, modeling and simulation, media technology, computer vision and pattern recognition, management information systems, business and international management	twin networks, smart grids, physical network, parallel system, matchings, industrial environments, transportation system, heterogeneous data, mechatronics, intelligent transportation systems, automated vehicles, parallel control, networkbased, humancentered computing, opensource, runtimes, intelligent operations, detection methods, construction method, management and controls
4	computer science, engineering, materials science, physics and astronomy, theoretical computer science, mathematics, fluid flow and transfer processes, process chemistry and technology, decision sciences, multidisciplinary	machine learning models, smart manufacturing, building information modelling, ontology's, blockchain, fault diagnosis, simulation model, augmented reality, autonomous systems, industrial internet of things, multiagent system, mathematical models, control systems, data models, integration, physical environments, drones, semantic web, security, kalman filter
5	computer networks and communications, information systems, instrumentation, and optics, atomic and molecular physics, hardware and architecture, biochemistry, signal processing, analytical chemistry, library and information sciences	edge computing, algorithms, industry, digital replicas, key technologies, production line, virtual worlds, cyber security, computing resource, aerial vehicle, edge networks, resource allocation, support vector machine, unmanned aerial vehicle, mixed reality, physical devices, industrial systems, industry sectors, metaverse, data collection
6	electrical and electronic engineering, optical and magnetic materials, electronic, applied mathematics, condensed matter physics, metals and alloys, materials chemistry, coatings and films, surfaces, acoustics and ultrasonics	optimisations, performance, digital holography, digital representation, power grids, reinforcement learning, inline digital holography, product life cycles, operational conditions, phase retrieval, additive manufacturing, information modeling, fault detection, state of the art, industrial processs, photomask, twin image, virtual simulation, digitisation, reconstructed image
7	chemical engineering, bioengineering, biomedical engineering, health informatics, biotechnology, chemistry, computational theory and mathematics, computational mathematics, health information management, medicine	ethical, personalized medicine, process analytical technology, technology, quality by design, deep learning, computer simulation, robotics, healthcare, biomanufacturing, process control, cardiovascular, drug industry, dementia, process modeling, manufacturing, sensors, biological products, female, genomics
8	strategy and management, management science and operations research, management and accounting, business, management of technology and innovation, industrial relations, polymers and plastics, probability and uncertainty, statistics, sensory systems	industry 4.0, production system, cyber physicals, cyber physical systems, sustainability, optimization, manufacturing industries, supply chain, discrete event simulation, physical objects, virtual tryon, decentralised, innovation, training, modelbased opc, body measurements, textile material, russian lubok, reverse engineering, business model
9	aerospace engineering, automotive engineering, space and planetary science, astronomy and astrophysics	mobile edge computing, methods: data analysis, prediction algorithms, property, remaining useful lives, aerospace systems, methods. data analysis, large scale structure of universe, largescale structure of universe, flight conditions, correction factors, trend prediction, telemetry systems, simulated flight, ground control, fault conditions, flight data, control monitoring, fault diagnosis model, performance modeling
10	education, sociology and political science, human factors and ergonomics, psychology	qualitative study, sustainable entrepreneurship, entrepreneurial attitude, dynamic panels, dynamic panel data approaches, national systems of entrepreneurship, attitudes, aspirations, abilities, is planning, indepth analysis, consumer response, business marketing, ebusinesses, digital marketing, digital innovations, consumer behavior, network educational reality, digitalization of education, professional competencies
11	diabetes and metabolism, endocrinology, internal medicine	type 2 diabetes, precision nutrition, hba1c reduction, diabetes medication elimination, continuous glucose monitoring

SOCI, PHYS, MATH, MATE, ENVI, ENER, DECI, COMP, CHEM, CENG and BUSI.

In order to obtain insights into higher-dimensional relations between subject areas, we extended the above analysis to the co-occurrence of three subject areas and visualized two-dimensional projects on particular subject areas. These results are shown in Figure 4 C and D where projections on engineering (ENGI) and computer science (COMP) are displayed because these two fields were by far the most frequent combinations we found from a pairwise analysis of subject areas (see Figure 4 B). It is important to note that for the analysis in Figure 4 C and D only publication with at least three subject areas have been used while for Figure 4 B publications with two or more subject areas have been considered. A consequence of this can be observed in the value of the significance threshold which was in Figure 4 B 94 while in Figure 4 C and D it is 23 and 19 respectively. This change in the value of the significance threshold leads to a number of differences. First, while the field Biochemistry, Genetics and Molecular Biology (BIOC) has no significant co-occurrence with any other field in Figure 4 B, it is significant for 3 respectively 2 fields in Figure 4 C and D. Interestingly, for environmental science (ENVI) the roles are reversed that means in Figure 4 B there are two significant fields while in Figure 4 C and D there are zero.

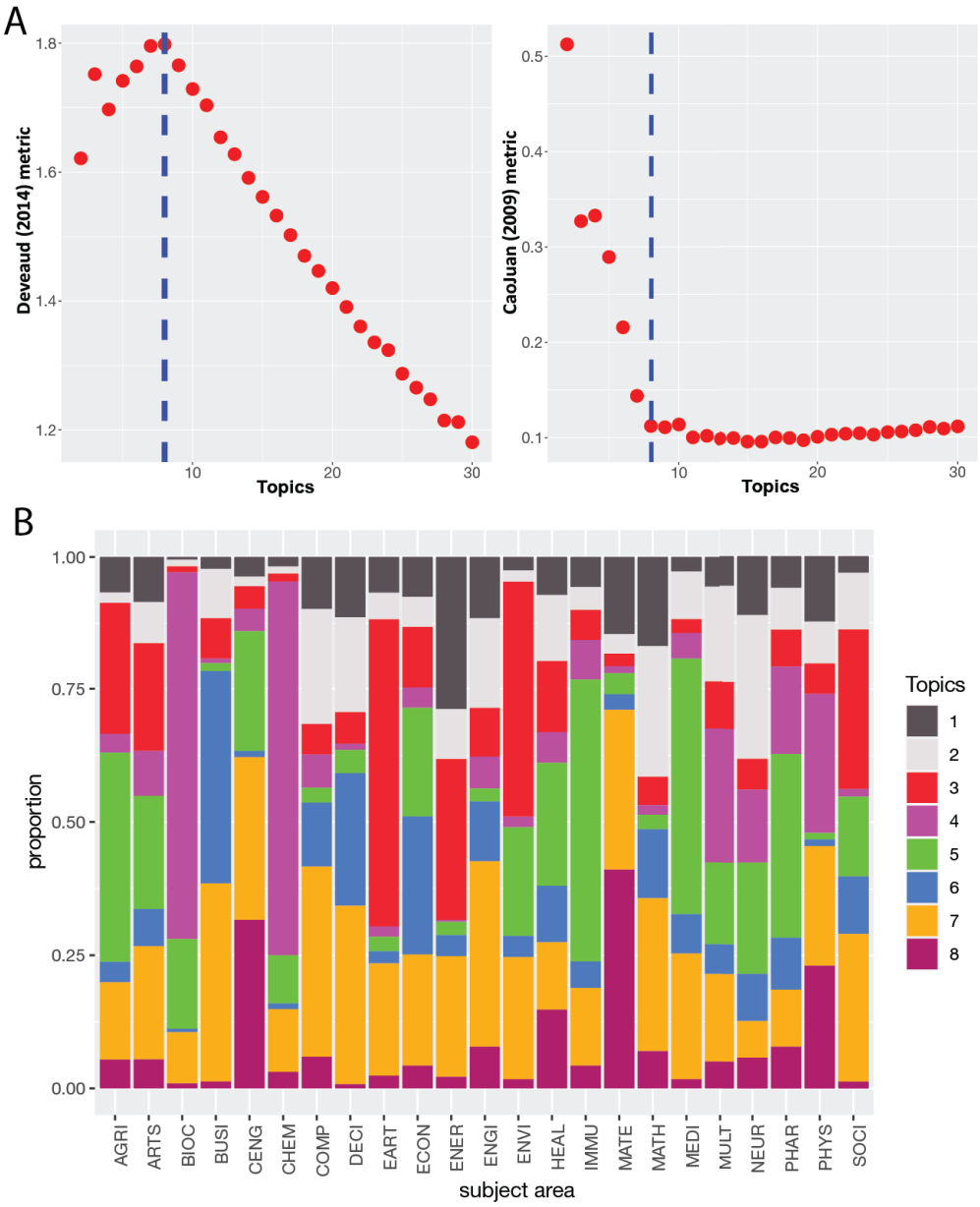


FIGURE 6. LSA analysis of the publication data. A: The Deveaud metric (left) and Cao metric (right) to select the optimal number of topics for the LDA. The vertical, dashed line indicates the optimal solution for 8 topics. B: Results of the LDA showing prevalence for different topics across subject areas. The corresponding topics are shown in Table 8.

Regarding higher-order subject areas, from Figure 4 C and D, we find decision sciences (DECI) and mathematics (MATH) on number three and four behind COMP and ENGI. It is interesting that also these subject areas are rather technical without a clear singleton application. This could indicate that so far literature about a digital twin is still in the experimental stage to find the best applications.

Finally, in the third part of our analysis we study a trend analysis and topic modeling. For the trend analysis, we investigated the frequency of author provided keywords over time. There are two notable observations arising from this analysis. First, it appears that each subject area is undergoing a consolidation process and, second, the keyword sets used to describe each subject area are diverse. The former observation suggests that these fields are maturing, with a growing consensus on what constitutes a digital twin; see Figure 5. However, this consensus is not universal; rather, it is subject-specific. This is reflected in the diversity of keyword sets, which vary depending on the subject area; see Table 5. Given that different subject areas have distinct focuses, this observation is unsurprising. Consequently, it seems that there is no single definition of a digital twin that can be applied across all

TABLE 8. Summary of the LDA analysis shown in Figure 6 B. Shown are the top 20 keywords of 8 topics and their overall prevalence in subject areas (proportion).

Topic No.	Proportion	Top 20 keywords
1	0.080	operation and maintenance, smart grids, power grids, neuralnetworks, energy, real time, neural networks, machinelearning, physical modelling, key technologies, physicsbased models, cloudbased, monitoring system, renewable energy, emerging technologies, monitor and control, distributed energy resources, whole life cycles, uncertainty quantifications, reduced order modelling
2	0.099	cybephysical systems, data driven, physical systems, cyber physical systems, fault diagnosis, edge computing, cyber security, intelligent manufacturing, production process, modelbase systems engineering, deep learning, cloudcomputing, ontology's, model and simulation, humanrobot collaboration, information and communication technology, complex products, virtual worlds, virtual simulation, federated learning
3	**0.135**	smart city, real time, building information modelling, digital transformation, digital technologies, optimization, 3d modelling, energy systems, deep learning, built environment, optimisations, machinelearning, sustainability, physicsbased, data driven, operation and maintenance, machine learning, energyconsumption, circular economy, energy
4	**0.122**	machine learning, technology, humans, algorithms, industry 4.0, industry, digital replicas, computer simulation, robotics, neural networks, artificial intelligence, cybephysical systems, deep learning, software, monitoring, support vector machine, sensors, computer vision, industrial production, edge networks
5	**0.167**	artificial intelligence, humans, modeling, digitalization, precision medicine, ethical, mathematical models, deep learning, database, conceptual frameworks, privacy, workflows, covid19, technology, process industry, data collection, personalized medicine, software, process control, prediction
6	0.092	blockchain, literature review, manufacturing companies, physical devices, virtual spaces, discrete event simulation, business process, supply chain, virtual commissioning, knowledge based, additive manufacturing, technology development, digital factories, construction projects, transportation infrastructures, transportation system, cyberphysical production system, business model, manufacturing resource, systems architecture
7	**0.221**	industry 4.0, cyberphysical systems, internet of things, simulations, smart manufacturing, real time, machine learning, artificial intelligence, optimisations, virtual representations, virtual reality, performance, production system, predictive maintenance, casestudies, digital transformation, digital modeling, manufacturing process, machinelearning, digital representation
8	0.080	machine learning, additive manufacturing, digital holography, modeling, fault diagnosis, collaborative robots, operational conditions, deep learning, inline digital holography, management is, computer models, condition monitoring, dynamic scheduling, manufacturing, semantic web, process modeling, development directions, system dynamics, hybrid model, 3d printing

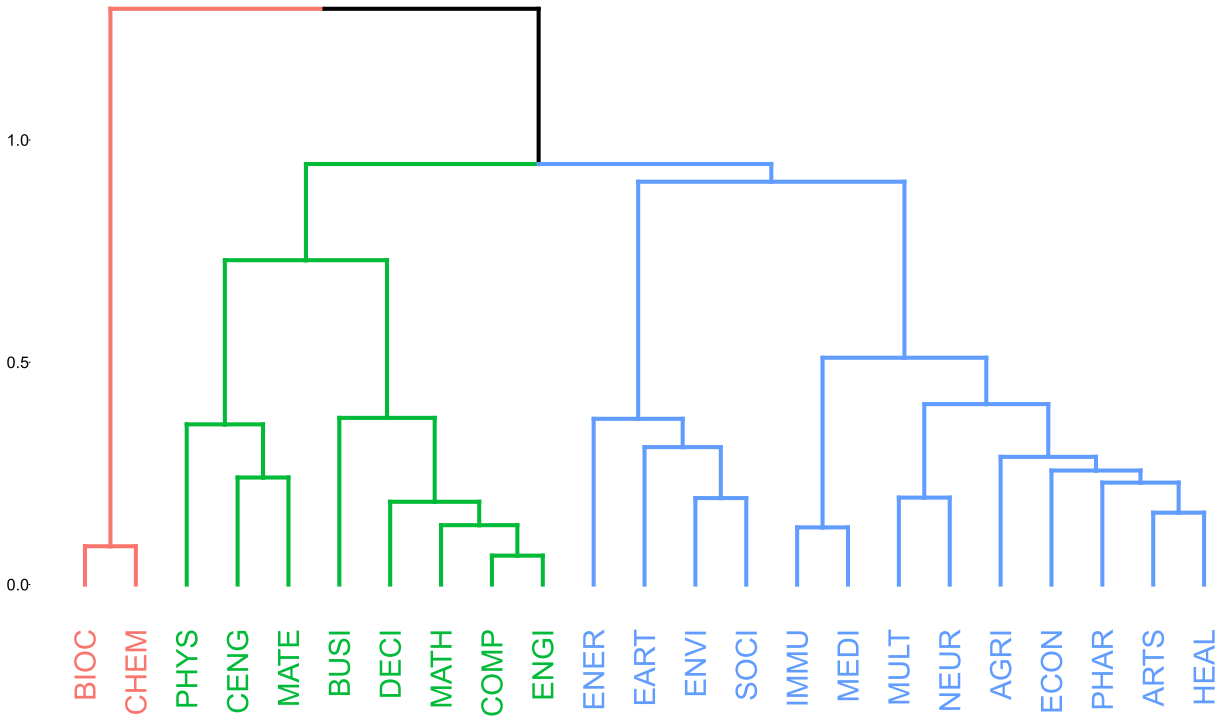


FIGURE 7. Dendrogram from the LDA analysis using the set of topic distributions over subject areas. It consists of 23 branches corresponding to 23 subject areas with an optimal number of 3 clusters identified with the *Ball & Hall* index. For the distance measure the Euclidean distance has been used and for the hierarchical clustering Ward's method.

subject areas and academic fields. Instead, its meaning is contingent on the specific context in which it is used.

This diversity is also confirmed by a significance analysis of keywords and a module analysis of a bipartite graph

considering sub-subject areas. From this analysis, we identified 11 modules representing common sub-subject areas and their associated keywords. Specifically, the bipartite graph consist of two sets of vertices - sub-subject areas and their corresponding closely connected keywords - and the module detection algorithm identifies closely connected non-overlapping sets of vertices of sub-subject areas and keywords. The high modularity (0.349) indicates that subject areas and keyword associations can be categorized into different sets where certain sub-subject areas and keyword associations are more tightly connected than other groups. Therefore, different modules indicate a preference of sub-subject areas, with the research topics diversifying different research subjects and keywords related to digital twin research. Thus, this approach can be used to describe subject-specific research themes compared to other sub-subject areas. Overall, the bipartite graph analysis and module detection approach provide insights into the organization and structure of research topics and their associated keywords. Therefore, it can be used to describe common subject-specific research themes compared to other sub-subject areas.

To gain even more detailed insights into the diversity of subject areas, we used topic modeling with a LDA model. A LDA is a quite complex model that estimates two sets of probability distributions. One is a set of topic distributions over subject areas and the other is a set of keyword distributions within each topic. That means in contrast to the frequency analysis of keywords the LDA estimates the optimal number of topics and the associated distributions simultaneously. Here it is important to note that the topics remain constant over the subject areas. Overall, by analyzing the frequency and co-occurrence of keywords across subject areas, the LDA extracts topics and provides insights into the most important themes in the field of digital twin research.

From this analysis, we obtain three main findings. First, all topics are not uniformly distributed across the subject areas, instead, the contribution of a topic various considerably; see Figure 6 B. For example, topic 6 makes the smallest contribution for CHEM while for BUSI it makes the largest contribution. Second, there is no dominating topic for any subject area but several topics are needed. Specifically, to obtain a coverage of 50% for most subject areas more than one topic is needed. This indicates that each subject area is a mixture of the underlying topics demonstrating the interdisciplinarity of the research problems. Third, the topic making the largest contribution to a subject area varies. Specifically, from Figure 6 B, we can see that topic 7 ranks first in *COMP, DECI ENGI, and MATH*, topic 5 ranks first in *AGRI, ARTS, HEAL, IMMU, MEDI, PHAR*, topic 3 ranks first in *EART, ENER, ENVI, SOCI*, topic 4 ranks first in *BIOC, CHEM, MULT, PHYS*, topic 2 ranks first in *NEUR*, topic 6 in *BUSI, ECON*, and topic 8 in *CENG, MATE*.

The interpretation of these results suggests that the meaning and the basic understanding of digital twins may differ between subject areas, as can be seen from the different proportions of different topics; see Figure 6 B. From the different topic proportions in the different subject areas, it is evident that research related to a digital twins needs to be understood in a domain-specific context. That means digital twin research needs to be understood in terms of the uniqueness of the research question and its interdisciplinary character. Another insight is that, besides the traditional digital twin-related keywords, Industry 4.0, smart manufacturing, digitalization, new technologies, and AI and ML-related keywords are top keywords significantly associated with different topics. The co-occurrence of these words in different topics indicates that the digital twin is not seen from a traditional perspective but an area dominated by technology, AI, and data science, with significant applications in industrial settings and smart manufacturing.

Finally, we would like to highlight that this diversity is not limitless but bound. This was demonstrated by a hierarchical clustering of the output from the LDA model. Using the conservative *Ball & Hall* index to identify the optimal number of clusters we found only 3 main clusters; see Figure 7. While there is clearly further sub-structure in the dendrogram suggesting possibly more than three clusters the identified 3 clusters are evidently the most dominating ones. This result suggests that among the 23 studied subject areas 3 main interpretations of a digital twin are prevalent in the literature which is surprisingly low dimensional considering that domain-specific elucidations are pronounced. Overall, we suggest to use this as a summary statistics and call it scientometric dimension (SD) of digital twin research.

The last point we would like to make is to note that no analysis is without limitations. For example, our study could be repeated for publication data from the Web of Science (WOS). While WOS and Scopus are authoritative and similar in many aspects they are not identical with respect to the listed information nor do they use the same definitions for subject areas. This could lead to variations in parts of our analysis. Hence, our results should be seen in the context of the used data from Scopus and all interpretations relate back to annotations and categorizations made for this bibliographic and citation resource.

V. CONCLUSION

Over the past few years, there has been a significant increase in the number of publications examining digital twins. To gain insights into trends and the structure of this literature, we conduct a scientometric analysis using large-scale bibliographic and citation data from Scopus.

From our analysis we obtain four key findings. First, the majority of articles on digital twins are published in the fields of engineering and computer science, which also receive the highest number of citations. However, publications in economics, econometrics, and finance receive the highest number of citations per publication. Additionally, while journal articles and conference papers are the most common type of publications, books have the highest number of citations per publication. Notably, papers in materials science receive the highest mean number of citations. Second, we observe

that most publications on digital twins are interdisciplinary involving two or more subject areas, with an average of 2.16 subject areas per publication and a maximum of 7. The two fields with the highest number of co-occurrences are engineering and computer science, followed by decision sciences and mathematics. Third, a trend analysis of keywords revealed a consolidation process after an initial phase, resulting in a stabilization of keywords over time. Importantly, each subject area has its own unique set of characteristic keywords, indicating a diversity in the interpretation of the concept of digital twins. Forth, the diversity of the digital twin concept as found in the literature is not limitless but bound. This is confirmed by a hierarchical clustering of the output from a LDA model identifying only 3 main clusters among all subject areas. This forms the bases of a summary statistics for the bound diversity of the scholarly literature, we call scientometric dimension (SD) of digital twin research.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

AUTHOR CONTRIBUTION

Frank Emmert-Streib conceived the study, conducted the analysis, interpreted the results and wrote the paper. Shailesh Tripathi conducted the analysis, interpreted the results and wrote the paper. Matthias Dehmer interpreted the results and wrote the paper. All authors approved the final version.

REFERENCES

[1] S. Aheleroff, R. Y. Zhong, and X. Xu, "A digital twin reference for mass personalization in industry 4.0," *Proc. CIRP*, vol. 93, pp. 228–233, Jan. 2020.

[2] G. H. Ball and D. J. Hall, "Isodata, a novel method of data analysis and pattern classification," Stanford Research Inst., Menlo Park, CA, USA, pp. 1–60, 1965, vol. 4.

[3] A. Bhatt and V. Karthikeyan, "Digital twin framework and its application for protection functions testing of relays," in *Proc. 3rd Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Aug. 2022, pp. 682–687.

[4] B. Björnsson, C. Borrebaeck, N. Elander, T. Gasslander, D. R. Gawel, M. Gustafsson, R. Jörnsten, E. J. Lee, X. Li, S. Lilja, D. Martínez-Enguita, A. Matussek, P. Sandström, S. Schäfer, M. Stenmarker, X. F. Sun, O. Sysoev, H. Zhang, and M. Benson, "Digital twins to personalize medicine," *Genome Med.*, vol. 12, no. 1, pp. 1–4, Dec. 2020.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

[7] K. Bruynseels, F. Santoni De Sio, and J. van den Hoven, "Digital twins in health care: Ethical implications of an emerging engineering paradigm," *Frontiers Genet.*, vol. 9, p. 31, Feb. 2018.

[8] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive LDA model selection," *Neurocomputing*, vol. 72, nos. 7–9, pp. 1775–1781, Mar. 2009.

[9] J. Corral-Acero et al., "The 'digital twin' to enable the vision of precision cardiology," *Eur. Heart J.*, vol. 41, no. 48, pp. 4556–4564, 2020.

[10] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," *Document Numérique*, vol. 17, no. 1, pp. 61–84, Apr. 2014.

[11] H. Duan, S. Gao, X. Yang, and Y. Li, "The development of a digital twin concept system," *Digit. Twin*, vol. 2, p. 10, Aug. 2022.

[12] N. Dugué and A. Perez, "Directed Louvain: Maximizing modularity in directed networks," Ph.D. dissertation, Université d'Orléans, Orléans, France, 2015.

[13] F. Emmert-Streib and M. Dehmer, "Large-scale simultaneous inference with hypothesis testing: Multiple testing procedures in practice," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 2, pp. 653–683, May 2019.

[14] F. Emmert-Streib and M. Dehmer, "Understanding statistical hypothesis testing: The logic of statistical inference," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 945–961, Aug. 2019.

[15] F. Emmert-Streib and O. Yli-Harja, "What is a digital twin? Experimental design for a data-centric machine learning perspective in health," *Int. J. Mol. Sci.*, vol. 23, no. 21, p. 13149, Oct. 2022.

[16] D. Gelernter, *Mirror Worlds: Or the Day Software Puts the Universe in a Shoebox…How it Will Happen and What it Will Mean*. Oxford, U.K.: Oxford Univ. Press, 1991.

[17] E. Glaessgen and D. Stargel, "The digital twin paradigm for future NASA and U.S. air force vehicles," in *Proc. 53rd AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn. Mater. Conf., 20th AIAA/ASME/AHS Adapt. Struct. Conf., 14th AIAA*, Apr. 2012, p. 1818.

[18] J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, 2017, pp. 85–113. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-38756-7_4

[19] W. Hood and C. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics*, vol. 52, no. 2, pp. 291–314, 2001.

[20] H. H. Hosamo, A. Imran, J. Cardenas-Cartagena, P. R. Svennevig, K. Svidt, and H. K. Nielsen, "A review of the digital twin technology in the AEC-FM industry," *Adv. Civil Eng.*, vol. 2022, pp. 1–17, Mar. 2022.

[21] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.

[22] R. Jha, A.-A. Jbara, V. Qazvinian, and D. R. Radev, "NLP-driven citation analysis for scientometrics," *Natural Lang. Eng.*, vol. 23, no. 1, pp. 93–130, Jan. 2017.

[23] T. Jin, Z. Sun, L. Li, Q. Zhang, M. Zhu, Z. Zhang, G. Yuan, T. Chen, Y. Tian, X. Hou, and C. Lee, "Triboelectric nanogenerator sensors for soft robotics aiming at digital twin applications," *Nature Commun.*, vol. 11, no. 1, p. 5381, Oct. 2020.

[24] A. Karanikola, C. M. Liapis, and S. Kotsiantis, "A comparative study of validity indices on estimating the optimal number of clusters," in *Proc. 12th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2021, pp. 1–8.

[25] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital twin in manufacturing: A categorical literature review and classification," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018.

[26] J. Mingers and L. Leydesdorff, "A review of theory and practice in scientometrics," *Eur. J. Oper. Res.*, vol. 246, no. 1, pp. 1–19, Oct. 2015.

[27] G. Moiceanu and G. Paraschiv, "Digital twin and smart manufacturing in industries: A bibliometric analysis with a focus on industry 4.0," *Sensors*, vol. 22, no. 4, p. 1388, Feb. 2022.

[28] E. Negri, L. Fumagalli, and M. Macchi, "A review of the roles of digital twin in CPS-based production systems," *Proc. Manuf.*, vol. 11, pp. 939–948, Jan. 2017.

[29] G. B. Ozturk, "Digital twin research in the AECO-FM industry," *J. Building Eng.*, vol. 40, Aug. 2021, Art. no. 102730.

[30] Q. Qi and F. Tao, "Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison," *IEEE Access*, vol. 6, pp. 3585–3593, 2018.

[31] *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, R Development Core Team, Vienna, Austria, 2008.

[32] R. Rosen, G. von Wichert, G. Lo, and K. D. Bettenhausen, "About the importance of autonomy and digital twins for the future of manufacturing," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 567–572, 2015.

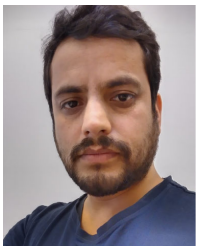
[33] B. Schleich, N. Anwer, L. Mathieu, and S. Wartzack, "Shaping the digital twin for design and production engineering," *CIRP Ann.*, vol. 66, no. 1, pp. 141–144, 2017.

[34] G. Schrotter and C. Hürzeler, "The digital twin of the city of Zurich for urban planning," *PFG-J. Photogramm., Remote Sens. Geoinf. Sci.*, vol. 88, no. 1, pp. 99–112, Feb. 2020.

[35] S. M. E. Sepasgozar, "Differentiating digital twin from digital shadow: Elucidating a paradigm shift to expedite a smart, sustainable built environment," *Buildings*, vol. 11, no. 4, p. 151, Apr. 2021.

[36] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2004.
[37] C.-S. Shim, N.-S. Dang, S. Lon, and C.-H. Jeon, "Development of a bridge maintenance system for prestressed concrete bridges using 3D digital twin model," *Struct. Infrastruct. Eng.*, vol. 15, no. 10, pp. 1319–1332, Oct. 2019.
[38] S. Shirowzhan, W. Tan, and S. M. E. Sepasgozar, "Digital twin and CyberGIS for improving connectivity and measuring the impact of infrastructure construction planning in smart cities," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 4, p. 240, Apr. 2020.
[39] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.
[40] F. Tao and M. Zhang, "Digital twin shop-floor: A new shop-floor paradigm towards smart manufacturing," *IEEE Access*, vol. 5, pp. 20418–20427, 2017.
[41] E. J. Tuegel, A. R. Ingraffea, T. G. Eason, and S. M. Spottswood, "Reengineering aircraft structural life prediction using a digital twin," *Int. J. Aerosp. Eng.*, vol. 2011, pp. 1–14, 2011.
[42] J. Vachálek, L. Bartalský, O. Rovný, D. Šišmišová, M. Morhác, and M. Lokšík, "The digital twin of an industrial production line within the industry 4.0 concept," in *Proc. 21st Int. Conf. Process Control (PC)*, Jun. 2017, pp. 258–262.
[43] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 3, no. 4, pp. 209–235, Aug. 2010.
[44] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *Int. J. Comput. Sci. Commun. Netw.*, vol. 5, no. 1, pp. 7–16, 2015.
[45] Z. Wan, Y. Dong, Z. Yu, H. Lv, and Z. Lv, "Semi-supervised support vector machine for digital twins based brain image fusion," *Frontiers Neurosci.*, vol. 15, Jul. 2021, Art. no. 705323.
[46] J. Wang, X. Li, P. Wang, and Q. Liu, "Bibliometric analysis of digital twin literature: A review of influencing factors and conceptual structure," *Technol. Anal. Strategic Manage.*, vol. 2022, pp. 1–15, Jan. 2022.
[47] H. Xie, M. Xin, C. Lu, and J. Xu, "Knowledge map and forecast of digital twin in the construction industry: State-of-the-art review using scientometric analysis," *J. Cleaner Prod.*, vol. 383, Jan. 2023, Art. no. 135231.
[48] J. Xie, X. Wang, Z. Yang, and S. Hao, "Virtual monitoring method for hydraulic supports based on digital twin theory," *Mining Technol.*, vol. 128, no. 2, pp. 77–87, Apr. 2019.
[49] F. Xue, W. Lu, Z. Chen, and C. J. Webster, "From LiDAR point cloud towards digital twin city: Clustering city objects based on gestalt principles," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 418–431, Sep. 2020.
[50] Q. Zhao and P. Fränti, "WB-index: A sum-of-squares based index for cluster validity," *Data Knowl. Eng.*, vol. 92, pp. 77–89, Jul. 2014.

FRANK EMMERT-STREIB received the B.Sc. and M.Sc. degrees in theoretical physics from the University of Siegen, Germany, and the Ph.D. degree in theoretical physics from the University of Bremen, Germany. He studied physics and mathematics. He is currently a Professor of data science with Tampere University, Finland, where he is leading the Predictive Society and Data Analytics Laboratory.



SHAILESH TRIPATHI received the Ph.D. degree in computational biology from Queen's University, Belfast. He is currently a Postdoctoral Researcher with the University of Applied Sciences Upper Austria. His research interests include data science, machine learning methods, data visualization, and network analysis. His research aims to extract and apply knowledge from industrial-manufacturing data and computational biology.

MATTHIAS DEHMER received the B.Sc. and M.Sc. degrees in mathematics from the University of Siegen, Germany, the Habiliation degree in applied mathematics from the Vienna University of Technology, Austria, and the Ph.D. degree in computer science from the University of Darmstadt, Germany. He is currently a Professor with the Department of Computer Science, Swiss Distance University of Applied Sciences, Brig, Switzerland, and Tyrolean Private University UMIT TIROL, Department of Mechatronics and Biomedical Computer Science, Hall, Austria.